Mitigating Bias in Machine Learning

Dr Leena Murgai

June 20, 2022

Contents

List of Figures

List of Tables

Notation and conventions

Mathematical notation

- $\mathbb{P}(A)$ denotes probability of event A
- $\mathbb{E}$ denotes expectation
- $\forall$ means for all
- $|$ means such that
- $\in$ means a member of
- $\Rightarrow$ means implies
- $\Leftrightarrow$ means if and only if
- square brackets are inclusive, round brackets are not, for example,

$$x \in [a, b) \quad \Leftrightarrow \quad a \leq x < b$$

- $\rightarrow$ means tends to. We use a superscript $+$ or $-$ to indicate if the limiting value is approached from above or below respectively,
  - $y(x) \rightarrow 0^+$ as $x \rightarrow \infty \Rightarrow y(x)$ tends to 0 from above as x tends to infinity.
  - $y(x) \rightarrow 0^-$ as $x \rightarrow \infty \Rightarrow y(x)$ tends to 0 from below as x tends to infinity.

Typographical conventions

In this book we mostly follow mathematical typographical conventions for variables. All variables are in italic. We use:

- lowercase letters for scalar variables, e.g. a,
- uppercase letters for random variables, e.g. X.
- lowercase bold typeface letters for (column) vectors, e.g. $\boldsymbol{y}$,
- uppercase bold typeface letters for matrices and vectors of random variables, e.g. $\boldsymbol{X}$.

Notice we have overloaded our notation slightly for matrices and random variables. If it is not clear from the context which we mean, it will be stated explicitly.

We may also use tensor notation where convenient. For the elements of a matrix $\boldsymbol{X}$ we use x_{ij} where i refers to the row and j to the column. Similarly for the elements of a vector $\boldsymbol{y}$, we use y_i where i refers to the row. For the transpose we use T in the superscript so that $\boldsymbol{y}^T$ would be a row vector. We also use $\boldsymbol{x}_i$ to denote the ith row of the matrix $\boldsymbol{X}$.

Data

We will use $\boldsymbol{X}$ to denote the (non-sensitive) feature matrix and $\boldsymbol{Z}$ to denote the sensitive feature matrix (features like gender and race for example). We use $\boldsymbol{y}$ to denote the target variable vector (a column vector with n elements, each corresponding to an sample) and $\hat{\boldsymbol{y}}(\boldsymbol{X}, \boldsymbol{Z})$ to denote the predicted target variable output by our model, which is a function of the features. We shall use $\mathcal{X}$ and $\mathcal{Z}$ to denote the set of possible values our feature vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ can take respectively and $\mathcal{Y}$ to denote the set of all possible values our outcome y can take. When more appropriate we will use set notation to denote our set of data points which we write as $(\boldsymbol{X}, \boldsymbol{Z}, Y) = \{\boldsymbol{x}_i, \boldsymbol{z}_i, y_i\}_{i=1}^n$

We shall use the same notation for our target and model output for both discrete (classification) and continuous (regression) variables. In the case where the target variable is discrete and derived from a continuous classifier (that is, one where we find the classification by applying a threshold to an underlying score), we denote the underlying score as $\boldsymbol{p}(\boldsymbol{X}, \boldsymbol{Z})$ (if y is a binary variable) in which case we can write,

$$\hat{y}_i(\boldsymbol{X}, \boldsymbol{Z}) = H\left(p_i(\boldsymbol{X}, \boldsymbol{Z}) - \tau\right) \quad \forall\ i$$

where $H(x)$ is the Heaviside step function:

$$H(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{otherwise} \end{cases}$$

and τ is the threshold.

Note that if there was a single sensitive feature (rather than multiple) we would use $\boldsymbol{z}$ (rather than $\boldsymbol{Z}$) to denote it (since it would be a vector) and if the target was multi-class rather binary we would use $\boldsymbol{P}$ (rather than $\boldsymbol{p}$) for the score since we would need a score for each class. If we have n examples, m_x non-sensitive features, and m_z sensitive features then, $\boldsymbol{X}$ is an $n \times m_x$ matrix, $\boldsymbol{Z}$ is an $n \times m_z$ and $\boldsymbol{y}$ and $\boldsymbol{p}$ are vectors with n elements. If $\boldsymbol{y}$ was a multi-class target variable with c possible classes, then $\boldsymbol{P}$ would be an $n \times c$ matrix.

For binary sensitive and target variables $\boldsymbol{z}$ and $\boldsymbol{y}$, we will set the advantaged group and advantageous target class to have the value one, the disadvantaged group and disadvantageous target class will then take the value zero.

Random variables

Following the typographical conventions described above, we use $\boldsymbol{X}$, $\boldsymbol{Z}$ (or Z for a single sensitive feature), Y, $\hat{Y}$ and P (or $\boldsymbol{P}$ for multi-class), to denote the the random variables corresponding to our non-sensitive features, sensitive features, target variable, model predicted target variable and model probability function respectively.

Special values

We will occasionally use $+$ or $-$ in the subscript of a binary variable to respectively denote the advantaged or disadvantaged outcome (or class). For example,

- $Y = y_+$ is the advantageous outcome
- $Y = y_-$ is the disadvantageous outcome
- $Z = z_+$ is the advantaged (privileged) class
- $Z = z_-$ is the disadvantaged (unprivileged) class

For brevity and readability, we shall (on occasion) omit the random variable in the event descriptor of a probability term (if it is obvious which random variable we are referring to). For example, for a binary target variable we might write,

$$\mathbb{P}(Y = y_+) = \mathbb{P}(y_+).$$

Probability density functions

As a shorthand we will use f_X to denote the probability density function for the random variable X. Note then that for a discrete random variable X, we can write,

$$\mathbb{P}(X = x) = f_X(x),$$

while for a continuous random variable we have

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x.$$

Expectations

We denote the expectation as,

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) = \int_{x \in \mathcal{X}} g(x) f_X(x)\,\mathrm{d}x$$

where we take the expectation of a multivariate function, we will use a subscript to indicate the variable the expectation is taken over, e.g. $\mathbb{E}_X[g(X, Y)]$.

Naming conventions

- n for number of examples or data points
- d for differences
- r for rates

Part I Introduction

Welcome to Mitigating Bias in Machine Learning. If you've made it here chances are you've worked with models and have some awareness of the problem of biased machine learning algorithms. You might be a student with a foundational course in machine learning under your belt, or a Data Scientist or Machine Learning Engineer, concerned about the impact your models might have on the world.

In this book we are going to learn and analyse a whole host of techniques for measuring and mitigating bias in machine learning models. We're going to compare them, in order to understand their strengths and weaknesses. Mathematics is an important part of modelling, and we won't shy away from it. Where possible, we will aim to take a mathematically rigorous approach to answering questions.

Mathematics, just like code, can contain bugs. In this book, each has been used to verify the other. The analysis in this book, was completed using Python. The Jupyter Notebooks are available on GitHub, for those who would like to see/use them. That said, this book is intended to be self contained, and does not contain code. We will focuses on the concepts, rather than the implementation.

Mitigating Bias in Machine Learning is ultimately about fairness. The goal of this book is to understand how we, as practicing model developers, might build fairer predictive systems and avoid causing harm (sometimes that might mean not building something at all). There are many facets to solving a problem like this, not all of them involve equations and code. The first two chapters (part I) are dedicated to discussing these.

In a sense, over the course of the book, we will zoom in on the problem. In chapter 1, we'll look at it from a broader perspective (philosophical, political, legal, technical and social). In chapter two we take a more practical view on the problem of ethical development (how to build and organise the development of models, with a view to reducing ethical risk).

In part II we will talk about how we quantify different notions of fairness.

In part III, we will look at methods for mitigating bias through model interventions and anlyse their impact.

Let's get started.

Chapter 1

Context

This chapter at a glance

- Problems with machine learning in socipolitical domains
- Contrasting socio-political theories of fairness in decision systems
- The history, application and interpretation of anti-discrimination law
- Association paradoxes and the difficulty in identifying bias
- The different types of harm caused by biased systems

The goal of this chapter is to shed light on the problem of bias in in machine learning, from a variety of different perspectives, in an effort to provide context. The word *bias* can mean many things but in this book, we use it interchangably with the term *unfairness*. We'll talk about why later.

Perhaps the biggest challenge in developing *sociotechnical systems* is that they inevitably involve questions which are social, philosophical, political, and legal in nature; questions to which there is often no definitive answer but rather competing viewpoints and trade-offs to be made. As we'll see, this does not change when we attempt to quantify the problem. There are many measures of fairness that each defined in a different way, and they cannot all be satisfied simultaneously. The problem of bias in sociotechnical systems is very much an interdisciplinary one and, in this chapter, we discuss them as such. We will make connections between concepts and language from the various subjects over the course of this book.

In this chapter we shall discuss some philosophical theories of fairness in sociopolitical systems and consider how they might relate to model training and fairness criteria. We'll take a legal perspective, looking at anti-discrimination laws in the US as an example. We'll discuss the history, practical application of them and tensions that exist in their interpretation. Data can be misleading. Correlation does not imply causation and so domain knowledge in building machine learning systems is important. We will discuss the technical difficulty in identifying bias in static data through illustrative examples of Simpson's paradox. Finally, why is it important to consider the fairness of our models? What happens if we don't? We'll finish the chapter by discussing some of the different types of harm caused by biased machine learning systems - some of which are difficult if not impossible to quantify.

Let's start by describing the types of problems we are interested in.

1.1 Bias in Machine Learning

Machine learning can be described as the study of computer algorithms that improve with (or learn) experience. It can be broadly subdivided into the fields of supervised, unsupervised and reinforcement learning.

Supervised learning For supervised learning problems, the experience come in the form of labelled training data. Given a set of features X and labels (or targets) Y, we want to learn a function or mapping f, such that $Y = f(X)$, where f generalizes to previously unseen data.

Unsupervised learning For unsupervised learning problems there are no labels Y, only features X. Instead we are interested in looking for patterns and structure in the data. For example, we might want to subdivide the data into clusters of points with similar (previously unknown) characteristics or we might want to reduce the dimensionality of the data (to be able to visualize it or simply to make a supervised learning algorithm more efficient). In other words, we are looking for a new feature Y and the mapping f from X to Y.

Reinforcement learning Reinforcement learning is concerned with the problem of optimally navigating a state space to reach a goal state. The problem is framed as an agent that takes actions, which result in rewards (or penalties). The task is then to maximize the cumulative reward. As with unsupervised learning, the agent is not given a set of examples of optimal actions in various states, but rather must learn them through trial and error. A key aspect of reinforcement learning is the existence of a trade-off between exploration (searching unexplored territory in the hope of finding a better choice) and exploitation (exploiting what has been learned so far).

In this we will focus on the first two categories (essentially algorithms that capture and or exploit patterns in data), primarily because these are the fields in which problems related to bias in machine learning are most pertinent (automation and prediction). As one would expect then, these are also the areas in which many of the technical developments in measuring and mitigating bias have been concentrated.

The idea that the kinds of technologies described above are *learning* is an interesting one. The analogy is clear, learning by example is certainly a way to learn. In less modern disciplines one might simply think of *training* a model as; solving an equation, interpolating data, or optimising model parameters. So where does the terminology come from? The term *machine learning* was coined by Arthur Samuel in the 1950's when, at IBM, he developed an algorithm capable of playing draughts (checkers). By the mid 70's his algorithm was competitive at amateur level. Though it was not called reinforcement learning at the time, the algorithm was one of the earliest implementations of such ideas. Samuel used the term *rote learning* to describe a memorisation technique he implemented where the machine remembered all the states it had visited and the corresponding reward function, in order to extend the search tree.

1.1.1 What is a Model?

Underlying every machine learning algorithm is a model (often several of them) and these have been around for millennia. Based on the discovery of palaeolithic tally sticks (animal bones carved with notches) it's believed that humans have kept numerical records for over 40,000 years. The earliest mathematical models (from around 4,000 BC) were geometric and used to advance the fields of astronomy and architecture. By 2,000 BC, mathematical models were being used in an algorithmic manner to solve specific problems by at least three civilizations (Babylon, Egypt and India).

A model is a simplified representation of some real world phenomena. It is an expression of the relationship between things; a function or mapping which, given a set of input variables (features), returns a decision or prediction (target). A model can be determined with the help of data, but it need not be. It can simply express an opinion as to how things should be related.

If we have a model which represents a theoretical understanding of the world (under a series of simplifying assumptions) we can test it by measuring and comparing the results to reality. Based on the results we can assess how accurate our understanding of the world was and update our model accordingly. In this way, making simplifying assumptions can be a means to iteratively improve our understanding of the world. Models play an incredibly important role in the pursuit of knowledge. They have provided a mechanism to understand the world around us, and explain why things behave as they do; to prove that the earth could not be flat, explain why the stars move and shift in brightness as they do or, (somewhat) more recently in

the case of my PhD, explain why supersonic flows behave uncharacteristically, when a shock wave encounters a vortex.

As the use of models has been adopted by industry, increasingly their purpose has been geared towards prediction and automation, as a way to monetize that knowledge. But the pursuit of profit inevitably creates conflicts of interests. If your goal is to learn more, finding out where your theory is wrong and fixing it is a core part of the game. In business, where the goal is to maximise profit and minimise cost, it need not be.

I recall a joke I heard at school describing how one could tell which field of science an experiment belonged to. If it changes colour, it's biology; if it explodes, it's chemistry and if it doesn't work, it's physics. Models of real world phenomena fail. They are, by their very nature, a reductive representation of an infinitely more complex real world system. Obtaining adequately rich and relevant data is a major limitation of machine learning models and yet they are increasingly being applied to problems where that kind of data simply doesn't exist.

1.1.2 Sociotechnical systems

We use the term sociotechnical systems to describe systems that involve algorithms that manage people. They make decisions for us, determine what we see, direct us and more. But managing large numbers of people inevitably exerts a level of authority and control. For many people that includes deciding the hours they work. Recent years have seen the adoption of just-in-time scheduling algorithms by large retailers to manage staffing needs. To predict footfall, the algorithms take into account everything from weather patterns to sports events. The cost of this efficiency is passed onto employees. The number of hours allocated are optimised to fall short of qualifying for costly health insurance. Employees are subjected to haphazard schedules with little notice that prevent them from being able to prioritise anything other than work and eliminating the possibility of any opportunity that might enable them to advance beyond the low-wage work pool.

Progress in the field of deep learning combined with increased availability and decreased cost of computational resources has led to an explosion in data and model use. Automation seemingly offers a path to making our lives easier, improving the efficiency and efficacy of the many industries we transact with day to day; but there are growing and legitimate concerns over how the benefit (and cost) of these efficiencies are distributed. Machine learning is already being used to automate decisions in just about every aspect of modern life; deciding which adverts to show to whom, deciding which transactions might be fraud when we shop, deciding who is able to access to financial services such as loans and credit cards, determining our treatment when sick, filtering candidates for education and employment opportunities, in detesrmining which neighbourhoods to police and even in the criminal justice system to decide what level bail should be set at, or the length of a given sentence. At almost every major life event, going to university, getting a job, buying a house, getting sick, decisions are being made by machines.

1.1.3 What Kind of Bias?

Bias can have numerous different meanings depending on the context even within the same subject. Let's talk about the kinds of biases that are relevant here. The word bias is used to describe systematic errors in variable estimation (predictions) from data. If the goal is to create systems that work similarly well for all types of people, we certainly want to avoid these. In a social context, bias is spoken of as prejudice or discrimination based on characteristics that we as a society deem to be unacceptable or unfair. Systemic discrimination exists and it shows up in the data in numerous different ways, in historical decisions, who is in it (and who is not). Bias need not be conscious, in reality it starts at the very inception of technology, in deciding which problems are worth solving in the first place. Bias exists in how we measure the cost and benefit of new technologies. For sociotechnical systems, these are all deeply intertwined.

Ultimately, mitigating bias in our models is about fairness and in this book we shall use the terms interchangeably. Machine learning models are capable of not only of proliferating existing societal biases, but amplifying them and are easily deployed at scale. But how do we even define fairness? And from whose perspective do we mean fair? The law can provide *some* context here. Laws, in many cases, define *protected*

characteristics and domains (we'll talk more about these later). We can potentially use these as a guide and we certainly have a responsibility to be law abiding citizens. A common approach historically has been to ignore protected characteristics. There's a few reasons for this. One is avoiding legal liability (we'll talk more about the law in the next section). Another is the false belief that, an algorithm cannot discriminate based on features not included in the data. This assumption is is easy to disprove with a counter example. A reasonably fool-proof way to systematically discriminate by race (without knowing it, is to discriminate by location/residence, that is, another variable that's strongly correlated and serves as a proxy. The legality of this practice depends on the domain. In truth, you don't need a feature, or a proxy, to discriminate based on it, you just need enough data, to be able to predict it. If it is predictable, the information there and the algorithm is likely using it.

1.2 A Philosophical Perspective

Cathy O'Neil describes models as "opinions embedded in code". Developing models is not an objective scientific process, it involves making a series subjective choices. One of the most fundamental ways in which we impose our opinion on a machine learning model is in deciding how we measure success. Which model is the *right* one so to speak? Let's look at the process of training a model. We start with some parametric representation (a family of models), which you hope is sufficiently in complex to be able to reflect the relationships between the variables in the data. The goal in training is to determine which model (in our chosen family) is *best*, the *best* model being the one that minimises the expected error (assuming our data is representive of some objective ground truth).

If all we are interested in doing is understanding how well our model represents the world, this approach might be valid. We don't concern ourselves with the direction of the error in training. Overestimation is as bad as underestimation, false positives are as bad as false negatives. In this case, understanding errors are a means to refine our model if it does not explain the particular phenomenon we are studying. But for sociotechnical systems, we're not simply trying to understand the world, we are making decisions off the back of it; decisions which result in a benefit or harm to those subjected to them. The very purpose of codifying a decision policy is often to cheaply deploy it at scale. The more people it processes, the more financial value there is in codifying the decision process. Another, way to look such models instead then, is as a system for distributing benefits (or conversely harms) among a population. In this section we briefly discuss some more philosphical theories relevant to these types of problems. We start with utilitarianism which is perhaps the easiest theory to draw parallels with when it comes to training a model.

1.2.1 Utilitarianism

Utilitarianism provides a framework for moral reasoning in decision making. Under this framework, the correct course of action, when faced with a dilemma, is the one that maximises the benefit for the greatest number of people. The doctrine demands that the benefits to all people are are counted equally. Variations of the theory have evolved over the years. Some differ in their notion of how benefits are understood. Others distinguish between the quality of various kinds of benefit. In a business context, one might consider it as financial benefit (and cost). Although, this in itself depends on one's perspective. Some doctrines advocate that the impact of the action in isolation should be considered, while others ask what the impact would be if everyone in the population took the same actions.

There are some practical problems with utilitarianism as the sole guiding principle for decision making. How do we measure benefit? How do we navigate the complexities of placing a value on immeasurable and vastly different consequences? What is a life, time, money or particular emotion worth and how do we compare and aggregate them? How can one even be certain of the consequences? Longer term consequences are hard if not impossible to predict. Perhaps the most significant flaw in utilitarianism for moral reasoning, is the omission of justice as a consideration.

Utilitarian reasoning judges actions based solely on consequences, and aggregates them over a population. So, if an action that unjustly harms a minority group happens to be the one that maximises the aggregate

benefit over a population, it is nevertheless the correct action to take. Under utilitarianism, theft or infidelity might be morally justified, if those it would harm are none the wiser. Or punishing an innocent person for a crime they did not commit could be justified, if it served to quell unrest among a population. For this reason it is widely accepted that utilitarianism is insufficient as a framework for decision making.

Utilitarianism is a flavour of consequentialism, a branch of ethical theory that holds that consequences are the yard stick against which we must judge the morality of our actions. In contrast deontological ethics judges the morality of actions against a set of rules that define our duties or obligations towards others. Here it is not the consequences of our actions that matter but rather intent.

The conception of utilitarianism is attributed to British philosopher Jeremy Bentham who authored the first major book on the topic *An Introduction to the Principles of Morals and Legislation* in 1780. In it Bentham argues that, it is the pursuit of pleasure and avoidance of pain alone that motivate individuals to act. Given this he saw utilitarianism as a principle by which to govern. Broadly speaking, the role of government, in his view, was to assign rewards or punishments to actions, in proportion to the happiness or suffering they produced among the governed. At the time, the idea that the well-being of all people should be counted equally, and that that morality of actions should be judged accordingly was revolutionary. Bentham was a progressive in his time, he advocated for women's rights (to vote, hold office and divorce), decriminalisation of homosexual acts, prison reform and the abolition of slavery and more. He argued many of his beliefs as a simple economic calculation of how much happiness they would produce. Importantly, he didn't claim that all people were equal, but rather only that their happiness mattered equally.

Times have changed. Over the last century, as civil rights have advanced, the weaknesses of utilitarianism in practice have been exposed time and time again. Utilitarian reasoning has increasingly been seen as hindering social progress, rather than advancing it. For example, utilitarian arguments were used by Whites in apartheid South Africa, who claimed that all South Africans were better-off under White rule, and that a mixed government would lead to social decline as it had in other African nations. Utilitarian reasoning has been used widely by capitalist nations in the form of trickle-down economics. The theory being that the benefits of tax-breaks for the wealthy drive economic growth and 'trickle-down' to the rest of the population. But evidence suggests that trickle-down economic policies in more recent decades have done more damage than good, increasing national debt and fueling income inequality. Utilitarian principles have also been tested in the debate over torture, capturing a rather callous conviction, one where the 'means justify the ends'.

Historian and author, Yuval Noah Harari has eloquently abstracted this problem. He argues that historically, decentralization of power and efficiency have aligned; so much so, that many of us cannot think of democracy as being capable of failing, to more totalitarian regimes. But in this new age, data is power. We can train enormous models, that require vast amounts of data, to process people en masse, organise and sort them. And importantly, one does not have to have a perfect system in order to have an impact because of the scale on which they can be deployed. The question Yuval poses is, *Might the benefits of centralised data, offer a great enough advantage, to tip the balance of efficiency, in favour of more centralised models of power?*

1.2.2 Justice as Fairness

In his theory Justice As Fairness[15], John Rawls takes a different approach. He describes an idealised democratic framework, based on liberal principles and explains how unified laws can be applied (in a free society made up of people with disparate world views) to create a stable sociopolitical system. One where citizens would not only freely co-operate, but further advocate. He described a political conception of justice which would:

1. grant all citizens a set of basic rights and liberties
2. give special priority to the aforementioned rights and liberties over demands to further the general good, e.g. increasing the national wealth
3. assure all citizens sufficient means to make use of their freedoms.

The special priority given to the basic rights and liberties in the political conception of justice contrasts with a utilitarian doctrine. Here constraints are placed on how benefits can be distributed among the population and a strategy for determining some minimum.

Principles of Justice as Fairness

1. **Liberty principle:** Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which is compatible with the same scheme of liberties for all;
2. **Equality principle:** Social and economic inequalities are to satisfy two conditions:
   (a) **Fair equality of opportunity:** The offices and positions to which they are attached are open to all, under conditions of fair equality of opportunity;
   (b) **Difference (maximin) principle** They must be of the greatest benefit to the least-advantaged members of society.

The principles of Justice as Fairness are ordered by priority so that fulfilment of the liberty principle takes precedence over the equality principles and fair equality of opportunity takes precedence over the difference principle.

The first principle grants basic rights and liberties to all citizens which are prioritised above all else and cannot be traded for other societal benefits. It's worth spending a moment thinking about what those rights and liberties look like. They are the the basic needs that are important for people to be free, to have choices and the means to pursue their aspirations. Today many of what Rawls considered to be basic rights and liberties are allocated algorithmically; education, employment, housing, healthcare, consistent treatment under the law to name a few.

The second principle requires positions to be allocated meritocratically, with all similarly talented (with respect to the skills and competencies required for the position) individuals having the same chance of attaining such positions i.e. that allocation of such positions should be independent of social class or background. We will return to the concept of *equality of opportunity* in chapter 3 when discussing *Group Fairness*.

The third principle acts to prevent redistribution of social and economic currency from the rich to the poor by requiring that inequalities are of maximal benefit to the least advantaged in a society, also described as the maximin principle. In this principle, Rawls does not take the simplistic view that inequality and fairness are mutually exclusive but rather concisely articulates when the existence of inequality becomes unfair. We shall return to maximin principle when we look at the use of *inequality indices* to measure algorithmic unfairness in a later chapter.

1.3 A Legal Perspective

It's important to remember that anti-discrimination laws are the result of long-standing and systemic discrimination against oppressed people. Their existance is a product of history; subjugation, genocide, civil war, mass displacement of entire communities, racial hierarchies and segregation, supremist policies (exclusive access to publicly funded initiatives), voter suppression and more. The law provides an important historical record of what we as a society deem fair and unfair, but without history there is no context. The law does not define the bechmark for fairness. Laws vary by durisdiction and change over time and in particular they often do not adequately recognise or address issues related to discrimination that are known and accepted by the sciences (social, mathematical, medical,...).

In this section we'll look at the history, practical application and interpretation of the law in the US (acknowledging the narrow scope of our discussion) Finally, we'll take a brief look at what might be on the legislative horizon for predictive algorithms, based on more recent global developments.

1.3.1 A Brief History of Anti-discrimination Law in the US

Anti-discrimination laws in the US rest on the 14th amendment to the constitution which grants citizens *equal protections of the law.* Class action law suit Brown v Board (of Education of Topeka, Kansas) was

a landmark case which in 1954, legally ended racial segregation in the US. Justices ruled unanimously that racial segregation of children in public schools was unconstitutional, establishing the precedent that "separate-but-equal" was, in fact, not equal at all. Though Brown v Board did not end segregation in practice, resistance to it in the south fuelled the civil rights movement. In the years that followed the NAACP (National Association for the Advancement of Coloured People) challenged segregation laws. In 1955, Rosa parks refusing to give up her seat on a bus in Montgomery (Alabama) led to sit ins and boycotts, many of them led by Martin Luther King Jr. The resulting Civil rights act of 1964 eventually brought an end to "Jim Crow" laws which barred Blacks from sharing buses, schools and other public facilities with Whites.

After the violent attack by Alabama state troopers on participants of a peaceful march from Selma to Montgomery was televised, The Voting Rights Act of 1965 was passed. It overcame many barriers (including literacy tests), at state and local level, used to prevent Black people from voting. Before this incidents of voting officials asking Black voters to "recite the entire Constitution or explain the most complex provisions of state laws"[10] in the south were common place.

In the years following the second world war, there were many attempts to pass an Equal Pay Act. Initial efforts were led by unions who feared men's salaries would be undercut by women who were paid less for doing their jobs during the war. By 1960, women made up 37% of the work force but earned on average 59 cents for each dollar earned by men. The Equal Pay Act was eventually passed in 1963 in a bill which endorsed "equal pay for equal work". Laws for gender equality were strengthened the following year by the Civil Rights Act of 1964.

Throughout the 1800's the American federal government displaced Native American communities to facilitate White settlement. In 1830 the Indian Removal Act was passed in order to relocate hundreds of thousands of Native Americans. Over the following two decades, thousands of those forced to march hundreds of miles west on the perilous "Trail of Tears" died. By the middle on the century, the term "manifest destiny" was popularlised to describe the belief that White settlement in North America was ordained by God. In 1887, the Dawes Act laid the groundwork for the seizing and redistribution of reservation lands from Native to White Americans. Between 1945 and 1968 the federal government terminated recognition of more than 100 tribal nations placing them under state jurisdiction. Once again Native Americans were relocated, this time from reservations to urban centres.

In addition to displacing people of colour, the federal government also enacted policies that reduced barriers to home ownership almost exclusively for White citizens - subsidizing the development of prosperous "White Caucasian" tenant/owner only suburbs, guaranteeing mortgages and enabling access to job opportunities by building highway systems for White commuters, often through communities of colour, simultaneously devaluing the properties in them. Even government initiatives aimed at helping veterans of World War II to obtain home loans accommodated Jim Crow laws allowing exclusion of Black people. In the wake of the Vietnam war, just days after the assassination of Martin Luther King J, the Fair Housing Act of 1968 was passed, prohibiting discrimination concerning the sale, rental and financing of housing based on race, religion, national origin or sex.

The Civil Rights Act of 1964 acted as a catalyst for many other civil rights movements, including those protecting people with disabilities. The Rehabilitation Act (1973) removed architectural, structural and transportation barriers and set up affirmative action programs. The Individuals with Disabilities Education Act (IDEA 1975) required free, appropriate public education in the least restrictive environment possible for children with disabilities. The Air Carrier Access Act (1988) which prohibited discrimination on the basis of disability in air travel and ensured equal access to air transportation services. The Fair Housing Amendments Act (1988) prohibited discrimination in housing against people with disabilities.

Title IX of the education amendments of 1972 prohibits federally funded educational institutions from discriminating against students or employees based on sex. The law ensured that schools (elementary to university level) that were recipients of federal funding (nearly all schools) provided fair and equal treatment of the sexes in all areas, including athletics. Before this few opportunities existed for female athletes. The National Collegiate Athletic Association (NCAA) offered no athletic scholarships for women and held no championships for women's teams. Since then the number of female college athletes has grown five fold. The amendment is credited with decreasing dropout rates and increasing the numbers of women gaining college

degrees.

The Equal Credit Opportunity Act was passed in 1974 when discrimination against women applying for credit in the US was rife. It was common practice for mortgage lenders to discount incomes of women that were of 'child bearing' age or simply deny credit to them. Two years later the law was amended to prohibit lending discrimination based on race, color, religion, national origin, age, the receipt of public assistance income, or exercising one's rights under consumer protection laws.

In 1978, congress passed the Pregnancy Discrimination Act in response to two Supreme Court cases that ruled that excluding pregnancy related disabilities from disability benefit coverage was not gender based discrimination, and did not violate the equal protection clause.

Table 1.1 shows a (far from exhaustive) summary of regulated domains with corresponding US legislation. Note that legislation in these domains extend to marketing and advertising not just the final decision. Table 1.2 provides a list of protected characteristics under US federal law with corresponding legislation (again not exhaustive).

Table 1.1: Regulated domains in the private sector under US federal law.

Domain	Legislation
Finance	Equal Credit Opportunity Act
Education	Civil Rights Act (1964) Education Amendment (1972) IDEA (1975)
Employment	Equal Pay Act(1963) Civil Rights Act (1964)
Housing	Fair Housing Act (1968) Fair Housing Amendments Act (1988)
Transport	Urban Mass Transit Act (1970) Rehabilitation Act (1973) Air Carrier Access Act (1988)
Public accommodation[a]	Civil Rights Act (1964)

[a]Prevents refusal of customers.

Table 1.2: Protected characteristics under US Federal Law.

Protected Characteristic	Legislation
Race	Civil Rights Act (1964)
Sex	Equal Pay Act (1963) Civil Rights Act (1964) Pregnancy Discrimination Act (1978)
Religion	Civil Rights Act (1964)
National Origin	Civil Rights Act (1964)
Citizenship	Immigration Reform & Control Act
Age	Age Discrimination in Employment Act (1967)
Familial status	Civil Rights Act (1968)
Disability status	Rehabilitation Act of 1973 American with Disabilities Act of 1990
Veteran status	Veterans' Readjustment Assistance Act 1974 Uniformed Services Employment & Reemployment Rights Act
Genetic Information	Civil Rights Act(1964)

1.3.2 Application and Interpretation of the Law

To get an idea of how anti-discrimination laws are be applied in practice and how they might translate to algorithmic decision making, we look at Title VII of the Civil rights act of 1964 in the context of employment discrimination[3]. Legal liability for discrimination against protected classes can be established as disparate treatment and/or disparate impact. Disparate treatment (also described as direct discrimination in Europe) refers to both differing treatment of individuals based on protected characteristics, and intent to discriminate. Disparate impact (or indirect discrimination in Europe) does not consider intent but addresses policies and practices that disproportionately impact protected classes.

Disparate Treatment

Disparate treatment effectively prohibits rational prejudice (backed by data showing the protected feature to be correlated) as well as denial of opportunities based on protected characteristics. For an algorithm, it effectively prevents the use of protected characteristics as inputs. It's noteworthy that in the case of disparate treatment, the actual impact of using the protected features on the outcome is irrelevant; so even if a company could show that the target variable produced by their model had zero correlation with the protected characteristic, the company would still be liable for disparate treatment. This fact is somewhat bizarre given that not using the protected feature in the algorithm provides no guarantee that the algorithm is not biased in relation to it. Indeed an organisation could very well use their data to predict the protected characteristic.

In an effort to avoid disparate treatment liability, many organisations do not even collect data relating to protected characteristics, leaving them unable to accurately measure, let alone address, bias in their algorithms, even if they might want to[1]. In summary, disparate treatment as applied today does not resolve the problem of unconscious discrimination against disadvantaged classes through their use of machine learning algorithms. Further it acts as a deterrent to ethically minded companies that might want to measure the biases in their algorithms.

Disparate treatment

Suppose a company predicts the sensitive feature and uses this as an input to its model. Should this be considered disparate treatment?

What about the case where the employer implements an algorithm, finds out that it has a disparate impact, and uses it anyway? Doesn't that become disparate treatment? No it doesn't and in fact, somewhat surprisingly, deciding not to apply it upon noting the disparate impact could result in a disparate treatment claim in the opposite direction[21]. We'll return to this later. Okay, so what about disparate impact?

Disparate Impact

In order to establish a violation, it is not enough to simply show that there is a disparate impact, but it must also be shown either that there is no business justification for it, or if there is, that the employer refuses to use another, less discriminatory, means of achieving the desired result. So how much of an impact is enough to warrant a disparate impact claim? There are no rules here only guidelines. The Uniform Guidelines on Employment selection procedures from the Equal Employment Opportunity Commission (EEOC) provides a guideline that if the selection rate from one protected group is less than four fifths of that from another, it will generally be regarded as evidence of adverse impact, though it also states that the threshold would depend on the circumstances.

Assuming the disparate impact is demonstrated, the issue becomes proving business justification. The requirement for business justification has softened in favour of the employer over the years; treated as

[1] In fact, I met a data scientist at a conference, who was working for a financial institution, that said her team was trying to predict sensitive features such as race and gender in order to measure bias in their algorithms.

"business necessity"[18] earlier on and later interpreted as "business justification"[19]. Today, it's generally accepted that business justification lies somewhere between the extremes of "job-relatedness" and "business necessity". As a concrete example of disparate impact and taking the extreme of job-relatedness - the EEOC along with several federal courts have determined that discrimination on the sole basis of a criminal record to be a violation under disparate impact unless the particular conviction is related to the role, because Non-White applicants are more likely to have a criminal conviction.

For a machine learning algorithm, business justification boils down to the question of job-relatedness of the target variable. If the target variable is improperly chosen, a disparate impact violation can be established. In practice however the courts will accept most plausible explanations of job-relatedness since not accepting it would set a precedent that it is determined discriminatory. Assuming the target variable to be proven job-related then, there is no requirement to validate the model's ability to predict said trait, only a guideline which sets a low bar (a statistical significance test showing that the target variable correlates with the trait) and which the court is free to ignore.

Assuming business justification is proven by the employer, the final burden then falls on the plaintiff to show that the employer refused to use a less discriminatory "alternative employment practice". If the less discriminatory alternative would incur additional cost (as is likely) would this be considered refusing? Likely not.

While on the surface, disparate impact might seem like a solution, the current framework of a weak business justification (in terms of a plausible target variable) and the employer refusing an alternative employment practice with no requirement to validate the model offers little resolve. Clearly there is need for reform.

Anti-classification versus Anti-subordination

Just as the meaning of fairness is subjective so is the interpretation of anti-discrimination laws. At one extreme, anti-classification holds the weaker interpretation, that the law is intended to prevent classification of people based on protected characteristics. At the other extreme, anti-subordination defines the stronger stance, that anti-discrimination laws exist to prevent social hierarchies, class or caste systems based on protected features and, that it should actively work to eliminate them where they exist. An important ideological difference between the two schools of thought is in the application of positive discrimination policies. Under anti-subordination principles, one might advocate for affirmative action as a means to bridge gaps in access to employment, housing, education and other such pursuits, that are a direct result of historical systemic discrimination against particular groups. A strict interpretation of the anti-classification principle would prohibit such actions. Both anti-classification and anti-subordination ideologies have been argued and upheld in landmark cases.

In 2003, the Supreme Court held that a student admissions process that favours "under-represented minority groups" does not violate the Fourteenth Amendment[20], provided it evaluated applicants holistically at an individual level. The same year, the New Haven Fire Department administered a two part test in order to fill 15 openings. Examinations were governed in part by the City of New Haven. Under the city charter, civil service positions must be filled by one of the top three scoring individuals. 118 (White, Black and Hispanic) fire fighters took the exams. Of the resulting 19 candidates who scored highest on the tests and could the considered for the positions, none were Black. After heated public debate and under threat of legal action either way, the city threw out the test results. This action was later determined to be a disparate treatment violation. In 2009, the court ruled that disparate treatment could not be used to avoid disparate impact without sufficient evidence of liability of the latter[21]. This landmark case was the first example of conflict between the two doctrines of disparate impact and disparate treatment or anti-classification and anti-subordination.

Disparate treatment seems to align well with anti-classification principles, seeking to prevent intentional discrimination based on protected characteristics. In the case of disparate impact, things are less clear. Is it a secondary 'line of defence' designed to weed out well masked intentional discrimination? Or is its intention to address inequity that exists as a direct result of historical injustice? One can draw parallels here with the 'business necessity' versus 'business justification' requirements discussed earlier.

1.3.3 Future Legislation

In May 2018, the European Union (EU) brought into action the General Data Protection (GDPR) a legal framework around the protection of personal data of EU citizens. The framework is divided into binding and non-binding recitals. The regulation sets provisions for processing of data in relation to decision making, described as 'profiling' under recital 71[8]. Though currently non-binding, it provides an indication of what's to come. The recital talks specifically about having the right not to be subject to decisions based solely on automated processing. It specifically talks about credit applications, e-recruiting and any system which analyses or predicts aspects of a persons performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements. The recital also talks about requirements around using "appropriate mathematical or statistical procedures" to prevent "discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation". More recently in 2021, the EU has proposed taking a risk based approach to the question of which technologies should be regulated, dividing it into four categories. Unacceptable risk, high risk, limited risk, minimal risk[7]. While things may change as the proposed law is debated but once agreed, it's not unlikely that it will serve as a prototype for legislation in the U.S. (and other countries around the world), as did GDPR.

In April 2019, the Algorithmic Accountability Act was proposed to the US Senate. The bill requires specified commercial entities to conduct impact assessments of automated decision systems and specifically states that assessments must include evaluations and risk assessment in relation to "accuracy, fairness, bias, discrimination, privacy, and security" not just for the model output but for the training data. The bill has cosponsors in 22 states and has been referred to the Committee on Commerce, Science, and Transportation for review. These examples are clear indications that the issues of fairness and bias in automated decision making systems are on the radar of regulators.

1.4 A Technical Perspective

The problem of distinguishing correlation from causation is an important one in identifying bias. Using illustrative examples of Simpson's paradox, we demonstrate the danger of assuming causal relationships based on observational data.

1.4.1 Simpson's Paradox

In 1973, University of California, Berkeley received approximately 15,000 applications for the fall quarter[4]. At the time it was made up of 101 departments. 12,763 applications reached the decision stage. Of these 8442 were male and 4321 were female. The acceptance rates for the applicants were 44% and 35% respectively (see Table 1.3).

Table 1.3: Graduate admissions data from Berkeley (fall 1973).

Gender	Admitted	Rejected	Total	Acceptance Rate
Male	3738	4704	8442	44.3%
Female	1494	2827	4321	34.6%
Aggregate	5232	7531	12763	41.0%

With a whopping 10% difference in acceptance rates, it seems a likely case of discrimination against women. Indeed, a χ^2 hypothesis test for independence between the variables (gender and application acceptance) reveals that the probability of observing such a result or worse, assuming they are independent, is 6×10^{-26}. A strong indication that they are not independent and therefore evidence of bias in favour of male applicants. Since admissions are determined by the individual departments, it's worth trying to understand which departments might be responsible. We focus on the data for the six largest departments, shown in

Table 1.4. Here again we see a similar pattern. There appears to be bias in favour of male applicants, and a χ^2 test shows that the probability of seeing this result under the assumption of independence is 1×10^{-21}. It looks like we have quickly narrowed down our search.

Table 1.4: Graduate admissions data from Berkeley (fall 1973) for the six largest departments.

Gender	Admitted	Rejected	Total	Acceptance Rate
Male	1198	1493	2691	44.5%
Female	557	1278	1835	30.4%
Aggregate	1755	2771	4526	38.8%

Figure 1.1 shows the acceptance rates for each department by gender, in decreasing order of acceptance rates. Performing χ^2 tests for each department reveals the only department where there is strong evidence of bias is A, but the bias is in favour of female applicants. The probability of observing the data for department A, under the assumption of independence, is 5×10^{-5}. So what's going on? Figure 1.2 shows the application

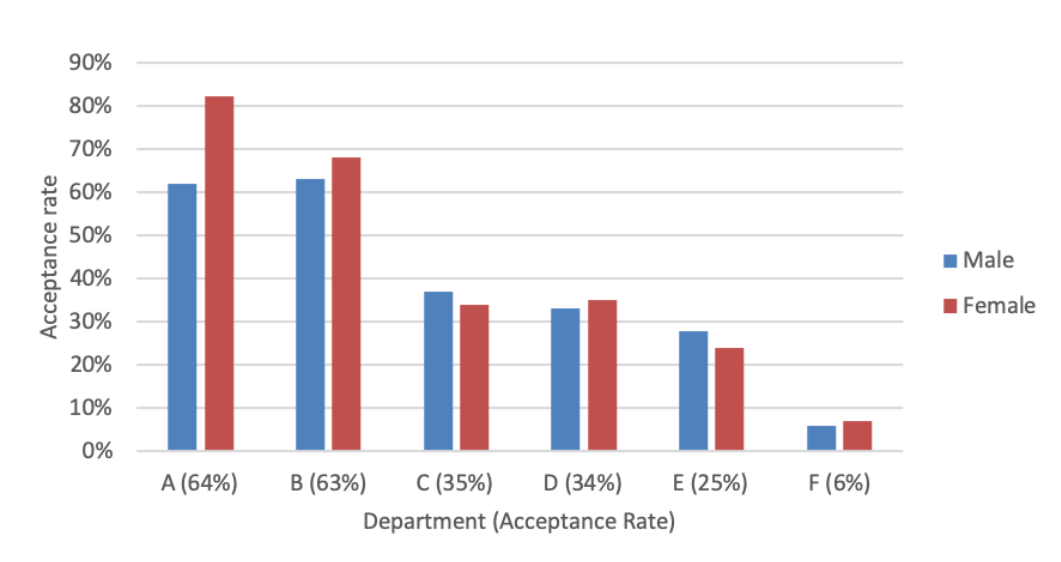


Figure 1.1: Acceptance rate distributions by department for male and female applicants.

distributions for male and female applicants for each of the six departments. From the plots we are able to see a pattern. Female applicants are more often applying for departments with a lower acceptance rate. In other words a larger proportion of the women are being filtered out overall, simply because they are applying to departments that are harder to get into.

This is a classic example of Simpson's Paradox (also known as the reversal paradox and Yule-Simpson effect). We have an observable relationship between two categorical variables (in this case gender and acceptance) which disappears or reverses, after controlling for one or more other variables (in this case department). Simpson's Paradox is a special case of so called association paradoxes (where the variables are categorical, and the relationship changes qualitatively), but the same rules also apply to continuous variables. The *marginal* (unconditional) measure of association (e.g. correlation) between two variables need not be bounded by the *partial* (conditional) measures of association (after controlling for one or more variables). Although Edward Hugh Simpson famously wrote about the paradox in 1951, it was not discovered by him. In fact, it was reported by George Udny Yule as early as 1903. The association paradox for continuous variables was demonstrated by Karl Pearson in 1899.

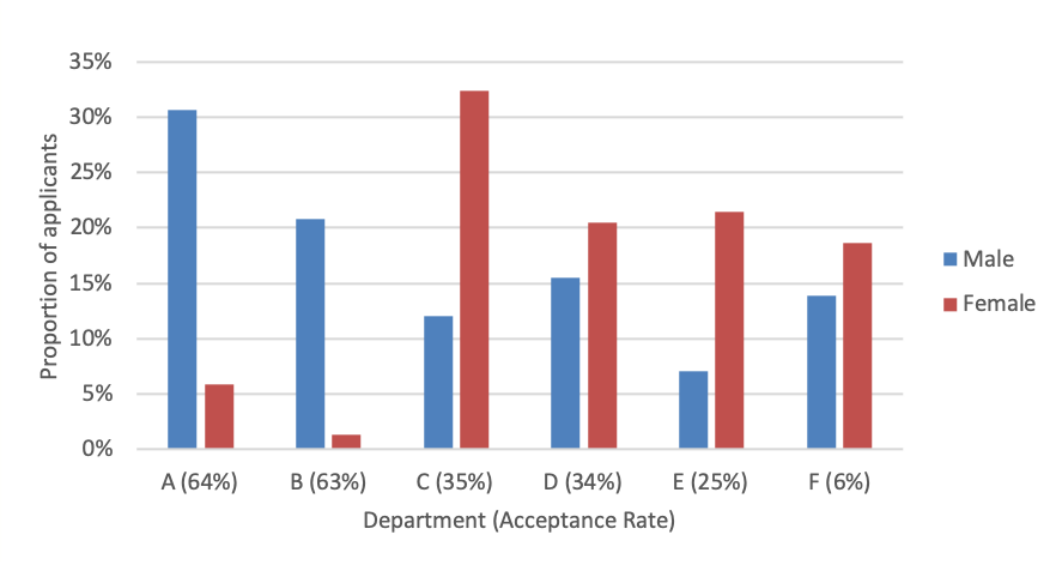


Figure 1.2: Application distributions by department for male and female applicants.

Let's discuss another quick example. A 1996 follow-up study on the effects of smoking recorded the mortality rate for the participants over a 20 year period. They found higher mortality rates among the non-smokers, 31.4% compared to 23.9% which, in itself, might imply a considerable protective affect from smoking. Clearly there's something fishy going on. Disaggregating the data by age group showed that the mortality rates were higher for smokers in all but one of them. Looking at the age distribution of the populations of smokers and non-smokers, it's apparent that the age distribution of the non-smoking group is more positively skewed, and so they are older on average. This concords with the rationale that non-smokers live longer - hence the difference in age distributions of the participants.

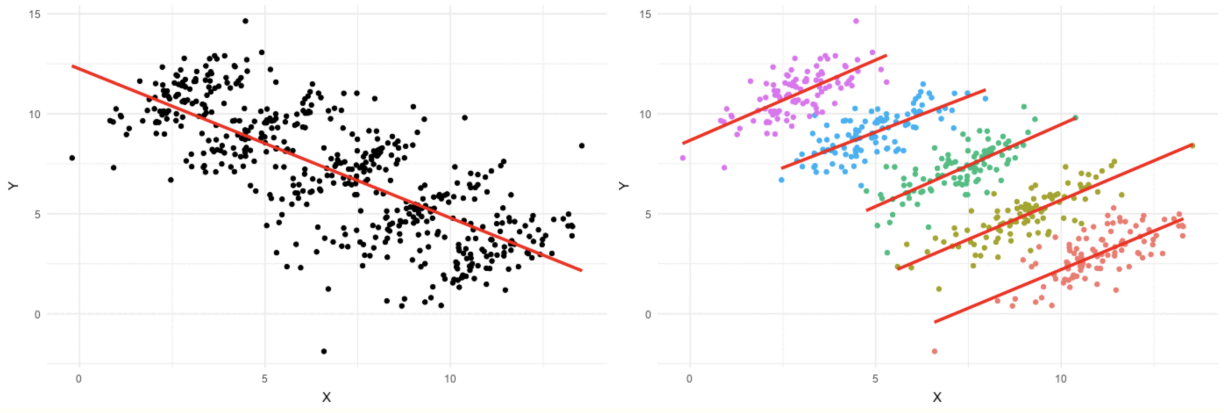


Figure 1.3: Visualisation of Simpsons Paradox. Wikipedia.

1.4.2 Causality

In both the above examples, it appears that the salient information is found in the disaggregated data (we'll come back to this later). In both cases it is the disaggregated data that enables us to understand the *true nature* of the relationship between the variables of interest. As we shall see in this section, this need not be

the case. To show this, we discuss two examples. In each case, the data is identical but the variables is not. The examples are those Simpson gave in his original 1951 paper[17].

Suppose we have three binary variables, A, B and C, and we are interested in understanding the relationship between A and B given a set of 52 data points. A summary of the data showing the association between variables A and B are shown in Table 1.5, first for all the data points and then stratified (separated) by the value of C (note the first table is the sum of the latter two). The first table indicates that A and B are unconditionally independent (since changing the value of one variable does not change the distribution of the other). The next two tables suggest A and B are conditionally dependent given C.

Table 1.5: Data summary showing the association between variables A and B, first for all the data points and then stratified by the value of C.

Black?/ Died?	Plain?/ Treated?	
	$A = 1$	$A = 0$
$B = 1$	20	6
$B = 0$	20	6
$\mathbb{P}(B\|A)$	50%	50%

	Stained? / Male?			
	$C = 1$		$C = 0$	
Black?/ Died?	Plain?/ Treated?			
	$A = 1$	$A = 0$	$A = 1$	$A = 0$
$B = 1$	5	3	15	3
$B = 0$	8	4	12	2
$\mathbb{P}(B\|A, C)$	38%	43%	56%	60%

[a]Each cell of the table shows the number of examples in the dataset satisfying the conditions given in the corresponding row and column headers.

Question: Which distribution gives us the most relevant understanding of the association between A and B, the marginal (i.e. unconditional) $\mathbb{P}(A, B)$ or conditional distribution $\mathbb{P}(A, B|C)$? To show that causal relationships matter, we consider two different examples.

Example a) Pack of Cards (Colliding Variable)

Suppose the population is a pack of cards. It so happens that baby Milen has been messing about with the cards and made some dirty in the process. Let's summarise our variables,

- A tells us the character of the card, either plain ($A = 1$) or royal (King, Queen, Jack; $A = 0$).
- B tells us the colour of the card, either black ($B = 1$) or red ($B = 0$).
- C tells us if the card is dirty ($C = 1$) or clean ($C = 0$).

In this case, the aggregated data showing $\mathbb{P}(A, B)$ is relevant since the cleanliness of the cards C has no bearing on the association between the character A and colour B of the cards.

Example b) Treatment Effect on Mortality Rate (Confounding Variable)

Next, suppose that the data relates to the results of medical trials for a drug on a potentially lethal illness. This time,

- A tells us if the subject was treated ($A = 1$) or not ($A = 0$).
- B tells us if the subject died ($B = 1$) or recovered ($B = 0$).
- C tells us the gender of the subject, either male ($C = 1$) or female ($C = 0$).

In this case the disaggregated data shows the more relevant association, $\mathbb{P}(A, B|C)$. From it, we can see that female patients are more likely to die than males overall; 56 and 60% versus 38 and 43%, depending on if they were treated or not. In both cases we see that treatment with the drug A reduces the mortality rate for both male and female participants, and the effect is obscured by aggregating the data over gender C.

Back to Causality

The key difference between these examples is the causal relationship between the variables rather than the statistical structure of the data. In the first example with the playing cards, the variable C is a *colliding* variable, in the second example looking at patient mortality, it is a *confounding* variable. Figure 1.4 a) and b) show the causal relationships between the variables in the two cases.

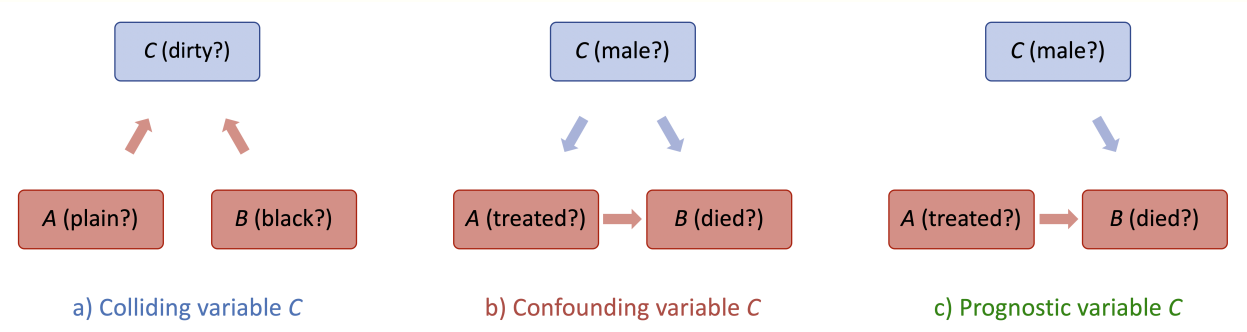


Figure 1.4: Causal diagrams for A, B and C when C is a colliding, confounding and prognostic variable.

The causal diagram in Figure 1.4 a) shows the variables A, B and C for the first example. The arrows exist both from card character and colour to cleanliness because apparently, baby Milen had a preference for royal cards over plain and red cards over black. Conditioning on a collider C generates an association (e.g. correlation) between A and B, even if they are unconditionally independent. This common effect is often observed as *selection* or *representation bias*. Representation bias can induce correlation between variables, even where there is none. For decision systems, this can lead to feedback loops that increase the extremity of the representation bias in future data. We'll come back to this in chapter 2, when we talk about common causes of bias.

The causal diagram in Figure 1.4 b) shows the variables A, B and C for the second example. The arrows exist from *gender* to treatment because men were less likely to be treated, and from gender to death because men were also less likely to die. The arrow from A to B represents the effect of treatment on mortality which is observable only by conditioning on gender. Note that there are two sources of association in opposite directions between variables A and B (treatment and death); a positive association, because men were less likely to be treated; and a negative association, because male patients are less likely to die. The two effects cancel each other out when the data is aggregated.

We see through the discussion of these two examples, that statistical reasoning is not sufficient to be able to determine which of the distributions (marginal or conditional) are relevant. Note that the above conclusions in relation to colliding and confounding variables does not generalize to complex time varying problems.

Before moving on from causality, we return to the example we discussed at the very start of this section. According to our analysis of the Berkeley admissions data, we concluded that the disaggregated data contained the *salient* information explaining the disparity in acceptance rates for male and female applicants. The problem is, we have only shown that application rates to be one of many possible *causes* of the differing acceptance rates (we cannot see outside of our data). In addition, we have not proven *gender discrimanation*, not to be the cause. What we have evidenced, is the existance of disparities in both acceptance rates and application rates across sex. One problem is that *gender discrimanation* is not a measurable thing in itself. It's complicated. It is made up of many components, most of which are not contained in the data. Beliefs, personal preferences, behaviours, actions, and more. A valid question we cannot answer is, *why do the application rates differ by sex?* How do we know that this is itself, is not a result of gender discrimnation. Perhaps some departments are less welcoming of women than others or, perhaps some are just much more welcoming of men than women? So how would we know if gender discrimination is at play here? We need to ask the right questions to collect the right data.

1.4.3 Collapsibility

We have demonstrated that correlation does not imply causation in the manifestation of Simpson's Paradox. But there is second factor that can have an impact; and that is the nature of the measure of association in question.

Example c) Treatment Effect on Mortality Rate (Prognostic Variable)

Suppose that in the study of the efficacy of the treatment (in Example 2 above), we remedy the selection bias so that male and female patients are equally likely to be treated. We remove the causal relationship between variables A and C (treatment and gender). In this case, the variable C becomes *prognostic* rather than confounding. See Figure 1.4 c). In this case the decision as to which distributions (marginal or conditional) are most relevant would depend only on the target population in question. In the absence of the confounding variable in our study one might reasonably expect the marginal measure of association to be bounded by the partial measures of association. Such intuition is correct only if the measure of association is *collapsible* (that is, it can be expressed as the weighted average of the partial measures), not otherwise. Some examples of collapsible measures of association are the risk ratio and risk difference. The odds ratio however is not collapsible. If you don't know what these are, don't worry, we'll return to them in chapter 3.

1.5 What's the Harm?

In this section we discuss the recent and broader societal concerns related to machine learning technologies.

1.5.1 The Illusion of Objectivity

One of the most concerning things about the machine learning revolution, is perception that these algorithms are somehow objective (unlike humans), and are therefore a better substitute for human judgement. This viewpoint is not just a belief of laymen but an idea that is also projected from within the machine learning community. There are often financial incentives to exaggerate the efficacy of such systems.

Automation Bias

The tendency for people to favour decisions made by automated systems despite contradictory information from non-automated sources, or *automation bias*, is a growing problem as we integrate more and more machines in our decision making processes especially in infrastructure - healthcare, transportation, communication, power plants and more.

It is important to be clear that in general, machine learning systems are not objective. Data is produced by a necessarily subjective set of decisions (how and who to sample, how to group events or characteristics, which features to collect). Modelling also involves making choices about how to process the data, what class of model to use and perhaps most importantly how sucess is determined. Finally, even if our model is calibrated to the data well, it says nothing about the distribution of errors across the population. The consistency of algorithms in decision making compared to humans (who individually make decisions on a case by case basis) is often described as a benefit[2], but it's their very consistency that makes them dangerous - capable of discriminating systematically and at scale.

Example: COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a "case management system for criminal justice practitioners". The system, produces recidivism risk scores. It has been used in New York, California and Florida, but most extensively in Wisconsin since 2012, at a variety of stages in the criminal justice, from sentencing to parole. The documentation for the software describes it as an "objective statistical risk assessment tool".

[2]One must not confuse consistency with objectivity. For algorithms, consistency also means consistently making the same errors.

In 2013, Paul Zilly was convicted of stealing a push lawnmower and some tools in Barron County, Wisconsin. The prosecutor recommended a year in county jail and follow-up supervision that could help Zilly with "staying on the right path." His lawyer agreed to a plea deal. But Judge James Babler upon seeing Zilly's COMPAS risk scores overturned the plea deal that had been agreed on by the prosecution and defense, and imposed two years in state prison and three years of supervision. At an appeals hearing later that year, Babler said "Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months"[2]. In other words the judge believed the risk scoring system to hold more insight that the prosecutor who had personally interacted with the defendant.

The Ethics of Classification

The appeal of classification is clear. It creates a sense of order and understanding. It enables us to formulate problems neatly and solve them. An email is spam or it's not; an x-ray shows tuberculosis or it doesn't; a treatment was effective or it wasn't. It can make finding things more efficient in a library or online. There are lots of useful applications of classification.

We tend to think of taxonomies as objective categorisations, but often they are not. They are snapshots in time, representative of the culture and biases of the creators. The very act of creating a taxonomy, can give life by existance to some individuals, while erasing others. Classifying people inevitably has the effect of reducing them to labels; labels that can result in people being treated as members of a group, rather than individuals; labels that can linger for much longer than they should (something it's easy to forget when creating them). The Dewey Decimal System for example, was developed in the late 1800's and widely adopted in the 1930's to classify books. Until 2015, it categorised homosexuality as a mental derangement.

Classification of People

From the 1930's until the second world war, machine classification systems were used by Nazi Germany to process census data in order to identify and locate Jews, determine what property and businesses they owned, find anything of value that could be seized and finally to send them to their deaths in concentration camps. Classification systems have often been entangled with political and social struggle across the world. In Aparteid South Africa, they were been used extensively in many parts of the world to enforce social and racial hierarchies that determined everything from where people could live and work to whom they could marry. In 2019 it was estimated that some half a million Uyghurs (and other minority Muslims) are being held in internment camps in China without charge for the purposes of countering extremism and promoting social integration.

Recent papers on detecting criminality"[27] and sexuality[26] and ethnicity[25] from facial images have sparked controversy in the academic community. The latter in particular looks for facial features that identify among others, Chinese Uyghurs. Physiognomy (judging character from the physical features of a persons face) and phrenology (judging a persons level of intelligence from the shape and dimensions of their cranium) have historically been used as pseudo-scientific tools of oppressors, to prove the inferiority races and justify subordination and genocide. it is not without merit then to ask if some technologies should be built at all. Machine gaydar might be a fun application to mess about with friends for some, but in the 70 countries where homosexuality is still illegal (some of which enforce the death penalty) it is something rather different.

1.5.2 Personalisation and the Filter Bubble

Many believed the internet would breath new life into democracy. The decreased cost and increased accessibility of information would result in greater decentralization of power and flatter social structures. In this new era, people would be able to connect, share ideas and organise grass roots movements at a such a scale that would enable a step change in the rate of social progress. Some of these ideas have been realised to an extent but the increased ability to create and distribute content and corresponding volume of data has created new problems. The amount of information available to us through the internet is overwhelming. Email, blog posts, Twitter, Facebook, Instagram, Linked In, What's App, You Tube, Netfix, TikTok and more. Today there are seemingly endless ways and places for us to communicate and share information. This

barrage of information has resulted in what has been described as the attention crash. There is simply too much information for us to attend to all of it meaningfully. The mechanisms through which we can acquire new information that demands our attention too have expanded. We carry our smart phones everywhere we go and sleep beside them. There is hardly a waking moment, when we are unplugged and inaccessible. The demands on our attention and focus have never been greater. Media producers themselves have adapted their content in order to accommodate our new shortened attention spans.

With so much information available it's easy to see the appeal of automatic filtering and curation. And of course, how good would said system really be if it didn't take into account our personal tastes and preferences? So what's the problem?! Over the last decade, personalisation has become entrenched in the systems we interact with day to day. Targeted advertising was just the beginning. Now it's not just the trainers you browsed once that follow you around the web until you buy them, it's everything. Since 2009, Google has returned personalised results every time someone queries their search engine, so two people who enter the same text don't get the same result. In 2021 You Tube had more than two billion logged-in monthly users. Three quaters of adults in the US use it (more than facebook and Instagram) and 80% of U.S. parents of children under 11 watch it. It is the second most visited site in the world, after Google with visitors checking on average just under 9 pages, and spending 42 minutes per day there. In 2018, 70% of the videos people watched on You Tube were recommended. Some 40% of Americans under thirty get their news through social networking sites such as twitter and Facebook but this may be happening without you even knowing. Since 2010, it's not the Washington Post that decides which news story you see in the prime real estate that is the top right hand corner of their home page, it's Facebook - the same goes for the New York Times. So the kinds of algorithms that once determined what we spent our money on now determine our very perception of the world around us. The only question is, what are they optimising for?

Ignoring, for a moment, the fact that having the power to shape people's perception of the world, in just a few powerful hands is in itself a problem. A question worth pondering on is what kind of citizens people who only ever see things they 'like', or feel the impulse to 'comment' on (or indeed any other proxy for interest/engagement/attention) would make. As Eli Pariser put it in his book The Filter Bubble, "what one seems to like may not be what one actually wants, let alone what one needs to know to be an informed member of their community or country". The internet has made the world smaller and with it we've seen great benefits. But the idea that, because anyone (regardless of their background) could be our neighbour, people would find common ground has not been realised to the extent people hoped. In some senses personalisation does the exact opposite. It risks us all living in a world full of mirrors, where we only ever hear the voices of people who see the world as we do, being deprived of differing perspectives. Of course we have always lived in our own filter bubble in some respects but the thing that has changed is that now we don't make the choice and often don't even know when we are in it. We don't know when or how decisions are made about what we should see. We are more alone in our bubbles than we have ever been before.

Social capital is created by the interpersonal bonds we build in shared identity, values, trust and reciprocity. It encourages people to collaborate in order to solve common problems for the common good. There are two kinds of social capital, bonding and bridging. Bonding capital is acquired through development of connections in groups that have high levels of similarity in demographics and attitudes - the kind you might build by, say socialising with collegues from work. Bridging capital is created when people from different backgrounds (race, religion, class) connect - something that might happen at a town hall meeting say. The problem with personalisation is that by construction it reduces opportunities to see the world through the eyes of people who don't necessarily look like us. It reduces bridging capital and that exactly the kind of social capital we need to solve wider problems that extend beyond our own narrow or short term self interests.

1.5.3 Disinformation

In June 2016, it was announced that Britain would be leaving the EU. 33.5 million people voted in the referendum of which 51.9% voted to leave. The decision that will impact the UK for, not just a term, but generations to come, rested on less than 2% of voters. Ebbw Vale is a small town in Wales where 62% of the electorate (the largest majority in the country) voted to leave. The town has a history in steel and coal

dating back to the late 1700's. By the 1930's the Ebbw Vale Steelworks was the largest in Europe by volume. In the 1960's it employed some 14,500 people. But, towards the end of the 1900's, after the collapse of the UK steel industry, the town suffered one of the highest unemployment rates in Britain. What was strange about the overwhelming support to leave was that Ebbw Vale was perhaps one of the largest recipients of EU development funding in the UK. A £350m regeneration project funded by the EU replaced the industrial wasteland left behind when the steelworks closed in 2002 with The Works (a housing, retail and office space, wetlands, learning campus and more). A further £33.5 in funding from the European Social Fund paid for a new college and apprenticeships, to help young people learn a trade. An additional £30 million for a new railway line, £80 million for road improvements and shortly before the vote a further £12.2 million for other upgrades and improvements were all from the EU.

When journalist Carole Cadwalladr returned to the small town where she had grown up to report on why residents had voted so overwhelmingly in favour of leaving the EU, she was no less confused. It was clear how much the town had benefited from being part of the EU. The new road, train station, college, leisure centre and enterprise zones (flagged an EU tier 1 area, eligible for the highest level of grant aid in the UK), everywhere she went she saw signs with proudly displayed EU flags saying so. So she wandered around town asking people and was no less perplexed by their answers. Time and time again people complained about immigration and foreigners. They wanted to take back control. But the immigrants were nowhere to be found, because Ebbw Vale had one of the lowest rates of immigration in the country. So how did this happen? How did a town with hundreds of millions of pounds of EU funding vote to leave the EU because of immigrants that didn't exist? In her emotive TED talk[23], Carole shows images of some the adverts on Facebook, people were targeted with as part of the leave campaign (see Figure 1.5). They were all centred around a lie - that Turkey was joining the EU.

Figure 1.5: Targeted disinformation adverts shown on Facebook[23].

Most people in the UK saw adverts on buses and billboards with false claims, for example that the National Health Service (NHS) would have an extra £350 million a week, if we left the EU. Although many believed them, those adverts circulated in the open for everyone to see, giving the mainstream media at the opportunity to debunk them. The same cannot be said for the adverts in Figure 1.5. They were targeted towards specific individuals, as part of an evolving stream of information displayed in their Facebook 'news'

feed. The leave campaign paid Cambridge Analytica (a company that had illegally gained access to the data of 87 million Facebook users) to identify individuals that could be manipulated into voting leave. In the UK, spending on elections in the is limited by law as a means to ensure fair elections. After a nine month investigation, the UK's Electoral Commission confirmed these spending limits had been breached by the leave campaign. There are ongoing criminal investigations into where the funds for the campaign originate (overseas funding of election campaigns is also illegal) but evidence suggests ties with Russia. Brexit was the precursor to the Trump administration winning the US election just a few months later that year. The same people and companies used the same strategies. It's become clear that current legislation protecting democracy is inadequate. Facebook, was able to profit from politically motivated money without recognizing any responsibility in ensuring the transactions were legal. Five years later, the full extent of the disinformation campaign on Facebook has yet to be understood. Who was shown what and when, how people were targeted, what other lies were told, who paid for the adverts or where the money came from.

Since then deep learning technology has advanced to the point of being able to pose as human in important ways that risk enabling disinformation not just through targeted advertising but machines impersonating humans. GANs can fabricate facial images, videos (deepfakes) and audio. Advancements in language models (Open AIs GPT-2 and more recently GPT-3) are capable of creating lengthy human like prose given just a few prompts. Deep learning now provides all the tools to fabricate human identities and target dissemination of false information at scale. There are growing concerns that in the future, bots will drown out actual human voices. As for the current state of play, it's difficult to know the exact numbers but in 2017, researchers estimated that between 9 and 15% of all twitter accounts were bots[24]. In 2020 a study by researchers at Carnegie Mellon University reported that 45% of the 200 million tweets they analysed discussing coronavirus came from accounts that behaved like bots[1]. For Facebook, things are less clear as we must rely on their own reporting. In mid-2019, Facebook estimated that only 5% of its 2.4 billion monthly active users were fake though its reporting raised some questions[12].

1.5.4 Harms of Representation

The interventions we'll talk about in most of this book are designed to measure and mitigate harms of allocation in machine learning systems.

Harms of Allocation

An allocative harm happens when a system allocates or withholds an opportunity or resource. Systems that approve or deny credit allocate financial resources; systems that decide who should and should not see adverts for high paying jobs allocate employment opportunities and systems that determine who will make a good tenant allocate housing resources. Harms of allocation happen as a result of discrete decisions at a given point in time, the immediate impact of which can be quantified. This makes it possible to challenge the justice and fairness of specific determinations and outcomes.

Increasingly however, machine learning systems are affecting us, not just through allocation, but are shaping our view of the world and society at large by deciding what we do and don't see. These harms are far more difficult to quantify.

Harms of Representation

Harms of representation occur when systems enforce the subordination of groups through characterizations that affect the perception of them. In contrast to harms of allocation, harms of representation have long-term effects on attitudes and beliefs. They create identities and labels for humans, societies and their cultures. Harms of representation don't just affect our perception of each other, they affect how we see ourselves. They are difficult to formalise and in turn difficult to quantify but the effect is real.

The Surgeon's Dilemma

A father and his son are involved in a horrific car crash and the man died at the scene. But when the child arrived at the hospital and was rushed into the operating theatre, the surgeon pulled away and said: "I can't operate on this boy, he's my son". How can this be?

Did you figure it out? How long did it take? There is, of course, no reason why the surgeon couldn't be the boy's mother. If it took you a while to figure out, or came to a different conclusion, you're not alone. More than half the people presented with this riddle do, and that includes women. The point of this riddle is to demonstrate the existence of unconscious bias. Representational harms are insidious. They silently fix ideas in peoples subconscious about what people of a particular gender, nationality, faith, race, occupation and more, are like. They draw boundaries between people and affect our perception of world. Below we describe five different harms of representation:

Stereotyping

Stereotyping occurs through excessively generalised portrayals of groups. In 2016, the Oxford English Dictionary was publicly criticised[13] for employing the phrase "rabid feminist" as a usage example for the word rabid. The dictionary included similarly sexist common usages for other words like shrill, nagging and bossy. But even before this, historical linguists observed that words referring to women undergo pejoration (when the meaning of a word deteriorates over time) far more often than those referring to men[16]. Consider words like mistress (once simply the female equivalent of master, now used to describe a woman in an illicit relationship with a married man); madam (once simply the female equivalent of sir, now also used to describe a woman who runs a brothel); hussy (once a neutral term for the head of a household, now used to describe an immoral or ill-behaved woman); and governess (female equivalent of governor, later used to describe a woman responsible for the care of children).

Unsurprisingly then, gender stereotyping is known to be a problem in natural language processing systems. In 2016 Bolukbasi et al. showed that word embeddings exhibited familiar gender biases in relation to occupations[5]. By performing arithmetic on word vectors, they were able to uncover relationships such as

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

In 2017 Caliskan et al. found that Google Translate contained similar gender biases.[6] In their research they found that "translations to English from many gender-neutral languages such as Finnish, Estonian, Hungarian, Persian, and Turkish led to gender-stereotyped sentences". So for example when they translated Turkish sentences with genderless pronouns: "O bir doktor. O bir hemişre.", the resulting English sentences were: "He is a doctor. She is a nurse." They performed these types of tests for 50 occupations and found that the stereotypical gender association of the word almost perfectly predicted the resulting pronoun in the English translation.

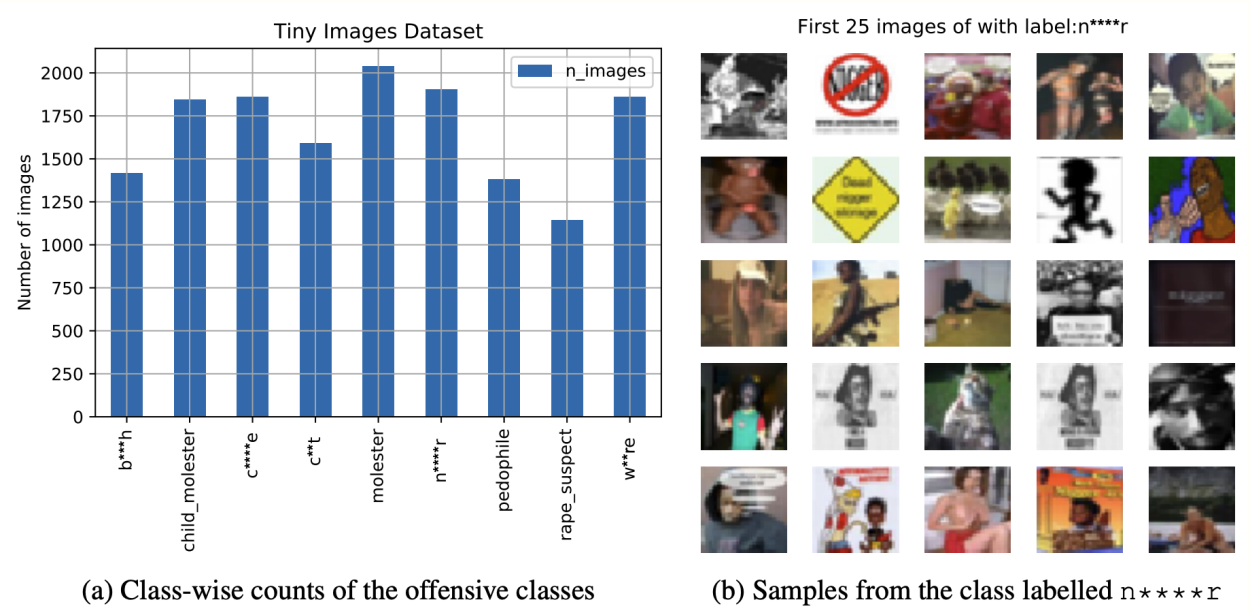
Recognition

Harms of recognition happen when groups of people are in some senses erased by a system through failure to recognise. In her TED Talk, Joy Buolamwini, talks about how as an undergraduate studying computer science she worked on social robots. One of her projects involved creating a robot which could play peek-a-boo, but she found that her robot (which used third party software for facial recognition) could not see her. She was forced to borrow her roommate's face to complete the project. After her work auditing several popular gender classification packages from IBM, Microsoft and Face++ in the project Gender Shades[9] in 2017 and seeing the failure of these technologies on the faces of some of the most recognizable Black women of her time, including Oprah Winfrey, Michelle Obama, and Serena Williams, she was prompted to echo the words of Sojourner Truth in asking "Ain't I a Woman?". Harms of recognition are failures in seeing humanity in people.

Denigration

In 2015, much to the horror of many people, it was reported that Google Photos had labelled a photo of a Black couple as Gorillas. It's hard to find the right words to describe just how offensive an error this is. It demonstrated how a machine, carrying out a seemingly benign task of labelling photos, could deliver an attack on a person's human dignity.

In 2020, an ethical audit of several large computer vision datasets[14], revealed some disturbing results. TinyImages (a dataset of 79 million 32 x 32 pixel colour photos compiled in 2006, by MIT's Computer Science and Artificial Intelligence Lab for image recognition tasks) contained racist, misogynistic and demeaning labels with corresponding images. Figure 1.6 shows a subset of the data found in TinyImages. The problem,



(a) Class-wise counts of the offensive classes

(b) Samples from the class labelled `n****r`

Figure 1.6: Subset of data in TinyImages exemplifying toxicity in both the images and labels[14].

unfortunately, does not end here. Many of the datasets used to train and benchmark, not just computer vision but natural language processing tasks, are related. Tiny Images was compiled by searching the internet for images associated with words in WordNet (a machine readable, lexical database, organised by meaning, developed at Princeton), which is where TinyImages inherited its labels from. ImageNet (widely considered to be a turning point in computer vision capabilities) is also based on WordNet and, Cifar-10 and Cifar-100 were derived from TinyImages.

Vision and language datasets are enormous. The time, effort and consideration in collecting the data that forms the foundation of these technologies (compared to that which has gone into advancing the models built on them), is questionable to say the least. Furthermore a dataset can have impact beyond the applications trained on it, because datasets often don't just die, they evolve. This calls into question the technologies that are in use today, capable of creating persistent representations of our world, and trained on datasets so large they are difficult and expensive to audit.

And there's plenty of evidence to suggest that this is a problem. For example, in 2013, a study found that Google searches were more likely to return personalised advertisements that were suggestive of arrest records for Black names[22] than White[3] This doesn't just result in allocative harms for people applying for jobs for example, it's denigrating. Google's Natural Language API for sentiment analysis is also known to have problems. In 2017, it was assigning negative sentiment to sentences such as "I'm a jew" and "I'm a

[3]Suggestive of an arrest record in the sense that they claim to have arrest records specifically for the name that you searched, regardless of whether they do in reality have them.

homosexual" and "I'm black"; neutral sentiment to the phrase "white power" and positive sentiment to the sentences "I'm christian" and "I'm sikh".

Under-representation

In 2015, the New York Times reported, that "Fewer women run big companies than men named John", despite this Google's image search still managed to under-represent women in search results for the word "CEO". Does this really matter? What difference would an alternate set of search results make? A study the same year found that "people rate search results higher when they are consistent with stereotypes for a career, and shifting the representation of gender in image search results can shift people's perceptions about real-world distributions."[11].

Ex-nomination

Ex-nomination occurs through invisible means and affects people's views of the norms within societies. It tends to happen through mechanisms which amplify the presence of some groups and suppress the presence of others. The cultures, beliefs, politics of ex-nominated groups over time become the default. The most obvious example is the ex-nomination of Whiteness and White culture in western society, which might sound like a bizarre statement - what is White culture? But such is the effect of ex-nomination, you can't describe it, because it is just the norm and everything else is not. Richard Dyer in his book White examines the reproduction and preservation of whiteness in visual media over five centuries, from the depiction of the crucifixion to modern day film. It's perhaps should not come as a surprise then, when facial recognition software can't see black faces; or when gender recognition software fails more often than not for black women; or that a generative model that improves the resolution of images, converted a pixelated picture of Barack Obama, into a high-resolution image of a white man.

The ex-nomination of White culture is evident in our language too, in terminology like whitelist and white lie. If you look up white in dictionary and or thesaurus and you'll find words like innocent and pure, light, transparent, immaculate, neutral. Doing the same for the word black on the other hand, returns very different associations, dirty, soiled, evil, wicked, black magic, black arts, black mark, black humour, blacklist and black is often used as a prefix in describing disastrous events. A similar assessment can be made for gender with women being under-represented in image data and feminine versions of words more often undergoing pejoration (when the meaning or status of a word deteriorates over time).

Members of ex-nominated groups experience a kind of privilege that it is easy to be unaware of. It is a power that comes from being the norm. They have advantages that are not earned, outside of their financial standing or effort, that the 'equivalent' person outside the ex-nominated group would not. Their hair type, skin tone, accent, food preferences and more are catered to by every store, product, service and system and it cost less to access them; they see themselves represented in the media and are more often represented in a positive light; they are not subject to profiling or stereotypes; they are more likely to be treated as individuals rather than as representative of (or as exceptions to) a group; they are more often humanised - more likely to be be given the benefit of the doubt, treated with compassion and kindness and thus recover from mistakes; they are less likely to be suspected of crimes; more likely to be trusted financially; they have greater access to opportunities, resources and power and are able to climb financial, social and professional ladders faster. The advantages enjoyed by ex-nominated groups accumulate over time and compound over generations.

Summary

Machine learning

- In this book we will use algorithm and model interchangeably. A model can be determined using data, but it need not be. It can simply express an opinion on the relationship between variables. In practice

the implementation is an algorithm either way. More precisely, a model is a function or mapping; given a set of input variables (features) it returns a decision or prediction (target).

- A model is a simplified representation of some real world phenomena. When trained on historic data they best capture the dense part of the distribution. Therefore, when faced with rare or unprecedented events they tend to perform poorly. Obtaining adequately rich and relevant data is a major limitation for most machine learning models.
- At almost every major life event, going to university, getting a job, buying a house, getting sick, decisions are increasingly being made by machines. By construction, these models encode existing societal biases. They not only proliferate but are capable of amplifying them and are easily deployed at scale. Understanding the shortcomings of these models and ensuring such technologies are deployed responsibly are essential if we are to safeguard social progress.

Discrimination, bias, fairness and ethics

- Building machine learning systems ethically is not about finding the perfect answer every time but rather expanding our perspectives on the technology we develop. It's looking for the cracks before deploying systems, preventing the foreseeable failures and doing the best we can on the ones we didn't see coming.
- According to uilitarian doctrine, the correct course of action (when faced with a dilemma) is the one that maximises the benefit for the greatest number of people. The doctrine demands that the benefits to all people are are counted equally.
- The standard, approach to training a model, which is essentially to minimise the aggregate error on the training data, is loosely justified in a utilitarian sense, in that we optimise our decision process which maximises utility (minimises the probability of error) for the greatest number of people (we aim to optimise over the target population, in the sense that we aim to reduce the generalisation error).
- Utilitarianism is a flavour of consequentialism, a branch of ethical theory that holds that consequences are the yardstick against which we must judge the morality of our actions. In contrast deontological ethics judges the morality of actions against a set of rules that define our duties or obligations towards others. Here it is not the consequences of our actions that matter but rather intent.
- There are some practical problems with utilitarianism but perhaps the most significant flaw in utilitarianism for moral reasoning is the omission of justice as a consideration.
- Principles of Justice as Fairness:

  1. **Liberty principle:** Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which is compatible with the same scheme of liberties for all;
  2. **Equality principle:** Social and economic inequalities are to satisfy two conditions:
     (a) **Fair equality of opportunity:** The offices and positions to which they are attached are open to all under conditions of fair equality of opportunity;
     (b) **Difference principle** They must be of the greatest benefit to the least-advantaged members of society.

  The principles of justice as fairness are ordered by priority so that fulfilment of the liberty principle takes precedence over the equality principles and fair equality of opportunity takes precedence over the difference principle. In contrast to utilitarianism, justice as fairness introduces a number of constraints that must be satisfied for a decision process to be fair. Applied to a machine learning one might interpret the liberty principle as a requirement of some minimum accuracy level (maximum probability of error) to be set for all members of the population, even if this means the algorithm is less accurate overall. Parallels can be drawn here in machine learning where there is a trade-off between fairness and utility of an algorithm.
- Anti-discrimination laws were born out of long-standing, vast and systemic discrimination against historically oppressed and disadvantaged classes. Such discrimination has contributed to disparities in all measures of prosperity (health, wealth, housing, crime, incarceration) that persist today.

- Legal liability for discrimination against protected classes may be established through both disparate treatment and disparate impact. Disparate treatment (also described as direct discrimination in Europe) refers to both formal differences in the treatment of individuals based on protected characteristics, and the intent to discriminate. Disparate impact (also described as indirect discrimination in Europe) does not consider intent but is concerned with policies and practices that disproportionately impact protected classes.
- Just as the meaning of fairness is subjective, so too is the interpretation of anti-discrimination laws. Two conflicting interpretations are anti-classification and anti-subordination. Anti-classification is a weaker interpretation, that the law is intended to prevent classification of people based on protected characteristics. Anti-subordination is the stronger interpretation that anti-discrimination laws exist to prevent social hierarchies, class or caste systems based on protected features and, that it should actively work to eliminate them where they exist.

Association paradoxes

- Identifying bias in data can be tricky. Data can be misleading. An association paradox is a phenomenon where an observable relationship between two variables disappears or reverses after controlling for one or more other variables. In order to know which associations (or distributions) are relevant, i.e. the marginal (unconditional) or partial associations (conditional distributions), one must understand the causal nature of the relationships. Association paradoxes can also occur for non-collapsible measures of association. Collapsible measures of association are those which can be expressed as the weighted average of the partial measures.

What's the harm?

- It is important to be clear that in general, machine learning systems are not objective. Data is produced by a necessarily subjective set of decisions. The consistency of algorithms in decision making compared to humans (who make decisions on a case by case basis) is often described as a benefit, but it's their very consistency that makes them dangerous - capable of discriminating systematically and at scale.
- Classification creates a sense of order and understanding. It enables us to find things more easily, formulate problems neatly and solve them. But classifying people inevitably has the effect of reducing people labels; labels that can result in people being treated as members of a group, rather than individuals.
- Personalisation algorithms that shape our perception of the world in a way that covertly mirror our beliefs can have the effect of trading bridging for bonding capital, the former kind is important in solving global problems that require collective action, such as global warming.
- Targeted political advertising and technologies that enable machines to impersonate humans are powerful tools that can be used as part of orchestrated campaigns of disinformation that manipulate perceptions at an individual level and yet at scale. They are capable of causing great harm to political and social institutions and pose a threat to security.
- An allocative harm happens when a system allocates or withholds an opportunity or resource. Harms of representation occur when systems enforce the subordination of groups through characterizations that affect the perception of them. In contrast to harms of allocation, harms of representation have long-term effects on attitudes and beliefs. They create identities and labels for humans, societies and their cultures. Harms of representation affect our perception of each other and even ourselves. Harms of representation are difficult to quantify. Some types of harms of representation are, stereotyping, (failure of) recognition, denigration, under-representation and ex-nomination.

References

[1] Bobby Allyn. Researchers: Nearly half of accounts tweeting about coronavirus are likely bots. *NPR*, May 2020. https://www.npr.org/sections/coronavirus-live-updates/2020/05/20/859814085/researchers-nearly-half-of-accounts-tweeting-about-coronavirus-are-likely-bots.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, March 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *Calif Law Rev.*, 104:671–732, 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

[4] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187, Issue 4175:398–404, February 1975. https://science.sciencemag.org/content/187/4175/398.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016. https://arxiv.org/abs/1607.06520.

[6] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186, April 2017. https://researchportal.bath.ac.uk/en/publications/semantics-derived-automatically-from-language-corpora-necessarily.

[7] Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence, April 2021. https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682.

[8] General Data Protection Regulation (GDPR): (EU) 2016/679 Recital 71, April 2016. https://gdpr-info.eu/recitals/no-71/.

[9] Sorelle A. Friedler and Christo Wilson, editors. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, volume 81. Proceedings of Machine Learning Research, 2018. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[10] President Lyndon B. Johnson. Speech to a joint session of congress on march 15, 1965. *Public Papers of the Presidents of the United States*, I, entry 107:281–287, March 1965. http://www.lbjlibrary.org/lyndon-baines-johnson/speeches-films/president-johnsons-special-message-to-the-congress-the-american-promise.

[11] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. Unequal representation and gender stereotypes in image search results for occupations. *ACM*, 2015. https://www.csee.umbc.edu/~cmat/Pubs/KayMatuszekMunsonCHI2015GenderImageSearch.pdf.

[12] Jack Nicas. Does facebook really know how many fake accounts it has? *The New York Times*, January 2019. https://www.nytimes.com/2019/01/30/technology/facebook-fake-accounts.html.

[13] Emer O'Toole. A dictionary entry citing 'rabid feminist' doesn't just reflect prejudice, it reinforces it. *The Guardian*, January 2016. https://www.theguardian.com/commentisfree/2016/jan/26/rabid-feminist-language-oxford-english-dictionary.

[14] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision?, 2020. https://arxiv.org/abs/2006.16923.

[15] John Rawls. *Justice As Fairness: a Restatement*. Harvard University Press, Cambridge, Mass., 2001. (1921-2002).

[16] David Shariatmadari. Eight words that reveal the sexism at the heart of the english language. *The Guardian*, January 2016. https://www.theguardian.com/commentisfree/2016/jan/27/eight-words-sexism-heart-english-language.

[17] E Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, Series B, 13:238–241, March 1951. https://www.jstor.org/stable/2984065.

[18] Griggs v. Duke Power Co., 401 U.S. 424, 1971. https://en.wikipedia.org/wiki/Griggs_v._Duke_Power_Co.

[19] Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 1989. https://en.wikipedia.org/wiki/Wards_Cove_Packing_Co._v._Atonio.

[20] Grutter v. Bollinger, 539 U.S. 306, 2003. https://en.wikipedia.org/wiki/Grutter_v._Bollinger.

[21] Ricci v. DeStefano, 557 U.S. 557, 2009. https://en.wikipedia.org/wiki/Ricci_v._DeStefano.

[22] Latanya Sweeney. Discrimination in online ad delivery. *SSRN*, 2013. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240.

[23] TED. *Facebook's role in Brexit - and the threat to democracy*, 2019. https://www.ted.com/talks/carole_cadwalladr_facebook_s_role_in_brexit_and_the_threat_to_democracy.

[24] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization, 2017. https://arxiv.org/abs/1703.03107.

[25] C. Wang, Q. Zhang, W. Liu, Y. Liu, and L. Miao. Facial feature discovery for ethnicity recognition. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018. https://espace.curtin.edu.au/handle/20.500.11937/71484.

[26] Y. Wang and M. Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 2018. https://www.gsb.stanford.edu/faculty-research/publications/deep-neural-networks-are-more-accurate-humans-detecting-sexual.

[27] Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images, 2017. https://arxiv.org/abs/1611.04135.

Chapter 2

Ethical development

This chapter at a glance

- The machine learning cycle - feedback from models to data
- The machine learning development and deployment life cycle
- A practical approach to ethical development and deployment
- A taxonomy of common causes of bias

In this chapter, we transition to a more systematic approach to understanding the problem of fairness in decisions making systems. In later chapters we will look at different measures of fairness and bias mitigation techniques but before we discuss and analyse these methods, we review some more practical aspects of responsible model development and deployment. None of the bias mitigation techniques that we will talk about in part three of this book will rectify a poorly formulated, discriminatory machine learning problem or remedy negligent deployment of a predictive algorithm. A model in itself is not the source of unfair or illegal discrimination, models are developed and deployed by people as part of a process. In order to address the problem of unfairness we need to look at the whole system, not just the data or the model.

We'll start by looking at the machine learning cycle and discuss the importance of how a model is used in the feedback effect it has on data. Where models can be harmful we should expect to have processes in place that aim to avoid common, foreseeable or catastrophic failures. We'll discuss how to take a proactive rather than reactive approach to managing risks associated with models. We'll discuss where in the machine learning model development cycle bias metrics and modelling interventions fit. Finally, we'll classify the most common causes of bias, identifying the parts of the workflow to which they relate.

Our goal is to present problems and interventions schematically, creating a set of references for building, reviewing, deploying and monitoring machine learning solutions that aim to avoid the common pitfalls that result in unfair models. We take a high enough view that the discussion remains applicable to many machine learning applications. The specifics of the framework, can be tailored for a particular use case. Indeed the goal is for the resources in this chapter can be used as a starting point for data science teams that want to develop their own set of standards. Together we will progress towards thinking critically about the whole machine learning cycle, development, validation, deployment and monitoring of machine learning systems. By the end of this chapter we will have a clearer picture of what due diligence in model development and deployment might look like from a practical perspective.

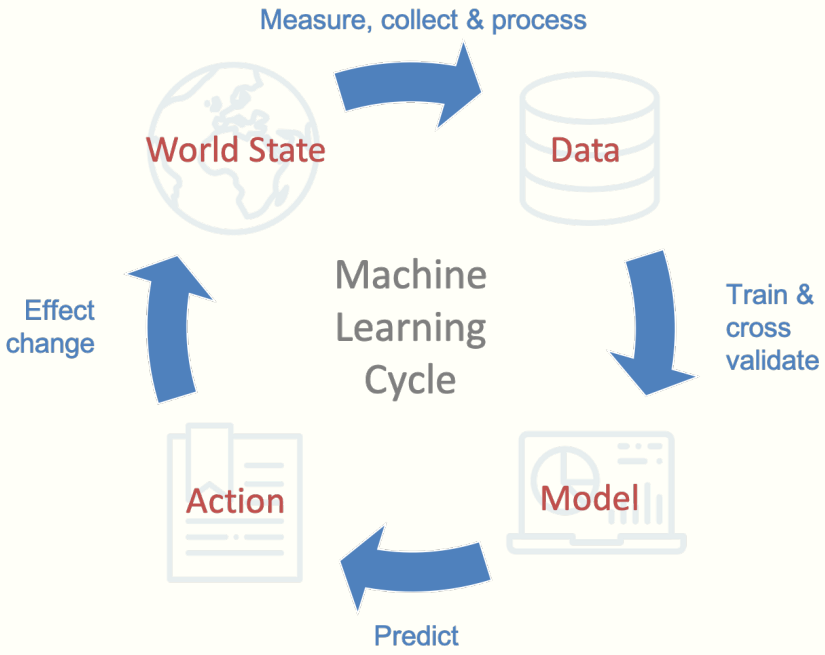


Figure 2.1: The machine learning cycle

2.1 Machine Learning Cycle

Machine learning systems can have longterm and compounding effects on the world around us. In this section we analyse the impact in a variety of different examples to breakdown the mechanisms that determine the nature and magnitude of the effect. In Figure 2.1, we present the machine learning cycle - a high-level depiction of the interaction between a machine learning solution and the real world. A machine learning system starts with a set of objectives. These can be achieved in a myriad of different ways. The translation of these objectives, into a tractable machine learning problem, consists of a series of subjective decisions; what data we collect to train a model on, what events we predict, what features we use, how we clean and process the data, how we evaluate the model and the decision policy are all choices. They determine the model we create, the actions we take and finally the resulting cycle of feedback on the data.

The most familiar parts of the cycle to most developers of machine learning solutions are on the right hand side; processing data, model selection, training and cross validation and prediction. Each action taken on the basis of our model prediction creates a new world state, which generates new data, which we collect and train our model on, and around it goes again. The actions we take based on our model predictions define how we use the model. The same model used in a different way can result in a very different feedback cycle.

Notice that the world state and data are distinct nodes in in the cycle. Most machine learning models rely on the assumption that the training data is accurate, rich and representative of the population, but this is often not the case. Data is a necessarily subjective representation of the world. The sample may be biased, contain an inadequate collection of features, subjective decisions around how to categorise features into groups, systematic errors or be tainted with prejudice decisions. We may not even be able to measure the true metric we wish to impact. Data collected for one purpose is often reused for another under the assumption that it represents the ground truth when it does not.

Stale data

Suppose a model trained on a dataset is used to make decisions but those decisions are never fed back into the model because the model is not retrained. Our model may continually impact the world state, but the resulting changes are not fed back into the model. What are the implicit assumptions one might be using here? Are there conditions under which this might be an acceptable practice? What are the factors that might affect the nature of the feedback. What makes this better, worse or no different to regular retraining?

2.1.1 Feedback from model to data

In cases where the ground truth assignment (target variable choice) systematically disadvantages certain classes, actions taken based on predictions from models trained on the data can reinforce the bias and even amplify it. Similarly, decisions made on the basis of results derived from machine learning algorithms, trained on data that under or over-represents disadvantaged classes, can have feedback effects that further skew the representation of those classes in future data. The cycle of training on biased data (which justifies inaccurate beliefs), taking actions in kind, and further generating data that reinforces those biases can become a kind of self-fulfilling prophecy. The good news is that just as we can create pernicious cycles that exaggerate disparities, we can create virtuous ones that have the effect of reducing them. Let's take two illustrative examples.

Predictive policing

In the United States, predictive policing has been implemented by police departments in several states including California, Washington, South Carolina, Alabama, Arizona, Tennessee, New York and Illinois. Such algorithms use data on the time, location and nature of past crimes, in order to determine how and where to patrol and thus improve the efficiency with which policing resources are allocated. A major flaw with these algorithms pertains to the data used to train them. It is not of where crimes occurred, but rather where there have been previous arrests. A proxy target variable (arrests) is used in place of the desired target variable (crime). Racial disparities in policing in the US is a well publicised problem. Figure 2.2 demonstrates this disparity for policing of drug related crimes. In 2015, an analysis by The Hamilton Project found that at the state level, Blacks were 6.5 times as Whites to be incarcerated for drug-related crimes[2] despite drug related crime being more prevalent among Whites. Taking actions based on predictions from an algorithm trained on arrest data will likely amplify existing disparities between under and over-policed neighbourhoods which correlate with race.

Car insurance

As a comparative example, let's consider car insurance. It is well publicised that car insurance companies discriminate against young male drivers (despite age and gender being legally protected characteristics in the countries where these insurance companies operate) since statistically, they are at higher risk of being involved in accidents. Insurance companies act on risk predictions by determining the price of insurance at an individual level - the higher the risk, the more expensive the cost of insurance. What is the feedback effect of this on the data? Of course young men are disadvantaged by having to pay more, but one can see how this pricing structure acts as an incentive to drive safely. It is in the drivers interest to avoid having an accident that would result in an increase in their car insurance premiums. For a high risk driver in particular, an accident could potentially make it prohibitively expensive for them to drive. The feedback effect on the data would be to reduce the disparity in incidents of road traffic accidents among high and low risk individuals.

Along with the difference in the direction of the feedback effects in the examples given above, there is another important distinction to be made in terms of the magnitude of the feedback effect. This is related to how much control the institution making decisions based on the predictions, has over the data. In the predictive policing example the data is entirely controlled by the police department. They decide where

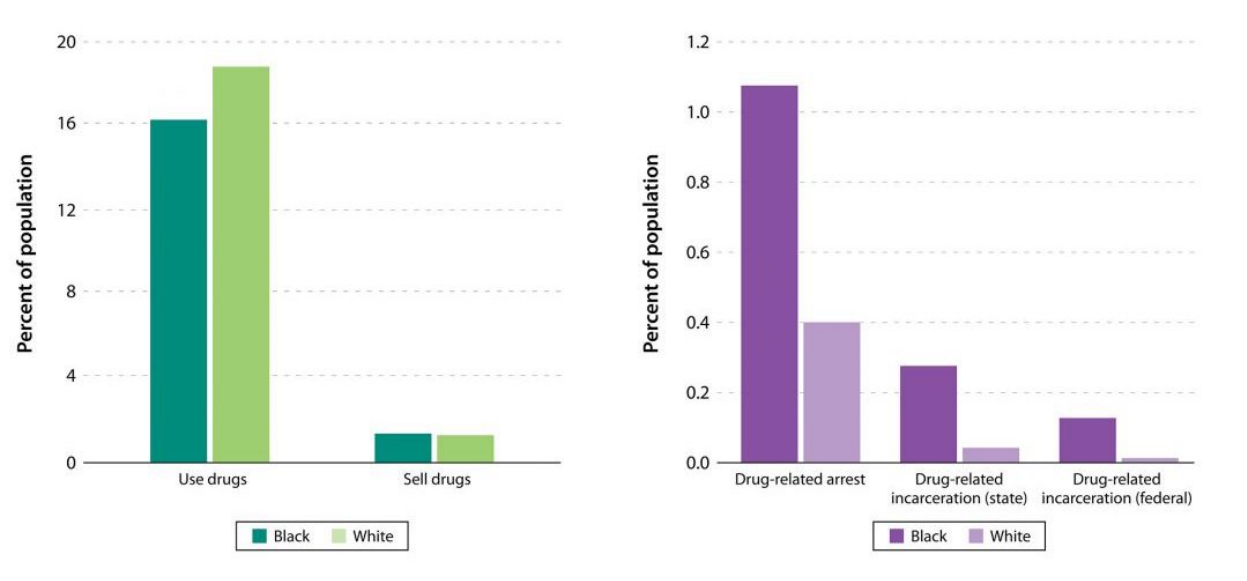


Figure 2.2: Rates of drug use and sales compared to criminal justice measures by race[2].

to police and who to arrest, ultimately determining the places and people that do (and don't) end up in the data. They produce the training data, in its entirety, as a result of their actions. Consequently, we would expect the feedback effect of acting on predictions based on the data to be strong and capable of dramatically shifting the distribution of data generated over time. Insurance companies by comparison, have far less influence over the data (consisting individuals involved in road traffic accidents). Though they can arguably encourage certain driving behaviours through pricing, they do not ultimately determine who is and who is not involved in a car accident. As such, feedback effects of risk-related pricing in car insurance are likely to be less strong in comparison.

Risk related pricing and discrimination

Do you think age and gender based discrimination in car insurance are fair? Why?

2.1.2 Model use

We've seen some examples illustrating how the strength and direction of feedback from models to (future) data can vary. In this section we'll demonstrate how the same model can have a very different feedback cycle depending on how it is used (i.e. the actions that are taken based on its predictions). A crucial part of responsible model development and deployment then should be clearly defining and documenting the way in which a model is intended to be used and relevant tests and checks that were performed. In addition, considering potential usecases for which one might be tempted to use the model but for which it is not suitable and documenting them can prevent misuse. Setting out the specific use case is an important part of enabling effective and focused analysis and testing in order to understand both its strengths and weaknesses.

The idea that the use case for a product, tool or model should be well understood before release; that it should be validated and thoroughly tested for that use case and further that the potential harms caused (even for unintended uses) should be mitigated is not novel. In fact, many industries have safety standards set by a regulatory body that enshrine these ideas in law. The motor vehicle industry has a rich history of regulation aimed at reducing risk of death or serious injury from road traffic accidents that continues to evolve today. In the early days, protruding knobs and controls on the dash would impale people in collisions. It was not

until the 1960s that seatbelts, collapsing steering columns and head restraints became a requirement. Safety testing and requirements have continued to expand to including rear brake lights, a variety of impact crash tests, ISOFIX child car seat anchors among others. There are many more such examples across different industries but it is perhaps more instructive to consider an example that involves the use of models.

Let's look at an example in the banking industry. Derivatives are financial products in the form of a contract that result in payments to the holder contingent on future events. The details, such as payment amounts, dates and events that lead to them are outlined in the contract. The simplest kinds of derivatives are called vanilla options; if at expiry, the underlying asset is above (call option) or below (put option) a specified limit, the holder receives the difference. In order to price them one must model the behaviour of the underlying asset over time. As the events which result in payments become more elaborate, so does the modelling required to be able to price them, as does the certainty with which they can be priced. In derivatives markets, it is a well understood fact that valuation models are product specific. A model that is suitable for pricing a simple financial instrument will not necessarily be appropriate for pricing a more complex one. For this reason, regulated banks that trade derivatives must validate models specifically for the instruments they will be used to price and document their testing. Furthermore they must track their product inventory (along with the models being used to price them) in order to ensure that they are not using models to price products for which the are inappropriate. Model suitability is determined via an approval process, where approved models have been validated as part of a model review process to some standard of due diligence has been carried out for specified the usecase.

Though machine learning models are not currently regulated in this way, it's easy to draw parallels when it comes to setting requirements around model suitability. But clear consideration of the use case for a machine learning model is not just about making sure that the model performs well for the intended use case. How a predictive model is used, ultimately determines the actions that are taken in kind, and thus the nature of the feedback it has on future data. Just as household appliances come with manuals and warnings against untested / inappropriate / dangerous uses, datasets and models could be required to be properly documented with descriptions, metrics, analysis around usecase specific performance and warnings.

It is worth noting that COMPAS[12] was not developed to be used in sentencing. Tim Brennan (the co-founder of Northpointe and co-creator of its COMPAS risk scoring system) himself stated in a court testimony that they "wanted to stay away from the courts". Documentation[18] for the software (dated 2015 two years later) describes it as a risk and needs assessment and case management system. It talks about it being used "to inform decisions regarding the placement, supervision and case management of offenders" and probation officers using the recidivism risk scales to "triage their case loads". There is no mention of its use in sentencing. Is it reasonable to assume that a model, developed as a case management tool for probation officers could be used to advise judges with regards to sentencing? Napa County, California, uses a similar risk scoring system in the courts. There a Superior Court Judge who trains other judges in evidence-based sentencing cautions colleagues in their interpretation of the scores. He outlines a concrete example of where the model falls short. "A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job. Meanwhile, a drunk guy will look high risk because he's homeless. These risk factors don't tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be."[12]

Propublica's review of COMPAS looked at recidivism risk for more than 10,000 criminal defendants in Broward County, Florida[11]. Their analysis found the distributions of risk scores for Black and White defendants to be markedly different, with White defendants being more likely to be scored low-risk - see Figure 2.3. Comparing predicted recidivism rates for over 7,000 of the defendants with the rate that actually occurred over a two-year period, they found the accuracy of the algorithm in predicting recidivism for Black and White defendants to be similar (59% for White and 63% for Black defendants), however the errors revealed a different pattern. They found that Blacks were almost twice as likely as Whites to be labelled as higher risk but not actually re-offend . The errors for White defendants were in the opposite direction; while being more likely to be labelled as low-risk, they more often went on to commit further crimes. See Table 2.1. How might different use cases for the model affect the feedback cycle? Let's consider some different usecases.

In the courts, the COMPAS recidivism risk score has been used by judges as an aid in determining

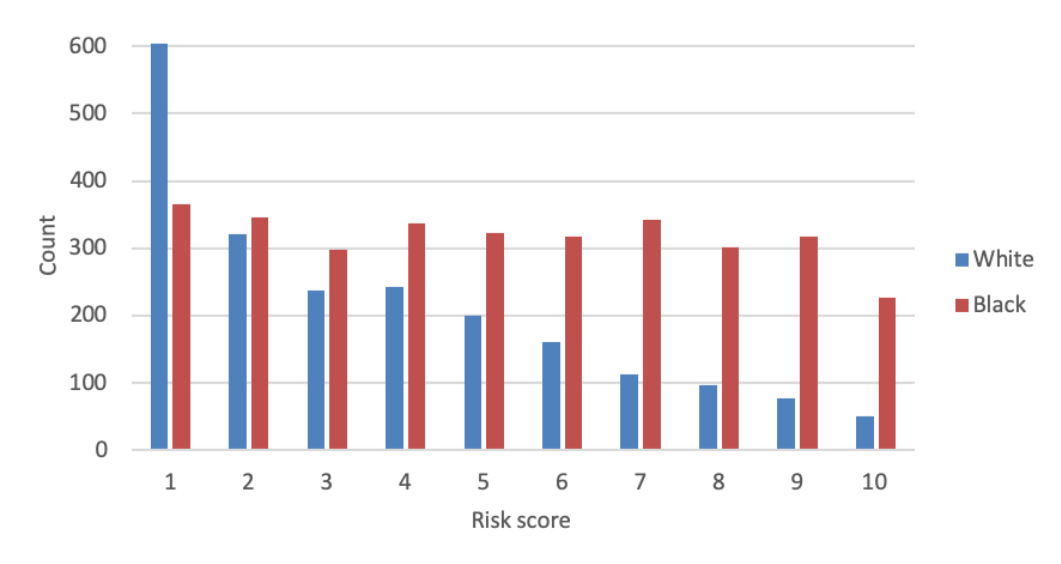


Figure 2.3: Comparison of recidivism risk scores for White and Black defendants[11]

Table 2.1: COMPAS comparison of risk score errors for White versus Black defendants

Error type	White	Black
Labelled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labelled Lower Risk, But Did Re-Offend	47.7%	28.0%

sentence length - the higher the risk, the longer the sentence. Of course being incarcerated limits ones ability to reoffend but unless the sentence is life, release is inevitable. What impact does a longer sentence have on recidivism? Current research suggests that "The longer and harsher the prison sentence – in terms of less freedom, choice and opportunity for safe, meaningful relationships – the more likely that prisoners' personalities will be changed in ways that make their reintegration difficult and that increase their risk of re-offending"[10]. Now in addition to this consider that as a Black defendent, you are more likely to be incorrectly flagged as high risk. If there was no racial disparity in recidivism rates in the data, we could expect the imbalance in errors to create one. What about crime rates - how do longer sentences impact those? Research shows that it is the certainty, rather than severity of punishment that acts as a deterrent to crime[17]. Long-term sentences are particularly ineffective for drug crimes as drug sellers are easily replaced in the community[14]. On balance, excessive incarceration has negative consequences for public safety because finite resources spent on prison are diverted from policing, drug treatment, preschool programs, or other interventions that might produce crime-reducing benefits.

Reducing incarceration rates

The US has the highest rate of incarceration in the world, at 0.7% of the population[22]. It's higher than countries with authoritarian governments, those that have recently been locked in civil war and those with murder rates more than twice that in the US. Comparing with countries that have stable democratic governments, the incarceration rate in the US is more than 5 times that of its closest peer - the UK. The US spends $57 billion a year on housing more than 2.2 million people in prison[13], almost half of which are private companies that spend significant sums on lobbying the federal government for policies that would further increase incarceration. Some have advocated for the use of risk scores in sentencing in order to reduce the rate of incarceration, the idea being that if the risk scores are low then defendants can be spared prison time. What might the feedback effect be for this usecase? What is the impact of the imbalance in error rates? What assumptions are you making to reach this conclusion?

Alternatively, suppose the software was used as a way to distribute limited rehabilitation resources, allocating them to those defendants that that were deemed to be at the highest risk of re-offending (and thus the most in need of intervention). Assuming the model to be accurate and that rehabilitation decreased the risk of reoffending, we can expect that using this model would serve to reduce existing disparities in recidivism rates between individuals. What about the imbalance in errors? Black defendents would more often erroneously be allocated rehabilitation resources and white defendents erroneously denied.

We have made numerous assumptions in our analysis of the feedback above; rehabilitation consistently reduces the risk of recidivism (regardless of the crime), that the relationship between sentence length and recidivism risk is monotonic and increasing. That two years is a long enough time horizon to consider. Without getting into the weeds, the point here is simply that the same model can have a very different feedback cycle if used in a different way. How a model is used is important and its *performance* cannot be evaluated in isolation from its usecase. A question to ask is, does the action taken on the back of the model serve to push extremes to the centre, or push them further apart? The relationships you have to understand to answer the question, will depend on the specifics of the problem.

2.2 Model development and deployment life cycle

In this section we cover the more practical aspects of ethical model development and deployment. We take a take a higher level view of the process by which machine learning systems are created and identify the stages at which we can build in safety considerations. We take inspiration from model risk management in finance where models are ubiquitous. In banking, processes and policies with regard to development, testing, documentation, review, monitoring and reporting of model related valuation risk, have been developed over decades, alongside regulation. Many of the ideas we discuss in this chapter were developed and implemented after the 2008 credit crisis in an effort to improve controls around valuation model risk for derivative products (more on this later).

Before we think about identifying and categorising common causes of harm in machine learning applications, it will be helpful to outline the workflow through which machine learning models might be developed and deployed responsibly. Figure 2.4 does exactly this.

2.2.1 Model governance standards

At the top, overarching the entire workflow, we have the model governance standards. These essentially outline the processes, roles and responsibilities that constitute the development, deployment and management of the machine learning system. It defines and documents a set of standards for the activities that constitute each stage of the depicted workflow. More on this later.

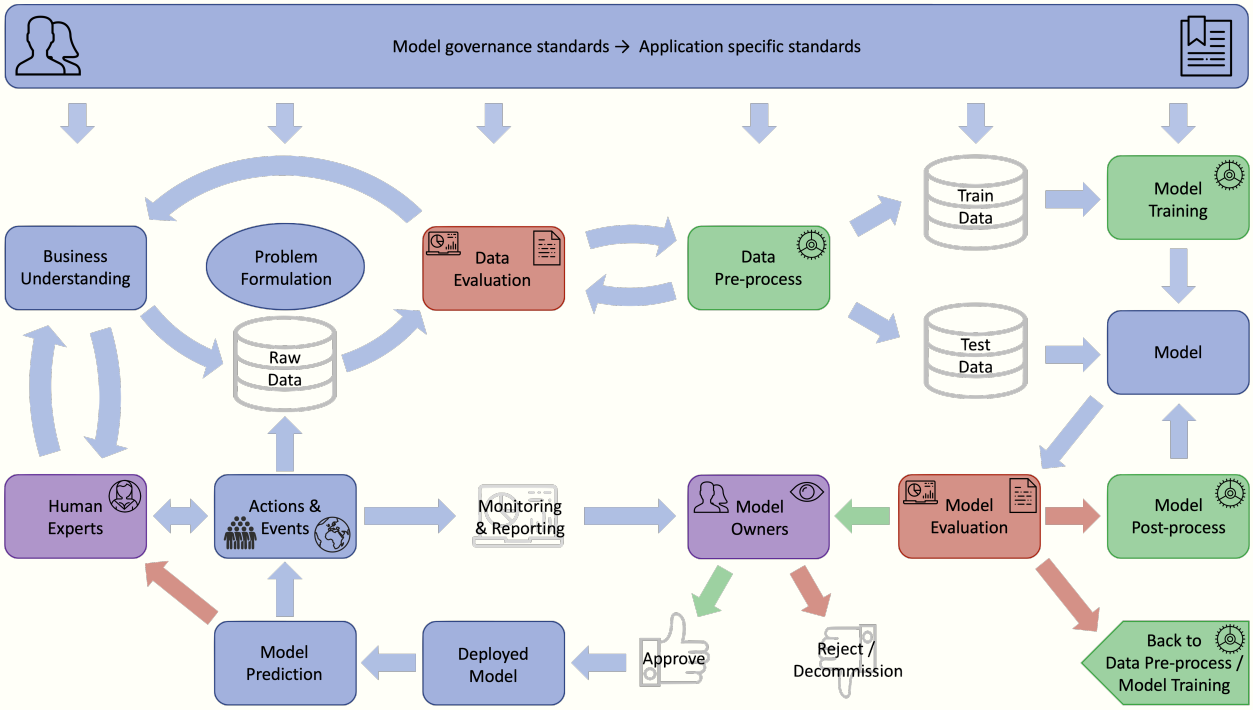


Figure 2.4: Fairness aware machine learning system development, deployment and management workflow.

2.2.2 Problem formulation

Below this, the life cycle of a machine learning system starts in top left corner with the formulation of the problem. This segment of the development process includes setting objectives, gathering data, analysing and processing it, determining a target variable, relevant features and metrics that indicate success (and failure) of the model (in training, evaluating and monitoring the deployed model). This process should include consulting with experts in the problem domain. The goal here is to understand the problem, data and impact of potential solutions for all the stakeholders. The arrows show that the problem formulation process is an iterative one where ideally domain experts, data collection and processing all inform each other in the creation of a tractable machine learning problem.

An assessment should be made with regards to how appropriate the data is for the model use case. Understanding the provenance of the data (who collected it, how it was collected and for what purpose) is important. Is it representative of the population the model built on it intends to serve? Exploratory data analysis (EDA) should include understanding if there is bias and or discrimination in the data. In particular understanding how is the target variable distributed for different subgroups of the population and what the nature of the resulting machine learning cycle might be for the intended and unintended use cases. Is there strong correlation between protected features and other variables?

Problem formulation should also consider the proposed materiality of the associated risk. What's the worst that can happen? How might the model be misused of misinterpreted? Would a disclaimer (what this model doesn't tell you...) be appropriate? How many individuals would be exposed to the model? Is the model within risk appetite (as defined in the model governance standards)? Having a way to understand and compare the risks posed by different models/applications is useful in ensuring the appropriate amount of resource and scrutiny is applied at all stages of the development, deployment and maintenance life cycle.

2.2.3 Model development

Once the problem is well understood and represented in the data the next broad segment is developing a model. This includes splitting the data into training and testing sets, evaluating the model against its objectives and consequently refining the data, model, evaluation metrics or other aspects. The splitting of data may be more complex, depending on the cross validation approach, but for simplicity we omit specific details in Figure 2.4. Part of model development and validation process should be to understand the model's limitations - where predictions might be unreliable, what it can and cannot be used for. The process of testing and analysing model output for performance should include analysis for discrimination and fairness. How are predictions and errors distributed for different subgroups of the population? How does the model output distribution differ from the training data? Again, model development is an iterative process and the data, metrics, training objectives, post-processing steps and more will evolve as the developers' understanding of the problem improves.

2.2.4 Model owners

For applications deemed ready for deployment, the documentation for the data and model analysis and implementation is submitted to the model owners for review. So who are these model owners? There are often many people involved in the development and deployment of a machine learning system (one would hope, at least two in general) and the model governance standards should specify which of them plays what role in deciding when a solution is ready to be deployed. Each of the model owners will have different (potentially conflicting) concerns. Model owners represent the different stakeholders of the risk associated with the model and collectively they are accountable, though for potentially differing aspects of it. These might include for example,

- **Product owners** that will use the system to make decisions.
- **Domain experts** that may have had input in the development of the solution (legal, domain or application specific council) and/or may be responsible for dealing with cases for which the model is deemed inappropriate (a radiologist for a pneumonia detector for example).
- **Model developers** that were involved in the construction of the model from collecting the data to building the model.
- **Independent model validators** that provide adversarial challenge around the modelling and implementation.
- **Engineers** that might be responsible for ensuring that infrastructure (for example, data collection, storage, post-deployment monitoring and reporting) requirements can be met.

2.2.5 Approval process

Together model owners determine if the model is approved for deployment or not. For the sake of brevity, and to emphasize the right of the model owners to reject proposed solutions, we describe the situation where the model is not approved, as it being rejected. In reality, rejecting a model need not mean that it is scrapped. Model owners may for example require further analysis or other changes to be made before it is resubmitted for approval. In any organisation, ideally the values, mission and objectives are well enough understood by the members, that a solution being scrapped at the last hurdle would be a rare event. The kinds of issues that would result in rejection should generally be caught at an earlier stage of the model development workflow. Model owners will also be responsible for monitoring the model post-deployment, periodic re-review of the risks and failure postmortums that determine what changes are required when issues arise, including amendments to the model governance standards themselves. The model governance standards might be interpreted as a contract between the model owners that describes their commitments, individually and collectively in managing the risk.

2.2.6 Management of deployed models

Ensuring the necessary reporting mechanisms are in place so the decision system can be monitored both for validity and exposure, should be a predeployment requirement. This kind of risk tracking can be used as a control, if say limits can be defined which reflect risk appetite. Limits might be set based on how well understood the risks associated with a product (the longer a model is monitored, the more information we have about it) are and what mitigation strategies might be in place, for example.

Importantly the post-deployment cycle of Figure 2.4 (like the machine learning cycle in Figure 2.1, at the start of the chapter) includes separate nodes for the model predictions and actions taken. Selbst et al.[20], describe five traps that one might fall into, even while attempting to create fair machine learning applications. In particular, they describe the *framing trap*, in which one might unwittingly ensure that an algorithm meets some narrow fairness criterion on outcomes or errors (over the *algorithmic frame*) but fail to consider its impact in the real world. For example, failing to be sufficiently transparent about the weaknesses of it which leads to it erroneously being prioritised over the judgement of human experts. Or we might fail to consider the longer term impacts on the sociopolitical landscape (over the *sociotechnical frame*) in determining something as complicated as fairness. If the actions taken off the back of the predictions include human judgement or interpretation, this should also be captured as part of monitoring the model. Are people using the model in ways that were not anticipated or is it having an adverse affect in some other way? Finally we include human experts in the loop again at the stage where predictions are acted upon. Human experts might for example be consulted in cases where the model is understood to produce less reliable predictions, or via an appeals process that is built into the decision system.

Processes and procedures for managing remedial work in the event of failures could be specified as part of the model governance standards. One of the issues with machine learning solutions is that when there are failures (say, a photo or sentence is labelled in an offensive way), the easiest response is an ad hoc rule based approach to 'fixing' the specific issue that occurred - the "if this, then do something else" solution, so to speak. But this kind of action isn't sufficient to address the root of the problem. Remedial work will typically require more resource and planning to fix. A failure should prompt a re-review. Having a more robust process around dealing with failures when they occur, should mean that not only is action is taken in a timely manner, but also that meaningful changes are made as a result of them and that work is appropriately prioritised.

Failure post-mortems that focus on understanding the weaknesses of the model governance process (not the failure of individuals) could also be a means for improving them. Once in production, periodic re-reviews of the model are a means to catch risks that may have been missed the first time around. The frequency of re-reviews can depend on the risk level of the model/application in question if these are being tracked.

2.2.7 Measuring fairness

Bias and fairness metrics are essentially calculated on data. There are two stages at which we'll be interested in measuring bias and or fairness in evaluating our machine learning system. The relevant nodes are coloured red in Figure 2.4.

1. **Model input:** The training data, during the *data evaluation* stage.
2. **Model output:** The predictions produced by our model, that is the *model evaluation* stage.

Our chosen fairness evaluation metrics calculated on the training data and model output will in general not be the same. By comparing the two, we can evaluate how well the model is replicating relationships in the data.

2.2.8 Bias mitigation techniques

There are three stages at which one can intervene in the development of machine learning model mapping to mitigate bias and they are categorised accordingly. Relevant nodes coloured green in Figure 2.4.

1. **Pre-processing** techniques modify the historical data on which the model is trained (at the *data pre-process* stage).
2. **In-processing** techniques alter the training process or objective (at the *model training* stage).
3. **Post-processing** techniques take a trained model/s and modify or combine the output (at the *model post-process* stage).

2.3 Responsible model development and deployment

In this section we examine a fairness aware development, deployment and management policies for a sociotechnical system. For the most part, the ideas are similar to those concerned with effective model risk management; one that acknowledges that models are fallible and accordingly sets standards for development, deployment, monitoring and maintenance. The intention being, to prevent foreseeable failures and mitigate the associated risks. The main difference is that we consider ethical risk as a central component of the risks that must be managed. Of course utility is an important consideration in being fair (it's hard to imagine a model that is no better than guessing, as being fair) but utility does not guarantee fairness. Viewing model evaluation through an ethical lens requires a more holistic assessment of the system, it's purpose, reliability and impact; not just for the business, but for all those exposed to or affected by it and society at large.

We'll address some of the problems that can't be solved through the kinds of model mapping interventions we'll talk about in this book. Another fair machine learning trap described by Selbst et al.[20] is the *formalism trap*, in which one fails to account for the full meaning of complex social concepts, such as fairness, which can't be formalised with mathematical equations. In chapter 3 we'll show that under such formalisms, a universally fair classifier is precluded by irreconcilable definitions. Fairness might more naturally be established *procedurally* (as often it is in law). Furthermore, social concepts are deeply *contextual*, and thus do not lend themselves well to abstraction (a core principal in mathematics which enables portability of solutions). Social concepts evolve over time, as cultural norms shift, therefore *contestability* is key, as it provides an avenue for change and challenge. These are qualities of a system rather than an equation and cannot be resolved through algorithmic interventions. They require people to do the right thing, and for organisations to define what they consider the right thing to be.

2.3.1 Policy

In industry, where innovation demands taking risks and time is money, how do we ensure the proper amount of care and attention is applied when creating products that have the potential for harm? Historically, the answer has been to impose rules that slow the process down, by requiring steps which prioritise safety over other concerns. In order to do this, one must first determine and define a safety standard. In Figure 2.4, overarching the whole process is a set of model governance standards. These essentially define that standard. They describe the process through which systems are developed and approved for deployment, and the standard to which systems are tested and evaluated.

In the financial sector, major banks (that are considered to be of systemic importance to a nations financial stability) are subjected to greater scrutiny by the central bank and regulators. An example of this might be requiring them to publish results of solvency stress tests. The currency might be social rather than financial for sociotechnical systems but the principal should be the same.

Prioritisation

Products which are of systemic importance to the sociopolitical landscape should have sufficient and appropriate resources (relative to those of the risk generating activities) to manage and mitigate their ethical risk. For applications that carry high risk of harm, risk functions should act as gatekeepers for model deployment and use.

Model governance standards

Though relatively new terminology in machine learning circles, the concept of model governance has existed for decades. For large financial institutions (which depend on vast numbers of proprietary models), operating and maintaining a model governance framework is a central part of model risk management and a regulatory requirement. The regulatory landscape of the financial sector is considerably more mature than that of other industries and the frameworks used to handle the associated risks have been developed and refined over time. It is therefore instructive to look at how such institutions manage their model risk and consider how these might be applied to sociotechnical systems.

So what does responsible and ethical machine learning development and deployment look like? In reality there is no one size fits all answer. As we've noted before, sociotechnical systems are context dependent. The answer can depend on a whole multitude of factors.

- **Domain:** Different domains will have different legal and ethical concerns for example employment versus say social media.
- **The number and complexity of the models being used by the business:** A large organisation that uses or tests hundreds of models and composes them to make decisions and create new products (such as Microsoft) would benefit greatly from infrastructure and methodologies for measuring the materiality of the associated risks that would enable prioritisation of work related to mitigating them. In contrast, for a business based on a single model that automates a specific task (such as taging images), this would be less of a concern.
- **Cost of errors:** Where the stakes are high, for example self driving cars, pre-deployment testing will need to be extensive and prescribed in order to reduce the probability of making mistakes. Well defined and mandatory processes will play an important role - checklists, contingency planning, detailed logging for postmortems and more. For these types of applications we would want authority over model use to be distributed to risk functions which determine when the product is approved for deployment and have the power to decommission them. For a wake word detector (think "Hey Siri", "Okay Google" and "Alexa") a lower standard would be accepted by most.

Given this, how does one approach the problem of responsible development? Step zero is to create a set of model governance standards, the purpose of which is to clearly define and communicate what responsible model development and deployment looks like for your specific application, use case, domain, business, principles and values.

What are the kinds of questions we might want our model governance standards to answer?

- Why is the work important? What kinds of events or uses of your models are you trying to avoid (or are outside of the organisation's risk appetite)? What legislation is the company subject to? What are the consequences of failures? What are the values of the company that you want to protect?
- Who is responsible? What are the roles that must be fullfilled to deploy monitor and manage the risks. Who are the stakeholders or model owners and what is their remit? Who is accountable?
- What are model owners responsible for? What technology is covered by the standard. What kind of expertise are required to be able to report, understand and manage the risks? What are the questions each stakeholder must answer? What are the responsibilities of those experts at the various stages of the model development and deployment life cycle? What authority do they have in relation to determining if the model is fit for deployment? Who decides what?
- How do you manage the risk? What are the rules, processes and requirements that ensure the companies values are maintained, people are treated fairly, the legal requirements are fulfilled and risks are appropriately managed? How do the stakeholders work together? For example some roles might need to be independent while others work alongside one another. What are the requirements around training data (documentation, review, storage, privacy, consent and such)? What are the requirements around modelling (documentation, testing, monitoring and such)? What are the processes around proposing, reviewing, testing, deploying, monitoring model related risks? For example, frequency of risk reviews, forums

for discussion and monitoring. What are the processes and requirements in place for (specific foreseeable types of) failures? Are there stakeholder specific templates or check-lists that ensure particular questions get answered at specific points in the model development and deployment life cycle?

The list of questions above is by no means exhaustive but a good starting point. Creating a set of model governance standards is about planning. Machine learning systems can be complicated and have many points of failure: problem formulation, data collection, data processing, modelling, implementation, interpretation. The only way to reduce the risk of failures is to be organised, deliberate and plan for them. Creating a set of standards does exactly that. Where the systems we build have real world consequences, the preparation, planning and process around development, review, analysis, deployment and monitoring of them should reflect that. Ensuring that the right questions get asked at the right time, knowing who is responsible for answering them and being prepared to address problems is a core part of developing and deploying models ethically.

Finally, we note that the benefits of having excellent model governance standards with well defined goals, processes, roles and responsibilities won't be realised if in practice they are not followed. In large organisations, consistency can be a challenge. The role of internal audit is to provide objective feedback on the risks, systems, processes and compliance at an executive level. From a model governance perspective the role of auditors is to ensure that there are good processes in place and that the processes are being followed. Internal audit's role is independent of the business up to the executive level. All functions within the business are required to cooperate with internal auditors and provide unfettered access to information requested. Internal audit does not contribute to the improvement of or compliance to processes directly. Their role is to , assess and report back to senior leadership. In a risk management context, internal audit are considered to be the *third line of defence*. We shall come to the first and second lines shortly.

Risk assessment

In order to manage risk it must be identified. Any algorithm, no matter how simple, carries the risk of implementation errors or bugs and thus should at the very least be subject to unit testing and independent code review before being deployed. For organisations with more complicated risk profiles, an important component of managing risk is having a system to measure and track it. Having a way to compare risk level across products and or product classes, even if comparisons are coarse, enables some degree of risk appropriate prioritisation and resource allocation in managing them. Risk can be estimated in many different ways and exactly how it is measured will depend on the details of the application. Broadly speaking it should consider both the severity of the event and liklihood. What's important is not the exact value but rather the ability to compare risks across products, applications or indeed any other lines along which a business is organised. Metrics that capture things like the scale on which the model is being used, utility, training data quality/representiveness, model complexity, potential for harm and more could potentially be used to coarsely judge the risk posed by different applications. Model governance standards can define risk bands or metrics if they are application specific enough.

2.3.2 Risk controls

In this section we return to the workflow and see how the policies, discussed above, feed into the development, deployment and management of a decisions system. Problem formulation is the first key step in developing a machine learning solution and an especially pivotal one in ethical risk assessment. The problem formulation stage plays perhaps the largest role in determining what the end product will actually be. It is the stage at which the model objectives, requirements, target variable and training data are determined.

Deployment bias

As part of problem formulation one should examine the machine learning cycle in the context of the biases in the data and consider the nature (direction and strength) of the feedback of resulting actions on future data. It's important to consider other ways in which the model might be used (other than that intended)

and understand the feedback cycle in those cases. How the model might be misused/misinterpreted? Are there ways in which it should not be used? Documenting these types of considerations is an essential step in preventing deployment bias; that is, systematic errors resulting through inapproriate model use or misinterpretation of model results. As creators of technologies which affect society at large, documenting our work might be interpreted as a civic duty. We consider documentation to be an essential part of a dataset and model without which it is incomplete and potentially harmful. As such we classify lack of documentation as a model issue.

Repurposing data of models is a risky thing to do and is often the source of bias in models. A good example of this was uncovered by researchers from Berkeley in 2019. They discovered racial bias in an algorithm used to make important health-care determinations for millions of Americans [19]. The algorithm was being used to identify patients that would benefit from high-risk care management programs, which improve patient outcomes and reduce healthcare costs for patients with complex healthcare needs. The researchers found that Black patients who had the same risk scores as White patients were far less healthier and thus less likely to be selected for the programs. The bias was the result of data documenting healthcare costs being used to predict healthcare needs.

A thorough examination of ethical issues demands consideration of a diversity of voices, which is well known to be lacking in technology. This is the stage at which it is important to consider who is affected by the technology, consult with them and ensure their views are incorporated in the understanding of the problem and design of a potential solution. Who are the human experts? People who would have valuable insight and opinions on the potential impact of the model you plan on building? Who does the model advantage and who does it disadvantage? Want to use machine learning to help manage diabetes? What are the interests of the health insurance company funding the development? Have you consulted with diabetics in addition to specialist physicians? What are their concerns? What is the problem from the different perspectives? Would a model be able to help or are there simpler solutions?

Independent model validation

In any system that is vulnerable to costly errors, unit testing and pre-deployment independent review is a well established method of preventing costly foreseeable failures. Whether it's a completely new solution built from scratch or a modification to an existing solution that's being deployed, an independent review process is an important element of responsible model development. Below we describe the responsibilities of two separate roles, the model developers and the model validators.

The model developers role is to translate the business problem into a tractable machine learning problem and create a solution. They will work with the business and receive input from other necessary domain experts relevant to the application to develop a possible solution. This will include tasks such as acquiring and interpreting data that is relevant for the problem, determining a target variable, model objectives, performance measures, fairness measures and more. In terms of preventing failures, model developers are considered the *first line of defence*. The responsibility of developing a model responsibly lies, in the first instance, with them. The model developers should aim to create a model they believe to be production ready and more specifically, fulfill the requirements specified in the model governance standards.

As part of the pre-deployment process, the model should be reviewed. Model validators will have a similar skill set to model developers but their goal is different to that of the model developers. Where the developers primary objective is to create a solution to the business problem that meets a standard which will be approved by model owners, the role of a model validator is to critique that solution and expose problems with it - the more the better. Their role is to adversarially challenge the solution. They might challenge performance claims (error, bias, fairness) by changing the data or metrics, or demonstrate problems with the model by comparing with an alternative solution. The goal is to expose model weaknesses and demonstrate the limits of its validity in testing and documentation. The model validator might devise mitigation strategies for identified risks. Such strategies might include setting model usage limits (that might trigger a re-review for example) or additional monitoring requirements. They might for example identify additional cases when human review might be required or reject the proposed solution entirely if the problems with the model are great enough. The role of the reviewer could be thought of as something akin to a hacker but with the

advantage of having the keys in the form of model documentation (provided by the developers). The model reviewer in pre-deployment can act as a gatekeeper.

Note that in our terminology, the model is simply a mapping. It need not be learned by calibration to historic data. Any algorithm where the decision being made is important enough should be treated as such and proper precautions should be taken. For an algorithm which will be used in production, no matter how simple, this should mean being subject to code review and unit testing that demonstrates its validity in some well chosen cases. A good example of where this would have been valuable came up in Decemeber 2020 when a bug in an algorithm, meant that Stanford Hospital Residents were not correctly prioritised for the COVID-19 vaccine, despite working with COVID-19 patients daily. The algorithm did not apparently account for the fact that Resident doctors had a blank 'location' field in the data. We might never know the details of how it was implemented and tested but it hard to imagine such a bungle passed any decent unit test.

The model review process acts as the *second line of defence*. To be effective, the model reviewer's role must be independent of the model developer's to some extent. What does independence mean? We mentioned the distinct goals of their roles and this is important. The validator should not drive the development of a solution approach or model but instead focus on critique. In reality, it's easy to see that the iterative nature of model development might mean that amendments adressing criticisms of the solution may get rolled into it's development at multiple stages, blurring the lines between critique and collaboration. From an efficiency perspective, it might make sense for the solution to be reviewed at several critical stages of the development process making the overall process indeed more collaborative. If there's a problem with the data that was missed, ideally the developer would want to fix it before going on to build and train a model on it. One of the challenges then is how to preserve independence between the roles, and ensure that the value of having adversarial criticism in preventing failures, is not lost in collaboration. How best to preserve independence will depend on the specifics and is something that should be determined within the model governance standards. In a bank, the model developers and validators are required (by the regulator) to serve under different business functions (the trading desk versus risk management). They have different reporting lines up to executive level, and work in physically separate locations.

Monitoring

Post-deployment monitoring is an important part of responsible model development and deployment. Analysis should not stop once the model is deployed. Decisions on what to monitor and necessary feedback mechanisms should be determined during development. It's important to understand if the model is performing in line with expectations (based on pre-deployment testing and analysis). Is the data coming out of the model more or less biased than the data going in? Distributional shifts should be of particular concern where the actions taken based on predictions have a strong impact on the composition of future data.

Domain expertise

In section 1.4 we spoke of the importance of domain knowledge in interpreting causal relationships in data. Consulting domain experts at the problem formulation stage can yield considerable ethical risk reducing benefits. Encorporating more diverse perpectives on a problem will surely result in a better design that will benefit a broader cross-section of society. Given that models are simplified representations of real world systems and we know that they will make errors, responsible development should build in processes for anticipating and dealing with such cases and, where appropriate, deferring to the judgement of a human expert.

2.4 Common causes of harm

There are many ways in which machine learning solutions can result in harm. In this section we present a taxonomy of common causes and provide examples. At the end of the section, we'll relate the causes in our taxonomy to the corresponding stages of the model development and deployment life cycle (discussed

earlier), indicating where consideration and intervention could prevent them from arising. The goal is for this to serve as a good starting point as a practical reference for developing fairer models. For practicing data scientists it could be helpful as a standard to compare our current practices against, avoid common pitfalls and hopefully help ensure we perform an appropriate level of due diligence before releasing our work. In our taxonomy, we aim to layout both the points at which issues arise and the various points at which one could assess and intervene. For this reason, the table may appear to contain duplications of the same problem viewed from different perspectives. This is intentional. Often different parts of an application are developed independently.[1] Taking this approach is beneficial since it provides multiple opportunities to see and remedy the same problems.

Before presenting this taxonomy, it's worth being clear that, in reality, there is no agreed upon terminology that describes the different types of issues that can arise or agreed upon framework for developing machine learning solutions that factor in ethical safety concerns (since regulation surrounding algorithmic decision systems is still in the process of being shaped). Indeed, developing one is the subject of recent research, [5], [21], [7], [16]. The word bias itself has many definitions and even in a given context can have multiple valid interpretations. Different practitioners would likely describe the same type of bias differently. Causes of bias in machine learning applications are often numerous and overlapping, thus difficult to attribute to a single source or prescribe a single solution for. The most appropriate remedy itself will be very much context dependent and different practitioners will choose different approaches.

In creating this taxonomy, we take inspiration from that described by d'Alessandro et. al.[5], in which the *model* or algorithm (function mapping f from features $(\boldsymbol{X}, \boldsymbol{Z})$ to predictions $\hat{Y}$), is distinguished from the larger *system* (people, infrastructure, processes, policies and risk controls) through which it is developed, deployed and managed. Evidence based medicine provides a rich terminology for different mechanisms through which systematic errors can be introduced in data and has perhaps the most comprehensive set of definitions and classification of bias types. This in itself can provide an important reference in determining which kinds of biases model developers should be aware of and we include some of them here. Table 2.2 summarises our taxonomy of common causes of harm in machine learning systems.

In section 2.3 we discussed a framework for responsible development and deployment of models. We summarise important elements of that discussion under **system issues** in our taxonomy of harms. The idea is that if having a process in place could avoid certain types of harms, then not having them is a failure of the system surrounding the model. In this section we discuss common causes of discrimination that relate directly to the model. We categorise these as originating from failures related to one of two sources:

1. **Data issues** refer to harms that arises as a direct result of issues with the data
2. **Misspecification** refers to harms that arise through misspecification of the underlying problem in the modelling of it.

The latter is an extension of the notion of model misspecification in statistics where the functional form of a model does not adequately reflect observed behaviour.

Before discussing our taxonomy for modelling issues, we address a point of contention in the machine learning community - that models are not biased, bias comes from data. The notion that bias is simply an artifact of data rather than a model is not uncommon among machine learning scholars and practitioners. In this book we've already discussed numerous examples of biased machine learning models, so where does this idea come from? In more theoretical disciplines a model is interpreted as being the parametric form. Under this definition of a model, different values of the parameters then don't change what we consider to be our model. For example, the term *linear model* describes a family of models. More practical disciplines view a model as a function mapping - provided with input, the model returns output. By this definition of a model, if the parameters change, so does the function and thus the model. From a practical perspective then it's clear that a model can discriminate since if the data documents historic discrimination, we would expect the trained model to reproduce it.

The idea that bias is a data problem, rather than a modelling one is at best a gross oversimplification of the problem and at worst misleading. It implies that in general, after training, a model will perfectly

[1] It's not uncommon for example (thanks to unprecidented growth in data markets), for a model to be built by one organisation, based on data collected by another.

Table 2.2: Taxonomy of common causes of harm in machine learning systems.

Element	Failure	Issue Type	Issue Description
System	Policy	Prioritisation	Failure to allocate appropriate/sufficient resource
			Failure to distribute power to manage conflicts of interest
		Governance	Failure to set or comply with application specific standards
		Risk assessment	Failure to identify and manage model related risk
	Controls	Deployment bias	Inappropriate model use / misinterpretation of model results
		Independent model validation	Data appropriateness and preparation
			Modelling approach and implementation
			Model evaluation metrics (pre and post deployment)
		Monitoring	Poor monitoring of model validity and impact
			Poor monitoring of risk exposure
		Domain expertise	Non deference to human domain expert
Model	Data	Historical bias	Data records wrongful discrimination
		Measurement bias	Quality of data varies across protected classes
			Measurement process varies across protected classes
			Recording proxies for immeasurable / ill defined variables
		Representation bias	Data not representative of target population
		Low support	Insufficient data for minority classes
		Documentation	Failure to adequately document
	Misspecification	Aggregation bias	Failure to model differences of type
		Target variable	Target variable subjectivity
			Proxy target variable learning
			Heterogeneous target variable
		Features	Inclusion of protected features without control variables
			Inclusion of protected feature proxies (redlining)
		Cost function	Failure to specify asymmetric error costs
			Omitted discrimination penalties
		Evaluation bias	Poor choice of evaluation metrics
			Test data not representative of the target
		Documentation	Failure to adequately document

reproduce the joint distribution of the variables in data. Anyone who's ever trained a model on real world data knows, is patently false. It suggests that models and data are independent when, in practice, they ought not be. Model development is an iterative process. The modelling choices we make can depend on the data and our model results should in turn influence our training data. Treating data and modelling as independent entities diminishes the responsibility of model developers in addressing the problem of biased and unfair applications. It ignores the very practical nature of developing models that serve real people and the societal impact they can have. For sociotechnical systems, the objectives must surely extend beyond utility. We consider defining those wider objectives and incorporating them part of the modelling process and thus failing to consider them a modelling problem.

2.4.1 Data issues

When it comes to bias, data driven medicine provides a rich vocabulary for the different types. We mention three here.

Historical bias

Historical bias arises as a result of differences between accepted societal values and cultural norms and those captured by data. These need not be a result of errors in the data. Even if data perfectly represents some world state, it can still capture a reality which society deems unfair. Training a model on such data will naturally lead to similary unfair predictions. Historical bias can manifest itself in data in numerous ways, through unfair outcomes recorded in the data, differing data quality across groups and under or over-representation of groups to name just a few. Take medical data where racial and gender disparities in diagnosis and treatment are well publicised as the *health gap*. There is a growing body of research across the US and Europe that exposes systematic under-treatment and misdiagnosis of pain in women ([4], [15], [9]) and Black patients (despite prescription drug abuse being more prevalent among White Americans), [8].

Measurement bias

Measurement bias refers to non-random noise in measurements across groups. This can occur if for example, there are geographic disparities in services provided by an institution or the quantity and quality of the measuring instruments that mean the accuracy and completeness of records vary by location (and other highly correlated variables like race). In some cases institutions can systematically fail to produce accurate and timely records for certain groups. For example, in medical data, where more frequent misdiagnosis of rare diseases for women leads to a longer lag before accurate diagnosis. In particular, 12 compared to 20 months for Crohn's disease (despite the disease being more prevalent among women) and 16 compared to 4 years for Ehlers-Danlos syndrome[1]. Systematic delays in diagnosis for protected groups mean that for any given snapshot in time, the medical records for more frequently misdiagnosed groups are less accurate.

Another way in which measurement bias can manifest is if the measurement process varies across groups, for example where the level of scrutiny varies across groups. Predictive policing discussed earlier provides an example of this where there are existing disparities in the level of policing across neighbourhoods. But in practice any process (algorithmic or otherwise) which seeks to identify a behaviour or property (good or bad), but where disproportionate attention is allocated to some subgroup will result in disproportionately more instances of that behaviour or property being observed among members of that group. The result is induced correlation in the data, even in cases where there may in reality be none. One must be careful of making the assumption that where no observation was made the behaviour or property did not exist. The result can be a cycle that continually amplifies the association. Since data often measures and records features which are in fact noisy proxies for the true variables of interest, measurement bias includes cases where use of proxies leads to systematic errors.

Representation bias

Representation bias occurs as a result of biased sampling from the target population. It can be observed as differences in the prevalence of groups when comparing the target population and the sample data. Under-represented classes are exposed to higher error rates; a problem which arises as a result of 'low support', that is a smaller pool of data points to train the model on. Looked at from the perspective of the majority class which dominates the aggregate error, the algorithm is naturally incentivised to focus learning characteristics of majority classes.

One of the drivers behind big data initiatives is the plummeting cost of collection and storage data. Companies and institutions are able to train models that better target individuals, reducing costs and boosting profits. However, data collection methods often fail to adequately capture historically disadvantaged classes of people that are less engaged in data generating ecosystems. A good example of this, given by Barocas & Selbst[3] is that of the phone app Street Bump, which was developed by the City of Boston to reduce the cost and time taken to find (and consequently repair) pot holes. The app uses data generated by the accelerometers and GPS of Boston residents' smart phones as they drive. Once a pothole is located it is automatically added to the city's system to schedule a repair. One can see easily see how this method of data collection might fail to adequately capture data from poorer neighbourhoods, where car and smart phone

ownership are less prevalent; neighbourhoods which probably correlate with race and are already likely to suffer from lack of investment.

In the extreme case of under-representation, there is no support, that is to say, no data points to train on at all. This can be a problem when say studies of symptoms or clinical trials for drugs have no representation for certain groups among which symptoms or drug effectiveness may well vary. A good example of this is diabetes, the impact of the disease and effectiveness of drugs for which have historically most often been measured on samples with few to no hispanic individuals in datasets at all.

Low support

Low support may lead to undesirably high errors for some groups even in the absence of representation bias, since minority classes naturally have fewer data points to train on. This is a particular problem for individuals belonging to multiple disadvantaged classes, for example Black women, which are often overlooked when studies seek to meet fairness metric targets.

Documentation

Documentation of datasets is an essential step in avoiding data misuse or misinterpretation of variables or relationships in the data due to lack of domain knowledge. Documentation should evidence that model governance standards were met. Summaries that explain the provenance of the data (who collected the data, for what purpose, what population was sampled from and how, limitations of the data, clear explanation of the target variables (including consideration of use cases for which it would not appropriate for), breakdown of the demographics and the variables by sensitive features pointing out classes that are not well represented. Documentation that is standardised through use of a template could ensure some level of consistency.

2.4.2 Misspecification

Aggregation bias

Aggregation bias occurs when heterogenous groups are modelled as homogeneous. In this case we are assuming the same model is approriate for all groups when in fact it is not, it is a failure to recognise differences in type. There are many examples of this is medical models for diagnosis or that measure the effectiveness of treatments. Historically much of medical research is based on data that over-represents White men. Diseases that manifest differently across gender or race are more often misdiagnosed or less effectively treated. Take autism spectrum disorder (AUD) for example, in 2016 research estimated that autism is four times more prevalent in boys than girls. However more recent research has suggested that a contributing factor maybe that autism more often goes undiagnosed in women because studies of the disorder have historically been focused on male subjects. The most notable difference between autistic males and females is how the social (rather than behavioural) symptoms manifest. It is thought that women, especially at the high-functioning end of the spectrum, are more likely to camouflage their symptoms.

Target variable selection

One of the challenges in developing a machine learning is the translation of the underlying problem by defining a target variable - something which can be observed, measured and recorded or obtained easily (from a third party vendor), and that accurately reflects the variable we wish to predict. While there are relatively uncontentious examples that machine learning solutions lend themselves well to (spam detection for emails or on-base or slugging percentage for major league baseball player valuation) for many problems the translation is non-trivial and subjective. Take a job applicant filter for example, that aims to find the most promising applicants. The attributes that one might consider to be held by an applicant that make them promising are likely to be described differently by different people even if they work in the same team. Even if two individuals agree on the attributes, it's likely they'll weigh the attributes differently based on

their experiences and preferences. Different choices will result in the different kinds of biases infiltrating our algorithm.

Often when data on the variable we want to affect doesn't really exist we use a proxy. In 2018, Amazon was forced to scrap a recruitment tool it spent four years developing. The algorithm rated resumes of potential employees and was trained on 10 years worth of resumes submitted by job applicants. The exact details of the algorithm were not publicised but based on the training data, it is likely that the proxy variable they used was some measure of how the candidates had performed in the hiring process previously. Thus predicting who they would have hired in the past (given their historical and existing biases) rather than who was the best applicant. The problem with such systems is that often they end up being how we define the thing that it's actually a proxy for.

Issues can also arise when defining a heterogeneous target variable, where a range of different events are coarsely grouped into a single outcome. This is a form of aggregation bias where the issue specifically concerns the target. This might happen for example where the event of particular interest is rare and by including more events in the target the predictive accuracy of the model increases as it has more data to learn from. D'Alessandro et. al[5] provide a useful example in predictive policing where the model developer is initially interested in predicting violent crime but ends up incorporating petty crimes (which happen much more frequently) in the target variable in pursuit of a more accurate model. The model then ends up trying to learn the features of a more nebulous concept of crime ignoring important differences between different types. Another example might be building a gender recognition system and only recognising people as one of two genders[6].

Feature selection

In an ideal world we would train a machine learning model on a sufficiently large dataset consisting of a rich set of features that actually influence the target variable rather than simply being correlated to it. More often than not, the reality is rather different. Comprehensive data can be expensive and difficult to collect. Factors that influence the target variable might not be easily measured or be measurable at all, while data containing more erroneous indicators might simply be cheaper to obtain or more readily available. This is a common way in which bias against protected classes can enter our model.

The inclusion of protected features without control variables might arise because a protected feature appears to be predictive of the target variable where explanatory variables are not known or available. Of course in cases where using protected characteristics as inputs to an algorithm would lead to disparate treatment liability, this is not a problem one is typically faced with, but it's worth reiterating the importance of controlling for confounding variables, in drawing conclusions about relationships between features from observational data (see section 1.4).

Inclusion of protected feature proxies, as is the case with redlining, is perhaps a more common problem. One where protected features are not used as inputs to the model, but features which are predictive of them are. Historically employers have taken the reputation of the university that applicants graduated from as a strong indicator of the calibre of the candidate. But many of the most reputable universities have very low rates of non-White/Asian students in attendance. A hiring process which is strongly influenced by the university from which the applicant graduated, can erroneously disadvantage racial groups that are less likely to have attended them. While the university an applicant graduated from, might correlate to some degree with success in a particular role, it is not in itself the driver. An algorithm that directly takes into account the skills and competencies required for the role would be more predictive and simultaneously less biased. Given the cost of collecting comprehensive data, one might argue that higher error rates for some classes would be financially justified (rational prejudice).

Cost function

A critical consideration in how we specify our model is the cost function. It is how we evaluate our model in training and essentially determines the model (parameters) we end up with. The cost function can be interpreted as an expression of our model objectives and so provides a natural route to addressing

discrimination concerns. A common failure in the design of classification models is proper accounting of the costs of the different types of classification errors (false negative versus false positives). If the harm caused by the different types of misclassification are asymmetric, the cost matrix should reflect this asymmetry.

More broadly (for both regression and classification), it is important to consider the contribution from each sample in the training data to the cost function in training. Upsampling (or simply up-weighting, depending on the learning algorithm you are using) is a valuable tool to keep in mind and can alleviate a number of the issues discussed above, that are common sources of bias. Let's take the issue of low support. By upsampling minority classes, one can increase the importance of reducing errors for those data points, relative to other more abundant classes, during learning. Though it's worth noting that it cannot resolve issues relating to a lack of richness of representation for classes with low support. Another case in which upsampling can help is that discussed in relation to definition of a heterogeneous target variable. By upsampling data points that correspond to the primary event of interest (violent crime in the example we discussed above), one can again increase the importance of the model fitting to those data points.

For an algorithm that solves a problem in a regulated domain, it would make sense for the absence of discrimination to be a model objective along with utility. This can be achieved by use of a penalty term in the cost function which relates to discrimination in the resulting predictions (just as we have terms that relate to the error or overfitting). Essentially the idea is similar to that of regularisation to avoid overfitting. We introduce an additional hyper-parameter to tune, which represents the strength of the penalty for discrimination in our cost. We will discuss this and upsampling in more detail when we discuss bias mitigation techniques, in part three of the book.

Evaluation bias

Evaluation bias arises when evaluating a model's performance. There are two main components here, the metrics chosen to describe the model's performance and the benchmark dataset on which they are calculated. Choosing either inappropriately will result in our evaluation metric inaccurately reflecting the efficacy of our model. For sociotechnical problems in particular choosing good metrics requires domain knowledge - the wider political, legal, social and historical context is important when defining what success and failure look like. For example, if building a gender recognition system, one should not simply think of the performance on the specific task but also the wider infrastructural systems which might find the technology useful. Where should we set the bar for such a technology? That should surely depend on how the technology is used after the prediction is made? Are there controls around model use? Should there be? What kinds of risk level does the model present? What might be the impact of the prediction being incorrect? When would an error be fair? What kind of examples would you expect your system to get wrong and why? What do they gave in common? Are they represented in the benchmark dataset? By asking these kinds of questions, when deciding what success looks like, it's hard to imagine thinking that minimising the mean squared error on a conveniently available dataset would be sufficient.

One approach might be to set accuracy thresholds across all (skin colour) phenotype and gender combinations [6]. This would be one way of thinking about success in a way that incorporates *some* of our societal values of equality. The gender recognition software we talked about in the previous chapter suffered from evaluation bias on both counts. The benchmark datasets used were not representative of the target population and the metrics that were chosen, failed to expose the models poor performance on darker skinned women. The problem of evaluation bias arising from poor choice of testing/benchmark data is often the result of trying to objectively compare performance across models and can lead to overfitting to said benchmark data.

Documentation

Documentation for models (as for datasets) can have a significant impact when it comes to avoiding model misuse (a model use it is not appropriate/approved for) and ensuring model limitations are well understood. It can reduce the risk of misinterpretation of variables as suitable proxies for other variables. Clear explanation of the model, testing that was performed, on what subgroups of the data can make it easier to know

which tests might be missing that would offer insight into the validity of the model. Documentation should evidence that the model governance standards have been met. Descriptions of the data and model, motivation behind subjective decisions that were made to arrive at the solution (how to process the data, what features were used/ignored and why, model type, cost function, sample weights, bias and success metrics), known data/model issues, how the model was tested, what it's limitations are, what it should and should not be used for with justification. Documentation of the model should provide enough detail to be able to re-implement the model, reproduce results and justify the solution approach. Documentation that is standardised through use of a template could ensure some level of consistency and efficiency across domains and applications. Recent research discusses the matter specifically for publicly released datasets[7] and machine learning models[16]. They suggest standardised analysis which for example demonstrates the performance of the algorithm for different subgroups of the population and requirements for proving efficacy for conjunctions of sensitive characteristics also.

2.5 Linking common causes of harm to the workflow

In Figure 2.5 we provide a visual summary of the taxonomy in Table 2.2, the goal being that it might be useful as a reference for teams developing machine learning technologies. Since failures of policy do not relate to any particular part of the model development and deployment life cycle but rather all of it, we omit these. At the top of Figure 2.5 we have a simplified version of the model development and deployment life

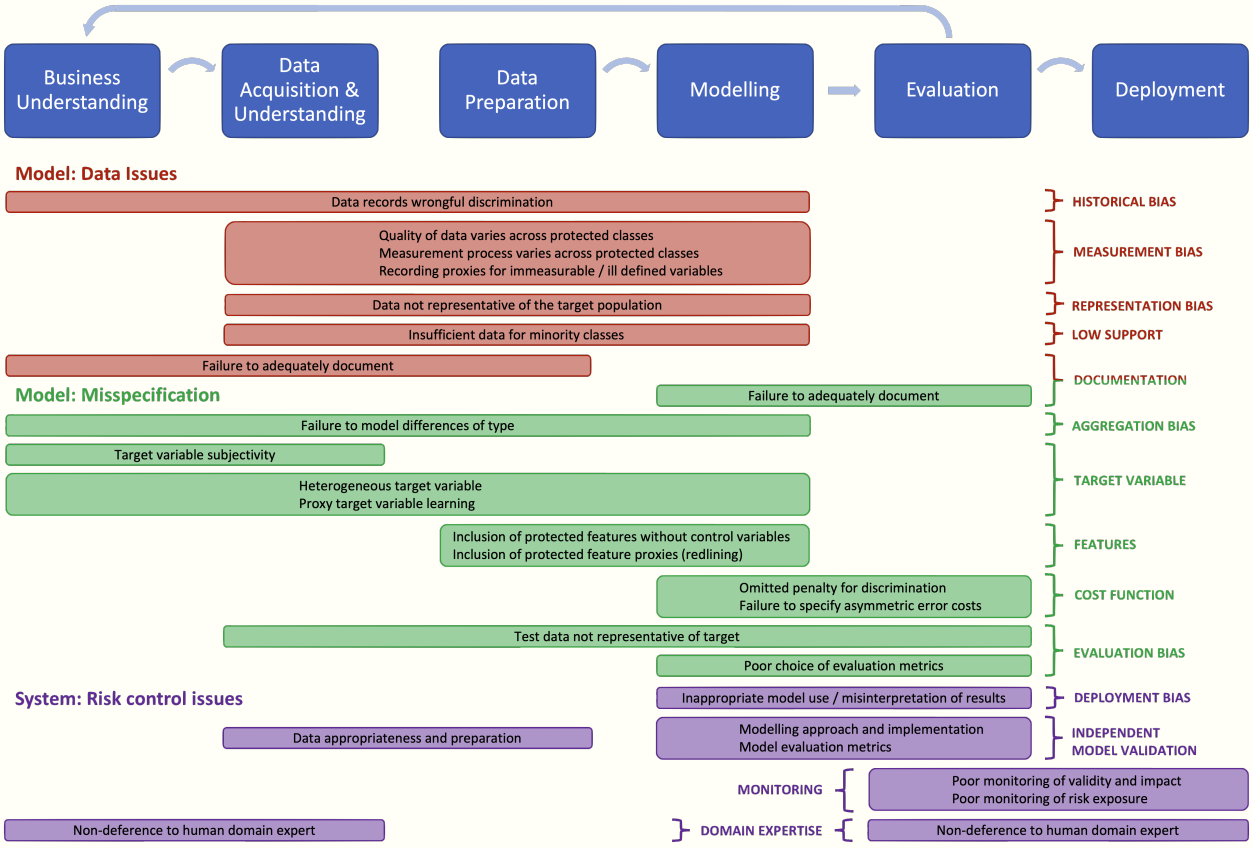


Figure 2.5: Taxonomy of common causes of bias in machine learning models together with the stages of the model development and deployment life cycle they relate to.

cycle. Below this, the causes of harm are displayed in boxes which span the parts of the lifecycle to which

they relate. We use colour to sepatate different categories of failures and curly brackets to group issues by type.

Summary

Machine learning cycle

- Machine learning solutions can have long-term and compounding effects on the world around us. Figure 2.1 illustrates the interaction between a machine learning solution and the real world.
- The translation of a given prblem and objectives into a tractable machine learning problem, requires a series of subjective choices. Choices around what data to train the model on, what events to predict, what features to use, how to clean and process the data, how to evaluate the model and what the decision policy should be will all determine the model we create, the actions we take and ultimately the cycle we end up with.
- Data is a necessarily subjective representation of the world. The sample may be biased, contain an inadequate collection of features, subjective decisions around how to categorise features into groups, systematic errors or be tainted with prejudice decisions. We may not even be able to measure the true metric we wish to impact. Data collected for one purpose is often reused for another under the assumption that it represents the ground truth when it does not.
- In cases where the ground truth (target variable) assignment systematically disadvantages certain classes, actions taken based on predictions from models trained on the data are capable of reinforcing and further amplifying the bias.
- Decisions made on the basis of results derived from machine learning algorithms trained on data that under or over-represents certain classes can have feedback effects that further skew the representation of those classes in future data.
- The actions we take based on our model predictions define how we use the model. The same model used in a different way can result in a very different feedback cycle.
- The magnitude of the feedback effect will depend how much control the institution making decisions based on the predictions, has over the data the training data.
- Just as we can create pernicious machine learning cycles that exaggerate disparities, we can also create virtuous ones that have the effect of reducing disparities. Therefore it's important to consider the whole machine learning cycle when formulating a machine learning problem

Model development and deployment life cycle

- Figure 2.4 depicts the model development, deployment and monitoring life cycle at a high level. Overarching the entire workflow, are the **model governance standards**. These essentially outline the processes, roles and responsibilities that constitute the development, deployment and management of the machine learning system. It defines and documents a set of standards for the activities that constitute each stage of the workflow.
- **Problem formulation:** Translating a business problem into a machine learning one.
    - The problem formulation stage plays a pivotal role in what the end product will actually be. It is the stage at which the model objectives, requirements, target variable and training data are determined and it is the stage at which perhaps the most important ethical question (whether the model should be built at all) must be answered.
    - Consider who is affected by the technology, consult with them and ensure their views are understood and incorporated in the understanding of the problem and design of a potential solution.
    - Assess the materiality of the risk. What's the worst that can happen? How likely is such a failure? How many people are exposed to the model?

- Examine the machine learning cycle in the context of the biases in the data and consider the nature (direction and strength) of the feedback of resulting actions on future data.
- Consider other ways in which the model might be used (other than that intended) and the corresponding feedback cycle in those cases. How the model might be misused?

- **Independent model validation:** An independent review process is an important element of responsible model development. This means that pre-deployment there are two separate data science roles, model development (designing a solution) and the model validation (critical assessment of the solution).
- **Model development:** The model developers role is to translate the business problem into a tractable machine learning problem and create a model solution.
  - The model developer will work with the business and receive input from other necessary domain experts relevant to the application to develop a possible solution.
  - The model developer should document the solution. Documentation should include descriptions of the data and model, justification of the approach, known issues and limitations, model testing (biases as well as performance), what the model should not be used for and why. Templates are a good way of standardising documentation.
  - In terms of preventing failures, the model developer is the *first line of defence*. The responsibility of developing a model responsibly and ethically lies, in the first instance, with them.
- **Model validation:** The role of a model validator is to criticise the proposed solution.
  - The model validator will identify and expose issues with the problem formulation, data and data processing. They will verify the model performance metrics (error, bias, fairness), look for model weaknesses and demonstrate them through testing. They may also devise mitigation strategies for identified risks.
  - The role of the reviewer might be thought of as a hacker but with the advantage of having access to the model documentation (provided by the model developer). They also act as a gate keeper.
  - The model reviewer must also document their analysis, testing and critique and recommendations regarding the solution.
  - The model reviewer acts as the *second line of defence*.
- **Model approval:** The model owners collectively determine if a solution is ready for deployment.
  - Model owners act as the final stage gate keepers before deployment. They will each have been involved in different aspects of the development and deployment of the machine learning system.
  - In effect, the model owners represent the different stakeholders of the risk associated with the model and collectively they are accountable, though for potentially differing aspects of it.
  - They will also be responsible for monitoring the model and risk materiality post-deployment and ensuring that periodic re-review, failure processes and post-mortems occur and are effective.
  - The model governance standards might be interpreted as a contract between the model owners that describes their commitments, individually and collectively in managing the risk.
- **Monitoring of deployed models:** The world is dynamic and the risk associated with models evolves with it. Deployed models should be monitored to understand if they are behaving in line with expectations. The metrics which should be reported to model owners should be identified pre-deployment by the model developer and validator.
- **Risk materiality tracking:** As model usage increases so does the associated risk. As part of monitoring, metrics that give an indication of the risk associated with the model is should be reported to the model owners.
- **Periodic re-review:** The pre-deployment independent review of the model is just the first. Thereafter, periodic re-reviews of the model are a means to catch risks that may have been missed the first time around. The frequency of re-reviews will depend on the risk level of the model/application in question.

- **Failure event process:** Processes and procedures in the event of failures should be specified as part of the model governance standards, in particular what steps should be taken by which model owner. Having a robust process around dealing with failures when they occur should mean that action is taken in a timely manner and that meaningful changes are made as a result of them.
- **Failure post-mortems:** A post-mortem should focus on understanding the weaknesses of the model governance process (not the failure of individuals) that contributed to it and appropriately prioritise any changes required to remedy them.
- **Measuring bias:** Bias and fairness metrics are essentially calculated on data; the data going into our model (training data) and the data coming out of it (the predictions produced by our model); the data evaluation and model evaluation stages.
- **Bias mitigation techniques:** There are three stages at which one can intervene to mitigate bias when developing a machine learning model labelled *data pre-process*, *model training* and *model post-process* in Figure 2.4. We categorise them accordingly:
    - **Pre-processing techniques** modify to the historical data on which the model is trained, the idea being that fair/unbiased data will result in a fair/unbiased model once trained.
    - **In-processing techniques** alter the training process or objective in order to create model with fairer/less biased predictions.
    - **Post-processing techniques** take a trained model and modify the output such that the resulting predictions are fairer/less biased.

Responsible model development and deployment

Model governance standards

- Machine learning systems can be complicated and have many points of failure: problem formulation, the data collection, data processing, modelling, implementation, deployment. The only way to reduce the risk of failures is to be organised, deliberate and plan for them. Creating a set of standards does exactly that. They make sure the right questions get asked at the right time and that there is clarity around who is responsible for what.
- The purpose of creating a set of model governance standards is to clearly define and communicate what responsible model development and deployment looks like for your specific application, domain, business, principles and values. It essentially documents and communicates the why, who, what and how of your model risk management approach.
    - **Why is the work important?** What kinds of events are you trying to avoid? What are the consequences of failures? What are the values of the company that you want to protect?
    - **Who is responsible?** Who are the stakeholders? Who is accountable for managing the various identified risks?
    - **What are they responsible for?** What are their roles/expertise? What authority do they have in relation to determining if the model is fit for deployment?
    - **How do you manage the risk?** What are the policies, processes and requirements that ensure the companies values are maintained, people are treated fairly, the legal requirements are fulfilled and the model risks are appropriately managed? How do the stakeholders work together?
- In large companies that carry lots of model risk it can be difficult to ensure there is consistency in standards of due diligence in model development and deployment across the board. The role of internal audit is to provide independent and objective feedback on the risks, systems, processes and compliance at an executive level. From a model governance perspective they determine if that there are good processes in place and that the processes are being followed. From a risk management perspective internal audit's role constitutes the *third line of defence*.

Common causes of harm

- Table 2.2 summarises the taxonomy of common causes of bias in a machine learning system.
- Figure 2.5 summarises common causes of bias in the context of the model development and deployment workflow, indicating both the stages of the workflow to which they relate and their categorisation within the taxonomy.

References

[1] The voice of 12,000 patients: Experiences and expectations of rare disease patients on diagnosis and care in europe, 2009. https://www.eurordis.org/IMG/pdf/voice_12000_patients/EURORDISCARE_FULLBOOKr.pdf.

[2] Rates of drug use and sales, by race; rates of drug related criminal justice measures, by race, 2015. Source: BLS n.d.c.; Carson 2015; Census Bureau n.d.; FBI 2015.

[3] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *Calif Law Rev.*, 104:671–732, 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

[4] Karen L. Calderone. The influence of gender on the frequency of pain and sedative medication administered to postoperative patients. *Sex Roles*, 23:713–725, 1990. https://link.springer.com/article/10.1007/BF00289259.

[5] Brian d' Alessandro, Cathy O'Neil, and Tom LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2):120–134, June 2017. https://arxiv.org/abs/1907.09013.

[6] Sorelle A. Friedler and Christo Wilson, editors. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, volume 81. Proceedings of Machine Learning Research, 2018. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2020. https://arxiv.org/abs/1803.09010.

[8] Kelly M. Hoffman, Sophie Trawalter, Jordan R. Axt, and M. Norman Oliver. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16):4296–4301, 2016. https://www.pnas.org/content/113/16/4296.

[9] Diane E. Hoffmann and Anita J. Tarzian. The girl who cried pain: A bias against women in the treatment of pain. *SSRN*, 2001. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=383803.

[10] Christian Jarrett. How prison changes people. *BBC Future*, May 2018. https://www.bbc.com/future/article/20180430-the-unexpected-ways-prison-time-changes-people.

[11] Jeff Larson. Propublica analysis of data from broward county, fla. Technical report, ProPublica, March 2016. https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb.

[12] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, March 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[13] Bryan Lufkin. The myth behind long prison sentences. *BBC Future*, May 2018. https://www.bbc.com/future/article/20180514-do-long-prison-sentences-deter-crime.

[14] Marc Mauer. Long-term sentences: Time to reconsider the scale of punishment. *The Sentencing Project*, November 2018. https://www.sentencingproject.org/publications/long-term-sentences-time-reconsider-scale-punishment/#:~:text=There%20are%20several%20reasons%20for,sentences%20add%20little%20to%20the.

[15] Esther H. Chen MD, Frances S. Shofer PhD, Anthony J. Dean MD, Judd E. Hollander MD, William G. Baxt MD, Jennifer L. Robey RN, Keara L. Sease MaEd, and Angela M. Mills MD. Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Academic Emergency Medicine*, 15:414–418, May 2008. https://pubmed.ncbi.nlm.nih.gov/18439195/.

[16] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, January 2019. http://dx.doi.org/10.1145/3287560.3287596.

[17] Daniel S. Nagin. Deterrence in the twenty-first century: A review of the evidence. *Crime and Justice*, 42, May 2018. https://www.journals.uchicago.edu/doi/abs/10.1086/670398.

[18] Northpointe. *Practitioners Guide to COMPAS Core*, March 2015. https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf.

[19] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453, 10 2019.

[20] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, January 2019. https://doi.org/10.1145/3287560.3287598.

[21] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. June 2021. https://https://arxiv.org/abs/1901.10002.

[22] Peter Wagner and Wendy Sawyer. States of incarceration: The global context. *Prison Policy Initiative*, June 2018. https://www.prisonpolicy.org/global/2018.html.

Part II Measuring Bias

“To measure is to know. If you cannot measure, you cannot improve it." Lord Kelvin.
“When a measure becomes a target, it ceases to be a measure." Goodhart’s Law

Chapter 3

Group Fairness

This chapter at a glance

- Group fairness concept and metrics
- Comparing different group fairness metrics
- Incompatibility of group fairness criteria

The term *group fairness* is used to describe a class of metrics that are used to measure discrimination or bias in a given decision process (algorithmic or otherwise). In this chapter we will introduce the different types of group fairness metrics in a structured way. We will become familiar with the terminology for well known metrics and we will compare and analyse them, in terms of their meaning and potential implications. We'll derive results that show how the various fairness metrics discussed, can in fact be incompatible in certain cases; that is to say, they cannot be satisfied simultaneously except in some degenerate cases. The goal of this chapter, is to develop a deeper understanding of group fairness and it's associated metrics, that will enable us to make more educated decisions about which metrics might offer valuable insights for a given problem.

Group fairness metrics all stem from the same high level notion of fairness; the idea that in a fair system, some *property* should be similar for different *subgroups* of a population. The *subgroups* are typically determined by the values of *protected characteristics* such as gender or ethnicity. We might also describe these as *sensitive features*. Partitions of the population could be defined by a single feature or logical conjunctions of multiple sensitive features if we are interested in measuring *intersectional* disparities. For example, if we were considering both race and gender simultaneously, one group of the partition might be Black women, another White men, and so on. The *property* we'll be interested in comparing will be some statistical measure; the particular kind, will depend on our beliefs about what fairness should mean in the context of the problem.

We broadly classify group fairness criteria into two types; those comparing *outcomes* across groups and those comparing *errors*. We discussed examples of both in chapter 1. Recall that in section 1.4, we compared outcomes (acceptance rates) for male and female applicants to Berkeley as an example of Simpson's rule. In section 1.5, we discussed Gender Shades, a project that compared the errors (or equivalently accuracy) of a set of gender recognition systems, across subgroups defined by skin tone and gender. We'll see how in general group fairness criterion can be understood as independence constraints on the joint distributions of the non-sensitive features X, sensitive features, Z, the target variable Y and predicted target $\hat{Y}$ (or rather P for a classification problem where we want our fairness criteria to hold for all thresholds). For brevity, we will express all constraints in terms of $\hat{Y}$, but keep in mind that for classification problems we might want to instead impose it on the score P. We will introduce the necessary mathematical notation as required

throughout this book. A helpful summary is provided in the preamble on page 1.

3.1 Comparing outcomes

First we look at fairness constraints on the relationship between the sensitive features Z, and the predicted target $\hat{Y}$ (or rather Y if we are interested in measuring the fairness of our data rather than our model). We'll discuss two fairness criteria. In the first we require the outcome $\hat{Y}$, to be marginally (unconditionally) independent of the sensitive features Z. In the second we are essentially trying to establish cause; we require the outcome $\hat{Y}$ to be independent of the sensitive features Z when conditioned on all other (non-sensitive) features X. We'll describe the latter as the twin test, that is $\hat{Y}$ and Z being independent ceteris paribus (all else, or rather all other variables X, being equal).

3.1.1 Independence

Of all the fairness criteria, independence is the most well known and that which (as we'll see later), imposes the strongest constraint. It requires the target variable to be marginally (unconditionally) independent of the sensitive feature, that is, $\hat{Y} \perp Z$. This is true, if and only if ($\iff$) the probability distribution of the target variable $f_{\hat{Y}}(y)$, is the same for all values of the sensitive feature Z, that is, $f_{\hat{Y}|Z}(y) = f_{\hat{Y}}(y)$. For a discrete target variable we can say

$$\hat{Y} \perp Z \quad \iff \quad \mathbb{P}(\hat{Y} = \hat{y} | Z = z) = \mathbb{P}(\hat{Y} = \hat{y}) \quad \forall \quad y \in \mathcal{Y}, \quad z \in \mathcal{Z},$$

or $\mathbb{P}(\hat{y}|z) = \mathbb{P}(\hat{y})$ in our abbreviated notation.

Independence might be viewed as addressing disparate impact, since we are only interested in the relationship between the outcome and sensitive feature. Recall that for the 1973 Berkeley admissions example in section 1.4, we looked at independence criterion, by comparing acceptance rates across the sensitive feature gender. Imposing independence is a strong expression of the view that fairness is equality. It might be interpreted as the notion that abilities (or features of importance X, in determining Y) in all groups are, or should be, similarly distributed; the belief that observed differences in the joint distributions in training data are a manifestation of unfair discrimination, rather than inherent differences in people belonging to groups that differ by their sensitive features Z.

Measures of independence

Below we will define a range of fairness metrics, which are all derived from the notion of independence. Along the way, we will familiarise ourselves with some of the terminology used to describe the various independence metrics that are commonly analysed. In each case the metric provides some measure, of how far from independent, the target variable and sensitive feature are. Notice that independence imposes a constraint on only two random variables, the predicted target $\hat{Y}$ and sensitive feature Z. In the equations that follow, we provide metrics that quantify the fairness of our model output $\hat{Y}$, but we could equally well replace the predicted target variable $\hat{Y}$, with the actual target variable Y to assess the fairness of our data under the same criterion.

Mutual information, denoted I, is popular in information theory for measuring dependence between random variables.

$$I(\hat{Y}, Z) = \int_{\hat{y} \in \mathcal{Y}} \int_{z \in \mathcal{Z}} f_{\hat{Y},Z}(\hat{y}, z) \log \frac{f_{\hat{Y},Z}(\hat{y}, z)}{f_{\hat{Y}}(\hat{y}) f_Z(z)} \, \mathrm{d}z \, \mathrm{d}\hat{y}. \tag{3.1}$$

It is equal to zero, if and only if the joint distribution of Z and $\hat{Y}$ is equal to the product of their marginal distributions, that is if $f_{\hat{Y},Z}(\hat{y}, z) = f_{\hat{Y}}(\hat{y}) f_Z(z)$. Therefore, two variables which have zero mutual information are independent. The **normalised prejudice index**[4] divides mutual information by a normalising factor

so that the resulting value falls between zero and one:

$$r_{\text{npi}} = \frac{I(\hat{Y}, Z)}{\sqrt{H(\hat{Y})H(Z)}}, \tag{3.2}$$

where

$$H(Y) = -\int_{y \in \mathcal{Y}} f_Y(y) \log f_Y(y) \, \mathrm{d}y, \tag{3.3}$$

is the entropy. We have provided the formula for continuous random variables, for discrete variables we simply replace the integrals with sums.

Exercise: Normalised prejudice index

Write a function that takes two arrays y and z of categorical features and returns the normalised prejudice index. Hint:

1. Compute the probability distributions $\mathbb{P}(y)$, $\mathbb{P}(z)$ and $\mathbb{P}(y, z)$. Note that these can be thought of as the frequency with which each event occurs.
2. Compute the entropies $H(y)$ and $H(z)$ shown in equation (3.3) and use these to compute the normalising factor, $\sqrt{H(y)H(z)}$.
3. Compute the mutual information $I(z, y)$ shown in equation (3.1) and divide by the normalising factor.

Test your implementation against scikit-learn's: sklearn.metrics.normalized_mutual_info_score. Solution in appendix D.1.

A simple relaxation of independence requires only the mean predicted target variable (rather than the full distribution) to be equal for all values of the sensitive feature, that is,

$$\mathbb{E}(\hat{Y}|Z = a) = \mathbb{E}(\hat{Y}|Z = b).$$

A popular metric derived from this for regression problems is called the **mean difference** (illustrated in Figure 3.1) which (as the name suggests) looks at the difference between the mean predictions for different partitions of the population based on the sensitive feature Z,

$$d = \mathbb{E}(\hat{Y}|Z = a) - \mathbb{E}(\hat{Y}|Z \neq a).$$

Taking the simplest example of discrete binary classifier where we have a binary sensitive feature. We can write the requirement of independence as,

$$\mathbb{P}(\hat{Y} = 1|Z = 1) = \mathbb{P}(\hat{Y} = 1|Z = 0).$$

This criterion goes by many names in research literature - **demographic parity**, **statistical parity** and **parity impact**, among others.

With this criterion for fairness of a classifier, we can quantify the disparity by looking at the difference or the ratio of the probabilities for each sensitive feature. Both are straightforward to calculate given the 2 $\times$ 2 contingency table (Table 3.1) summarising the observed relationship between the sensitive feature and outcome. Each cell of the contingency table shows the number of examples satisfying the conditions given in the corresponding row and column headers. So for example, n_{01} is the number of data points for which $Z = 0$ and $\hat{Y} = 1$. In bio-medical sciences, the **risk difference**:

$$d = \mathbb{P}(\hat{Y} = 1|Z = 1) - \mathbb{P}(\hat{Y} = 1|Z = 0) = \frac{n_{11}}{n_{Z=1}} - \frac{n_{01}}{n_{Z=0}},$$

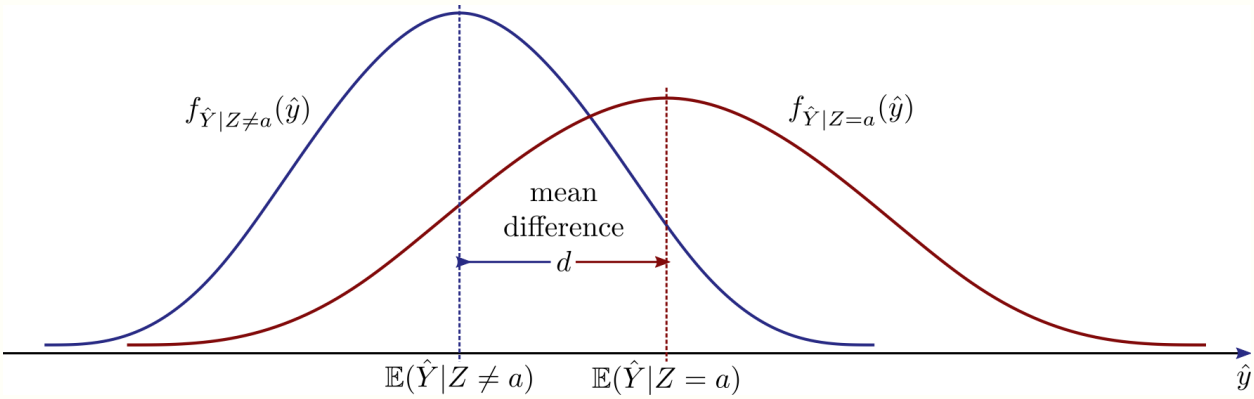


Figure 3.1: Visualisation of the mean difference for a continuous target variable.

Table 3.1: Contingency table for prediction against the sensitive feature.

	$\hat{Y}=1$	$\hat{Y}=0$	Total
$Z=1$	n_{11}	n_{10}	$n_{Z=1}$
$Z=0$	n_{01}	n_{00}	$n_{Z=0}$
Total	$n_{\hat{Y}=1}$	$n_{\hat{Y}=0}$	n

measures the impact of treatment (or risk factors), Z on outcome, $\hat{Y}$. In discrimination literature, it has been described as the **discrimination score** and **statistical parity difference** among others. Note that if $\hat{Y}=1$ is the advantageous outcome and $Z=1$ is the advantaged group, we would expect d to be non-negative. The algorithm is fair when $d=0$. The further from zero, the more unfair. A modified version of this metric is the **normalised difference**[9] which divides the difference by,

$$d_{\max} = \min\left\{\frac{\mathbb{P}(\hat{Y}=1)}{\mathbb{P}(Z=1)}, \frac{\mathbb{P}(\hat{Y}=0)}{\mathbb{P}(Z=0)}\right\} = \min\left\{\frac{n_{\hat{Y}=1}}{n_{Z=1}}, \frac{n_{\hat{Y}=0}}{n_{Z=0}}\right\}, \tag{3.4}$$

thus ensuring the normalised difference is bounded between plus and minus one.

Exercise: Statistical parity difference maximum

Show that

$$d_{\max} = \min\left\{\frac{\mathbb{P}(\hat{Y}=1)}{\mathbb{P}(Z=1)}, \frac{\mathbb{P}(\hat{Y}=0)}{\mathbb{P}(Z=0)}\right\}.$$

Solution in appendix D.1.

Alternatively, we could instead take the ratio as a measure of discrimination:

$$r = \frac{\mathbb{P}(\hat{Y}=1|Z=1)}{\mathbb{P}(\hat{Y}=1|Z=0)} = \frac{n_{11}/n_{Z=1}}{n_{01}/n_{Z=0}}.$$

In biomedical sciences this measure is called the **risk ratio**. It is used to measure the strength of association between treatment (or risk factors), Z, and outcome, $\hat{Y}$. It has been described in discrimination aware machine learning literature as the **impact ratio** or **disparate impact ratio**. The algorithm is fair if $r=1$. The further from one r is, the more unfair. The Equal Employment Opportunity Commission (EEOC) have

used this measure in their guidelines for identifying discrimination in employment selection processes[3]. As a rule of thumb, the EEOC determine that a company's selection system is having an adverse impact on a particular group if the selection rate for that group is less than four-fifths (or 80%) that of the most advantaged group, that is, the impact ratio is less than 0.8 where $Z = 0$ is the most advantaged group (for which the acceptance rate is the highest).

The **elift ratio**[7] is similar to the impact ratio but instead of comparing acceptance rates for protected groups to each other, we compare them to the overall acceptance rate:

$$r_{\text{elift}} = \frac{\mathbb{P}(\hat{Y} = 1|Z = 0)}{\mathbb{P}(\hat{Y} = 1)}.$$

In theory, any measure of association suitable for the data types can be used as a metric to understand the magnitude of discrimination in our data or predictions. The **odds ratio** (popular in natural, social and biomedical sciences) is the ratio of the odds of a positive prediction for each group. We can write it as:

$$r_{\text{odds}} = \frac{\mathbb{P}(\hat{Y} = 1|Z = 1)\mathbb{P}(\hat{Y} = 0|Z = 0)}{\mathbb{P}(\hat{Y} = 0|Z = 1)\mathbb{P}(\hat{Y} = 1|Z = 0)} = \frac{n_{11}n_{00}}{n_{10}n_{01}}.$$

The odds ratio is equal to one when there is no discrimination. Recall that the odds ratio is not a collapsible measure (see section 1.4.3).

Exercise: Odds ratio

Show that the odds ratio is always greater than or equal to one in the case where $\hat{Y} = 1$ in the advantaged outcome and $Z = 1$ is the privileged group. Solution in appendix D.1.

As mentioned earlier, independence metrics is they can be evaluated on both the data and the model. A common problem in machine learning is that existing biases in the data can be exaggerated if protected groups are minorities in the population. By comparing bias metrics for the data with those of our model output, we can understand if our model is inadvertently introducing biases that do not originate from the data.

It might seem intuitive already, that independence can only be satisfied by a model (optimising for utility), if the target variable Y and sensitive feature Z are in fact independent. [1] If this is not the case, then satisfying independence for your model, will not permit the theoretically 'perfect' solution $\hat{Y} = Y$ (should your model be able to achieve it). We would also then naturally, expect that the stronger the relationship between the sensitive feature and target, the greater the trade-off between fairness and utility in satisfying independence criterion.

A major shortcoming of independence (discussed in section 1.4) is that it doesn't consider that there may be confounding variables. It assumes that all relevant features, are distributed among all protected groups, equally. This is a strong assumption, and erronously enforcing it does not guarantee fairness in a broader sense. Consider a simple hypothetical example where, there are discrepancies between credit card approval rates for men and women at the population level, which disappear once you control for (the confounding variable) income. It could be argued then that the real issue with respect to fairness here, appears to be the fact that women, generally earn less than men. If the lender was to enforce independence between gender and its loan approval rate by, for example, setting lower income requirements for women than men, this might feasibly lead to higher default rates among women. Clearly a less than desirable solution which, arguably, doesn't address the actual underlying problem. Furthermore, it might be argued that enforcing independence could lead to less fair outcomes, on an individual level; in the sense that a man and woman who were the same in all other respects (features) would receive different outcomes, violating the twin test, which we'll talk about in the next section. We'll talk about individual fairness in the next chapter.

[1] We'll prove this to be true in section 3.3, for the case where our variables are binary.

Suppose we want to measure the relationship between the sensitive feature and outcome using one of the above metrics. A natural solution to the problem of confounding variables, is to control for them, that is if, you have them recorded in your dataset. Next, we consider the case where we condition on all the non-sensitive variables X.

3.1.2 The twin test

The twin test tries to establish cause, by conditioning on all non-sensitive features. In this case, our fairness criterion requires the predicted target variable to be independent of the sensitive features when conditioned on all other features. This is true, if and only if, the probability distribution of Y conditioned on X is the same, for all values of the sensitive feature;

$$\hat{Y} \perp Z|X \quad \iff \quad f_{\hat{Y}|X}(\hat{y}, z; x) = f_{\hat{Y}|X}(\hat{y}; x).$$

Suppose we wish to establish a causal connection between the decision or outcome and an individual's membership in some protected group. Typically, in a human decision process which is subjective, there are a number of unobserved variables. Take a job interview for example, the factors that determine who gets hired are typically subjective (which means that two people might rate an interview candiate differently even on the same feature, or, a candidate might rate differently on the same feature if they were given an alternative test. This makes proving a causal connection difficult. But in the case where a decision is made purely on the basis of an algorithm, making this causal connection becomes trivial. We simply perform a so called 'twin test'. Imagine a 'counterfactual' world in which for every individual in this world (say John Doe) there exists an 'identical twin' in the counterfactual world which differers only by the sensitive feature gender (Jane Doe). If a model produces predictions that are different for for John and Jane, we have established the individual's gender as being the reason for it.

Taking this approach to establishing cause with a model is pretty straight forward. We simply conduct a randomized experiment. The individuals for which we check the model output, need not exist, we can simply fabricate them, and compare model predictions with those of the counterfactual twin. Performing the twin test on a dataset (i.e. where you do not have access to the algorithm, only the outcomes recorded in the data) is less trivial, since the conterfactual twin for any given example, need not exist; and we have no way of producing them without the algorithm. Furthermore, for any given point in the non-sensitive feature space, the number of data points will likely be too small, to justify the use of statistical methods in establishing cause. Barring these limitations, using the counterfactual approach to establishing the fairness of our model, we can consider all the metrics we have above with independence as our fairness criterion but conditioned on X as well as Z. So for example we define the **causal mean difference** as

$$d = \mathbb{E}(\hat{Y}|Z = 1, X = x) - \mathbb{E}(\hat{Y}|Z = 0, X = x).$$

and the **observed mean difference** as

$$d = \mathbb{E}(Y|Z = 1, X = x) - \mathbb{E}(Y|Z = 0, X = x).$$

Calculating the mean difference is a popular way to establish disparate treatment in an algorithmic decision process because it exposes differing treatment of individuals based on protected class membership.

3.2 Comparing errors

In this section we learn about fairness criteria which seek to compare model errors across groups, rather than outcomes. A fundamental assumption here is that the training data is fair; it represents the the ground truth; the target variable is the that which we wish to affect, the data is accurate and representative of the population, and include the features related related to the target. Assuming we have said data, under these criteria, for our model to be fair, we require the errors, not only to be sufficiently small, but also distributed similarly for different subgroups of the population (defined by the values of the sensitive features).

Expressed differently, we want the errors to be independent of protected characteristics, that is, $(\hat{Y} - Y) \perp Z$. We discussed earlier in the chapter how independence and twin test constraints have been interpreted as avoiding disparate impact and disparate treatment respectively. Analogously, criteria on model errors have been described as avoiding **disparate mistreatment**[8].

3.2.1 Regression

Once again, relaxation of this criterion compares the mean error for the groups (rather than comparing the full distributions). **Balanced residuals**[2] takes the difference of the mean errors as a measure of fairness:

$$d_{\text{err}} = \mathbb{E}(\hat{Y} - Y | Z = 1) - \mathbb{E}(\hat{Y} - Y | Z = 0).$$

This can be calculated for $n = n_0 + n_1$ data points as,

$$d_{\text{err}} = \frac{1}{n_0} \sum_{i|z_i=0} (y_i - \hat{y}_i) - \frac{1}{n_1} \sum_{i|z_i=1} (y_i - \hat{y}_i).$$

Here $d_{\text{err}} = 0$ would be considered fair.

3.2.2 Classification

For a classification problem the most obvious relaxation would be to ensure equal error rates (or equivalently accuracy) for all groups. As an example, recall the project Gender Shades we discussed in section 1.5, that audited several commercial gender classification packages measured their accuracy for different protected groups. To derive a measure of fairness from this criterion we could (as before) take the difference, or the ratio. The **error rate difference** is given by,

$$d_{\text{err}} = \mathbb{P}(\hat{Y} \neq Y | Z = 1) - \mathbb{P}(\hat{Y} \neq Y | Z = 0).$$

Again here $d_{\text{err}} = 0$ would be considered fair. **error rate ratio** is given by

$$r_{\text{err}} = \frac{\mathbb{P}(\hat{Y} \neq Y | Z = 1)}{\mathbb{P}(\hat{Y} \neq Y | Z = 0)}$$

in which case $r_{\text{err}} = 1$ would be considered fair.

Table 3.2: Summary of error rate types for a binary classifier

		Ground Truth		
		$y = 1$	$y = 0$	Error Rate Type
Prediction	$\hat{y} = 1$	True Positive	False Positive Type I Error	False Discovery Rate $\mathbb{P}(\hat{y} \neq y \mid \hat{y} = 1)$
	$\hat{y} = 0$	False Negative Type II Error	True Negative	False Omission Rate $\mathbb{P}(\hat{y} \neq y \mid \hat{y} = 0)$
Error Rate Type		False Negative Rate $\mathbb{P}(\hat{y} \neq y \mid y = 1)$	False Positive Rate $\mathbb{P}(\hat{y} \neq y \mid y = 0)$	Error Rate $\mathbb{P}(\hat{y} \neq y)$

A binary classification model can make two different types of errors (false positives and false negatives), the costs of which will typically not be equal. Table 3.2 summarises terminology for the different types of error rates for a binary classification model.

Fairness criteria that compare errors (or equivalently performance metrics - see Table 3.3 for a summary) across groups can be broken down into two conditional independence constraints on the joint distributions of the sensitive features, Z, the target feature Y and predicted target $\hat{Y}$. These can be described as **separation**

Table 3.3: Summary of performance metrics for a binary classifier

		Ground Truth		Metric
		$y = 1$	$y = 0$	
Prediction	$\hat{y} = 1$	True Positive	False Positive Type I Error	Positive Predictive Value[a] $\mathbb{P}(\hat{y} = y \vert \hat{y} = 1)$
	$\hat{y} = 0$	False Negative Type II Error	True Negative	Negative Predictive Value $\mathbb{P}(\hat{y} = y \vert \hat{y} = 0)$
Metric		True Positive Rate[b] $\mathbb{P}(\hat{y} = y \vert y = 1)$	True Negative Rate $\mathbb{P}(\hat{y} = y \vert y = 0)$	Accuracy $\mathbb{P}(\hat{y} = y)$

[a] Positive Predictive Value = Precision

[b] True Positive Rate = Recall

$(\hat{Y} \perp Z|Y)$ and **sufficiency**[1] $(Y \perp Z|\hat{Y})$. Each of these criteria can be defined as a conditional independence constraint on the joint distributions of the sensitive features, Z, the target feature Y and predicted target $\hat{Y}$. Separation requires equal error (false negative and false positive) rates along the columns (conditioning on Y), while sufficiency requires equal error (false discovery and false omission) rates along the rows (conditioning on $\hat{Y}$) of the confusion matrix (Table 3.2).

Separation

Separation requires the predicted target variable to be independent of the sensitive feature, conditioned on the target variable, that is, $\hat{Y} \perp (Z|Y)$. We can say that the predicted target $\hat{Y}$, is 'separated' from the sensitive feature Z, by the target variable Y. The corresponding graphical model for separation criteria is shown in Figure 3.2. Essentially we are saying that for a fixed value of the target variable, there should be

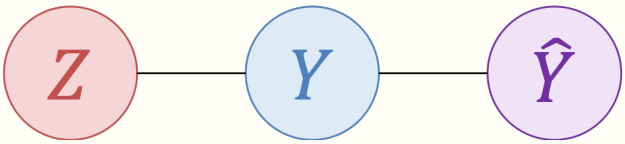


Figure 3.2: Graphical model for separation.

no difference in the distribution of the predicted target variable, for different values of the sensitive feature. That is,

$$\mathbb{P}(\hat{y}|y, z) = \mathbb{P}(\hat{y}|y).$$

Unlike independence, separation, allows for dependence between the predicted target variable and the sensitive feature but only to the extent that it exists between the actual target variable and the sensitive feature.

Once again let's take the simplest example of discrete binary classifier where we have a single sensitive binary feature. We can write this requirement (most well known as **equalised odds**[5]) as two conditions,

$$\mathbb{P}(\hat{Y} = 1|Z = 1, Y = 1) = \mathbb{P}(\hat{Y} = 1|Z = 0, Y = 1),$$
$$\mathbb{P}(\hat{Y} = 1|Z = 1, Y = 0) = \mathbb{P}(\hat{Y} = 1|Z = 0, Y = 0).$$

Recall that $\mathbb{P}(\hat{Y} = 1|Y = 1)$ is the true positive rate (TPR) of the classifier and $\mathbb{P}(\hat{Y} = 1|Y = 0)$ is the false positive rate (FPR). We see then that separation requires the true positive rate, and the false positive rate, to be the same for all values of the sensitive feature. Note that the true positive rate is equal if and only if the false negative rate is equal, so thinking in terms of error metrics only, separation requires the false negative and false positive rates to be equal for all values of the sensitive feature.

Two related metrics are the average odds difference and average odds error. The **average odds difference** measures the magnitude of unfairness as the average of the difference in true positive rate and false positive rate, that is,

$$d_{\text{av-odds}} = \frac{1}{2}[TPR_{Z=0} - TPR_{Z=1} + FPR_{Z=0} - FPR_{Z=1}].$$

The **average odds error** measures the magnitude of unfairness as the average of the absolute difference in true positive rate and false positive rate, that is,

$$d_{\text{av-odds-err}} = \frac{1}{2}[|TPR_{Z=0} - TPR_{Z=1}| + |FPR_{Z=0} - FPR_{Z=1}|].$$

A relaxed version of equalised odds, called **equal opportunity**[5], requires only the true positive rates to be the same across all groups (assuming a positive prediction is the more advantageous ourcome). A metric which uses this as a criterion to measure unfairness is **equal opportunity difference** which takes the difference in true positive rates across groups, that is,

$$d_{\text{eq-op}} = TPR_{Z=0} - TPR_{Z=1}.$$

Exercise: Fair equality of opportunity

Can you see how the metric *equal opportunity* relates to the second principle of justice as fairness (Fair equality of opportunity) discussed in section 1.2?

Sufficiency

Sufficiency requires the sensitive feature Z and target variable Y to be independent, conditional on the predicted target variable $\hat{Y}$, that is, $Y \perp (Z|\hat{Y})$. We can say that the predicted target $\hat{Y}$ is 'sufficient' for the sensitive feature Z. That is to say, given $\hat{Y}$, Z provides no additional information. The corresponding graphical model for sufficiency criteria is shown in Figure 3.3. Comparing sufficiency to separation we

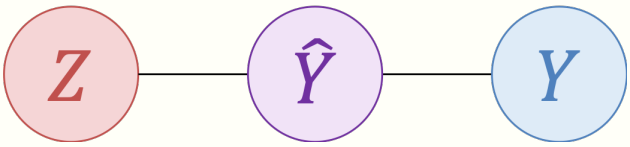


Figure 3.3: Graphical model for sufficiency.

note that Y and $\hat{Y}$ are reversed in the graphical model and conditional independence constraint. It should hopefully be straightforward to see then that sufficiency requires the false omission rate and false discovery rate (see Table 3.2) to be equal across protected groups.

Exercise: Sufficiency

Show that sufficiency is satisfied if and only if the false omission rate and false discovery rate are equal for all groups. Solution in appendix D.1.

Comparing group fairness metrics

There are some nice properties of separation and sufficiency criteria. Note that unlike criteria comparing outcomes they do not preclude the theoretically 'perfect' solution, $\hat{Y} = Y$. The criteria also preclude large differences in error rates for different groups that are typical when disadvantaged classes are minorities

suffering from low support. It's worth reiterating that unlike independence, separation and sufficiency criteria assume that the relationship between Y and Z prescribed by the training data is fair. In such cases, error comparison criteria allow flexibility in the choice which types of errors are important to equalize, based on the cost and our values. For example, in pretrial risk assessment we might choose to prioritise ensuring equal false positive rates if we believe that it is preferable to set free a guilty defendant than incarcerate an innocent one. As another example, let's take the infamous NYPD stop-and-frisk program where pedestrians were stopped, interrogated and searched on 'reasonable' suspicion of carrying contraband. In this case we might want to ensure false discovery rates are equal across groups to ensure we are not disproportionately targeting particular minority groups.

Exercise: Stop-and-frisk

- Why might we choose to compare false discovery rates for stop-and-frisk, rather than say false omission, false negative or false positive rates?
- Is it fair if false discovery rates are similar?
- How might we go about measuring the false omission rate if we wanted to compare them?

Of our two fairness criteria, separation and sufficiency, the latter imposes a weaker constraint on our model. To understand why, we explore another interpretation of sufficiency which intuitively explains why it might be satisfied implicitly through the training process[6]. Let us look at sufficiency criteria in terms of the classification score P,

$$\mathbb{P}(Y = 1|P = p, Z = 1) = \mathbb{P}(Y = 1|P = p, Z = 0) \quad \forall p$$

We say that a classifier score is calibrated if

$$\mathbb{P}(Y = 1|P = p) = p \quad \forall p.$$

Essentially, this is the requirement that the proportion of data points assigned the score p, which did in fact have a positive outcome $Y = 1$, should be equal to the score p. The score p can then be interpreted, at the population level, as the probability that the a positive prediction $\hat{Y} = 1$ would be correct[2].

From the definitions above we can see that if our classifier scores are calibrated for all groups, sufficiency is automatically satisfied. If our model satisfies sufficiency but not calibration by group, we can calibrate our model score through a simple transformation. We simply pick a value for Z, $Z = 1$ say, and then calculate the mapping,

$$\mathbb{P}(Y = 1|P = p, Z = 1) = f(p).$$

We then transform all our scores to new scores (which satisfy calibration by group) by applying the inverse mapping $f^{-1}(P)$. The resulting model is both sufficient and the model score is calibrated.

3.3 Incompatibility between fairness criteria

So far in this chapter we have learned a range of different group fairness criteria and seen how each of them can be viewed as imposing different constraints on the joint distributions of our variables X, Z, Y and $\hat{Y}$. In this section we will prove that these fairness criteria can be restrictive enough to mean that satisfying more than one of them is impossible, except in some degenerate cases. For a useful recap of the rules of probability (which we will use in our proofs), see in Appendix C.

[2]For the score to be interpretable as this probability at the individual level, we would need to satisfy the stronger criteria $P = \mathbb{E}[Y|X]$.

3.3.1 Independence versus Sufficiency

Independence versus Sufficiency

Independence ($Z \perp \hat{Y}$) and sufficiency ($Z \perp Y | \hat{Y}$) can only be simultaneously satisfied if the sensitive feature, Z and the target variable $\hat{Y}$ are independent ($Z \perp Y$).

To prove this we consider the conditional distribution $Z|Y, \hat{Y}$.

$$\begin{aligned} \text{Independence: } Z \perp \hat{Y} \quad &\Rightarrow \quad \mathbb{P}(z|y, \hat{y}) = \mathbb{P}(z|y) \\ \text{Product rule} \quad &\Rightarrow \quad \mathbb{P}(z|y) = \frac{\mathbb{P}(z, y)}{\mathbb{P}(y)} \\ &\Rightarrow \quad \mathbb{P}(z|y, \hat{y}) = \frac{\mathbb{P}(z, y)}{\mathbb{P}(y)}. \end{aligned} \tag{3.5}$$

Applying Sufficiency, followed by independence gives,

$$\begin{aligned} \text{Sufficiency: } Z \perp Y | \hat{Y} \quad &\Rightarrow \quad \mathbb{P}(z|y, \hat{y}) = \mathbb{P}(z|\hat{y}) \\ \text{Independence: } Z \perp \hat{Y} \quad &\Rightarrow \quad \mathbb{P}(z|\hat{y}) = \mathbb{P}(z) \\ &\Rightarrow \quad \mathbb{P}(z|y, \hat{y}) = \mathbb{P}(z). \end{aligned} \tag{3.6}$$

Equating (3.5) and (3.6) and rearranging gives,

$$\mathbb{P}(z, y) = \mathbb{P}(z)\mathbb{P}(y).$$

Thus, Z and Y must be independent.

3.3.2 Independence versus Separation

Independence versus Separation

In the case that Y is binary, independence ($Z \perp \hat{Y}$) and separation ($Z \perp \hat{Y} | Y$) criteria can only be simultaneously satisfied if either $\hat{Y} \perp Y$ or $Y \perp Z$.

To prove this we consider the distribution of $\hat{Y}$.

$$\begin{aligned} \text{Sum rule:} \quad &\Rightarrow \quad \mathbb{P}(\hat{y}) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}, y). \\ \text{Product rule} \quad &\Rightarrow \quad \mathbb{P}(\hat{y}) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}|y)\mathbb{P}(y). \end{aligned} \tag{3.7}$$

$$\begin{aligned} \text{Conditioning on } Z \quad &\Rightarrow \quad \mathbb{P}(\hat{y}|z) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}|y, z)\mathbb{P}(y|z). \\ \text{Independence: } \hat{Y} \perp Z \quad &\Rightarrow \quad \mathbb{P}(\hat{y}) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}|y)\mathbb{P}(y|z). \end{aligned} \tag{3.8}$$

Equating (3.7) and (3.8) and rearranging gives,

$$\sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{y}|y)[\mathbb{P}(y) - \mathbb{P}(y|z)] = 0 \tag{3.9}$$

For binary Y, $\mathcal{Y} = \{0, 1\}$. Denoting $\mathbb{P}(y) = p_y$ and $\mathbb{P}(y|z) = q_y$, then $p_1 = 1 - p_0$ and $q_1 = 1 - q_0$. Substituting into (3.9) gives,

$$\begin{aligned} & \mathbb{P}(\hat{y}|Y=0)(p_0 - q_0) + \mathbb{P}(\hat{y}|Y=1)[1 - p_0 - (1 - q_0)] = 0 \\ \Leftrightarrow \quad & [\mathbb{P}(\hat{y}|Y=0) - \mathbb{P}(\hat{y}|Y=1)](p_0 - q_0) = 0 \end{aligned}$$

which is true if and only if,

$$\begin{array}{lllll} \text{either} & & \mathbb{P}(\hat{y}|Y=0) = \mathbb{P}(\hat{y}|Y=1) & \Leftrightarrow & \hat{Y} \perp Y, \\ \text{or} & p_0 = q_0 \quad \Leftrightarrow & \mathbb{P}(Y=0) = \mathbb{P}(Y=0|z) & \Leftrightarrow & Y \perp Z. \end{array}$$

3.3.3 Separation versus Sufficiency

Separation versus Sufficiency I

In the case where all events in the joint distribution of Z, Y and $\hat{Y}$ have non zero probability, separation ($Z \perp \hat{Y}|Y$) and sufficiency ($Z \perp Y|\hat{Y}$) can only be simultaneously be satisfied if the sensitive feature, Z is independent of both the target variable Y and the predicted target $\hat{Y}$, that is if $Z \perp Y$ and $Z \perp \hat{Y}$.

To prove this we consider the conditional distribution $\mathbb{P}(z|y, \hat{y})$.

$$\begin{aligned} \text{Separation: } Z \perp \hat{Y}|Y \quad & \Rightarrow \quad \mathbb{P}(z|y, \hat{y}) = \mathbb{P}(z|y) \\ \text{Sufficiency: } Z \perp Y|\hat{Y} \quad & \Rightarrow \quad \mathbb{P}(z|y, \hat{y}) = \mathbb{P}(z|\hat{y}) \\ & \Rightarrow \quad \mathbb{P}(z|y) = \mathbb{P}(z|\hat{y}). \end{aligned} \tag{3.10}$$

$$\begin{aligned} \text{Product rule:} \quad & \mathbb{P}(z, y) = \mathbb{P}(z|y)\mathbb{P}(y) \\ (3.10) \quad \Rightarrow \quad & \mathbb{P}(z, y) = \mathbb{P}(z|\hat{y})\mathbb{P}(y). \end{aligned} \tag{3.11}$$

$$\begin{aligned} \text{Sum rule:} \quad & \mathbb{P}(z) = \sum_{y \in \mathcal{Y}} \mathbb{P}(z, y) \\ (3.11) \quad \Rightarrow \quad & \mathbb{P}(z) = \sum_{y \in \mathcal{Y}} \mathbb{P}(z|\hat{y})\mathbb{P}(y) \end{aligned}$$

If all events have non-zero probability, we can move $\mathbb{P}(z|\hat{y})$ outside of the summation,

$$\mathbb{P}(z) = \mathbb{P}(z|\hat{y}) \tag{3.12}$$

Thus showing that Z and $\hat{Y}$ must be independent. Equating (3.10) and (3.12) shows that Z and Y must also be independent.

Separation versus Sufficiency II

In the case where Y is binary, separation and sufficiency can only be satisfied simultaneously if the sensitive feature is independent of the target variable, or the model has an accuracy of 100% ($\hat{Y} = Y$) or 0% ($\hat{Y} = 1 - Y$).

Consider the case where Y is binary. Separation requires all groups to have the same true positive rate (recall or TPR) and the same false positive rate (FPR). On the other hand, sufficiency requires all groups to

have the same positive predictive value (precision or PPV) and the same negative predictive value (NPV). A problem is evident at this point. For a fixed number of data points, the confusion matrix for a binary classifier only has three degrees of freedom but satisfying both separation and sufficiency introduces four constraints which requires four degrees of freedom in order be able to satisfy them. We can write the positive and negative predictive values in terms of the true positive and false positive rates as follows:

$$PPV = \frac{pTPR}{pTPR + (1-p)FPR} \tag{3.13}$$

and

$$NPV = \frac{(1-p)(1-FPR)}{p(1-TPR) + (1-p)(1-FPR)} \tag{3.14}$$

where $p = \mathbb{P}(Y = 1)$.

Exercise: Predictive values

Prove the results given in equations (3.13) and (3.14). Refer to Table 3.3 for a summary of model performance metrics for a binary classifier. Solution in appendix D.1.

Denote $p_z = \mathbb{P}(Y = 1|Z = z)$ then we can show from equations (3.13) and (3.14) that for any distinct pair of groups $Z = a$ and $Z = b$ for both separation and sufficiency to hold we must have

$$FPR(p_a - p_b)TPR = 0 \tag{3.15}$$

and

$$(1 - FPR)(p_a - p_b)(1 - TPR) = 0 \tag{3.16}$$

respectively.

Exercise: Separation versus sufficiency

Show that for separation and sufficiency to hold equations (3.15) and (3.16) must hold for for any pair of groups $Z = a$ and $Z = b$. Solution in appendix D.1.

Equations (3.15) and (3.16) can only be simultaneously satisfied in 3 cases:

1. $p_a = p_b \, \forall \, a, b$ in which case $Y \perp Z$,
2. $FPR = 0$ and $TPR = 1$ in which case $Y = \hat{Y}$,
3. $FPR = 1$ and $TPR = 0$ in which case $Y = 1 - \hat{Y}$.

Summary

Group fairness

- The term group fairness is used to describe a class of metrics that all stem from the same high level idea; the notion that some property should be equally distributed across different subgroups of a population.
- Group fairness metrics are often used to measure discrimination or bias in a given decision process.
- In general group fairness criterion and measures can be derived from independence constraints on the joint distributions of the non-sensitive features X, sensitive features, Z, the target feature Y and predicted target $\hat{Y}$.
- Group fairness criteria can be broadly classified into two types; those seeking to compare outcomes across groups and those comparing errors.

Table 3.4: Group fairness metrics summary.

Comparing	Outcomes		Errors	
Criterion	Independence	Twin test	Separation	Sufficiency
Constraint	$Y \perp Z$	$Y \perp (Z \mid X)$	$\hat{Y} \perp (Z \mid Y)$	$Y \perp (Z \mid \hat{Y})$
Preventing	Disparate impact	Disparate treatment	Disparate mistreatment	

Comparing Outcomes

Independence

- Independence is a strong expression of the view that fairness is equality. It might be interpreted as the notion that abilities (or features) in all groups are, or should be, equally distributed; the belief that observed differences in the distributions in training data are a manifestation of unfair discrimination, errors in data collection, or both, rather than inherent differences in the abilities of people belonging to one group or another.
- Independence metrics is they can be evaluated on both the data and the model. A common problem in machine learning is that existing biases in the data can be exaggerated if protected groups are minorities in the population. By comparing independence metrics for the data and with those of our model output we can understand if our model is inadvertently introducing biases that do not originate from the data.
- If the target variable Y and sensitive feature Z are not independent then satisfying independence for your model will not permit the theoretically 'perfect' solution $Y = \hat{Y}$. We would naturally expect that the stronger the relationship between the sensitive feature and target, the greater the trade-off between fairness and utility in satisfying the independence criterion.
- A major shortcoming of independence is that it doesn't consider that there may be confounding variables. It assumes that all relevant features are held by all protected groups equally and where there are differences it assumes unfairness and passes the task of correcting for it to the decision maker.

The twin test

- The twin test has been interpreted as addressing disparate treatment, since it exposes differing treatment of individuals based on protected class membership. In reality, while it would be sufficient to demonstrate disparate treatment, it is not necessary. Using protected features in the algorithm would be enough to result in disparate treatment liability in the US, the impact of using the feature is irrelevant.
- In the case where a decision is made purely on the basis of an algorithm and there are no unobserved variables, we can perform a 'twin test' to establish disparate treatment. We conduct a randomized experiment and compute the causal mean difference. If the value is non-zero, we have established the existence of disparate treatment.

Comparing errors

- Criteria comparing errors assume that the relationship between the target variable and sensitive feature prescribed by the training data is fair so only make sense if the target variable is reliable as the ground truth. Under this assumption that our data is fair, for our model to be fair, we require errors to be distributed similarly for different subgroups of the population (defined by the values of sensitive features).
- Ensuring equal errors has been described as avoiding disparate mistreatment.
- For a regression model balanced residuals takes the difference of the mean errors for each group as a measure of fairness.
- For a classification problem we could could use the error rate difference or the error rate ratio as a measure of fairness.

- Unlike criteria comparing outcomes, criteria comparing errors do not preclude the theoretically 'perfect' solution, $\hat{Y} = Y$.

Separation

- Separation, allows for dependence between the predicted target variable and the sensitive feature but only to the extent that it exists between the actual target variable and the sensitive feature.
- For a binary classification model, separation requires both the false negative and false positive rates to be equal across groups. This criterion is known as equalised odds
- Equal opportunity criterion requires only the true positive rates to be the same across all groups (assuming a positive prediction is the more advantageous outcome).

Sufficiency

- For a binary classification model, sufficiency requires both the false omission rate and false discovery rates to be equal across protected groups.
- Sufficiency is is a weaker model constraint compared to separation as it is satisfied implicitly through the training process.

Incompatibility between fairness criteria

- Independence ($Z \perp \hat{Y}$) and sufficiency ($Z \perp Y | \hat{Y}$) can only be simultaneously be satisfied if the sensitive feature Z, and the target variable $\hat{Y}$, are independent ($Z \perp Y$).
- In the case that Y is binary, independence ($Z \perp \hat{Y}$) and separation ($Z \perp \hat{Y} | Y$) criteria can only be simultaneously satisfied if either $\hat{Y} \perp Y$ or $Y \perp Z$.
- Separation ($Z \perp \hat{Y} | Y$) and sufficiency ($Z \perp Y | \hat{Y}$) can only be simultaneously be satisfied if the sensitive feature, Z is independent of both the target variable Y and the predicted target $\hat{Y}$, that is if $Z \perp Y$ and $Z \perp \hat{Y}$.
- In the case where Y is binary, separation and sufficiency can only be satisfied simultaneously if the sensitive feature is independent of the target variable, or the model has an accuracy of 100% ($\hat{Y} = Y$) or the model has an accuracy of 0% ($\hat{Y} = 1 - Y$).

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning.* fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[2] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, December 2013. `https://www.researchgate.net/publication/261637367_Controlling_Attribute_Effect_in_Linear_Regression`.

[3] U.S. Equal Employment Opportunity Commission. Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *Federal Register*, 44(43), March 1979. `https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines`.

[4] Kazuto Fukuchi, Jun Sakuma, and Toshihiro Kamishima. Prediction with model-based neutrality. *IEICE TRANS. INF. & SYS.*, E98-D(8), August 2015. `https://www.kamishima.net/archive/2015-t-ieice-print.pdf`.

[5] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016. https://arxiv.org/abs/1610.02413.

[6] Lydia T. Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning, 2019. https://arxiv.org/abs/1808.10013.

[7] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568, August 2008. https://www.researchgate.net/publication/221654695_Discrimination-aware_data_mining.

[8] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, April 2017. http://dx.doi.org/10.1145/3038912.3052660.

[9] Indre Zliobaite. On the relation between accuracy and fairness in binary classification, 2015. https://arxiv.org/abs/1505.05723.

Appendix A

AIF360

A.1 Installing AIF360

1. In this book we will use Python in Jupyter notebooks from the Anaconda Python distribution platform. If you don't already have it download and install it.

2. Create an environment named `mbml`. Using the command line interface (CLI):

   ```
   \$ conda create --name mbml python=3.7
   ```

3. Activate your new environment:

   ```
   $ conda activate mbml
   ```

4. This book is a work in progress. As part of analysing the metrics and methods it uses code that is not yet available with the library[1]. Once it is merged, you will just be able to just pip install the aif360 library. Until then you must clone this fork of AIF360:

   ```
   $ git clone https://github.com/leenamurgai/AIF360.git
   ```

5. Download the notebook `mbml_german.ipynb` from Manning's GitLab repository and save it in the "AIF360/examples" folder.

6. You should now be able to open and run the notebook from the CLI as you usually would:

   ```
   $ jupyter notebook mbml_german.ipynb
   ```

[1] If you're interested, here is the open pull request.

A.2 Group fairness in AIF360

A.2.1 Comparing outcomes

Now that we have covered some measures of fairness, let's dive into calculating them. In this book we are going to use IBM's AI Fairness 360 (AIF360). AIF360 is currently the most comprehensive open source library available for measuring and mitigating bias in machine learning models. The Python package includes an extensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models many of which we will cover in this book. The system has been designed to be extensible, adopted software engineering best practices to maintain code quality, and is well documented. The package implements techniques from at-least eight published papers and includes over 71 bias detection metrics and nine bias mitigation algorithms. These techniques can all be called in a standard way, similar to scikit-learn's fit/transform/predict paradigm.

In this section we're going to use AIF360 to calculate some of the metrics we've talked about in the previous section as a means to get started working with it. For calculating the metrics we've talked about so far, using AIF360 might seem to add unnecessary overhead as they are reasonably straightforward to code up directly once you have your data in a Pandas DataFrame. But remember, the library contains implementations of more complicated metrics and bias mitigations algorithms that we'll cover later on in this book. Before we can use the library, we need to install it. Instructions are provided in Appendix A.1.

Statlog (German Credit Data) Data Set

The Jupyter Notebook, `mbml_german.ipynb`, contains an example calculating some of the above fairness metrics on both a dataset and model output. It uses the Statlog (German Credit Data) Data Set, in which one thousand loan applicants are classified as representing 'good' or 'bad' credit risks based on features such as loan term, loan amount, age, gender, marital status and more.

Exercise: Statlog (German Credit Data) Data Set

Sections 1-3 in the Jupyter Notebook, `mbml_german.ipynb`, load the data and perform some exploratory data analysis (EDA), looking at correlation heat maps (using a variety of different measures of association) and comparing distributions of the target for different values of the features. Open the notebook and run the code up to section four. You should be able to answer the following questions by working through the notebook.

1. What proportion of the population is classified as male/female?
2. What proportion of the population have good credit vs bad?
3. How many continuous variables are there? What are they? Do any of them appear to be related? If so how?
4. How many categorical variables are there? What are they? Do any of them appear to be related? If so how?

Calculating independence metrics

In order to calculate our metrics on the data using AIF360, we must have it in the correct format; that is, in a Pandas DataFrame (`data_df`) containing only numeric data types. In code listing A.1, we calculate the rate at which male and female applicants are classified as being good credit risks (`base_rate`) along with the difference (`mean_difference`) and the ratio (`disparate_impact`) of these rates.

Listing A.1: Calculating independence metrics for the data using AIF360

```
# Create a DataFrame to store results in
outcomes_df = pd.DataFrame(columns=['female', 'male',
```

```
                                    'difference', 'ratio'],
                          index=['data', 'model',
                                 'train data', 'train model',
                                 'test data', 'test model'])

# Define privileged and unprivileged groups
privileged_groups = [{'sex_male':1}]
unprivileged_groups = [{'sex_male':0}]

# Create an instance of BinaryLabelDataset
data_ds = BinaryLabelDataset(df = data_df,
    label_names = ['goodcredit'],
    protected_attribute_names = ['sex'])

# Create an instance of BinaryLabelDatasetMetric
data_metric = BinaryLabelDatasetMetric(data_ds,
    privileged_groups = privileged_groups,
    unprivileged_groups = unprivileged_groups)

# Compute the metrics with data_metric and store them in outcomes_df
outcomes_df.at['data', 'female'] = data_metric.base_rate(privileged=0)
outcomes_df.at['data', 'male'] = data_metric.base_rate(privileged=1)
outcomes_df.at['data', 'difference'] = data_metric.mean_difference()
outcomes_df.at['data', 'ratio'] = data_metric.disparate_impact()
```

In the notebook we look at these metrics on both the data and the model output for three different sets of the data (the full dataset, the train set and the test set) with two different models (one trained on the full dataset and another trained only on a subset of the data - the training set). In code listing A.1, we create a DataFrame to display the results in (`outcomes_df`) and populate the first row of it. First we define our privileged and unprivileged groups.

Defining privileged and unprivileged groups

The format for these is a list of dictionaries. Each dictionary in the list defines a group, the key being a feature and the value being the value of the feature for members of the group. The key, value pairs in the dictionaries are joined with an intersection (AND operator) and the dictionaries in the list are joined with a union (OR operator). So for example,

```
[{'sex': 1, 'age>=30': 1}, {'sex': 0}]
```

corresponds to individuals such that,

```
(data_df['sex']==1 AND data_df['age>=30']==1)  OR (data_df['sex']==0)
```

Next we create a `BinaryLabelDataset` object (`data_ds`) which in turn is used to create a `BinaryLabelDatasetMetric` object (`data_metric`). We then calculate the fairness metrics from `data_metric` and store the results in `outcomes_df`.

Exercise: Multiple sensitive features

Calculate independence metrics (base rates, difference and ratio) for the full dataset in the case where the privileged group is males age 30 and over, and the unprivileged group is females under the age of 30. Do this two ways, using AIF360 and using Pandas. Compare your results to make sure they match.

Once we have trained a model and made predictions, similar code can be written to calculate independence metrics on the model predictions for the full dataset. Code listing A.2 shows how we do this using the predictions from the trained model `clf`.

Listing A.2: Calculating independence metrics for the model using AIF360

```
# Create a DataFrame with the features and model predicted target
model_df = pd.concat([X, pd.Series(clf.predict(X), name='goodcredit')],
        axis=1)

# Create an instance of BinaryLabelDataset
model_ds = BinaryLabelDataset(df = model_df,
    label_names = ['goodcredit'],
    protected_attribute_names = ['sex_male'])

# Create an instance of BinaryLabelDatasetMetric
model_metric = BinaryLabelDatasetMetric(model_ds,
    privileged_groups = privileged_groups,
    unprivileged_groups = unprivileged_groups)

# Compute the metrics with model_metric and store them in outcomes_df
outcomes_df.at['model', 'female'] = model_metric.base_rate(privileged=0)
outcomes_df.at['model', 'male'] = model_metric.base_rate(privileged=1)
outcomes_df.at['model', 'difference'] = model_metric.mean_difference()
outcomes_df.at['model', 'ratio'] = model_metric.disparate_impact()
```

Table A.1 shows the results of the calculations stored in `outcomes_df` from the notebook. From Table A.1 we note some variation in the rates at which men and women are predicted to present good credit risks for the model versus the data. In particular, the model acceptance rates are higher for both male and female applicants than those observed in the data. There are particularly big differences when we compare results for the test data versus the model on the test data (test model), which is not surprising since the mean difference and impact ratio for the train data and test data are markedly different. In addition we are aware that our model is overfitting. Without intervention, our model appears to be reducing the bias present in the data for the test set (as measured by our independence metrics).

Exercise: Twin test

Implement the twin test (described in section 3.1.2) for the model trained on the full dataset. Calculate the causal mean difference between male and female applicants using 2000 data points (1000 male and 1000 female applicants) i.e. the full dataset together with the 'twin' of the opposite gender.

A.2.2 Comparing errors

In order to calculate balanced error metrics with AIF360, we need to create an object of type `ClassificationMetric`. Returning to our example working with the German Credit Data, code listing A.3 calculates a series of bal-

Table A.1: Acceptance rates for the Statlog (German Credit) Data Set.

	Female	Male	Difference	Ratio
Data	0.648	0.723	-0.0748	0.897
Model[a]	0.674	0.749	-0.0751	0.900
Train data	0.659	0.719	-0.0601	0.916
Train model[b]	0.667	0.731	-0.0647	0.911
Test data	0.607	0.741	-0.1345	0.819
Test model[b]	0.705	0.820	-0.1152	0.860

[a]Model trained on the full dataset.

[b]Model trained on the train dataset only.

anced error metrics and populates the DataFrame `errors_df` with them. Note that `data_ds` and `model_ds` were created, and `privileged_groups` and `unprivileged_groups` were defined in earlier code listings.

Listing A.3: Calculating balanced error metrics with AIF360

```
# Create a DataFrame to store results in
errors_df = pd.DataFrame(columns=['female', 'male',
                                  'difference', 'ratio'],
                         index=['ERR', 'FPR', 'FNR', 'FDR', 'FOR'])

# Create an instance of ClassificationMetric
clf_metric = ClassificationMetric(data_ds,
    model_ds,
    privileged_groups = privileged_groups,
    unprivileged_groups = unprivileged_groups)

# Compute the metrics with clf_metric and store them in errors_df
# Error rates for the unprivileged group
errors_df.at['ERR', 'female'] = clf_metric.error_rate(privileged=False)
errors_df.at['FPR', 'female'] =
    clf_metric.false_positive_rate(privileged=False)
errors_df.at['FNR', 'female'] =
    clf_metric.false_negative_rate(privileged=False)
errors_df.at['FDR', 'female'] =
    clf_metric.false_discovery_rate(privileged=False)
errors_df.at['FOR', 'female'] =
    clf_metric.false_omission_rate(privileged=False)

# Error rates for the privileged group
errors_df.at['ERR', 'male'] = clf_metric.error_rate(privileged=True)
errors_df.at['FPR', 'male'] =
    clf_metric.false_positive_rate(privileged=True)
errors_df.at['FNR', 'male'] =
    clf_metric.false_negative_rate(privileged=True)
errors_df.at['FDR', 'male'] =
    clf_metric.false_discovery_rate(privileged=True)
errors_df.at['FOR', 'male'] =
    clf_metric.false_omission_rate(privileged=True)
```

```
# Differences in error rates
errors_df.at['ERR', 'difference'] = clf_metric.error_rate_difference()
errors_df.at['FPR', 'difference'] =
    clf_metric.false_positive_rate_difference()
errors_df.at['FNR', 'difference'] =
    clf_metric.false_negative_rate_difference()
errors_df.at['FDR', 'difference'] =
    clf_metric.false_discovery_rate_difference()
errors_df.at['FOR', 'difference'] =
    clf_metric.false_omission_rate_difference()

# Ratios of error rates
errors_df.at['ERR', 'ratio'] = clf_metric.error_rate_ratio()
errors_df.at['FPR', 'ratio'] = clf_metric.false_positive_rate_ratio()
errors_df.at['FNR', 'ratio'] = clf_metric.false_negative_rate_ratio()
errors_df.at['FDR', 'ratio'] = clf_metric.false_discovery_rate_ratio()
errors_df.at['FOR', 'ratio'] = clf_metric.false_omission_rate_ratio()

display(errors_df)
```

The DataFrame `error_df` is shown in Table A.2. This time we just look at the metrics for the model trained

Table A.2: Error metrics for the Statlog (German Credit Data) Data Set.

Error metric[a]	Female	Male	Difference	Ratio
ERR	0.246	0.180	0.066	1.37
FPR	0.458	0.472	-0.014	0.97
FNR	0.108	0.078	0.030	1.39
FDR	0.250	0.152	0.098	1.65
FOR	0.235	0.296	-0.061	0.79

[a]We abbreviate error rate (ERR), false positive rate (FPR), false negative rate (FNR), false discovery rate (FDR) and false omission rate (FOR). See appendix B for detailed descriptions of confusion matrix metrics.

on the training set and calculated on the test set. We note that the overall error rate is 37% higher for female applicants. The false negative rate is 39% higher for female applicants, that is for female applicants we more often incorrectly predict that they represent bad credit risks when they are in fact good credit risks. We also note that the false discovery rate is 65% higher for female applicants which means that when we do predict women to be credit worthy they are more often not. The false omission rate is 21% lower for female applicants which means we are more often correct when we predict that they are not credit worthy. Our findings are not surprising given the difference in prevalence of credit worthy male and female applicants between our training and test sets shown in Table A.1.

Recall that when we compared fairness metrics under the independence criterion, it appeared that our model was reducing the level of bias in the data. Note that comparing balanced error metrics (in addition to independence metrics) gives us a richer understanding of the behaviour of our model in relation to protected groups.

Appendix B

Performance Metrics

Confusion Matrix Metrics

Performance Metrics

Table B.1: Summary of performance metrics for a binary classifier

		Ground Truth		
		$y = 1$	$y = 0$	Metric
Prediction	$\hat{y} = 1$	True Positive	False Positive Type I Error	Positive Predictive Value[a] $\mathbb{P}(\hat{y} = y \mid \hat{y} = 1)$
	$\hat{y} = 0$	False Negative Type II Error	True Negative	Negative Predictive Value $\mathbb{P}(\hat{y} = y \mid \hat{y} = 0)$
Metric		True Positive Rate[b] $\mathbb{P}(\hat{y} = y \mid y = 1)$	True Negative Rate $\mathbb{P}(\hat{y} = y \mid y = 0)$	Accuracy $\mathbb{P}(\hat{y} = y)$

[a] Positive Predictive Value = Precision

[b] True Positive Rate = Recall

Error Metrics

Table B.2: Summary of error rate types for a binary classifier

		Ground Truth		
		$y = 1$	$y = 0$	Error Rate Type
Prediction	$\hat{y} = 1$	True Positive	False Positive Type I Error	False Discovery Rate $\mathbb{P}(\hat{y} \neq y \mid \hat{y} = 1)$
	$\hat{y} = 0$	False Negative Type II Error	True Negative	False Omission Rate $\mathbb{P}(\hat{y} \neq y \mid \hat{y} = 0)$
Error Rate Type		False Negative Rate $\mathbb{P}(\hat{y} \neq y \mid y = 1)$	False Positive Rate $\mathbb{P}(\hat{y} \neq y \mid y = 0)$	Error Rate $\mathbb{P}(\hat{y} \neq y)$

Combined table

Table B.3: Summary of performance metrics for a binary classifier

Prediction	Ground Truth		Performance	Error rate
	$y = 1$	$y = 0$		
$\hat{y} = 1$	True Positive	False Positive Type I Error	Positive Predictive Value[a] $\mathbb{P}(\hat{y} = y \mid \hat{y} = 1)$	False Discovery Rate $\mathbb{P}(\hat{y} \neq y \mid \hat{y} = 1)$
$\hat{y} = 0$	False Negative Type II Error	True Negative	Negative Predictive Value $\mathbb{P}(\hat{y} = y \mid \hat{y} = 0)$	False Omission Rate $\mathbb{P}(\hat{y} \neq y \mid \hat{y} = 0)$
Performance	True Positive Rate[b] $\mathbb{P}(\hat{y} = y \mid y = 1)$	True Negative Rate $\mathbb{P}(\hat{y} = y \mid y = 0)$	Accuracy $\mathbb{P}(\hat{y} = y)$	
Error Rate	False Negative Rate $\mathbb{P}(\hat{y} \neq y \mid y = 1)$	False Positive Rate $\mathbb{P}(\hat{y} \neq y \mid y = 0)$		Error rate $\mathbb{P}(\hat{y} \neq y)$

[a] Positive Predictive Value = Precision

[b] True Positive Rate = Recall

Appendix C

Rules of Probability

Table C.1: Rules of probability

Rule	**Continuous Variables**	**Discrete Variables**
Sum rule	$f_X(x) = \int_{y \in \mathcal{Y}} f_{X,Y}(x,y)\,\mathrm{d}y$	$\mathbb{P}(x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(x,y)$
Product rule	$f_{X,Y}(x,y) = f_{Y\|X}(x,y) f_X(x)$	$\mathbb{P}(x,y) = \mathbb{P}(y\|x)\mathbb{P}(x)$
Bayes' rule	$f_{Y\|X}(x,y) = \dfrac{f_{X\|Y}(x,y) f_Y(y)}{f_X(x)}$	$\mathbb{P}(y\|x) = \dfrac{\mathbb{P}(x\|y)\mathbb{P}(y)}{\mathbb{P}(x)}$
Independence		
$X \perp Y$	$f_{Y\|X}(x,y) = f_Y(y)$	$\mathbb{P}(y\|x) = \mathbb{P}(y)$
From the product rule	$f_{X,Y}(x,y) = f_X(x) f_Y(y)$	$\mathbb{P}(x,y) = \mathbb{P}(x)\mathbb{P}(y)$
Conditional Independence		
$X \perp Y \| Z$	$f_{Y\|X,Z}(x,y,z) = f_{Y\|Z}(y,z)$	$\mathbb{P}(y\|x,z) = \mathbb{P}(y\|z)$
Using the product rule	$f_{X,Y\|Z}(x,y,z) = f_{Y\|X,Z}(x,y,z) f_{X\|Z}(x,z)$	$\mathbb{P}(x,y\|z) = \mathbb{P}(y\|x,z)\mathbb{P}(x\|z)$
Substituting for $Y\|X,Z$	$= f_{Y\|Z}(y,z) f_{X\|Z}(x,y)$	$= \mathbb{P}(y\|z)\mathbb{P}(x\|z)$

Appendix D

Solutions to Exercises

D.1 Chapter 3: Group Fairness

D.1.1 Comparing outcomes

Exercise: Normalised prejudice index

Write a function that takes two arrays y and z of categorical features and returns the normalised prejudice index. Hint:

1. Compute the probability distributions $\mathbb{P}(y)$, $\mathbb{P}(z)$ and $\mathbb{P}(y, z)$. Note that these can be thought of as the frequency with which each event occurs.
2. Compute the entropies $H(y)$ and $H(z)$ shown in equation (3.3) and use these to compute the normalising factor, $\sqrt{H(y)H(z)}$.
3. Compute the mutual information $I(z, y)$ shown in equation (3.1) and divide by the normalising factor.

You can test your implementation against scikit-learn's:
sklearn.metrics.normalized_mutual_info_score.

Listing D.1: Calculating the normalised prejudice index

```
# Import the necessary classes
import pandas as pd
import scipy.stats as ss

def normalised_mutual_information(x, y):
    """normalised mutual information between x and y"""

    # Compute the probability distributions
    px   = x.value_counts(normalize=True)
    py   = y.value_counts(normalize=True)
    pxy  = pd.Series(zip(x,y)).value_counts(normalize=True)

    # Compute the normalising factor
    norm = math.sqrt( ss.entropy(px) * ss.entropy(py)  )

    # Compute mutual information, divide by the normalising factor
```

```
    # and return the result
    return sum([p * math.log(p / (px[xy[0]] * py[xy[1]]))
                for xy, p in p_xy.items()]) / norm
```

Exercise: Statistical parity difference maximum

Show that

$$d_{\max} = \min\left\{\frac{\mathbb{P}(\hat{Y}=1)}{\mathbb{P}(Z=1)}, \frac{\mathbb{P}(\hat{Y}=0)}{\mathbb{P}(Z=0)}\right\}.$$

We can write statistical parity difference as

$$d = \mathbb{P}(\hat{Y}=1|Z=1) - \mathbb{P}(\hat{Y}=1|Z=0).$$

Let's rewrite this with advantaged and disadvantaged outcomes and groups to make it more concrete,

$$d = \mathbb{P}(y^+|z^+) - \mathbb{P}(y^+|z^-) = \frac{\mathbb{P}(y^+,z^+)}{\mathbb{P}(z^+)} - \frac{\mathbb{P}(y^+,z^-)}{\mathbb{P}(z^-)} \leq \frac{\mathbb{P}(y^+)}{\mathbb{P}(z^+)}.$$

This maximal value occurs when

$$\mathbb{P}(y^+,z^+) = \mathbb{P}(y^+) \quad \text{and} \quad \mathbb{P}(y^+,z^-) = 0;$$

that is, when all members of the advantaged class, receive the advantaged outcome. We can also write,

$$\begin{aligned} d = \mathbb{P}(y^+|z^+) - \mathbb{P}(y^+|z^-) &= \mathbb{P}(y^-|z^-) - \mathbb{P}(y^-|z^+) \\ &= \frac{\mathbb{P}(y^-,z^-)}{\mathbb{P}(z^-)} - \frac{\mathbb{P}(y^-,z^+)}{\mathbb{P}(z^+)} \leq \frac{\mathbb{P}(y^-)}{\mathbb{P}(z^-)}. \end{aligned}$$

Here the maximal value occurs when

$$\mathbb{P}(y^-,z^-) = \mathbb{P}(y^-) \quad \text{and} \quad \mathbb{P}(y^-,z^+) = 0;$$

that is, when all members of the disadvantaged class, receive the disadvantaged outcome. Thus,

$$d_{max} = \min\left\{\frac{\mathbb{P}(y^+)}{\mathbb{P}(z^+)}, \frac{\mathbb{P}(y^-)}{\mathbb{P}(z^-)}\right\}.$$

Note that,

$$\frac{\mathbb{P}(y^+)}{\mathbb{P}(z^+)} = \frac{\mathbb{P}(y^-)}{\mathbb{P}(z^-)} \quad \Leftrightarrow \quad \mathbb{P}(y_+) = \mathbb{P}(z_+);$$

that is, when all members of the advantaged class, receive the advantaged outcome and all members of the disadvantaged class, receive the disadvantaged outcome.

Exercise: Odds ratio

Show that the odds ratio is always greater than or equal to one in the case where $\hat{Y} = 1$ in the advantaged outcome and $Z = 1$ is the privileged group.

$$r_{\text{odds}} = \frac{\mathbb{P}(\hat{Y}=1|Z=1)\mathbb{P}(\hat{Y}=0|Z=0)}{\mathbb{P}(\hat{Y}=0|Z=1)\mathbb{P}(\hat{Y}=1|Z=0)}.$$

Let $\hat{Y} = 1$ be the advantaged outcome and $Z = 1$ be the privileged group, then we can write,

$$r_{\text{odds}} = \frac{\mathbb{P}(\hat{y}_+|z_+)\mathbb{P}(\hat{y}_-|z_-)}{\mathbb{P}(\hat{y}_-|z_+)\mathbb{P}(\hat{y}_+|z_-)}$$

In this case, since $\mathbb{P}(\hat{y}_+|z_+) > \mathbb{P}(\hat{y}_+|z_-)$ and $\mathbb{P}(\hat{y}_-|z_-) > \mathbb{P}(\hat{y}_-|z_+)$, the numerator is always greater than the denominator and the odds ratio will be greater than one.

D.1.2 Comparing errors

Exercise: Sufficiency

Show that sufficiency is satisfied if and only if the false omission rate and false discovery rate are equal for all groups.

Sufficiency implies

$$\mathbb{P}(y|\hat{y}, z) = \mathbb{P}(y|\hat{y}).$$

For the simplest case of a binary classifier where we have a single sensitive binary feature. We can write this requirement as two conditions,

$$\mathbb{P}(Y = 1|Z = 1, \hat{Y} = 1) = \mathbb{P}(Y = 1|Z = 0, \hat{Y} = 1),$$
$$\mathbb{P}(Y = 1|Z = 1, \hat{Y} = 0) = \mathbb{P}(Y = 1|Z = 0, \hat{Y} = 0).$$

Recall that $\mathbb{P}(Y = 1|\hat{Y} = 1)$ is the positive predictive value (PPV) of the classifier and $\mathbb{P}(Y = 1|\hat{Y} = 0)$ is the false omission rate (FOR). We see then that sufficiency requires the positive predictive value to be the same for all values of the sensitive feature and the false omission rate to be the same for all values of the sensitive feature. Note that the positive predictive value is balanced if and only if the false discovery rate is balanced, so thinking in terms of error metrics only, separation requires the false discovery and false omission rates to be balanced.

D.1.3 Incompatibility of fairness criteria

Separation versus Sufficiency

Exercise: Predictive values

Prove the results given in equations (3.13) and (3.14).

We want to write the positive and negative predictive values (PPV and NPV respectively) in terms of the true positive, false positive and acceptance rates (TPR, FPR and p respectively). We start by looking at some relationships between the elements of a confusion matrix shown in Table D.1. where $n = TP + FP + FN + TN$ denotes the total number of data points. Using the equations in the final row of the table we can write,

$$pTPR = \frac{TP}{n}, \qquad (1-p)FPR = \frac{FP}{n},$$
$$p(1 - TPR) = \frac{FN}{n}, \qquad (1-p)(1 - FPR) = \frac{TP}{n}.$$

Table D.1: Confusion matrix

		Ground Truth		
		$y = 1$	$y = 0$	
Prediction	$\hat{y} = 1$	True Positive (TP)	False Positive (FP)	$PPV = \dfrac{TP}{TP + FP}$
	$\hat{y} = 0$	False Negative (FN)	True Negative (TN)	$NPV = \dfrac{TN}{FN + TN}$
		$TPR = \dfrac{TP}{TP + FN}$ $1 - TPR = \dfrac{FN}{TP + FN}$ $p = \dfrac{TP + FN}{n}$	$FPR = \dfrac{FP}{FP + TN}$ $1 - FPR = \dfrac{TN}{FP + TN}$ $1 - p = \dfrac{FP + TN}{n}$	

Finally, we can substitute these into our expressions for PPV and NPV in the right hand column of Table D.1 to find the relationships in equations (3.13) and (3.14).

$$PPV = \frac{pTPR}{pTPR + (1-p)FPR}$$
$$NPV = \frac{(1-p)(1-FPR)}{p(1-TPR) + (1-p)(1-FPR)}.$$

Exercise: Separation versus sufficiency

Show that for separation and sufficiency to hold equations (3.15) and (3.16) must hold for for any pair of groups $Z = a$ and $Z = b$.

For separation to hold the true positive rate (TPR) and false positive rate (FPR) must be constant across all values of the sensitive features. For sufficiency to hold the positive predictive value (PPV) and negative predictive value (NPV) must be constant across all values of the sensitive features. For brevity we shal use a subscript to denote conditioning on Z, for example $p_z = \mathbb{P}(Y = 1|Z = z)$. For both separation and sufficiency to hold, we must have

$$\begin{aligned}
& PPV_a = PPV_b \\
\Leftrightarrow \quad & \frac{p_a TPR}{p_a TPR + (1-p_a)FPR} = \frac{p_b TPR}{p_b TPR + (1-p_b)FPR} \\
\Leftrightarrow \quad & p_b TPR[p_a TPR + (1-p_a)FPR] = p_a TPR[p_b TPR + (1-p_b)FPR] \\
\Leftrightarrow \quad & p_b TPR(1-p_a)FPR = p_a TPR(1-p_b)FPR \\
\Leftrightarrow \quad & TPR(p_b - p_a)FPR = 0.
\end{aligned}$$

Similarly,

$$
\begin{aligned}
& NPV_a = NPV_b \\
\Leftrightarrow \quad & \frac{(1-p_a)(1-FPR)}{p_a(1-TPR)+(1-p_a)(1-FPR)} = \frac{(1-p_b)(1-FPR)}{p_b(1-TPR)+(1-p_b)(1-FPR)} \\
\Leftrightarrow \quad & (1-p_b)(1-FPR)[p_a(1-TPR)+(1-p_a)(1-FPR)] \\
& \qquad = (1-p_a)(1-FPR)[p_b(1-TPR)+(1-p_b)(1-FPR)] \\
\Leftrightarrow \quad & (1-p_b)(1-FPR)p_a(1-TPR) = (1-p_a)(1-FPR)p_b(1-TPR). \\
\Leftrightarrow \quad & (1-FPR)(p_b-p_a)(1-TPR) = 0.
\end{aligned}
$$