National Research University Higher School of Economics
Faculty of Computer Science
Programme 'Master of Data Science'

MASTER'S THESIS

Causal Bayesian Networks for correction of potential bias in ML models

English title

Причинно-следственные байесовские сети для коррекции потенциальной систематической ошибки в моделях машинного обучения

Russian Title

Qualification paper – Master of Data Science Dissertation Field of study 01.04.02 «Applied Mathematics and Computer Science»

Carl Bylin

Student

Alexey Lugovoy

Supervisor

Moscow, 2021

Abstract

It is now public knowledge that systems created for automated decision-making tend to absorb the bias of the underlying datasets used for training or from the practitioners creating these systems.

Identifying and correcting bias in machine learning is not always an easy task. The main purpose of this work is to 1) briefly describe the problem of fairness in ML models and the work that has been done in this area, 2) offer a deep explanation on how Causal Bayesian networks can be used to identify and correct unfair path-specific bias, and 3) present practical examples of this proposed solution in action.

We will start with a motivating example before going through the concepts of fairness in ML, causality and Bayesian Networks. By the end of this work we will have the necessary context to put these concepts into practice with concrete examples using synthetic data and a publicly available dataset.

A main guiding principle of this work is that a causal lens through the explicit definition of a causal graph to model the underlying data generating process is necessary in most cases to accurately measure and counteract bias in complex processes.

Abstract

Сегодня общеизвестно, что системы, созданные для автоматизированного принятия решений, как правило, учитывают предвзятость самих базовых наборов данных, используемых для обучения или тех, кто их генерирует.

Выявление и исправление предвзятости в машинном обучении не всегда является простой задачей. Основной целью данной работы является 1) краткое описание проблемы предвзятости в моделях ML и работы, которая была проделана в этой области, 2) предложить чёткое объяснение того, как причинно-следственные байесовские сети могут быть использованы для идентификации и коррекции предвзятых специфических путей смещения, и 3) представить практические примеры этого предложенного решения в действии.

Начнем с мотивирующего примера, а затем рассмотрим понятия справедливости в ML, причинности и байесовских сетей. К концу этой работы мы будем иметь необходимый контекст для применения этих концепций на практике с помощью конкретных примеров с использованием синтетических данных и общедоступного набора данных.

Основным руководящим принципом данной работы является четкое определение причинно-следственного графа для моделирования основного процесса генерирования данных, необходимого в большинстве случаев для точного измерения и противодействия смещению в сложных процессах.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and significance of the topic

Traditionally, only human agents like presidents, generals, managers, executives, and other leaders made decisions that could impact thousands or millions of lives. This doesn't guarantee that they have acted correctly, but they stand accountable for their decisions. In today's technological landscape we're seeing organizations that interact with millions of users through automated decision-making agents. As these agents become more complex, they are being increasingly used to make high-stakes decisions [1] that could have important consequences on people's lives.

Unfortunately, the underlying components of the algorithms making these important decisions have been shown [2] to often inherit bias from the training data or from the people creating the system, which can potentially have a negative influence on the decisions made by the system. Even more problematic is the fact that these systems are usually made up of multiple components managed by different teams such that identifying and correcting bias is extremely difficult and accountability for the results is diluted.

The purpose of this work is to build on top of existing material in this topic to facilitate understanding of how bias in machine learning systems is produced and which tools can be used to accurately identify and correct this bias.

1.2 Structure of the work

In chapter 2 we start with a review regarding how fairness has been approached in the domain of machine learning up until now. This includes a closer look into how bias is generated and the proposed methods for measuring bias. The chapter closes by explaining the connection between bias and causality, which is one of the central themes of this report.

This connects directly with chapter 3 on causality. We will go through a motivating example and explain why this is such a critical topic in the context of bias in machine learning systems.

Chapter 4 explains the main framework we will use to connect causality to bias. In this part of the report we will understand how Bayesian Networks serve as a form of knowledge representation encoding our understanding of uncertainty and relationships between variables in the data-generating process.

After this, in chapter 5 we will present a practical application of Causal Bayesian Networks to quantify and correct bias in trained machine learning models using both generated and publicly available data.

The report ends with closing remarks and opportunities for further research.

1.3 Related work

The main inspiration of this report is a blog post by DeepMind researchers Silvia Chiappa and William Isaac titled "Causal Bayesian Networks: A flexible tool to enable fairer machine learning" [3]. This post was in turn based on the papers "Path-Specific Counterfactual Fairness" [4] and "A Causal Networks Viewpoint on Fairness" [5].

Chapter 2

2. Background on Fairness in Machine Learning

It is impossible to escape the fact that the concept of fairness is highly philosophical and no universally accepted definition exists at this point in time. Multiple definitions and perspectives on fairness have been presented over the centuries. Some argue that justice and fairness are closely connected [6] while others try to establish if fairness is equivalent or not to equality [7]. Invariably, many connect the principle of fairness to a notion of moral obligation [8]. We will not assume a particular definition for this work. Instead we will explore how the concept of fairness has been used in computer science and other relevant details we should keep in mind when analyzing bias through a causal lens.

As an additional note, it is important to mention that in this work we will use the term "bias" interchangeably with "unfairness" in a machine learning model as there are many other types of bias not related to the concept of fairness.

2.1 Sources of bias in machine learning

The authors in [1] have thoroughly investigated multiple real-world applications to better understand how bias is introduced into AI applications. On top of this they have also collected an in-depth list of 23 different types of bias that are often found in AI systems, as well as in which stages of the AI life-cycle they are introduced. Finally, the authors also present a taxonomy of different types of discrimination and

a list of proposed solutions by domain and sub-domain.

Two of the main components of any AI application where bias is generally introduced is through the algorithm and through the data. For example, one of the simplest tools used to present content to users is a popularity-based recommendation algorithm. By definition, this type of algorithm will end up favoring the opinion of the masses, which can often be radically different to the opinion of the minorities in the community. Another example is the choice of optimization function. This function is the means through which decision makers communicate to the machine how the ideal solution should be. An optimization function that has been designed to only take profits into account might optimize for this goal and unwillingly introduce bias towards a group of users even if this bias was not originally present in the training data.

There are also many ways in which bias can be introduced into the system through the data used to train the model. This could for instance be due to unfairness in society causing unfair treatment of certain groups that then is learned and promoted by the machine learning system. But it could also be caused by an incorrect collection of data that only focuses on a particular subset of the population such that members of other subsets are treated unfairly.

There are clearly also cases when bias is produced by both data and algorithm. During training it might be discovered that a certain subgroup is treated unfairly, but trying to remove the features that identify this subgroup can potentially severely reduce the model's predictive power such that the team decides to allow the algorithm to optimize for the business goal without applying a constraint to ensure fairness.

Still, given that most of the biases identified by the authors of the survey are directly related to the data we conclude that it is important to build an understanding on how the variables interact with each other to produce the data used for training. Pearl describes in [9] that Simpson's paradox can be resolved by either using an aggregated or segregated view of the data and this can not be decided from the data alone. Only the understanding of the data generating process and its underlying causal structure permits a clear decision on which of the two options to use.

2.2 Existing methods for evaluating unfairness

This section does not aim to be an exhaustive presentation of fairness measures, this is much better done in [10]. More broadly, three main groups are shown: statistical measures, similarity-based measures, and methods based on causal reasoning.

Out of the reviewed measures we will take a look at some of the more widely used in order to set the context for the rest of this current work.

2.2.1 Fairness through unawareness

This definition is based on the intuition that a decision should not be unfair if the decision-maker is blind to the sensitive attribute. We present this option first because it is widely enforced legally to reduce unfairness in automated systems.

Unfortunately, multiple researchers have shown [11] that simply removing protected attributes from the data in order to obtain "fairness through unawareness" does not always have the intended effect. We will come back to this topic in the context of causality.

2.2.2 Statistical parity

Statistical (or demographic) parity means that the prediction should be statistically independent of the sensitive attribute [12]. In the binary case:

$$P(\hat{Y} = y|A = 1) = P(\hat{Y} = y|A = 0) \tag{2.1}$$

This means, that information about the sensitive attribute should not tell us anything new about the potential outcome of the model.

A common objection to this measure is that even if it does help to keep a model fair at a group level, it is often blind to unfairness at the individual level. This is shown in [13] through various examples where data can be modified in such ways that statistical parity is maintained but there is still unfairness in the predictions. One example is precisely unfairness through unawareness as in the domain of advertising. The authors suggest that ads can be created in such a way that showing the add obeys statistical parity while the next action of clicking is highly correlated with the sensitive attribute.

2.2.3 Predictive parity

In this case the actual outcome should be conditionally independent of the sensitive attribute given a positive prediction:

$$P(Y|\hat{Y}=1,A=1) = P(Y|\hat{Y}=1,A=0) \tag{2.2}$$

For instance, a technical analysis from Northepointe was done in [14] to evaluate claims by ProPublica in [15] that the COMPAS risk scales, created for predicting criminals, is biased against blacks. The technical analysis uses the measure of predictive parity to defend the fairness of the tool.

2.2.4 Calibration

It is said [16] that $\hat{Y}$ satisfies the calibration criterion if given the predicted outcome, the actual outcomes and the sensitive attribute are conditionally independent:

$$Y \perp\!\!\!\perp A|\hat{Y} \tag{2.3}$$

This means that given a predicted outcome (i.e. recommend to hire), the actual outcomes should be the same for members of different groups.

It is argued in [17] that the calibration criteria is impossible to satisfy at the same time as other criteria like Equalized Odds (except in very constrained cases) and [5] mentions that EFPRs/EFNRs and calibration are incompatible in the COMPAS case due to differences in base rates across groups.

2.2.5 Equal False Positive Rates and Equal False Negative Rates

Another common intuition or way of thinking of fairness is that if the model makes a mistake then the rate of mistakes should be the same among the different groups defined by the sensitive attribute. In other words, police should not falsely arrest more blacks than whites if we assume that the arrested are actually not criminals. The opposite case would indicate that the police are maybe arresting more of one group than the other.

Still, EFPRs/EFNRs have fallen out of favor for some because it also fails to

identify certain cases of unfairness. A perfect model, for example, would have no false positives or false negatives and this criteria would be satisfied even if there actually is unfairness in the underlying system.

In [18], the author goes even further and argues that equal false positive rates and equal false negative rates actually don't provide any information about fairness and that the confusion is rooted in a philosophical misunderstanding such that it is mistakenly considered a type of equality of opportunity.

2.2.6 Equality of opportunity

Khan et al. [19] is one of the most recent attempts to understand fairness in the context of Automated Decision Systems. The authors argue that the concept of fairness in this context should be rooted in the doctrines of fairness as Equality of Opportunity. This paper explores some of the real difficulties that explain why it is so hard to identify a lack of fairness when the concept is not clearly defined. For example, "formal equality of opportunity" is one extreme of the spectrum that considers only the qualifications that are relevant to an opportunity as the basis for a fair decision. Nevertheless, this particular mindset would not correct for inherent biases in society that unfairly distribute the possibility to obtain those qualifications in the first place. This is a first glimpse into the main theme of this report. Some notions of fairness require that we understand the causal dependencies in the data generating process to be able to identify bias.

2.2.7 Counterfactual fairness

Back in 2011, Dwork et al. [13] decided to analyze fairness specifically in supervised classification tasks with the purpose of avoiding the discrimination of certain individuals specifically due to their affiliation to some group (i.e. gender, religion, age, etc.) while balancing the predictive power of the model. The authors review multiple definitions of fairness, analyze how some measures like statistical parity fail to identify bias in some cases and propose a hypothetical metric based on similarity that can be used as a constraint during training to guarantee fairness. The justification is based on the notion that two individuals that are similar to each other

should also be treated similarly by the machine learning system.

A few years ago, Kusner et al. [20] took this concept even further to establish the notion of "counterfactual fairness". This concept is deeply rooted in the language of causality. The question of bias is no longer centered on similar individuals, but on the same individual in the real world compared to a counterfactual world where the person belongs to a different demographic group or the sensitive attribute is different. Through the use of interventions on certain variables the authors argue that it is possible to create a counterfactual such that for any individual the predictor produces the same result independently of the value of the protected attribute.

Building on top of the concept of counterfactual fairness, in [4] a new definition of path-specific counterfactual fairness is presented in response to some of the limitations of a general counterfactual fairness concept. The authors argue that there are multiple paths through which information can flow between a protected attribute and the predictor, some of which are unfair while others can be perfectly fair. Counterfactual fairness fails to make this distinction.

Path-specific counterfactual fairness also differs in the way it is used to eliminate bias by correcting only the descendants of the sensitive attributes through unfair paths. In contrast, the authors of counterfactual fairness propose constraints during training that usually have a more negative impact on the model's predictive power and interpretation.

2.2.8 Other measures

In [21], the authors propose using Empirical Risk Minimization subject to a fairness constraint defined in the loss function such that the model is considered fair if the error produced on the positive class is the same across all possible values of the sensitive attribute.

The way bias is introduced into an AI system is also highly dependent on how data is encoded and presented. For example, [22] tackles the issue of bias in the domain of Natural Language Processing, and more specifically in word embeddings. As seen in [23], words of similar semantic meaning tend to posses ge-

ometric similarity in the embedded vector space. Surprisingly, this similarity is also maintained after algebraic operations as seen in the now classical example: vector(”King”) − vector(”Man”) + vector(”Woman”) = vector(”Queen”).

This intuition has produced work like [24] and [25] where the authors use different applications of cosine similarity to identify and correct bias in word embeddings. A biased solution might produce the following result: vector(”Doctor”) − vector(”Man”) + vector(”Woman”) = vector(”Nurse”). Decisions made by this system would probably perpetuate a historical bias where being a doctor was considered a profession for men and being a nurse a profession for women. This method seems intuitive and easy to correct, but [22] proves that cosine-similarity based metrics often fail in reporting bias and propose an alternative metric. Still, as in other machine learning domains, this is far from being a solved task.

As we can see, it has not been easy to find a direct, reliable and consistent way of measuring fairness of a model during and after training. For this reason, [22] identifies that lately more effort has also been put into measuring bias in downstream tasks were the identification of bias can be more noticeable. Clearly, this method should be used more as a safeguard in addition to previous models because it assumes that bias is already present in upstream tasks.

2.3 The relationship between bias and causality in ML models

We have now seen that there is certain division between practitioners and researchers regarding which measures of fairness to use and how these should be interpreted. In this context, causal reasoning is not presented as a silver bullet but it is considered to solve some of the problems when purely observational criteria fall short.

As an example of this, [12] presents two different worlds that admit the same joint distribution, but accept multiple social interpretations. Trying different types of observational measures of fairness will not provide a definitive answer, instead it becomes necessary to form a set of assumptions regarding the data generating

process.

A clear understanding of the data generating process also permits us to find proxies for sensitive attributes such that we can intervene on the proxies when the sensitive attributes are not observed.

Indeed, [4] takes it even further and shows that under a specific data generating process modeled as a graph there can be cases where information flows from the sensitive attribute to the predicted outcome through both fair and unfair paths.

In [26], the authors clearly articulate that we have been wrong in using simple metrics as proxies for how humans treat fairness in their decisions. Human beings make their decisions within the context of philosophical, ethical and cultural frameworks. These are all external to the data we collect and can't be deduced from the data.

From here we will now explore the tools we will use in this paper to be able to clearly encode our assumptions and the causal relationships into a model that can be used to evaluate and correct potential bias in ML models.

Chapter 3

Causality

"Everything that begins to exist has a cause of its existence". This is the first premise of *The Kalam Cosmological Argument* as presented in [27].

Clearly, the topic of causality is even older and greater than the topic of fairness and we will by no means go into the philosophical ramifications of causality. Still, the premise above was shared to show that at least intuitively, we humans think in terms of "cause and effect".

In *The Book of Why* [28], Judea Pearl explains how using a causal lens has allowed him to see scientific questions differently instead of blindly applying the classical "Correlation is not causation" mantra that for many decades seemed to imply a prohibition on talking about causality. Pearl describes [29] the causal lens as the underlying assumption that "there exists an unknown but true Data Generating Process (DGP) that explains the world". This means that we can ask questions to the DGP and get answers about the world.

As the true DGP is unknown we can try to model it with data in combination with our assumptions about the inherent dependencies or relationships between relevant variables.

Next we will illustrate with a motivating example the difference between viewing only data and viewing the data through a causal lens by also analyzing the true data generating process.

3.1 Motivating example

Imagine this hypothetical scenario. Your country has been receiving a high number of immigrants these last few years due to political turmoil in their home countries. Many think this is a good thing as they believe it is a collective moral responsibility to help those in need. Still, some are worried about the effect this has on the job market as the rate of unemployment has been slightly higher than normal.

An independent organization collects data about hiring rates among locals and foreigners. The government has also been promoting the use of a public employment agency for all in need to try to reduce overall unemployment rates. As the first results come out, opinions about the incoming wave of immigrants become polarized. It seems like foreigners have a higher employment rate than locals:

Foreigner	0	1
Job	0.64	0.68

Figure 3.1: Employment rates

We can read this table as $P(Job|Foreigner = 1) = 0.68$ and $P(Job|Foreigner = 0) = 0.64$. This is extremely polarizing because some think the foreigners are leaving locals without jobs. At the same time foreigners have started to vocally express that their experience is the complete opposite. To them it feels like the hiring managers usually prefer hiring locals instead of foreigners. To make things worse, the independent organization releases the following segmented view of the same data used to produce the previous results separating whether the person used the public employment agency or not to find the job:

Foreigner	0	1
P(Job)	0.64	0.68
P(Job|Agency=1)	0.81	0.75
P(Job|Agency=0)	0.60	0.40

Figure 3.2: Segmented employment rates by usage of employment agency

It seems like both the group of people proclaiming unfairness against locals and the group of people proclaiming unfairness against foreigners are correct. This

is indeed paradoxical because in this hypothetical example we assume only two possible states: foreigner or local. How can their be bias against both groups? No other variables than the ones mentioned until now (foreigner status and usage of an employment agency) have been used to produce the set of data for this example. Additional to this, how can there be an overall bias in favor of foreigners at the same time that there exists bias against foreigners that use the employment agency and foreigners that don't use the employment agency when both these options account for the whole population of foreigners?

It seems almost impossible from the data alone. This is a case of what is known as Simpson's Paradox as described in [9]. Now we will try a small trick to see if this changes. We can simply change the labels while still using the same data. In this first case we will evaluate the probability of high cholesterol levels in young and elderly. We will also segment by the tendency to exercise:

Exercise	0	1
P(Cholesterol)	0.64	0.68
P(Cholesterol|Young=1)	0.81	0.75
P(Cholesterol|Young=0)	0.60	0.40

Figure 3.3: Cholesterol rates by tendency to exercise

Higher cholesterol for people that exercise compared to people who don't exercise feels intuitively wrong because in school we usually learn about the effect exercise has on the human body. Instead, if we focus on just the young people we notice that they have a lower chance of having high cholesterol levels if they exercise. The same holds true for the elderly. It seems like the segmented view is the correct one, but is it always so? We can change the labels again:

No exercise	0	1
P(Cholesterol)	0.64	0.68
P(Cholesterol|Sunny=1)	0.81	0.75
P(Cholesterol|Sunny=0)	0.60	0.40

Figure 3.4: Cholesterol rates by weather

In this case higher cholesterol for people that don't exercise (No exercise=True) compared to people who do exercise feels like the correct order of nature. We can also see that now the segmentation is based on weather. Once again, what we have learned about how the world works immediately raises a red flag when we see that "not exercising" has a different impact on cholesterol levels depending on the weather. In this case, the aggregated view definitely seems correct. Next we will attempt to understand why.

3.2 Causal graphs and the data generating process

In all these cases we have used the same data but the conclusions have been different. What made the exercise so much easier in the last two examples was the previous knowledge we have about how the world works. In a sense, we carry our own mental models of the data generating processes. The order of cause and effect tells us something with respect to how we should look at the data.

An objection can be made here saying that the tables actually seemed to indicate that weather does give us information about varying cholesterol levels. In this case it is important to remember that weather and exercise often have a causal relationship, people tend to exercise more when it is sunny. This relationship means that we can use "inverse probability" to get information about the cause from the effect. But more on that in chapter 4.

All of the examples we have shown so far follow this causal graph:

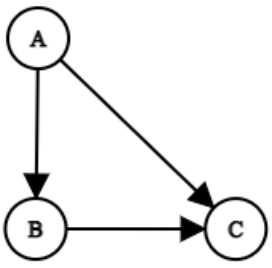

Figure 3.5: Simple Causal Graph

In scenario 1: {'A': 'Foreigner', 'B': 'Employment Agency', 'C': 'Job'}

In scenario 2: {'A': 'Youth', 'B': 'Exercise', 'C': 'Cholesterol'}

In scenario 3: {'A': 'Sunny', 'B': 'No exercise', 'C': 'Cholesterol'}

Looking close at the three examples, we can see that scenario 1 and 2 follow the same causal order, while scenario 3 is inverted. This would seem to indicate that the segmented view might be correct in the fairness case.

Indeed, this can be confirmed by analyzing the predefined parameters in the true data generating process:

F	F(foreigner)	F(foreigner)	F(local)	F(local)
A	A(agency)	A(no agency)	A(agency)	A(no agency)
J(hired)	0.75	0.4	0.8	0.6

Figure 3.6: Fairness DGP parameters

Here we clearly see the underlying assumptions of the data generating process. This hypothetical scenario was designed assuming a strong bias from hiring managers against foreigners (we do not assume this is necessarily the case in the real world). Going through an employment agency practically eliminates this bias because in this imaginary world we assume that the hiring managers have decided to hire applicants from the employment agency without access to sensitive attributes like place of origin (although a slight bias was included by the agency itself which could be attributed to slightly more paperwork for foreigners).

This example is consistent with Simpson's paradox because there is a reversal in hiring rates when analyzing segmented and aggregated data. The reason this happens is because of different underlying rates in the population and different preference rates for using the public employment service or not:

F(foreigner)	0.2
F(local)	0.8

Figure 3.7: Place of origin rates

F	F(foreigner)	F(local)
A(agency)	0.8	0.2

Figure 3.8: Usage of employment agency rates

The data generating process assumes that that 20% of the population are foreigners and 80% of foreigners use the public employment agency. On the other hand, the majority of people are locals and the majority of locals prefer to apply directly without going through the employment agency.

3.2.1 Causality and subjective beliefs

The examples in this chapter have shown us that we are limited in the type of conclusions we can make if we don't have a causal lens that we can use.

In [28], Pearl explains that causal analysis can't be done with data alone. While the vast majority of work that has been done in statistics relies on objective analysis of hard facts, Pearl explains that causal analysis requires that the researcher makes a subjective commitment by drawing a diagram that represents, to the best of his or her knowledge, the nature of the causal processes at work.

This idea of working with subjective beliefs immediately reminds us of a very polemical domain in statistics that we will explore in the next chapter.

We now have a general understanding of why causality is important, but in order to move towards more complex and important topics we need to find a method to explicitly represent causal relationships between variables. The tool we will be using for this is the Bayesian network.

Chapter 4

Bayesian networks

A major topic of discussion in AI is how to encode the structure of the problems we have in such a way that the machine can find a solution.

One such attempt that has been widely used is known as constraint satisfaction problems [30] consisting in the definition of variables (with their respective domains) and constraints that the proposed solution has to satisfy.

Many of the languages and frameworks used to encode problems over the years have fallen short. One of the main reasons for this is because some of these frameworks treat modeling as an essentially deterministic problem. Hard-coding the unimaginable number of states of a world quickly makes defining and solving the problem intractable.

For this reason, it becomes essential to use a framework of knowledge representation that allows us to define and model uncertainty in addition to the variables and their relationships. In this work we will focus on Bayesian networks.

4.1 Basics of Bayesian Networks

Bayesian networks are in one sense the marriage between graph theory and probabilistic reasoning. More formally, a Bayesian network represents a joint probability model over a set of variables [31] where a joint probability is considered to be the probability of a particular state of the world (a unique combination of events happening together).

4.1.1 Joint and conditional probabilities

Probability distributions assign specific probabilities for each possible value of a random variable. It is also worth noting that we don't necessarily have to store all of these in a table as most computations would be prohibitive when we add more variables. The relationship between event and probability in a random variable can be defined through a mathematical function.

As mentioned above, we can compute the probability of two or more events happening together by evaluating their joint probability:

$$p(x,y) = P(X = x, Y = y) \tag{4.1}$$

Analyzing the relationship between two variables opens up new possibilities, like reasoning with partial information [32]. Even though we might not have full information about a problem, knowing the value of one variable can effectively change the probability distribution of another variable. For example, our belief of someone having a particular disease will probably be different before and after we have access to the result of a medical test.

In general, a conditional probability distribution can be defined as:

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} \tag{4.2}$$

It becomes then trivial to see that we can actually use conditional probabilities to specify unconditional probabilities (with simplified notation):

$$P(X,Y) = P(Y|X)P(X) \tag{4.3}$$

This relationship can be extended to even more variables through the use of the chain rule so that the joint probability can be computed as the product of a series of conditional probabilities:

$$P(X_1,\ldots,X_n) = P(X_1|X_2,\ldots,X_n)P(X_2|X_3,\ldots,X_n)\ldots P(X_n) = \prod P(X_i|X_{i+1},\ldots,X_n) \tag{4.4}$$

With this we have now found a way of encoding uncertainty into our modeling process, but the previous equations do very little to help us understand how the variables actually influence each other.

Indeed, if we know that there is independence between certain variables we can greatly simplify the computations we need to make.

4.1.2 Directed acyclic graphs

Graph theory helps us to better formalize the relationships between the random variables by allowing us to represent each variable as a node in a graph. Directed edges between nodes can then be used to explicitly state the conditional dependencies between the variables such that every node X_i is defined by the conditional distribution $p(X_i|\text{pa}_i)$ considering pa_i as the set of parents of X_i.

In other words, the probability of a node taking a particular value given the values of it's ancestors can be obtained by multiplying the conditional probabilities leading up to that node through equation 4.4.

For instance, take the following simple graph:

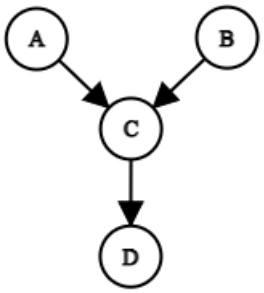

Figure 4.1: Directed Acyclic Graph

Here C is a parent of D (that is equivalently a child of C). Both A and B are parents of C and ancestors of D (also known as a descendant of A and B). With this

information we can now state that:

$$P(D) = P(D|A,B,C)P(C|A,B)P(A)P(B) \tag{4.5}$$

It is worth noting here that in practice these graphs have a set of root nodes (as seen above) that don't have any parents. If this were not the case, we would notice a major limitation known as the Qualification problem [33]: trying to fully define our model with all possible preconditions quickly becomes prohibitive.

Finally, by definition we do not allow cycles in a directed acyclic graph, which would allow for a node to potentially be its own ancestor. This is a simplifying assumption which can be relaxed as in the case of Markov networks.

4.2 Bayes' rule

Now that we have reviewed some of the basics, we can pick up where we left of in chapter 3. The final statement implied that causal analysis works beyond the data alone. Practitioners have to use prior knowledge and expertise to create a model that represents their subjective beliefs about how the world works. They do this by drawing the variables and their relationships using a directed acyclic graph, as seen in the previous section.

This leads us to Bayes' theorem that has a strong relationship with the concepts of joint and conditional probabilities that we have already discussed. By substituting and re-ordering in equations 4.1, 4.2 and 4.3 we can easily arrive at the following equality:

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A) \tag{4.6}$$

Knowing this we can discard the joint probability altogether and only focus on the relationship between the conditionals. From this we obtain the following relationship that has been so polemical and at the same time so benefitial in many domains throughout the years:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4.7}$$

This is the famous Bayes' rule, which we assume the reader already knows so we can divert back to one of our main topics of this work: causality.

In equation 4.7 variables A and B are placeholders for practically any event we might want to model. We can now change our interpretation of this formula to explicitly state a powerful concept:

$$P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})} \tag{4.8}$$

There are multiple components in this equation that we can now break down to build an even better intuition for the coming sections of this work.

To illustrate this we can go back to our scenario of foreigners arriving in a new country. In this case the foreigner needs to apply for a residency to be able to stay in the country. Let's say that there are only three available options: a student visa, a work visa or a visa to move to a family member. Each type has a different set of prerequisites and the resulting visa has varying degrees of rights that the foreigner receives. The study visa is the most limited and only gives the holder the right to study during the length of the educational program. The third type is the least restrictive and gives the holder almost equal rights as a citizen. This also means that the time it takes to get an approval or rejection is different for each type.

We are now ready to talk about the probability of a cause from an effect. When a person who has already applied for a visa calls the migration office to ask what the expected time for an answer is, the person on the other line might naturally start by asking what type of visa it is. This follows a natural understanding that a more restrictive visa with very few prerequisites is very easy and quick to review while others are more slow. In other words, their is a causal relationship between the type of application (cause) and the expected time to get an answer (effect). This relationship is relatively easy to compute by taking the average time to answer by type of visa for the last year. This is what we would think of as forward probability.

Reverend Thomas Bayes attacked the opposite problem, which was known as *inverse probability* at the time. Imagine that an intern at the migration office accidentally erases the type classification of all visa applications, is there a way to recover this information? Bayes' rule seems to imply that we can calculate the probability of an application being of a certain type if we know how long the application has been open. This makes intuitive sense, if we know that the application has been open for a very short amount of time it is likely that it is a student visa (although it could still be the early stages of another type). As time increases it becomes less and less likely that it is a student visa and more likely one of the more complex types.

Clearly, we could also take into account other types of information. If we see that the person has included a marriage certificate, then the probability of the process being of the third type also increases. Drawing this as a graph gives us nothing less than a Bayesian network.

In essence, the Bayesian network can be used to identify the most likely cause or explanation given the evidence or effect at hand and a model of the underlying data generating process. The corresponding conditional probabilities $p(X_i|\text{pa}_i)$ can be conditionally updated through observational data [34]. In other words, Bayesian reasoning is used to propagate information up and down the network [35] and with certain additional assumptions we can convert it into a Casual Bayesian network which we will explore now.

4.3 Causal Bayesian Networks

In the first sections of this chapter we talked about the relationships between variables simply in terms of ancestors and descendants. Keeping the same structure, we can change the interpretation of the graph such that a parent of a node is considered to be its immediate cause. We would also consider other ancestors of a node as potential causes of that node.

It is worth noting that a Bayesian network can be defined with edges that point in directions that don't follow the causal directions and still be valid. For example,

$A \rightarrow B$ and $B \leftarrow A$ (where both A and B are random variables) can both be considered Bayesian networks. But if we say that $A =$ food and $B =$ satisfaction, we can clearly see that the first graph is causal as the act of eating food can cause satisfaction, while the second graph is clearly not causal because a person's satisfaction doesn't magically generate food.

Another perspective on this that we hinted towards in the last section, is shared by Pearl in [28]: "While probabilities encode our beliefs about a static world, causality tells us whether and how probabilities change when the world changes, be it by intervention or by act of imagination".

4.4 The basic components of Bayesian networks

We have already used causal graphs in this work but it is time to officially present the basic components.

First of all, it is worth noting that Bayes' rule actually encodes the passing of information in the most simple Bayesian network: two nodes and one edge.

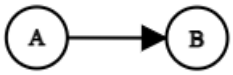

Figure 4.2: Bayes' rule DAG

Pearl explains in [28] that he was inspired by work done by Rumelhart on how information is passed in the human brain. This led him to believe that artificial intelligence systems should follow the same message-passing architecture as in human neural information processing. After a long time thinking about this topic his conclusion was that in machine reasoning the messages that are passed along the network are conditional probabilities in one direction and likelihood ratios in the other.

Indeed, let's remember that another way to explain 4.7 is that we have a posterior conditional distribution on one side, which is calculated by multiplying a prior by a likelihood (the probability of seeing a particular piece of evidence given a hypothesis) normalized by the probability of the evidence.

By adding an additional node and edge we can produce the three basic components of Bayesian networks as seen in [9]: the chain, the fork and the collider.

4.4.1 The Chain

We obtain this structure by "chaining" together three nodes sequentially:

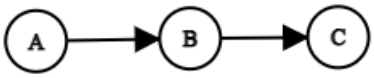

Figure 4.3: Chain structure

This is a very interesting structure, because A is an ancestor of C, which means that there is a causal relationship between the two mediated by variable B (the direct cause of C).

Pearl uses the example of a fire alarm in [28]. He notes that fire alarms are actually triggered by smoke not alarms. If we imagine a world where smoke is only produced by fire we arrive at the structure in 4.3 such that {'A': Fire, 'B': Smoke, 'C': Alarm}.

If we collect data on fire alarms we will find that the data shows a clear correlation between fire and a triggered alarm such that $P(\text{Alarm}|\text{Fire=1}) > (\text{Alarm}|\text{Fire=0})$, because we expect alarms to have been triggered more often when there was a fire and stay silent when there was no fire.

Now to the interesting part, it turns out that $P(\text{Alarm}|\text{Smoke},\text{Fire=1}) = P(\text{Alarm}|\text{Smoke},\text{Fire=0})$. Suddenly, knowing that there was a fire or not does not give us any new information about the alarm being triggered or not because we have already received all the information we need from the "Smoke" variable.

If we then calculate the correlation between fire and alarm conditioning on smoke we find that the two have now become conditionally independent from each other and no longer seem related. We represent this as:

$$P(A \perp\!\!\!\perp C|B) \tag{4.9}$$

We will see this idea of conditional independence appearing again in the next two structures and after that we will understand why it is important for our case.

4.4.2 The Fork

We obtain this structure by defining one variable as a parent of two other variables:

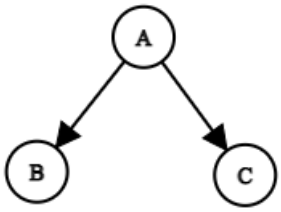

Figure 4.4: Fork structure

This is the most famous structure and is easily recognized under the name of a "confounder". It is also one of the main reasons why practitioners are so careful about not concluding that there is a causal relationship just because two variables are correlated.

To illustrate this we can once again think about our visa application example. Suppose we analyze public data and find that the number of approved visa applications is strongly correlated with the number of sold flight tickets. This would be extremely convenient for someone seeking to study in that country. Contrary to following normal procedure of applying for a visa from a home country it would seem better to buy a flight ticket because this would in theory increase the probability of an approved visa application.

Clearly this logic is faulty. The reason the two are correlated is because the migration office might expect to receive a higher number of applications the months leading up to spring and autumn semesters. This period coincides with vacation seasons (Christmas holidays and summer). If we condition by the time of the year (the confounding variable) we'll find that visa applications and number of flight tickets become completely independent. This is conditional independence in action again.

4.4.3 The Collider

The third basic structure is the collider and follows this pattern:

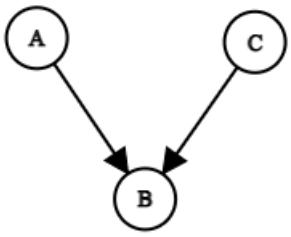

Figure 4.5: Collider structure

Following the topic on traveling to countries, imagine this hypothetical analysis about tourist destinations inspired on the example used in [36]. Suppose a country becomes a famous tourist destination (B) due to its beautiful nature (A) or for its historical significance (C).

Here we see a different pattern from before. From the outset A and C are independent and in the graph we confirm that they don't have any parents in common. Countries can have both beautiful nature and historical significance, or just one of the two and still be a tourist destination. Knowing that that a country has beautiful nature does not give us any information that helps us identify if it also has historical significance or not.

This changes when we condition on B and analyze only the subset of the data that is considered a tourist destination and we discard all countries that are not tourist destinations. Suddenly, A and C have an almost perfect negative correlation. This happens because a country only needs one of the two characteristics to become a tourist destination. So, when we know that a country actually is a tourist destination and that it for instance has beautiful nature, then the probability that it also has historical significance drops.

We can also see this from the other perspective, if you see a tourist destination and you know that it has no beautiful nature whatsoever, then the only reasonable explanation (under our simple model) is that it has to have historical significance. If this were not the case then it would not have been considered a tourist destination in the first place.

In this case we conclude that two variables that share the same collider are statistically independent but conditionally dependent on the collider.

With just these three simple structures we can now build extremely complex Bayesian networks, but the general rules regarding conditional independence and dependence are still valid in more complex networks.

This is extremely fortunate, because added complexity brings us to a new problem. Computing joint and conditional probabilities through the 4.4 can become computationally very expensive. Here conditional independence becomes our secret weapon. In [37], we learn that the tabular representation of the joint distribution of a Bayesian network grows exponentially with n (the number of variables) while conditional independence can reduce the size so it becomes linear in n.

We will now move on to show how we can use this information to operate on Bayesian networks to get the results we are interested in.

4.5 Intervention calculus for Bayesian networks

The idea of messages or information flowing up and down Bayesian networks like water flowing through pipes is very powerful. We have talked about this information like conditional probabilities in one direction and likelihood ratios in the other direction.

While analyzing the three fundamental structures we also discussed another powerful idea: conditional independence. In practice conditional independence is like blocking the flow of information between some of the variables.

For a simple motivating example consider the following network:

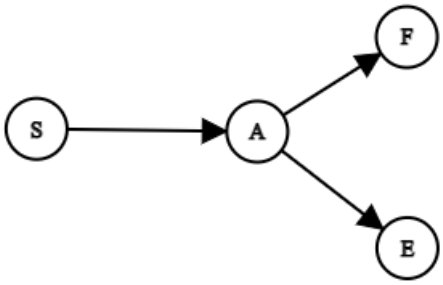

Figure 4.6: Control outside causal path

For this example we will revisit the smoke and alarm scenario. Here we can say that {'S': Smoke, 'A': Alarm, 'E': Evacuation, 'F': Firefighters}. This model assumes that smoke triggers the fire alarm which has two effects: the firefighters are called to extinguish a potential fire and the people in the building evacuate.

This means that there should clearly be a correlation between smoke and firefighters coming to the rescue.

Let's look at the following numbers:

	Smoke	Alarm	Firefighter	Evacuation
Smoke	1.000	0.635	0.555	0.526
Alarm	0.635	1.000	0.877	0.829
Firefighter	0.555	0.877	1.000	0.727
Evacuation	0.526	0.829	0.727	1.000

Figure 4.7: Correlation between variables

These numbers have been created sampling from a data generating process that follow graph 4.6. Here we can confirm that there is a clear positive correlation between smoke and firefighters (0.555). The reason this correlation is perfect is that we assume certain randomness in the process (i.e. alarm misfires, firefighters fail to respond, etc.).

We also learned by analyzing the three simple structures that controlling for A in this case makes S and F conditionally independent. Instead, we will focus on a more interesting case by asking ourselves what would happen if we instead control for E. Surely, people leaving the building doesn't cause the firefighters to come. They are called to the scene by the alarm (or equivalently someone calling emergencies). We confirm this visually by noting that E is not on the causal path between S and F. Still, if we condition by Evacuation=1 we now see on figure 4.8 that suddenly the correlation between smoke and firefighters is practically 0.

This might seem a bit counter-intuitive at first but opens our eyes to a new set of tools at our disposal. We learned that in the chain structure the variables on either side of a mediator are usually correlated but become conditionally independent when conditioning on the mediator because this variable already tells us all we need

	Smoke	Alarm	Firefighter
Smoke	1.000	0.181	0.042
Alarm	0.181	1.000	0.231
Firefighter	0.042	0.231	1.000

Figure 4.8: Adjusted correlation between variables

to know and makes the ancestors redundant.

The same happens here but to a lesser degree. Evacuations can serve as a proxy for alarms because most of the time they have the same values: people evacuate when the fire alarm sounds. So when we condition on evacuation we are also already receiving most of the information related to smoke, making it redundant as well.

4.5.1 Directed separation

Judea Pearl compares this to water flowing through pipes [28]. When a path is closed, information can't "flow" across two nodes in the path as we have seen in the previous examples.

This analogy is beneficial to us because the models we use in real life are rarely as simple as the examples we have seen so far.

This leads us to the concept of *d-separation* (directed separation) as seen in [9]. If two sets of nodes are d-separated then they are independent. On the other hand, if two sets are d-connected then we know that they are most likely dependent. Using the water and pipe analogy we would say that two d-separated sets of nodes don't have any open pipes between them, such that no water (or information) can flow from one set to the other.

From the simple structures we learned that colliders block the path between to variables (and conditioning on the blocker unblocks it). This means that two sets are d-separated if they are only connected by a collider. Chains and forks do allow information to pass, so we would need to condition on a mediator to block the information (although we have to be careful to not unblock another path if the conditioned set of nodes also contains a collider).

4.5.2 Intervention vs conditioning

Now it's time to explain why all of this is important. Knowing how to produce d-separated or d-connected sets is of great importance for us when trying to evaluate counterfactuals in our model when the possibility to naturally see a counterfactual does not exist.

For this we need to first understand the difference between conditioning on a variable and doing an intervention on that variable.

We already learned about the concept of conditional probabilities with equation 4.2 where the probability of $P(A|B)$ is equal to finding the subset of data where both A and B hold certain predefined values (the joint probability $P(A,B)$) normalized by the probability of event B ($P(B)$). In a sense, we are just filtering the dataset and focusing on part of it without changing anything else.

Intervention implies an actual change. We not only filter data by a certain value in a variable, instead we manually set that variable to take the value that we want. When we do this we can expect its descendants to also change accordingly even if that combination of values does not currently exist in the sample of data we have access to.

To be able to work with this idea of intervention we need to use a new notation to separate the two and we have decided to use the one described by Pearl [28]. We are already familiar with conditional notation ($P(Y|X = x)$). We will use what is known as the *do* operator to clearly identify interventions. This means that the probability of Y given our intervention to set X to x is defined as $P(Y|do(X = x))$.

The impact of this is not only in the notation. Interventions actually change the structure of our graph. To explain this we can now go back to our original motivating example. For clarity we can reproduce the graph representing the underlying data generating process in figure 4.9.

We can also remember here that the underlying parameters of this graph assume that there does exist an unfair bias from hiring managers against foreigners. The path going through employment agencies also has a bias against foreigners potentially due to more paperwork or other hurdles. Just like the example in [4] this

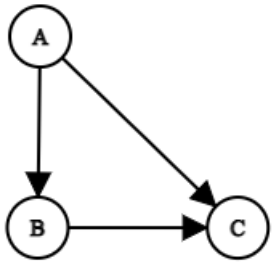

Figure 4.9: Bias against foreigners

second path can be considered fair or unfair under different circumstances. It can be considered fair as nobody is actually trying to be unfair against the foreigners on purpose. But it could also be considered unfair if we assume that it is society's responsibility to make sure that there should be equal procedures for everyone to avoid future propagation of injustice.

This time we are more prepared to understand this example. First of all, we notice that the path $B \leftarrow A \rightarrow C$ is a fork and the path $A \rightarrow B \rightarrow C$ is a chain. This means that trying to identify the direct cause of employment agency on the probability of getting a job gets mixed with variable A acting as a confounder (this is known as an open back-door path).

To try to solve this problem we can try to use intervention and with this we will illustrate how it changes the causal graph. Say we want to calculate $P(C = c|do(B = b))$. An intervention will effectively produce the following modified graph:

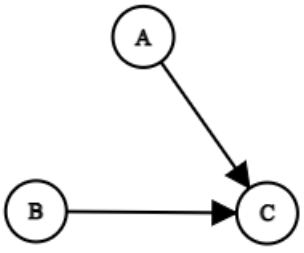

Figure 4.10: Intervention on employment agency

It is clear that the arrow between A and B has disappeared. This is explained by the fact that B no longer depends on A for it's value. The causal graph shows causal

dependencies, but we are arbitrarily deciding to set B to a specific value without taking into account any other variables. This removes all incoming arrows into the variable we are intervening on.

This change has another important effect. We are already familiar with operations on conditional probabilities, but how do we get the answer we need from $P(C = c|do(B = b))$ without actually running a real intervention? In this new graph, finding the direct graph seems much easier. There are no back-door paths that can leak information so we should be able to use $P_m(C = c|B = b)$.

Note that 1) we are now dealing with a simple conditional probability and 2) this probability has a subscript m indicating a modified model so it is not immediately comparable to the original probability we are interested in.

Luckily, we can use information about the underlying data generating process to solve this problem. First of all, the probability of a person being a foreigner or local actually doesn't change with the intervention as this was a root node and did not depend on the variable we intervened on. This means that:

$$P_m(A) = P(A) \tag{4.10}$$

On top of this, we can also conclude that the mechanism through which C receives its value from A and B doesn't change, we have only changed the value of B. From this we conclude that:

$$P_m(C = c|A = a, B = b) = P(C = c|A = a, B = b) \tag{4.11}$$

We will now also use the marginalization principle as explained in [38] to declare that:

$$P_m(C = c|B = b) = \sum_a P_m(C = c|A = a, B = b)P_m(A = a|B = b) \tag{4.12}$$

Now, a critical step in this process comes specifically from the concept of d-separation. We already declared that if two sets of nodes are d-separated then they

are independent. We see in graph 4.10 that C serves as a collider blocking any information between A and B. With this simple information (and remembering that if X and Y are independent then $P(Y|X) = P(X)$) we can rewrite 4.12 as:

$$P_m(C = c|B = b) = \sum_a P_m(C = c|A = a, B = b)P_m(A = a) \tag{4.13}$$

It becomes immediately clear that both right-hand components correspond to the equations in 4.10 and 4.11. This means that we can now conclude:

$$P(C = c|do(B = b)) = \sum_a P(C = c|A = a, B = b)P(A = a) \tag{4.14}$$

This is almost magical, because using information we get from the causal graph (not from the data itself) we have been able to compute a formula that allows us to use observational data to calculate the effect of an intervention without running a real equivalent experiment. The value of causal graphs, conditional independence and d-separation becomes much clearer.

We can compute this with the same data used for the motivational example:

$$P(C = 1|do(B = 1)) = 0.75 * 0.2 + 0.8 * 0.8 = 0.79 \tag{4.15}$$

This is the same result we get by doing the intervention directly with the library *pgmpy*:

```
intervention = CausalInference(model)
print(intervention.query(variables=['J'], do={'A': 'agency'}, show_progress=False))
```

```
+--------------+----------+
| J            |   phi(J) |
+==============+==========+
| J(hired)     |   0.7900 |
+--------------+----------+
| J(not hired) |   0.2100 |
+--------------+----------+
```

Figure 4.11: Intervention with *pgmpy*

It is important to note here that simply conditioning results in $P(C = 1|B = 1) = 0.775$.

According to Pearl in [9], we can use interventions, *do*-expressions and graph

surgery to extract actual causal relationships and separate them from simple correlation, even without performing a randomized experiment.

If we compute $P(C = 1|do(B = 0))$ we can calculate the Average Causal Effect (ACE) as:

$$ACE = P(C = 1|do(B = 1)) - P(C = 1|do(B = 0)) \tag{4.16}$$

In our case, $ACE = 0.79 - 0.56 = 0.23$. This result shows us that using the employment agency has a positive causal effect in our hypothetical job market (which is consistent with the assumptions of the data generating process).

We were able to arrive at this result without a randomized controlled experiment where people are randomly assigned to usage or not of the employment agency to evaluate if it has a positive causal effect or not.

4.6 Path-specific computations of fairness

In previous sections we have learned about using d-separation, do-calculus and our assumptions about the data generating process to compute causal effects.

We illustrated this by computing the causal effect of using the employment agency in our motivating example through an intervention. Trying to do the same for place of origin seems easy at first:

$$P(C = c|do(A = a)) = P(C = c|A = a) \tag{4.17}$$

This makes sense because according to our data generating process, node A corresponding to place of origin is the root node and affects the outcome (C) directly through one path and indirectly through employment agency.

Still, this is problematic when we want to separate both effects. As we have noticed earlier, it could be that one path is considered unfair and the other fair. So for the case of measuring bias in our data/model we don't want to focus on the overall causal effect but on the effect of the sensitive attribute on the outcome through unfair paths.

Once again, the causal diagram helps us establish how to untangle this direct

effect. To "block" the indirect path, the rules of d-separation in chains tell us that we can condition on usage of the employment agency (assuming independent error terms not currently drawn in the graph), such that the only open bath between place of origin and hiring is the direct path.

4.6.1 Counterfactual definition of direct and indirect effects

Unfortunately, this does not yet allow us to establish unfair bias defined by the concept of counterfactual fairness covered in chapter 2. Under that definition, we want to evaluate if the bias of a model is fair by comparing the actual outcome of the model to the outcome in a counterfactual world where the sensitive attribute has been modified.

An intervention allows the intervened attribute to modify its descendants accordingly, but this might significantly change the characteristic traits of the analyzed individuals. Concretely, in the Berkeley admission example explored in [9] we find that gender has a causal effect on the department a person applies to. Intervening on gender would then potentially change not only the gender but also the department applied to such that it is no longer a comparable counterfactual to the real-world situation. According to the counterfactual definition we want to know if the female applicant would have been admitted if she were a male while still maintaining the same decision architecture for any mediating variables.

In [9], Pearl clearly defines the different types of effects.

The *total effect* is defined by

$$TE = E[Y_1 - Y_0] = E[Y|do(X = 1)] - E[Y|do(X = 0)] \tag{4.18}$$

This is the simplest effect we have seen so far. If X is place of origin we can intervene on this variable setting it to foreigner (1) and local (0) such that the total effect is defined by the difference of these two values. Unfortunately, this does not directly serve our purpose to understand counterfactual fairness.

The next type of effect we can compute is the *controlled direct effect*:

$$CDE = E[Y_{1,m} - Y_{0,m}] = E[Y|do(X = 1, M = m)] - E[Y|do(X = 0, M = m)] \quad (4.19)$$

This takes us a bit closer because it allows us to intervene on the mediator such that we can set it to a specific value over the whole population. This closes any back-door paths through the mediator, but still does not satisfy the counterfactual fairness definition as we are manually changing the decision architecture for the whole population.

To compute the effect of intervening on the sensitive attribute while allowing each individual to maintain their original values for the mediator we can use the *natural direct effect*:

$$NDE = E[Y_{1,M_0} - Y_{0,M_0}] \quad (4.20)$$

This equation does a better job at establishing a counterfactual comparison by intervening and changing the sensitive attribute, but keeping the rest of the decision architecture the same by setting the mediator to the value it had before the intervention.

Similarly, we can also calculate the *natural indirect effect*:

$$NIE = E[Y_{0,M_1} - Y_{0,M_0}] \quad (4.21)$$

In this case we are not interested in changing the sensitive attribute, so we intervene to keep it at it's original value (e.g. foreigner). To compute the indirect effect we do change the mediator so that it take the value as if we had intervened on the sensitive attribute (i.e. local). This way, the direct causal effect between the sensitive variable and output gets cancelled such that any variation has to be the result of a change in the mediator as if we had changed the sensitive attribute.

These concepts are equivalent to the *average direct effect* (ADE) and *average indirect effect* (AIE) described in [5]. Chiappa et al. then generalize these formulas to produce the *path-specific effect* (PSE).

To explore the difference we can add an additional variable to our original DGP, such that place of origin also affect qualifications (through different educational systems among countries) which also affects the probability of being hired. This indirect path is considered fair because hiring managers are interested in highly qualified individuals whether they are locals or foreigners. This might cause long-term unfairness though that could be revised through public policies implementing special training programs for foreigners and their families to avoid propagating this inequality of opportunity.

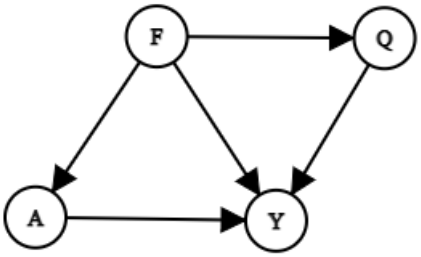

Figure 4.12: Path-specific unfairness

Under this data generating process and assuming that the direct path $F \rightarrow Y$ and the indirect path $F \rightarrow A \rightarrow Y$ are unfair but the path $F \rightarrow Q \rightarrow Y$ is fair would be defined as):

$$PSE_{10} = E[Y_{0,Q_1,A_0} - Y_{1,Q_1,A_1}] \tag{4.22}$$

We can read this as the path-specific potential outcome of $F = 1$ (foreigner) with respect to $F = 0$ (local) restricted to causal paths such that $F = 0$ in the direct path between F and Y, as well as the indirect path $F \rightarrow A \rightarrow Y$; and keeping $F = 1$ in the indirect path $F \rightarrow Q \rightarrow Y$.

Once again, intuitively this means that we want to compare the counterfactual ($E[Y_{0,Q_1,A_0}$) where we are changing the sensitive attribute in unfair paths, to the real-world case ($E[Y_{1,Q_1,A_1}$) where all variables keep their original values (the person in question is a foreigner).

Chapter 5

5. Experimental application

We can now apply this to a slightly more complex version of the example we have been working with and a public dataset.

5.1 Synthetic example

We will now start by showing a new graph describing the Causal Bayesian network of our motivating example with some additional variables:

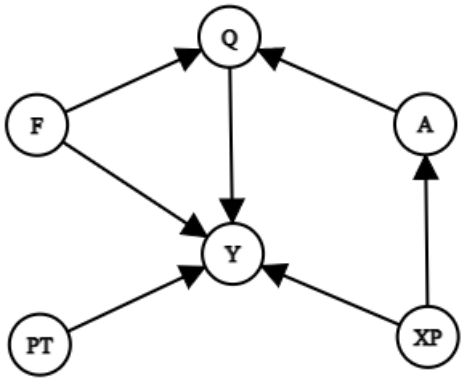

Figure 5.1: Bias against place of origin

A brief description of each of the variables depicted above (the model also assumes error terms for each variable that have not been mapped out for simplicity):

- F: This is the sensitive attribute we are interested in understanding. It is a binary variable that takes the value 1 to indicate a foreigner and the value 0 to indicate a local person.

- XP (or X): This variables indicates the years of experience a person has. This does not depend on the place of origin. It affects ability due to the acquired mastery of an art over the years, and salary as this is one characteristic hiring managers often explicitly ask for during the screening process.

- A: Ability can be a bit abstract, but in this case we take it as the skills of a person that can be tested during the interview process (if this were not possible we would be forced to set it as an unobserved variable).

- Q: Qualifications are directly affected by the natural or acquired ability of a person. In this case it represents formal qualifications like education or professional degrees. It also depends on the place of origin of the person as the educational system in some countries is better than others.

- PT (or P): This variable represents the type of contract and takes the value 1 for part-time jobs or 0 for full-time jobs. It has a negative relationship with salary as a part-time job pays less than a full-time contract.

- Y: This is our outcome variable and corresponds to the monthly salary of a person. This can either be $\hat{Y}$ if we are using the output of a ML model as proxy or Y if we have access to the real data.

The exercise we are doing here can be thought of as an attempt to measure, understand and reduce the migrant pay gap in a country.

For this example we will also assume that the data generating process follows a linear structure according to:

$$
\begin{aligned}
F &\sim Bern(\pi) \\
P &= \varepsilon_p \\
X &= \varepsilon_x \\
A &= \theta^a + \theta^a_x X + \varepsilon_a \\
Q &= \theta^q + \theta^q_a A + \theta^q_f F + \varepsilon_q \\
Y &= \theta^y + \theta^y_f F + \theta^y_q Q + \theta^y_x X + \theta^y_p P + \varepsilon_y
\end{aligned} \tag{5.1}
$$

This is the underlying data generating process which we have used to sample data. Researchers analyzing the data for bias would state their assumptions about the process as we have done with the causal diagram in 5.3. They would not have access to the real values of the parameters.

For the sake of this example we will also assume that the team working on this project has trained a Random Forest regressor on the collected data. So, we will be using $\hat{Y}$ instead of Y. We also assume that $\hat{Y}$ is a proxy for Y with a distribution that gets closer to the true distribution of Y as the number of training samples increases. For this reason we don't need to do any additional corrections for this example.

After sampling we can now look at the average salary grouped by the place of origin:

	salary
foreigner	
0	7856.455340
1	7005.666296

Figure 5.2: Average salary by place of origin

This clearly raises a red flag as we can see that foreigners have a lower salary on average than locals. The output of this model would not satisfy demographic parity as $E[\hat{Y}|F=1] - E[\hat{Y}|F=0] = 7856.56 - 7005.67 = -850.79 \neq 0$.

Still, it is not as simple as stating that unfair bias in this process is by this magnitude. We are assuming in our Causal Bayesian network that there is a causal

path $F \rightarrow Q \rightarrow Y$ that is not unfair. This is caused by differences in educational systems at country level and not a type of direct discrimination against foreigners. Keeping everything else constant, a less qualified person can expect a lower salary than a more qualified person. Still, it can also be the case that the government should help these families with additional training or subsidies to avoid propagating the pay gap across generations.

In chapter 4 we used equation 4.20 (the natural direct effect) as a way to measure the fairness from the counterfactual perspective. Here the question becomes: what is the expected difference between locals and foreigners if foreigners were treated as locals but still maintained the rest of the decision architecture intact?

We can compute this value manually, but at this point we will show the power of tools like *doWhy* as found in [39].

With this tool we can import the graph representing the Causal Bayesian network in gml format with a sample of the data and apply causal operations on the data with its graph. For this particular example we have imported the graph into *doWhy* and used it to first use a 'nonparametric-nde' estimand to identify causal effects. Using this estimand we have used a two-stage regression process to compute the NDE:

*** Causal Estimate ***

Identified estimand Estimand type: nonparametric-nde

Estimate Mean value: -498.23877234180037

With this we can now reveal that in reality $\theta_f^y = -500$. The estimate found by *doWhy* is very close keeping in mind that it used the causal structure and a data sample, without access to the original parameters. Now we know that there is clearly an unfair bias against foreigners along the direct causal path $F \rightarrow Y$ with a magnitude of approximately -500. This means that a foreigner in the real world of this example is loosing an expected value of USD 500 every month simply by not being a local.

Interestingly, in [4] Chiappa et al. explain how the unfair bias in the output of a model can be corrected and this is what we will do next.

5.1.1 Correction of unfair bias

In equation 4.22 we saw one of example of how we can compute the path-specific effect. There isn't one single way to do it as it depends on knowledge of the Causal Bayesian network and the unfair dependencies in the network.

In this case we are assuming that the only unfair bias is produced in the path $F \rightarrow \hat{Y}$. The path $F \rightarrow Q \rightarrow \hat{Y}$ is considered fair. This means that the PSE in this case is defined (using the same notation we used previously) as:

$$PSE = E[Y_{0,Q_1} - Y_{1,Q_1}] \tag{5.2}$$

This small equation carries a lot of information. We are interested here in finding the effect produced by changing the treatment variable (in this cases the sensitive attribute place of origin) from the value $F = 1$ to $F = 0$, or equivalently from foreigner to local. This is done while the mediator variables are set to the values they would have received before the change. Although we have multiple variables, the only mediator directly caused by the sensitive attribute is qualifications, so it is set to Q_1.

From another perspective, this is equivalent to saying that the person still has the same qualifications as before with all that it entails, but the hiring company no longer perceives them as foreigners but as locals.

Here it is important to remember that when we defined the equations of the data generating process we assumed unobserved errors. This would seem like one of the first hurdles to overcome. Still, even though these error are completely random in general, for each given sample n the randomness has already been applied and it is the same in both the real and in the counterfactual world. We can use this fact to rewrite the error terms by reordering the equations in 5.1 and changing the variables for specific values.

First we can re-write Y_{0,Q_1} as:

$$Y_{0,Q_1} = \theta^y + \theta^y_{\mathrm{f}}\bar{\mathrm{f}} + \theta^y_q Q_{\bar{\mathrm{f}}} + \theta^y_x X + \theta^y_p P + \varepsilon_y \tag{5.3}$$

Using the same notation as in [4] we find that:

$$\begin{aligned}
\delta_{\varepsilon_x} &= x^n \\
\delta_{\varepsilon_p} &= p^n \\
\delta_{\varepsilon_a} &= a^n - \theta^a - \theta^a_x x^n \\
\delta_{\varepsilon_q} &= q^n - \theta^q - \theta^q_a a^n - \theta^q_f f^n \\
\delta_{\varepsilon_y} &= y^n - \theta^y - \theta^y_f f^n - \theta^y_q q^n - \theta^y_x x^n - \theta^y_p p^n
\end{aligned} \tag{5.4}$$

This usage of the error terms is very useful under the idea of *twin Bayesian networks* described by Pearl in [40]. This idea plots out the graphical structure of the real world and the counterfactual almost like mirror images of each other, connected by the error terms. An example of this can also be found in [5]. Applying the rules of d-separation we can see that $Y_{0,Q_1} \perp\!\!\!\perp \{F,Q,A,P,N,Y\} | \{\varepsilon_q, \varepsilon_a, \varepsilon_p, \varepsilon_n, \varepsilon_y\}$. This means that, just like in [4], we can just the fact that:

$$P(Y_{0,Q_1}|\mathrm{f}, a^n, q^n, x^n, p^n, y^n) = \int_{\varepsilon} P(Y_{0,Q_1}|P(\varepsilon|\mathrm{f}, a^n, q^n, x^n, p^n, y^n) \tag{5.5}$$

The right-hand term factorizes over the ε terms. So we can use this with 5.3, 5.4 and 5.5 to compute:

$$\begin{aligned}
Y_{0,Q_1} &= \theta^y + \theta^y_{\mathrm{f}}\bar{\mathrm{f}} + \theta^y_q(\theta^q + \theta^q_a a^n + \theta^y_f f + q^n - \theta^q - \theta^q_a a^n - \theta^y_f f) + \theta^y_x x^n + \theta^y_p p^n + \\
&\quad y^n - \theta^y - \theta^y_{\mathrm{f}}\mathrm{f} - \theta^y_q(q^n) - \theta^y_x x^n - \theta^y_p p^n \\
&= \theta^y_{\mathrm{f}}\bar{\mathrm{f}} + y^n - \theta^y_{\mathrm{f}}\mathrm{f} \\
&= y^n + \theta^y_{\mathrm{f}}(\bar{\mathrm{f}} - \mathrm{f}) \\
&= y^n + PSE_{\mathrm{f}\bar{\mathrm{f}}}
\end{aligned}$$

Interestingly enough, in this particular case we can see that because we are assuming that only the direct path between F and Y is unfair, our calculations end

up in terms of the coefficient θ_{f}^{y} which corresponds to the direct effect between F and Y.

This result also tells us how to correct the predictions of our model, more concretely:

$$Y_{0,Q_1} = y^n + \theta_{\mathrm{f}}^{y}(\bar{\mathrm{f}} - \mathrm{f}) = y^n + (-498.24)(0-1) = y^n + 498.24 \tag{5.6}$$

Here we are using the estimated coefficient instead of the true coefficient because we usually don't have access to the true value. It is also worth noting that if we had included more unfair paths the correction would have been different.

Remember that the salary values we have are the predictions of a ML model trained on sample data. Now we want to reduce the unfair bias by applying the correction we obtained to all our predictions $\hat{Y}$ where $F = 1$. Below we can see the new result grouped by place of origin:

foreigner	adj_salary
0	7856.455340
1	7503.906296

Figure 5.3: Adjusted average salary by place of origin

The pay gap has been greatly reduced and the remaining gap can be attributed to the difference in qualifications between the groups.

5.1.2 UCI Adult Income

Now we will move on to analyze the Adult Data Set (also known as Census Income), a public benchmark from the UCI Machine Learning Repository found in [41]. The main declared task for this dataset is to predict whether the income of a person exceeds $50K yearly based on census data.

This dataset was also used in [4] and [42]. In both cases the focus is on sex as the sensitive attribute. The original dataset also contains a feature related to the race of the person which was not included in the implementation of the mentioned papers (country of origin is used).

This is the causal graph we will be working with:

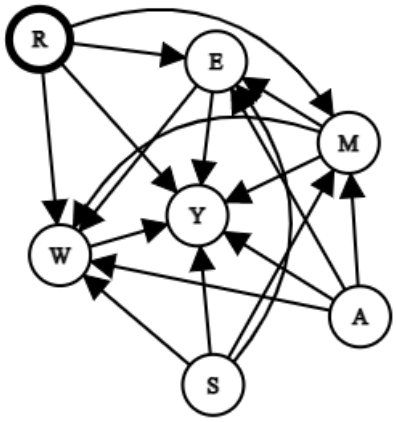

Figure 5.4: Causal graph for the Adult dataset

These are the variables:

- R: This is the sensitive attribute we are interested in understanding. It is a binary variable that takes the value 1 to indicate a white person and 0 to indicate black or another race. It is considered a root node as color of skin and other physical attributes are defined at birth.
- A: This variable stands for the age of a person. It is considered a root node as age is a function of the number of years passed since birth.
- S: This attribute represents Sex of the person and is the one used in other work as the sensitive attribute. It is considered as a root node as we assume that biological sex is defined at birth and not caused by any of the other variables in this model. Here 1 represents male and 0 represents female.
- M: Marital status is clearly affected by age as it is more probable that a person is married the older the person becomes. We also assume that marital status is affected by sex. One reason is because marital status includes "widowed" and women tend to live longer than men. It is also affected by race as in some cases race could also act as a proxy for different cultural factors.
- E: Education is assumed to have causal influence from age because a teenager rarely is in the position to start a master's degree. The influence of sex and

bias can be considered part of the fair/unfair influences in this causal graph. Access to different levels of education can be influenced by these two variables due to direct bias or due to a more abstract influence due to social factors (i.e. race and neighborhood could be highly related and therefor affect access to education, or cultural pressure might guide women towards certain degrees and not others).

- W1: The variable W actually represents three different variables with the same causal paths related with work. It would be possible to draw additional dependencies across these variables but that would break the acyclical nature of the graph. So we assume they don't have dependencies between each other but have the same ingoing and outgoing causal paths. W1 stands for the occupation or work a person does (i.e. technical support, armed forces, etc.)

- W2: Hours per week.

- W3: Work class (i.e. private, federal government, self-employed, etc.).

- Y: This is our outcome variable and corresponds to a binary variable set to 1 if the yearly salary of a person is higher or equal than 50.000 US dollars and set to 0 if not. This can either be $\hat{Y}$ if we are using the output of a ML model as proxy or Y if we have access to the real data.

Just as the authors did in [42], we will assume linear underlying patterns between these variables (which is clearly debatable) and defined by these functions (a logistic regression is used to model the output variable):

$$
\begin{aligned}
R &\sim Bern(\pi) \\
A &= \varepsilon_a \\
S &\sim Bern(\pi_s) = \varepsilon_s \\
M &= \theta^m + \theta^m_r R + \theta^m_a A + \theta^m_s S + \varepsilon_m \\
E &= \theta^e + \theta^e_r R + \theta^e_a A + \theta^e_s S + \theta^e_m M + \varepsilon_e \\
Wi &= \theta^{wi} + \theta^{wi}_r R + \theta^{wi}_a A + \theta^{wi}_s S + \theta^{wi}_m M + \theta^{wi}_e E + \varepsilon_{wi} \\
Y &= \exp(\theta^y + \theta^y_r R + \theta^y_a A + \theta^y_s S + \theta^y_m M + \theta^y_e E + \theta^y_{w1} W1 + \theta^y_{w2} W2 + \theta^y_{w3} W3 + \varepsilon_y)
\end{aligned} \tag{5.7}
$$

We can now also look at the average outcome grouped by race:

race	high_income
0	0.152582
1	0.255860

Figure 5.5: Average outcome by race

Due to the outcome being a binary variable we can think of this as the probability of a person belonging to each group having a salary of USD 50.000 or higher. We can clearly see that in this dataset being white ($Y = 1$) gives a person a higher probability of a high income than being black or another race ($Y = 0$).

We now also have to define the unfair paths we want to focus on. As mentioned, race is the sensitive attribute in this case making the direct path $R \to Y$ unfair. We also consider partially unfair the influence of race through education (race should not influence opportunities of education) and the work-related attributes (race should not affect the type of job or the characteristics of the job). These correspond to the paths defined by $R \to E \to Y$, $R \to E \to Wi \to Y$, and $R \to Wi \to Y$.

This allows us to define the path-specific effect as follows:

$$PSE = E[Y_1(M_0, E_1(M_0), Wi_1(M_0, E_1(M_0))) - Y_0] \tag{5.8}$$

Because we are using a logistic regression to model the output we can use the odds ratio scale instead to simplify computations and improve intuition:

$$PSE = E[\frac{Y_1(M_0, E_1(M_0), Wi_1(M_0, E_1(M_0)))}{Y_0}] \tag{5.9}$$

To clarify the notation we can analyze the $E_1(M_0)$ component. This means that we want to understand the effect if the value of race where set to 1 (white) in the context of education, but set to 0 in the case of marital status that education depends on. This is because we consider the effect of education (partially) unfair due to the unfair effect of race on education. But the effect of marital status on education is not considered unfair due to reasons described earlier.

The joint distribution of this causal graph is defined by:

$$\begin{aligned} P(Y,R,A,S,M,E,W1,W2,W3) = P(Y|R,A,S,M,E,W1,W2,W3) \\ P(W1|R,A,S,M,E)P(W2|R,A,S,M,E)P(W3|R,A,S,M,E) \\ P(E|R,A,S,M)P(M|R,A,S)P(S)P(A)P(R) \end{aligned}$$

This comes from the definition of the Causal Bayesian Network which allows us to use the properties of d-separation and conditional independence to simplify the graph instead of having to define the conditional dependencies between all variables.

This means that we can declare an intervention on race as follows:

$$\begin{aligned} P(Y|do(R=r)) = \sum_{A,S,M,E,Wi} P(Y|r,A,S,M,E,W1,W2,W3)P(W1|r,A,S,M,E) \\ P(W2|r,A,S,M,E)P(W3|r,A,S,M,E)P(E|r,A,S,M)P(M|r,A,S)P(S)P(A) \end{aligned}$$

We can now use the underlying functions we defined in 5.7 and compute this expression by expressing Y as a function of $R = r$ (as done in [5]) through recursive substitutions:

$$\begin{aligned} Y_0 = \exp(&\theta^y + \theta_r^y 0 + \theta_a^y \varepsilon_a + \theta_s^y \varepsilon_s + \theta_m^y(\theta^m + \theta_r^m 0 + \theta_a^m \varepsilon_a + \theta_s^m \varepsilon_s + \varepsilon_m) \\ &+ \theta_e^y(\theta^e + \theta_r^e 0 + \theta_a^e \varepsilon_a + \theta_s^e \varepsilon_s + \theta_m^e(\theta^m + \theta_r^m 0 + \theta_a^m \varepsilon_a + \theta_s^m \varepsilon_s + \varepsilon_m) + \varepsilon_e) \\ &+ \sum_1^3(\theta_{wi}^y(\theta^{wi} + \theta_r^{wi} 0 + \theta_a^{wi} \varepsilon_a + \theta_s^{wi} \varepsilon_s + \theta_m^{wi}(\theta^m + \theta_r^m 0 + \theta_a^m \varepsilon_a + \theta_s^m \varepsilon_s + \varepsilon_m) \\ &+ \theta_e^{wi}(\theta^e + \theta_r^e 0 + \theta_a^e \varepsilon_a + \theta_s^e \varepsilon_s + \theta_m^e(\theta^m + \theta_r^m 0 + \theta_a^m \varepsilon_a + \theta_s^m \varepsilon_s + \varepsilon_m) \\ &+ \varepsilon_e) + \varepsilon_{wi})) + \varepsilon_y) \end{aligned}$$

We now do the same thing for $Y_1(M_0, E_1(M_0), Wi_1(M_0, E_1(M_0)))$ setting R to 1 or 0 where it corresponds to set up the counterfactual and we take the expectation over the whole expression. For this it is worth noting that the function of each variable has independent zero-mean Gaussian variables used as error terms. Age and Sex become constants corresponding to the average of each variable and are cancelled out (like most parameters in the expression) because we use the fact that when we are dividing two exponents with the same base, we can subtract the exponent of the denominator from the exponent of the numerator. The outcome is a new exponent. We will not write out all of the operations here but it is trivial from the expression above and the counterfactual to find that:

$$PSE = \exp(\theta_r^y + \theta_e^y \theta_r^e + \theta_{w1}^y(\theta_e^{w1} \theta_r^{w1}) + \theta_{w2}^y(\theta_e^{w2} \theta_r^{w2}) + \theta_{w3}^y(\theta_e^{w3} \theta_r^{w3})) \quad (5.10)$$

Using the Adult dataset to compute the parameters of the linear regressions corresponding to each of the variables and the logistic regression corresponding to the output we find that this expression results in:

$$PSE = 1.344 \quad (5.11)$$

As this is in the odds ratio scale we reach the following conclusion: *The odds of a non-white person having a high income would have been 1,3 higher if the person were white in the context of education, work and direct definition of income, but still considered non-white in the context of marital status*.

We also went though this same process using the same assumptions of fairness as [42] and calculated the PSE between females and males to be 3,4 which matches the result obtained in their paper.

From the linearity assumption we also assume that the bias is the same for all members of the population (i.e. the bias does not depend on the specific characteristics of a specific sample). For this reason we have applied a naive correction to the predictions by computing $d = 0.5 - \frac{\text{odds}}{1+\text{odds}}$ and then adjusting the predictions such that: $\hat{Y_{adj}} = \hat{Y} - d$.

This comes from the fact that odds of 1 would indicate that the probability of the counterfactual having a high income would be the same as the real world probability of a non-white person through the paths we defined. The output of the model corresponds to the probability of a given person having a high income so the predictions of non-white members of the dataset are boosted by the difference computed above (predictions were obtained from a test set not used during training of the ML model or to obtain the parameters of the causal graph).

Doing this we see the following results for the raw predictions:

- Probability of high income for non-white: 0.11
- Probability of high income for white: 0.192
- Accuracy raw predictions: 0.848

Our predictions seem to have a slightly less pronounced income disparity compared to the proportions found directly in the dataset, but we can clearly see that the difference is still there. This confirms one of the dangers mentioned in the beginning of this work. ML models tend to pick up the underlying bias in the datasets used for training.

We also obtained the following results for the adjusted predictions:

- Probability of high income for non-white: 0.149
- Probability of high income for white: 0.192
- Accuracy adjusted predictions: 0.847

It is interesting to see that this naive correction reduces the general income disparity between groups in the sensitive attribute with minimal difference to the prediction accuracy. It is worth noting that in [4] and [5] Monte Carlo approximation is used as a more complex relationship between variables and non-linearity is assumed.

The code used for these examples can be found in: `https://github.com/KalleBylin/cbn__for_fairness_in_ml/`

Chapter 6

6. Conclusion and future research

This document started with a brief introduction to the topic of bias in the context of training and using machine learning models. This was followed by a brief introduction to different definitions of fairness and metrics used to measure fairness in ML models.

This was followed by a review of Causality and how using a causal lens by explicitly defining our assumptions about the data generating process allows us to answer questions that can't be answered from the data alone.

In practice, Bayesian networks can be used to model the dependencies between variables and if we assume causal meaning for these dependencies they become Causal Bayesian networks. Chapter 4 presented the basics of these tools that we would need to apply them for our purpose. This includes the structure of Causal Bayesian networks as directed acyclic graphs that can be decomposed into three simple structures. These in turn allow us to understand the flow of information between the variables so that we can close or open this flow according to our requirements.

At this point we used the concepts reviewed in all the previous sections to better understand the direct and indirect effect of a sensitive attribute on the outcome variable.

All of this made it possible for us to apply what we have learned on an example of synthetic data, as well as a public dataset used in previous related work.

In both cases we used the information given to us by the structure of the

(assumed) underlying data generating process to calculate the path-specific effect (PSE) of the sensitive attribute through unfair paths (assuming that the effect through other paths can be considered fair). We also used the PSE to then adjust the predictions obtained from a ML model trained on the data that had clearly absorbed the underlying unfair bias.

6.1 Limitations

Before reaching our conclusions it is important to explicilty state some of the limitations of this work.

First of all, we have assumed strictly linear patterns for the data generating process in both processes that we worked with. This would imply that the causal effect is the same for all individuals of the population. For example, if we define a very simple model where $Y_i = \alpha * i + \beta * x$ we can easily see that $Y_1 - Y_0 = \alpha * 1 + \beta * x - (\alpha * 0 + \beta * x) = a$. In other words, the PSE does not depend on specific attributes x of a particular sample in the dataset. This is convenient for adjusting the predictions, but will often be too restrictive to model real-world processes. In this case, [5] proposes using Monte Carlo approximation to obtain the counterfactual distributions needed to compute the PSE and for correcting the unfair bias.

Additional to this, it is important to note that the process described in this work requires the property of identifiability in the Causal Bayesian network. More information about this can be found in [43]. If we don't have access to the right data or the causal network does not allow us to compute the distributions we need then we would need to perform a randomized experiment to estimate the underlying fairness in the data/model.

A third limitation is that this process also requires a properly defined data generating process through a Causal Bayesian network. The causal graph is used to understand the dependencies between the graph and clearly establish the definition of the path-specific effect. Without a notion of these paths they can't be computed and a network that is incorrect will lead to incorrect results.

Finally, Bayesian networks are known run into scalability issues as the number

of variables and the number of dependencies between variables increase. We can already see a hint of this comparing the causal graphs of the first and second examples in chapter 5. The second graph is highly interconnected with many variables having a direct relationship with most of the other variables in the graph. Here simplifying assumptions and approximation methods can currently be used to reduce the complexity of the problem and reduce the operations needed to obtain an answer.

6.2 Conclusions

Having a better understanding of the limitations of this work it is still clear to us that there is very important work being done in this area. Especially through the use of causal models to identify and correct unfair bias in data and machine learning models.

We have found it possible to very precisely pinpoint the unfair effect that certain sensitive variables can have on the outcome variable. This means that we can separate fair and unfair effects that would normally be considered as one and the same when arguing that a model is biased.

More importantly, one main principle learned is that it is very difficult and in some cases impossible to clearly identify unfair bias from data alone. We used an example of Simpson's paradox to show that one dataset viewed from different perspectives produces different answers. In our case it seemed like foreigners were being victims of unfair bias at the same time as they were the beneficiaries of positive bias. It was only through the definition of a causal graph that we were able to untangle the dependencies and find the correct way to view the data.

In our daily life we often run into this problem and resolve it unconsciously by using our built-in sense of causality (i.e. rain causes wet roads, wet roads don't cause rain). But there are often cases where this is not so simple like in the discrimination cases analyzed in similar work. We conclude that even though better tools or means of identifying unfair bias might be used, it will still be important for practitioners to explicitly state and share their assumptions regarding the causal dependencies in their data to guarantee a fairer world.

6.3 Future work

For a wider adoption of the methods and tools described in this work it is important for practitioners to have a clearer understanding of causality in statistics and how to explicitly define their assumptions about the causal dependencies in their domain of expertise. Additional work clarifying these concepts and facilitating practical use is needed.

Additional work is also needed to better understand how to correctly compute PSE and correct bias when linearity cannot be assumed. There are equalities and relationships in this work that don't hold in the non-linear setting as described in [5].

Bibliography

[1] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[2] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[3] Silvia Chiappa and William Isaac. Causal bayesian networks: A flexible tool to enable fairer machine learning, Oct 2019.

[4] Silvia Chiappa and Thomas P. S. Gillam. Path-specific counterfactual fairness, 2018.

[5] Silvia Chiappa and William S. Isaac. A causal bayesian networks viewpoint on fairness. *IFIP Advances in Information and Communication Technology*, page 3–20, 2019.

[6] Manuel Velasquez, Clair Andre, Thomas Shanks, S. J., and Michael J. Meyer. Justice and fairness, Aug 2014.

[7] Sorin Baiasu. Why fairness matters more than equality – three ways to think philosophically about justice, Jun 2020.

[8] Idil Boran. Benefits, intentions, and the principle of fairness. *Canadian Journal of Philosophy*, 36(1):95–115, 2006.

[9] Judea Pearl. *Causal inference in statistics : a primer*. Wiley, Chichester, West Sussex, 2016 - 2016.

[10] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[11] Boris Ruf and Marcin Detyniecki. Active fairness instead of unawareness. *CoRR*, abs/2009.06251, 2020.

[12] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

[14] Compas risk scales : Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county, 2016.

[15] Machine bias, May 2016.

[16] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

[17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. *CoRR*, abs/1709.02012, 2017.

[18] Robert Long. Fairness in machine learning: against false positive rate equality as a measure of fairness. *CoRR*, abs/2007.02890, 2020.

[19] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. Fairness as equality of opportunity: Normative guidance from political philosophy, 2021.

[20] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018.

[21] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints, 2020.

[22] Sarah Schröder, Alexander Schulz, Philip Kenneweg, Robert Feldhans, Fabian Hinder, and Barbara Hammer. Evaluating metrics for bias in word embeddings. *CoRR*, abs/2111.07864, 2021.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[24] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.

[25] Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187, 2016.

[26] Luca Oneto and Silvia Chiappa. Fairness in machine learning. *CoRR*, abs/2012.15816, 2020.

[27] William L. Craig. *The Kalām Cosmological Argument*. Macmillan Press, 1979.

[28] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, 2018.

[29] IPAMUCLA, Nov 2019.

[30] Sally C. Brailsford, Chris N. Potts, and Barbara M. Smith. Constraint satisfaction problems: Algorithms and applications. *Eur. J. Oper. Res.*, 119:557–581, 1999.

[31] Timo Koski and John M. Noble. *Bayesian Networks: An Introduction*. John Wiley & Sons, Ltd., 2009.

[32] Dimitri P. Bertsekas and John N. Tsitsiklis. *Introduction to Proabability*. Athena Scientific, 2002.

[33] Pat Hayes. The frame problem and related problems in artificial intelligence. In A. Elithorn and D. Jones, editors, *Artificial and Human Thinking*, pages 45–59. Jossey-Bass, Inc. and Elsevier Scientific Publishing Company, 1973.

[34] Judea Pearl and Stuart Russell. Bayesian networks, 2000.

[35] Judea Pearl. A personal journey into bayesian networks, 2018.

[36] Felix Elwert and Christopher Winship. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40(1):31–53, 2014. PMID: 30111904.

[37] Mitch Marcus. Probability, conditional probability & bayes rule. `https://www.seas.upenn.edu/~cis391/Lectures/probability-bayes-2015.pdf`, 2015. Accessed: 2021–11-23.

[38] Harri Valpola, Oct 2000.

[39] Amit Sharma, Emre Kiciman, et al. DoWhy: A Python package for causal inference. https://github.com/microsoft/dowhy, 2019.

[40] Judea Pearl. *Causality*. Springer, 2000.

[41] Ronny Kohavi and Barry Becker. Adult data set, 1996.

[42] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes, 2018.

[43] Oliver J. Maclaren and Ruanui Nicholson. What can be estimated? identifiability, estimability, causal inference and ill-posed inverse problems, 2020.