Causal Fairness of Machine Learning

Bridging Individual- and Population-Level Fairness Methods

by

Cecilia Casolo

to obtain the degree of Master of Science Applied Mathematics
at the Delft University of Technology,
to be defended publicly on November 26, 2021.

Student number:	5152534	
Thesis committee:	Dr. Geurt Jongbloed,	TU Delft (supervisor)
	Dr. Marloes Maathuis,	ETH (supervisor)
	Dr. Tina Nane,	TU Delft

TU Delft

Preface

This is a key moment in our society for researching the ethical implications of AI deployment, especially focusing on the fairness and interpretability of these algorithms' decisions. I believe that, as scientists, we have the honor and the responsibility of guiding the industry in a responsible application of algorithms that could potentially lead to the enhancement of current societal biases.
During this thesis I was also introduced to the field of Causality. It was a great discovery, I am looking forward to continuing researching on it. All the rising open causal questions applied to a vast range of topics have contributed to the dynamism and curiosity of this research field.

The months working on this thesis have been incredibly enriching, both from a personal and a research level. The support I received from the exceptional people I worked with had a great impact on my work and on my personal growth. I am grateful for the trust Geurt and Marloes had in me, enabling this collaboration to take place. Their suggestions, on-going feedback, enthusiasm, priceless availability and understanding have created a solid and enriching research environment, allowing me to range in my curiosity and passions. They have helped me in developing a critical research approach and they supported me in understanding the next steps of my academic path. Thank you, I truly had a great time collaborating with you. I would also like to thank Tina for eagerly helping me form the thesis committee.

The love and support I received from the people around me have been essential for this achievement. I would hence like to spend a few words for thanking them. I am extremely grateful for the unconditional love that my family has given me. Mom and dad, thank you for always supporting me, even in choices that implied being far away from you. Any achievement in my life is and will always be thanks to you. Thank you Lorenzo for the positive energy you gave me and for always believing in me. Giacomo, thanks for your contagious and resourceful curiosity that has always inspired me. I also thank all of my friends, for filling my life with joy and enthusiasm, always being there for my successes as well as failures.

Cecilia Casolo
Delft, November 2021

Abstract

As Machine Learning models are being applied to a wide range of fields, the potential impact that these algorithms can have on people's lives is increasing. In a growing number of applications, such as criminal justice, financial assessments, job and college applications, the data points are indeed people's profiles. Therefore, in the presence of such sensitive attributes, the risk for algorithmic predictions leading to discrimination should be carefully addressed. Among the state-of-the-art methods aiming at solving such complex problems by taking fairness into account, path-specific causality-based methods are selected in this work. In fact, causality-based fairness metrics are acclaimed in the literature for satisfactorily capturing unfairness in Machine Learning models. In this work, selected state-of-the-art causality-based methods and metrics are compared, emphasizing the methodological and experimental differences between individual- and population-level fairness approaches. Based on these results, a novel method Conditional Path-Specific Effect (CPSE) is proposed, with the goal of bridging the two different approaches by leveraging the properties of conditional Path-Specific Effects. CPSE is tested in comparison with other state-of-the-art methods, both on simulated and empirical datasets. The results suggest a high potential of CPSE for successfully detecting and correcting unfairness.

Contents

List of Figures

List of Tables

Acronyms

ML	Machine Learning	3
PSCF	Path-Specific Counterfactual Fairness	42
FIO	Fair Inference on Outcomes	45
VAE	Variational Auto-Encoder	48
PSE	Path-Specific Effect	30
NDE	Natural Direct Effect	32
NIE	Natural Indirect Effect	32
PS	Propensity Score	63
IDC	Compete Identification Algorithm for Conditional Effects	64
PS-IDC	Compete Identification Algorithm for Conditional Path-Specific Effects	64
CPSF	Conditional Path-Specific Fairness	62
MSE	Mean-Squared Error	55
ATE	Average Treatment Effect	55
MMD	Maximum Mean Discrepancy	56
DAG	Directed Acyclic Graph	9
SWIG	Single-World Intervention Graph	16
SCM	Structural Causal Model	11

1 Introduction

In the last decade, the interest in ethical decision making has risen relevantly [17]. The supposed impartiality of Machine Learning compared to the mutable nature of human decision making has been largely explored, as expressed in the words of Dr. Andrew McAfee: “If you want the bias out, get the algorithms in.” [1]. Biases in human decision making are hardly verifiable. Indeed, bias has been shown to be inherently present in human-made decisions [13]; these biases are mainly caused by unconscious preferences, gut-feelings and personal background motifs. The main obstacle of human decision making in applications with a high societal impact (such as human resources or criminal justice) is the difficulty in probing and validating the reasons behind a certain decision, as most of the times the inherent human bias is unconscious. Most of the time, humans are unaware of their biases and the reasoning behind their decisions, leading to possible inconsistency and logical gaps. These unconscious biases can be generally linked with discriminatory behaviour.

In the last decades, most of the applications have seen a steep increase of the use of Machine Learning algorithms, mainly because of time, cost and accuracy efficiency. Since these algorithms are a human creation, it is expected that these will decrease humans’ subjective interpretation of the data and will be more easily controllable, as represented by McAfee’s quote. Nevertheless, it has been proven that these algorithms can perpetuate existing social and cultural biases and introduce new ones [13], [14]. Furthermore, these algorithms are often not highly interpretable, hence also not controllable.

In the past years, numerous applications of Machine Learning leading to ethically arguable decisions have been observed in practice; this phenomenon increased the awareness of the academic and industrial environments. In fact, the majority of applications of Machine Learning algorithms can potentially lead to detrimental societal consequences, enhancing existing inequities and creating new ones. The source of this algorithmic bias can occur at all the stages of the learning mechanism, from data collection to the interpretation of the model. In order to face this truly pervasive and severe issue, it is fundamentally important to come up with strategies for identifying and mitigating the algorithmic biases. This topic has developed in a wide range of concomitantly investigated research questions among the so named AI principles, such as fairness, accountability, interpretability, privacy, trust, safety,... [29].

This work will focus on Machine Learning algorithms having a societal impact. Specifically, in the analysed applications, the data instances are people’s profiles and there is a set of sensitive attributes that can potentially lead to discrimination. Some examples of these fields are loan and job applications, criminal justice, exam grading, health-care decisions, insurance policies, ...
This project explores the fairness topic, specifically on the detection and correction of potential societal discrimination by using a causality-based approach. The topic of fairness is highly subjective, rising the attention towards the mathematical formulation of discrimination and fairness. The reader should be aware that there is no straightforward objective definition of fair/unfair, thus the definitions used in this work are the most reasonable and logical in the authors’ perspectives in order to cover many

possible discrimination scenarios.
Two main different approaches prevail in the field of fairness of Machine Learning, focusing respectively on individual- and population-level fairness. The former aims at ensuring fairness towards all the single individuals of the population, while the latter has the scope of evaluating fairness along the entire population with respect to the sensitive attribute. Although these approaches might seem similar, they include substantial differences in terms of methods and assumptions, especially when referring to causality-based metrics. The scope of this project is to compare the current state-of-the-art methods and to bridge the gap between population-level fairness and the individual-level one, by leveraging the advantages of each of these approaches. The newly proposed method will hence aim at predicting individually fair outcomes, or at least *more* individually fair outcomes, by using the weaker level of assumptions characterising population-level metrics.

The document is structured in the following way. In Chapter 2, a background on the fairness of Machine Learning research is given, emphasizing the relevance and reasoning of causality applied to fairness. Chapter 3 includes an introduction on causality; this chapter introduces the main definitions and mathematical formulations that will be used in the project. The introduction of the most commonly used causality-based fairness metrics is covered in Chapter 4. Furthermore, the state-of-the-art methods based on these metrics are compared in Chapter 5, both on simulated and real-world datasets. Based on these results, the newly proposed method Conditional Path-Specific Effect is introduced in Chapter 6, along with its application to various experiments in Chapter 7. This work is concluded in Chapter 8, along with the addition of possible future works.
It is important to notice that the background on causality and fairness presented in this work has the scope of helping the reader in understanding our novel method its research motivations. Hence, only a selected part of the current literature content will be presented, according to the scope of this project.

2

Background on Fairness

The definition itself of fairness, and the terminology related to it, is a complex matter. Indeed, it has been discussed about for years in ethical and philosophic terms ([28], [106], [10], [80]). In this chapter, we will outline what the foundation of fairness is and which approach was used in this work.

Fairness is defined in [56] as "the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics". More specifically, a decision can be considered unfair towards an individual or group of people if it is based on grounds that are unreasonable or inappropriate [32]. Albeit the above definition is suitable for the context of decision making, the concept of fairness can have different technical meanings across various sectors. For example, the authors of [39] define fairness in economics as avoiding wage cuts in periods of high unemployment rates. This work will be focused on fairness in decision-making, thus the former definition will be used. Specifically, in the context of algorithmic fairness in Machine Learning.
Based on the previous definition, in this work it will be assumed that the concept of unfairness coincides with that of discrimination. This is an assumption that could be argued against, as a non-discriminatory decision is not necessarily a fair one. Nevertheless, a fair decision is necessarily non-discriminatory. I deem that connecting unfairness and discrimination is a necessary first step in order to quantify certain decisions' fairness and to progress in the research of fairness of Machine Learning (ML).

The proposed definition of fairness is indeed highly subjective as it is based on common sense and is context-specific. For this reason, people should be encouraged in working on domain-specific algorithmic fairness to collaborate with teams specialized in the sector. In fact, when working on fairness of Machine Learning algorithms, it is important to bear in mind the purpose of the research itself, as focusing too much on the mathematical formulation of the metrics could take attention away from the broader meaning of fairness in the specific context. In particular, the authors of [32] point out that there is a common assumption that the "fair Machine Learning" community seems to share: the existence of fairness definitions (i.e. metrics) that can be operationalized and, as long as these criteria are satisfied, the fairness of the decision making algorithm is guaranteed. The authors of [32] argue that this assumption could actually lead to a loss of attention on the substantive notions of fairness in the real world. Nevertheless, I believe that this approach is a first and necessary step in order to make further improvements in the research. Once scalable and reliable methods will be developed, it will be possible to relax the underlying strict assumptions.

As a first step in this field, it is essential to take into account the possible bias in the data. Predictive models indeed highly depend on the data they have been trained on. Therefore, biased data can relevantly affect the learning process of Machine Learning algorithms, causing them to be unfair and biased [40], [9], [16].People might naturally assume that predictive algorithms used to support the human in making decisions will do it by avoiding human bias. Nevertheless, it has been proven that these models might lead to unfair decisions, judging subjects by their gender, race, or class [25]. In order to understand and evaluate the fairness of a model, it is important to identify the causes behind the bias of the data. in the first place. In this analysis the bias of data is split into *statistical bias* and *societal*

bias, following the structure of [18] and [59]. In [56], a detailed summary of the different types of bias is given.
Statistical bias corresponds to a mismatch between the real world and the data used to represent it when training the model. In particular, among statistical bias, *sampling bias* is caused by a sampling procedure that does not fully represent the population For instance, minorities could be under-represented in the dataset. Another possible challenge regarding statistical bias is *measurement bias*. This bias refers to datasets in which, due to historical or cultural reasons, the measurement error is higher for certain sub-groups compared to others [100]. For instance, referring to dataset regarding loan applications, immigrant communities were historically involved with informal lending systems [23]. This could be a possible cause of measurement bias due to an under-estimation of the number of loans re-payed by the immigrant community. Indeed, in this scenario the variable of re-payed loans would have a higher measurement error for the immigrant community since it does not take into consideration background historical factors.
Societal bias is usually harder to identify in a dataset, since it includes instances of discrimination that are inherent in the societal system [94]. Societal bias is very complicated to address in Machine Learning models, as it is a measure of an underlying social environment that is biased at first. It is then difficult to recognize what characteristics of the environment can be taken as biased or unbiased.

In the research on fairness of Machine Learning, generally no distinction between classifications/predictions and decisions is made. I deem this distinction highly relevant, as it is a direct consequence of the setting, scope and dataset to which the Machine Learning algorithm is applied. The differences between these two scenarios will now be presented.
When referring to classification and prediction tasks, the goal consists in maximising accuracy (or a specific performance measure) with respect to the target variable. Indeed, the outcome predicted by the Machine Learning algorithm represents an observed values, perceived as the true realization of the target variable. By analysing this scenario under the fairness lens, the algorithm should ensure accurate and fair predictions. It can be intuitively noticed that the balance between the accuracy and fairness metric might lead to a conflict that has been referred to in the literature as the “accuracy-fairness trade-off” [19], [104]. Hence on one side, the accuracy of the algorithm with respect to the observed target should be satisfactorily. On the other side, it is important to take the label bias into account; this is a result of a society that involves discrimination, stereotypes and social disparities. The balance between these two objectives should be defined depending on the task and on the domain-related information. The purpose of this work is to provide fairness-compliant algorithms, meaning that the accuracy of the target label will follow as a secondary target. In the research on fairness of ML it has been argued against the common misconception of data scientists that an accurate prediction is a good decision. By considering fairness, not predicting the observed target correctly can be coherent with the purpose of the prediction [99].
The second scenario for Machine Learning algorithms regards the decisions setting. This is fundamentally different from the previous classification/prediction task because it involves the presence of selective labels [45],[16], [24], [59]. The decision can indeed be interpreted as a middle-step between the observed features of an individual and his/her respective target label. Therefore, the presence/absence of a target label will be subject to the previously taken decision. For instance, in the loan applications example, the label referring to the repayment of the loan will be present in the dataset only if the individual was granted the loan at first (positive decision outcome). These settings likely involve a high number of missing values and highly biased selective labels. The fairness risk of training a ML algorithm on this dataset involves achieving a sub-optimal solution and leveraging a positive feedback-loop that could potentially be highly biased towards certain subgroups of the population. This scenario will not be investigated in this work and is left for future work.

2.1. Notation

The variables notation that will be used throughout the document is the following:

- A is the sensitive attribute, potentially leading to discrimination against individuals or groups of people (e.g. gender, ethnicity, disability, ...).
- X is the set of non-sensitive features.
- Y is the true outcome. The true outcome is not necessarily fair but it has been observed or measured, hence can be used as a training set for the algorithm. Some examples of the outcome could be the choice made to assign a loan to a candidate or not, of hiring an applicant, of predicting the potential re-offending of a parole,...
- $\hat{Y}$ is the predicted outcome from Y. Differently form Y, the predicted outcome is considered to be a fair version of Y, or at least a 'fairer' one.

The respective state spaces of these variables will be called with the following notation: $\mathcal{X}_{(\cdot)}$. For instance, the state space of A is $\mathcal{X}_A$.
In the next section, different real world examples related to the research on fairness will be explored.

2.2. Real-World Examples

The following section includes some examples that have been analysed in the fairness of Machine Learning community. They are particularly important because they are subject of ethical debates on decision making. In all of these applications, artificial agents are used to optimize processes that would normally take a long time to be assessed or to simply support the human decisions.
These examples will be referred to throughout the report in order to have some practical references that can support the reader along the theoretical insights.

2.2.1. Parole Prediction

One of the main sectors used in analysing fairness of Machine Learning is the criminal justice system [63]. In particular, algorithms are used in the sentencing and parole phase. Given a defendant, these decision support tools should decide whether to release or detain a defendant before his/her trial. A dataset often used in this setting is the COMPAS dataset [6].

2.2.2. Loan Applications

The fairness of Machine Learning algorithms is a very important topic also in the finance industry. Indeed, it might be necessary to assess whether a person would be able to pay back a loan in order to decide if it should be granted to him/her; it is useful in these cases to predict the future income of the person applying for the loan. One of the most commonly used datasets in this application is the Adult dataset from the UCI repository [50].

2.2.3. College Applications

Among the fields where Machine Learning is applied, human resources is one of the sector that involve decisions affecting individuals. In general, when people are applying for working positions or for university courses, individuals belonging to specific societal groups (by race and gender) might be discriminated. In the fairness research the dataset regarding Berkeley College admissions is the most popular one [12].

2.3. Non-Causal Fairness Metrics

In the past, fairness has been mainly analysed through metrics that are based on observational criteria, which means that they only depend on the joint probability distribution $P(X, Y, A, \hat{Y})$. In this scenario, X is the set of observed variables excluding the sensitive attribute A, Y is the possibly unfair outcome of an algorithm, $\hat{Y}$ is the estimated “fairer” outcome of the proposed algorithms. In the next definitions a classification problem is considered, where the target label Y is binary. It is immediate to generalize the following notions for prediction problems with continuous outcomes.

The below metrics have the purpose of quantifying the unfairness of a model. Since there is no objective definition of fairness, formulating suitable metric involves subjectivity of the ethical interpretation of the subject. For this reason, defining a suitable fairness metric is one of the most commonly debated challenges of the topic. In many applications it might be convenient to create a domain-specific fairness metric that is highly suitable for the targeted unfairness challenges. Nevertheless, this work will be focused on using a general fairness metric that can ideally be applied to various applications. Further, fairness metrics depend on the targeted subjects in different ways. Specifically, they can be categorized in individual- or population-level metrics. The former has the aim to avoid discrimination towards individuals, meanwhile the latter focuses on discrimination of groups of people (for example gender or a racial groups).

In the state of the art, many fairness metrics have been defined. Specifically, in this section the most popular non-causal metrics will be presented. It is interesting to point out that these metrics have been proven to be incompatible among each others ([81], [44], [16]). It will be further explained that non-causal fairness metrics fail to fully represent our idea of fairness. Indeed, in this work's perspective, solely causal-based metrics can achieve this result.

2.3.1. Demographic parity

A predictive algorithm is considered to satisfy demographic parity if, on average, it gives the same predictions to the different groups [107]. In mathematical formulas, this corresponds to having $\hat{Y} \perp\!\!\!\perp A$ [26]. In our binary scenario, given $y \in \mathcal{X}_Y$, then:

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

so the likelihood of a positive outcome should be the same across the different groups [72].

Demographic parity has been criticized, as it only contains information regarding the output and the group. This is considered to be a criterion that is not suitable in some cases, which will now be presented. First of all, demographic parity is always satisfied, as long as the groups of the sensitive attribute are given the same opportunity (positive outcomes with the same rate); however, no information is given on how the fractions are chosen in each group, risking to fall into positive discrimination [35]. Secondly, in some examples there might be a legitimate correlation between A and $\hat{Y}$, for example the choice of the department in the college admission example or the job qualifications in the loan applications example. Since these are the most common scenarios in real world datasets, this work will relevantly focus on these; a more detailed explanation will be given in the next sections.

2.3.2. Equalised odds / Equal opportunity

Equalised odds has the scope of giving the same positive predictor to the individuals in each group [35]. Then, when referring to a binary outcome, the rate of true positives and true negatives should be the same across each group in a sensitive attribute. For a predictor $\hat{Y}$, this metric can be mathematically formalized as:

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

for all $y \in \mathcal{X}_Y$, equivalently requiring $\hat{Y} \perp\!\!\!\perp A|Y$. With a binary classifier, this means that the false positive rates and false negative rates should be the same along the different groups; thus, equalized odds requires equality of error rates among the sub-groups defined by A. The origin of the name indeed comes from the scope of matching the false positive rates and true positive rates for different values of the sensitive attribute.

Equal opportunity is a specific version of equalized odds, mainly focusing on false positive rates compared to false negative rates:

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

The motivation behind this metric is that typically, with a binary classifier, there is more interested in not wrongly denying a desirable opportunity rather than providing an undeserved one; thus, the probability of a subject being assigned a positive outcome, given that he/she truly belongs to the positive class, should be equal for all the groups [101].
This metric has been criticized for not being able to describe possible societal bias in the dataset. For instance, an algorithm that grants loans to individuals that have been able to repay previous loans in the past can be considered. Regarding this, equal opportunity does not take into account that for some societal groups it is harder to find a high-paying job, and hence to repay the loans. Without capturing similar societal unfairness situations, equal opportunity and equalized odds might not be suitable metrics in some contexts.

2.3.3. Fairness through Unawareness

The most intuitive procedure to avoid unfairness towards individuals from a specific group is to omit the sensitive attribute A from the inputs of the model [34]. For example, when training the algorithm that has as output the loan granting decision, data on ethnic (or gender) origin would be removed. Thus, the estimate output $\hat{Y}$ that ensures fairness through unawareness would be defined as $\hat{Y} : \mathcal{X}_X \Rightarrow \mathcal{X}_Y$ (in our notation X does not include A). This method cannot actually guarantee a fair outcome. Indeed, in most real world examples, sensitive attributes are correlated with other features in the dataset. For instance, an algorithm could actually learn information regarding the gender of a person from his/her name. Moreover, if historically a neighborhood is populated by people with the same ethnicity background, the address of the house could potentially give information regarding the race of the subject. These covariates are called *proxy variables*, potentially causing proxy discrimination [42]. Albeit fairness through unawareness is commonly used in the industry because it is immediate and intuitive, most of the times this method is not sufficient for removing discrimination.

2.4. Why Causality in Fairness?

In recent years, the use of causality in defining fairness has been increasing more and more. Causality-based fairness seems to be the most promising pathway, as only by truly understanding the causes of discrimination we can actually build a fair algorithm that is able to avoid it. Normally, when analysing a human-made decision, there are certain questions that should be answered in order to evaluate the fairness of the decision process; for instance, “was the gender the cause of her rejection?”. The analysis of discrimination is commonly done by questioning the causal relations between the sensitive attribute and the decision. The same reasoning can be applied to fairness in Machine Learning algorithms, by formalising these questions as causal fairness metrics. It is indeed necessary to use causal reasoning in order to justify the influence that certain individuals’ characteristics have on the model, specifically by analysing the causal effect that the sensitive attribute might have on the outcome. Furthermore, causal models can be helpful for modelling the underlying data generation mechanism [43].

In order to show the importance of causality in the evaluation of fairness, the example of Berkeley college admissions is used. With an analysis on the overall dataset, it was shown that women had on average an admission rate 10% lower than men. However, decisions were made on a departmental basis, where the acceptance among females and males was the same. Therefore, when conditioning on the department choice, the trend of females-males rates acceptance is different from the one of all the applicants. This is commonly known as Simpson’s Paradox; it describes datasets where the subgroups

of subjects do not represent the trends of the whole population [89]. In this specific case, the fairness through unawareness method would not detect the unfairness of this dataset, as it would not recognize the causal relationship between the gender and the choice of the department. Indeed, even though the acceptance rate in each department is the same among men and women, women might be influenced by their socialization to choose departments that are generally more selective, more crowded, leading to an overall acceptance rate lower for women [12]. In [66], a detailed analysis of the Simpson's Paradox on a causal basis is given.
This is only one of the examples that shows how fairness cannot be assesses based only on the concepts of association or correlation, but causality is a useful tool that can be used.

The above example shows the usefulness of considering the underlying causal relations among the data. In this scenario, if the causal structure of the model had been taken into account, the causes of unfairness could have been revealed. Nevertheless, since distinct causal mechanisms can lead to the same observational distribution, non-causal fairness metrics will generally fail in identifying the true causes of discrimination. In conclusion, the causal framework is the most appropriate one for evaluating unfairness.

3

Introduction to Causality

In this chapter, an introduction to causality will be given. Causal models will be introduced, along with the notation that will used in the document. Furthermore, the notion of identifiability will be explored with its respective conditions depending on the graphical structures.

3.1. Notation and Definitions

Directed graphs can be useful for visualizing causal models. These are defined as:

Definition 3.1.0.1. ***Directed Graph*** *[2] A directed graph $\mathcal{G}$ on the set of vertices $W = \{w_1, w_2, ..., w_n\}$ is a subset of $W \times W$, where the members of $\mathcal{G}$ are called edges $\mathcal{E}$.*

A *directed path* in $\mathcal{G}$ from vertex u to vertex v is a sequence of distinct edges $e_1, e_2, ..., e_p$ with $p \geq 1$ such that there exists a corresponding sequence of vertices $u = w_0, w_1, w_2, ..., w_p = v$ satisfying $(w_k, w_{k+1}) \in \mathcal{G}$, for $0 \leq k \leq p-1$. In other words, a directed path is a sequence of two or more unique nodes among which each successive node has an edge with its predecessor in the sequence [2]. In this work, by $(w_0 w_1)_\rightarrow$ is meant the edge going from w_0 to w_1; similarly, the notation $w_0 \rightarrow w_1 \rightarrow w_2$ (or equivalently $(w_0 w_1 w_2)_\rightarrow$) is the path starting in w_0, ending in w_2 and passing by w_1.
Among the useful definitions regarding directed graphs, cycles need to be introduced. A *cycle* is a directed path beginning and ending at the same vertex and which passes through at least one other vertex. This notion is useful in order to define a specific subgroup of graphs:

Definition 3.1.0.2. ***Directed Acyclic Graph*** *[2] A Directed Acyclic Graph (DAG) is a Directed Graph which does not contain cyclic connections of directed edges.*

In the setting of this work, a DAG $\mathcal{G}$ is defined by the duple $(W, \mathcal{E})$, composed by a set of vertices (nodes) W and edges $\mathcal{E}$. The nodes represent the different random variables used to describe the model, meanwhile the edges denote a certain relationship, depending on the application of the DAG. It will be shown that in the causal interpretation of DAGs the edges represent causal relations among the variables.

In general, the set of variables W can be split into a set of observed variables V and unobserved variables U, such that $W = V \cup U$. The observed variables can be measured/observed in an experimental or real-world setting, meanwhile the unobserved variables cannot be measured in practice. Referring to the notation previously used when introducing fairness, the set of observable variables V is the union of the set of the non-sensitive attributes X, the outcome Y and the sensitive attribute A, $V = X \cup A \cup Y$. To have a better understanding of unobserved variables in the fairness context, these might represent the background factors that cannot be measured (for instance the influence of the socio-economic background on the covariates of an individual). While the set of $V = \{V_i\}_{i=1}^n$ are the endogenous variables, hence measurable, U is the set of exogenous variables (also referred to as noise variables). As the name suggests, the exogenous variables are root nodes (see Definition 3.1.0.7) by construction, as they are not descendants of any other observed variables. A common assumption that is made in many methods is that U_i for all i are independently distributed. As an example of this notation, in Figure 3.1 a directed graph is represented, where the set of observable variables V is $V = A \cup M \cup Y$ and the set of unobserved variables is $U = U_A \cup U_M \cup U_Y$.

It will be useful for the next chapters to define the notions of parents, children, ancestors, descendant, topological ordering, root nodes in DAGs.

Definition 3.1.0.3. ***Parents, Children of a Node*** *The parents and children of a node $V_i \in V$ in the Directed Acyclic Graph $\mathcal{G}$ are respectively defined as:*

$$Pa(V_i) = \{B \in V : (BV_i)_{\rightarrow} \in \mathcal{G}\}, \tag{3.1}$$

$$Ch(V_i) = \{B \in V : (V_iB)_{\rightarrow} \in \mathcal{G}\}. \tag{3.2}$$

The parents $Pa(V_i)$ will also be referred to as Pa_i. Furthermore, it is possible to introduce the set of ancestors and descendants.

Definition 3.1.0.4. ***Ancestors*** *Given a node V_i in the Directed Acyclic Graph $\mathcal{G}$, the set of ancestors of V_i is defined as $An_{\mathcal{G}}(V_i) = \{X : \{X \rightarrow ... \rightarrow V_i\} \in \mathcal{P}\}$, where $\mathcal{P}$ is the set of all the paths contained in $\mathcal{G}$.*

Definition 3.1.0.5. ***Descendants*** *Given a node V_i in the Directed Acyclic Graph $\mathcal{G}$, the set of descendants of V_i is defined as $De_{\mathcal{G}}(V_i) = \{X : \{V_i \rightarrow ... \rightarrow X\} \in \mathcal{P}\}$, where $\mathcal{P}$ is the set of all the paths contained in $\mathcal{G}$.*

Once the ancestors and descendants have been defined, the topological order of a graph $\mathcal{G}$ is introduced in [86] as:

Definition 3.1.0.6. ***Topological Order and Predecessors*** *A topological order $\prec_{\mathcal{G}}$ on the set of nodes V in $\mathcal{G}$ is such that, given $V_1, V_2 \in V$, if $V_1 \prec_{\mathcal{G}} V_2$ then $V_2 \notin An_{\mathcal{G}}(V_1)$. Given an order $\prec_{\mathcal{G}}$ defined on V, for any $V_1 \in V$, $pre_{\prec_{\mathcal{G}}}(V_1) := \{W \in V \backslash \{V_1\} : W \prec_{\mathcal{G}} V\}$.*

As previously mentioned, unobserved variables are root nodes by construction. Root nodes are defined in the following way:

Definition 3.1.0.7. ***Root Node*** *Given a causal graph $\mathcal{G}$ with nodes V, a node $V_i \in V$ is a root node, $V_i \in Ro(\mathcal{G})$, if it has no observed parents.*

Analysing different graphical structures, it can be noticed that each variable has a different role in the model also depending on the number of emitted/admitted edges. In order to clearly define the most typical and significant structures that can be met in DAGs, it can be helpful to introduce specific terminology for the next sections.

The presence of a node with two or more edges emitted towards different children is a common graphical structure; this network structure is called *fork* and it usually corresponds to three nodes, with two edges emanated from the initial node. For instance, in Figure 3.2, the triplet (C, A, Y) represents a fork. Differently, a *chain* structure corresponds to three nodes with one arrow directed into and one arrow directed out of the middle variable, called *mediator*. For instance, the mediator variable of Figure 3.2 is M. Finally, a *collision* node is a variable that has two or more edges directed into it; for instance, in Figure 3.2 Y is a collision node.

In the next section, Causal Models will be presented, specifically uncovering the link between DAGs and causal models.

3.2. Causal Models

In this section, causal models will be introduced. In particular, the section will focus on key concepts of causality that will be used throughout the document. DAGs will be presented as a useful tool for representing causal models, specifically in the analysis of causal assumptions and in facilitating observational-based causal inference [69].

3.2.1. Causal Graphs and Notation

Causal graphs can describe the causal assumptions regarding the data generating process through structural causal models (also called structural equations model). Structural causal models are here defined as:

Definition 3.2.0.1. ***Structural Causal Model (SCM)*** *A structural causal model can be defined as a triple* (U, V, F) *[66] such that:*

- V *is the set of observed variables;*
- U *is the set of unobserved variables;*
- F *is a set of functions* $\{f_1, ..., f_n\}$ *for each of the observed variables* $V_i \in V$*, such that* $V_i = f_i(Pa_i, U_{Pa_i})$*. These equations are known as structural equations;*
- $P(U)$ *is a distribution over the unobserved variables* U*.*

By Pa_i it is meant the set of nodes among the observed variables that are direct causes of the variable V_i, also called observed parents of V_i. Meanwhile, U_{Pa_i} is the set of direct parent of V_i among the set of unobserved variables U. Albeit the unobserved variables are omitted from the model, these are considered to be relevant for describing the causal mechanisms; they are also referred to as "noise terms" or "errors".
Intuitively, the structural equations are useful for representing how the variable V_i changes in response to its direct causes $Pa_i \cup U_{Pa_i}$. These functions are called *structural* because they are actually invariant from arbitrary changes in the other functions, as they describe which values the variables assumed based on nature [66] [88]. For instance, the structural function f_i corresponding to V_i will not change, even if the function f_j corresponding to the other variable V_j changes form. Indeed, as Pearl states in [69], these functions actually describe the underlying real-world physical processes.

An example of a structural causal model can be the following. Given the causal graph in Figure 3.1, its respective structural equations could have the following formulation:

$$\begin{aligned} A &= f_A(U_A) \\ M &= f_M(A, U_M) \\ Y &= f_Y(M, A, U_Y). \end{aligned}$$

As previously explained, these equations are functions f_i of the parents of the respective variable V_i, both observed and unobserved. In this example, three different unobserved variables were used for modelling the causal structure.

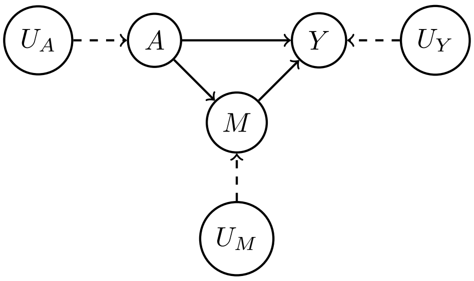


Figure 3.1: Example causal graph with mediator and common cause for A and Y

The edges of causal graphs have a duplicate interpretation: a probabilistic and a causal one. In the former, the edges give us information regarding the conditional dependencies among the variables; in the latter the edges represent the causal relations in the data-generating process. Analysing the causal perspective, it is easier to understand why the exogenous nodes are defined as root nodes: when modelling the causal structure of the dataset, it is chosen not to explain how unobserved variables are caused.
It is important to notice that in a causal graph the absence of an edge between two variables is a stronger condition than an edge itself. Indeed, drawing an edge among two variables could mean that there is a possible causal relation. Differently, not drawing the edge ensures a lack of causal relation.

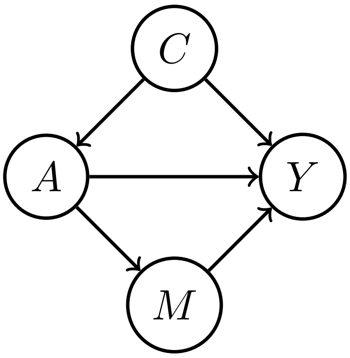


Figure 3.2: Example causal graph with unobserved variables

Properties of Causal Graphs

In this section, different characteristics of causal graphs will be analysed. It will indeed be shown how DAGs are useful for deriving conditional independence relations among the variables.

By assigning a causal meaning to DAGs, an edge between two nodes can be interpreted as the source variable being direct cause of the latter one. Two of the main properties of DAGs will be presented; specifically, the Markov property and the d-separation one will be introduced. It was previously mentioned that these graphical objects can have two (among many) functions: representing causal structures and probability distributions. The following properties connect DAGs with probability distributions, without involving any causal interpretation of the models. Nevertheless, by interpreting these properties with causal lenses, a connection between the probabilistic relations and the causal structure of the model can be made.
The Causal Markov Condition, as stated in [69], is the following:

Definition 3.2.0.2. ***Causal Markov Condition*** *Any distribution generated by a Markovian model $\mathcal{M}$ can be factorized as:*

$$P(V_1 = v_1, V_2 = v_2, ..., V_n = v_n) = \prod_i P(V_i = v_i | Pa_i = pa_i) \tag{3.3}$$

where $V_1, V_2, ..., V_n$ are the endogenous variables in $\mathcal{M}$, and pa_i are the values obtained by the endogenous parents of V_i in the causal diagram associated with $\mathcal{M}$.

This condition is equivalent to stating that each variable V_i is independent on its descendants conditioned on its parents Pa_i in $\mathcal{G}$. Markovian models are associated with acyclic graphs with jointly independent error terms. The property of Markovian models in Equation 3.3 is also called Factorization Property.

Analysing the properties of Directed Acyclic Graphs, the concept of *d-separation* can be introduced, as proposed in [69]. It is important to notice that the following proposed properties characterize DAGs, independently on the causal meaning assigned. This property was explained with causal lenses in

order to make the future content more intuitive to the reader. The notion of d-separation uncovers information on the dependencies corresponding to all the datasets generated by the same given graph, independently on the form of the structural equations.
The concept of *blocking* underlies the one of d-separation:

Definition 3.2.0.3. ***Blocking**: A set S of nodes is said to block a path p if either*

- *among the nodes in the set S, at least one of the edges emitted by one of them is part of p; or*
- *p contains at least one collision node that is outside S and has no descendant in S*

For instance, in Figure 3.1, the path $(U_A AMY)_{\rightarrow}$ is blocked by both the sets $S = \{A\}$ and $S = \{M\}$, as each of these variables emits an arrow along the path. The intuition of *blocking* will become clearer in the next paragraph.
Analysing the probability distributions induced by DAGs, it can be useful to formulate a criterion in order to predict conditional independence relations, for simple or more complex graphical structures. In a DAG, two variables are said to be connected if there exists a connecting path (not necessarily directed) among them. Meanwhile, two variables are d-separated if there exists no such path between them. The former means that the two variables are possibly dependent, meanwhile the latter means that these are independent. This can be intuitively understood based on the definition of the causal models and the corresponding structural equations. Usually, two nodes in a causal graph are always connected, which means that rarely two variables belonging to the same causal graph are independent among each others. Nevertheless, when conditioning on a set of nodes S, it is possible to get further relevant information regarding the independence relations conditioned on the set of variables S.

Definition 3.2.0.4. ***D-separation**: If the set of variables S blocks all paths from the variable X to the variable Y, it is said to "d-separate X and Y", and then $X \perp\!\!\!\perp Y|S$.*

In order to understand the concept of d-separation, it can be useful to analyse the connection between blocking and conditional independence when only one path is present between the two variables. Analysing the first condition of blocking, non-collider nodes with at least one emitting edge part of p could potentially be nodes with two emitted arrows or with one entering and one emitting arrow (transition nodes). In both of these scenarios, it can be shown that the extreme variables of the path, following the notation of Definition 3.2.0.4 X and Y, are independent among each others when conditioning on the non-collider node. For instance, analysing the path $(AMY)_{\rightarrow}$ in Figure 3.5a, the joint probability of A and Y conditioned on M (assuming that $P(M = m) > 0 \quad \forall m \in \mathcal{X}_M$) can be written as:

$$\begin{aligned}
P(Y = y, A = a|M = m) &= \frac{P(Y = y, A = a, M = m)}{P(M = m)} \\
&= \frac{P(A = a|M = m)P(M = m)P(Y = y|M = m)}{P(M = m)} \\
&= P(Y = y|M)P(A = a|M = m) \quad \forall m \in \mathcal{X}_M, a \in \mathcal{X}_A, y \in \mathcal{X}_Y \implies Y \perp\!\!\!\perp A|M.
\end{aligned}$$

The same procedure can be applied to scenarios where M has two emitted edges towards A and Y.
The intuition behind the second condition of the blocking definition follows from the fact that a collision node blocks the dependence among its parents, hence blocking the respective path. When conditioning on a set of nodes S that is part of the collision node or one of its descendants, the parents are dependent among each others. Nevertheless, if the conditioned set is not part of the collision node or of its descendants, the parents are actually independent. For instance, the following ideal model can be considered. Both the weather conditions and the drivers' sobriety are direct causes of the number of accidents (in this example a collision node). The sobriety of a driver and the weather conditions are independent. Nevertheless, by knowing if an accident has happened or not leads the sobriety state and the weather conditions not to be independent anymore.
These consideration can be generalised to more complex causal structures, where X and Y are set of nodes connected by multiple paths. Looking back at the example in Figure 3.1 and at the proposed variables U_A and Y, it was previously explained why the set $S = \{M\}$ blocks the path $(U_A AMY)_{\rightarrow}$. Nevertheless, this set $S = \{M\}$ does not block the path $p = \{(U_A AY)_{\rightarrow}\}$, as M has no emitting arrow in p and there is no collision node in p. Meanwhile, $S = \{A\}$ block all the paths from U_A to Y; indeed,

in both of the cases it presents an arrow emitting edge in the path. This means that the random variable A d-separates U_A from Y, leading to $U_A \perp\!\!\!\perp Y|A$. In the next paragraphs, it will become clear why the above conditioned independence relations are important when referring to identifiability conditions.

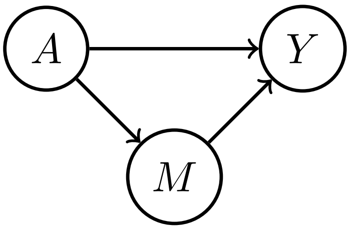


Figure 3.3: Example causal graph with mediator for A and Y

It can be useful to consider models with unobserved common causes of multiple observed variables.In these scenarios, the Causal Markov Condition (Definition 3.2.0.2) fails, and it is necessary to introduce latent variables in order to describe the causal structure [55], [92]. Semi-Markovian causal models can be defined as:

Definition 3.2.0.5. ***Semi-Markovian Model*** *A model $\mathcal{M}$ is semi-Markovian if the* DAG *$\mathcal{G}$ associated to the model includes a set of observable variables $V = \{V_1, ..., V_n\}$ and latent variables $U = \{U_1, ..., U_d\}$. By construction, every latent variable is a root node.*

As no restrictions are imposed on the co-dependencies of the unobserved variables, it can be noticed that Markovian models are a subgroup of semi-Markovian models, specifically characterized by jointly independent error terms of SCMs [69]. In Semi-Markovian models dependence relations among the unobserved variables lead to useful insights regarding causal relations among the observed variables. These models can be graphically illustrated by representing hidden variables in DAGs with nodes U_i that have one or more children nodes. In order to simplify the graphical structure of semi-Markovian models, it is possible to analyse a semi-Markovian projection instead. This is defined in [97] like:

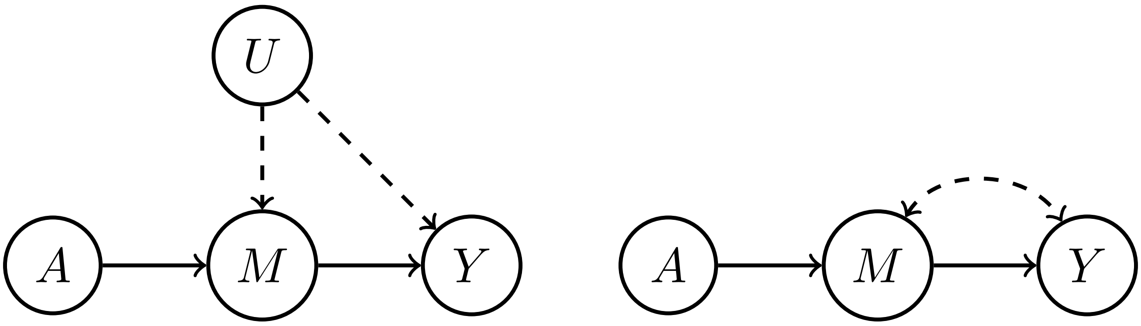
Definition 3.2.0.6. ***Projection*** *The projection of a* DAG *$\mathcal{G}$ over $V \cup U$ on the set V, denoted by $PJ(\mathcal{G}, V)$, is a* DAG *over V with bi-directed edges constructed as follows:*

- *Add each variable in V as a node of $PJ(\mathcal{G}, V)$.*
- *For each pair of variables $Z, Y \in V$, if there is an edge between them in $\mathcal{G}$, add the edge to $PJ(\mathcal{G}, V)$.*
- *For each pair of variables $Z, Y \in V$, if there exists a directed path from Z to Y in $\mathcal{G}$ such that every internal node on the path is in U, add edge $(ZY)_{\rightarrow}$ to $PJ(\mathcal{G}, V)$ (if it does not exist yet).*
- *For each pair of variables $Z, Y \in V$, if there exists a divergent path between Z and Y in $\mathcal{G}$ such that every internal node on the path is in U $(X \leftarrow U_i \rightarrow Y\ )$,add a bi-directed edge $Z \longleftrightarrow X$ to $PJ(\mathcal{G}, V)$.*

This projection is called Acyclic Directed Mixed Graphs (ADMGs), as it can be used in support of DAGs for analysing semi-Markovian models [102]. An ADMG is a mixed graph without directed cycles, containing both directed edges $\rightarrow$ and bi-directed arrows $\leftrightarrow$. Bi-directed arrows between two nodes V_i and V_j are used to represent graphical causal structures like $V_i \leftarrow U \rightarrow V_j$, where U is a hidden confounder. For instance, Figure 3.4b is the ADMG of Figure 3.4a. The bi-directed arrow between M and Y hence represents the co-dependency between the unobserved variables U_M and U_Y, also possibly represented as a common unobserved cause of M and Y.
When considering ADMGs, it is useful to introduce the notion of *district*. Following the definition in [87]:

Definition 3.2.0.7. ***District*** *A district $\mathcal{D}$ in an ADMG $\mathcal{G}$ with vertex set V is any maximal set of nodes in V that are mutually connected by bi-directed paths which are themselves entirely in $\mathcal{D}$.*

The variables in a district are thus connected by confounded paths. These are represented with bidirected arrows in ADMGs, while in DAGs as paths where all directed arrowheads point at observed nodes, and never away from observed nodes. The districts of a graph are disjoint, and can thus be either composed by single observed nodes or set of observed nodes that are connected by confounded paths. More intuitively, any two observed variables that share a common unobserved parent belong to the same district. For instance, in Figure 3.4b the disjoint districts are $\mathcal{D}_1 = \{M, Y\}$ and $\mathcal{D}_2 = \{A\}$.In a Markovian model, each node in the causal graph is a distinct district.



(a) Example Causal Graph with Unobserved Variables

(b) Respective Acyclic Directed Mixed Graph

Figure 3.4: Example causal graph with its respective ADMG

3.2.2. Interventions and Counterfactuals

Intervention is a manipulation used for analysing causal relations among variables. This is a core concept of causal inference. The main idea behind interventions is that by manipulating a variable A as $a \in \mathcal{X}_A$, it is possible to analyse the consequence it has on a different variable Y. Hence, intervening on a variable A as $A = a$ with respect to Y means that the variable A is forced to take the value a and Y assumes the value that it would naturally have following the intervention on A. Going back to causal models, this means that all the children of A will also naturally change according to the new value of A. This intervention procedure is useful, since the changes of the value of Y happen only in virtue of the changes in the attribute A.

Using the approaches of [66] and [93], interventions can be formalised with both structural equations and causal graphs. Starting from the structural equations, intervening on a variable V_i with the arbitrary value $v_i \in \mathcal{X}_{V_i}$ corresponds to substituting the equation $V_i = f_i(Pa_i, U_{Pa_i})$ with the equation $V_i = v_i$. This corresponds to manipulating the model by fixing the value of V_i to v_i, regardless of the causal structure. As an example, analysing the graph in Figure 3.1 and the respective previously presented structural equations, after intervening on A as $a \in \mathcal{X}_A$ the structural equations are the following:

$$\begin{aligned} A &= a \\ M &= f_M(A, U_M) \\ Y &= f_Y(M, A, U_Y). \end{aligned}$$

The formalization of interventions with causal graphs follows the same intuition. As a matter of fact, it can be noticed that in this example the structural equations also correspond to a causal graph with the same structure as Figure 3.1 but with the edges directed towards the intervened variable A eliminated. Indeed, the structural equation corresponding to the variable (node) A is substituted with the fixed value a. This procedure can be generalized for different causal structures and interventions, by eliminating the edges entering into the intervened variable. This new causal graph representing the intervention on A as a is referred to in the literature as $\mathcal{G}_{\bar{a}}$.

An intervention on V_i is denoted by the operator $do(v_i)$ [66], [71]; based on this concept, a do-calculus has been proposed by [66], the foundations of which will be presented in the later sections.

In the research on Machine Learning fairness, it is useful to analyse the output Y when the sensitive attribute has been intervened upon. Indeed, this can be quantified with the interventional distribution denoted by $P(Y|do(a))$ with $a \in \mathcal{X}_A$. In our notation, we will refer to $P(Y|do(a))$ as $P(Y(a))$ in order to simplify further steps and to avoid confusion with conditioning of traditional statistics. The quantity $Y(a)$ is also denoted as *potential outcome*. The concept of conditional probability of Y given A should not be confused with its interventional distribution. There is a relevant difference between the quantities $P(Y(a))$ and $P(Y|A = a)$, as the former represents the distribution of the entire population of Y if everyone had their A set to a, while the latter corresponds to the distribution of the individuals that had the value of A equal to a.

The potential of structural equations as a modelling tool for interventions is supported by the modularity property. This feature of invariance is useful for linking the probability distributions of interventional and observed quantities. In [22], the Modularity property is defined in the following way:

Property 3.2.1. ***Modularity*** *The pair $(\mathcal{G}, P(V))$ obeys modularity if, for any node $V_i \in V$ in $\mathcal{G}$, its conditional distribution given its* DAG *parents Pa_i is the same, no matter which variables in the system (other than V_i itself) are intervened upon.*

It is important to notice that this holds for the concept of manipulation (intervention) and not observation. The Modularity property is coherent with the previous definition of structural equations. As it was mentioned in the introduction of SCMs (refer to Definition 3.2.0.1), these functions reflect underlying relations among the variables that will remain unchanged, independently of changes in the other variables (like interventions).

Another key property of interventional distributions is consistency. This is defined in [77] as:

Property 3.2.2. ***Consistency*** *Given two (sets of) variables $V_i, V_j \in V$, the following holds:*

$$V_j = v_j \implies V_i(v_j) = V_i, \tag{3.4}$$

where $v_j \in \mathcal{X}_j$.

This property can be intuitively understood by analysing the meaning of interventions. Indeed, the random variable $V_i(v_j)$ is the potential outcome of V_i if the variable V_j was set to be v_j. In the scenario where V_j is observed and is equal to v_j itself, then it follows that the potential outcome $V_i(v_j)$ will be the random variable V_i itself. The equality $P(V_i(v_j) = v_i|V_j = v_j) = P(V_i = v_i|V_j = v_j)$ is hence a consequence of the consistency property. The intuitive reason is that an intervention corresponding to an event that has already happened has no consequence on the rest of the variables of the model.

In the causality research, different approaches have been developed. Specifically, two main formalisms have been developed for reasoning about causal questions: potential outcomes and DAGs. Single-World Intervention Graphs (Single-World Intervention Graph (SWIG)s) are introduced in [74] and will be applied in order to combine the graphical approach with the potential outcomes one. In these graphs, as it can be noticed in Figure 3.5b, a "node" operation is made. This consists in splitting the node corresponding to the intervention into the random variables itself and the particular value it has been set through the intervention, which will be referred to as $a \in \mathcal{M}_A$. Thus, by considering $Y(a)$ in the respective SWIG, the node A would be split into a random half and a fixed half, where the fixed half inherits the outgoing edges, while the random part the incoming edges. More formally,

Definition 3.2.2.1. ***Single-World Intervention Graph*** *[74] The Single-World Intervention Graph (SWIG) $\mathcal{G}(v_j)$ for the intervention $V_j = v_j$ is constructed from $\mathcal{G}$ via the following steps:*

- ***Node splitting**: For every variable in V_j, the node V_j is split into a random and a fixed component, respectively labelled V_j and v_j. The random half inherits all edges into V_j in $\mathcal{G}$ and the fixed half inherits all edges out of V_j in $\mathcal{G}$.*
- ***Labelling**: For every random node $V_i \in De_{\mathcal{G}}(V_j) \backslash V_j$ in $\mathcal{G}$, the label is $V_i(v_j)$.*

These graphical representations of potential outcome densities can be useful for understanding the independence assumptions that interventional quantities involve, as these are needed in order to test

their identifiability. SWIGs help unify the potential outcome and graphical formalisms, as these tools allow us to graphically interpret conditional independencies involving both factual and counterfactual quantities. As potential outcomes analysis is based upon statements on counterfactual quantities, these can be interpreted graphically through Single World Intervention Graphs.
In [74], multiple properties of SWIGs are presented. Given a SWIG $\mathcal{G}(a)$ with respective counterfactual

distribution $P(V(a))$, the Factorization and Modularity (respectively Equation 3.3 and Property 3.2.1) characteristics can be extended to SWIG causal structures. Specifically, the following proposition holds:

Proposition 3.2.3. *Under the FFRCISTG model (see Definition 4.2.0.3) associated with $\mathcal{G}$, the distribution $P(V(a))$ over the variables in $\mathcal{G}(a)$ factorizes according to $\mathcal{G}(a)$.*

Furthermore, if a model satisfies the Causal Markov Condition (Definition 3.2.0.2), then the d-separation implies statistical independence on SWIGs. Additionally, the Modularity property for SWIGs holds according to the following proposition of [74]:

Proposition 3.2.4. *Under the FFRCISTG model associated with $\mathcal{G}$, the pairs $(\mathcal{G}, P(V))$ and $(\mathcal{G}(a), P(V(a)))$ obey modularity.*

The above property shows that the conditional density associated with $P(V(a))$ in $\mathcal{G}(a)$ corresponds to the one associated with $\mathcal{G}$ after substituting to the random variable A the value a.
Both the Modularity and Factorization properties applied to SWIGs are important; the former one allows to link the two sets of distributions on the original graph $\mathcal{G}$ and the SWIG $\mathcal{G}(a)$, while the latter one enables the use of the d-separation criterion for SWIGs. The proof of these two properties can be found in Appendix B.1 of [74].

Differently from simple interventional quantities, which could be expressed by the marginal interventional distributions, for instance $Y(a_0)$ and $Y(a_1)$ with $a_0, a_1 \in \mathcal{X}_A$, counterfactuals imply a joint distribution between the factual and the counterfactual. The models only involving the marginal distributions of the interventional quantities are also called *single-world models*, meanwhile the ones involving a joint probability distribution like $P(X, Y(a_0), Y(a_1))$ are referred to as *cross-worlds models* [74].
Single-world models are the most used ones in the causality research, as cross-worlds models involve assumptions that are experimentally untestable and that are not intuitive [3]. Nevertheless, it will be explained in the next sections how cross-world models can potentially be useful in certain fairness metrics. Differently from standard interventions, in cross-world models all the involved variables are not measured in the same world.

In causal inference, the concept of unit-level *counterfactuals*, based on interventions, is useful for quantifying single unrealized or untrue events. In [69], counterfactuals are presented as values that answer *if* questions, such as an hypothetical condition that has not happened in the real world. Unit-level counterfactuals are a special type of intervention that is applied at an unit-level, answering what would have happened if one of the unit features happened to be different. The computation of counterfactuals is a common causal inference task, as they can be useful when counterfactual versions of the same unit of a datasets are needed.

The computation of unit-level counterfactuals is different from the previously introduced interventional quantities. An example that will make this distinction easier is now presented. Analysing the loan application example (see Section 2.2), given $\mathcal{X}_A = \{a_0, a_1\}$, having a_0 corresponding to males and a_1 to females, the interventional distribution of the outcome setting $A = a_0$ is defined as $P(Y(a_0) = y)$ and it describes the population distribution of Y as if *everyone* in the population had their gender value set to males. Differently, when analysing the unit-level counterfactual distribution of a female subject i, the goal is to assess the outcome that the same subject at that particular instant of time would have if she was a male.
Formally speaking, setting the snapshot of a particular individual corresponds to setting a specific instantiation of the back-ground variables U as the realization u. This set of unmeasured variables can then determine all the other observable ones V through the structural equations, except for the ones that have been intervened upon. Indeed, each instantiation $U = u$ ideally corresponds to a single

member in the population (the unit of the fairness setting) and can hence uniquely determine the values of all the observable variables V [48], [52]. As previously mentioned, the unobserved variables describe the background environment. For example, when the units of a dataset are people, U could include knowledge, family situation, propensity to engage in physical activity, and many others; even though these cannot be included in the model, they can potentially uniquely define each subject.
Unit-level counterfactuals are defined in [69] as the following:

Definition 3.2.4.1. ***Unit-Level Counterfactuals*** *Let $\mathcal{M}$ be a SCM (Definition 3.2.0.1) with $W = V \cup U$ and $A \in V$. Let $\mathcal{M}_a$ be a modified version of $\mathcal{M}$, with the equation of A replaced by $A = a$. Denote the solution for Y in the equations of $\mathcal{M}_a$ for the instantiation $U = u$by the symbol $Y_{\mathcal{M}_a}(u)$. The counterfactual $Y_a(u)$ (Read: "The value of Y in unit u, had A been a") is given by:*

$$Y_a(u) = Y_{\mathcal{M}_a}(u). \tag{3.5}$$

In the above definition, $\mathcal{M}_a$ represents the SCM subsequent to the intervention on A as a. This definition reflects the intuition that the units are defined by the vector of background variables U, which determines the response and snapshot of a specific behaviour of Y. Following the definition, the reasoning behind the name of these quantities in [69] as *unit-level counterfactuals* is intuitive. The identification of these unit-level counterfactuals will be presented in Section 3.3.2.

3.3. Identifiability

In causal inference, the identification of interventional quantities is a relevant topic. It might indeed be useful to express interventional and counterfactual quantities as functions of the joint distribution of observed quantities [98]. It is then necessary to first assess whether these distributions can be uniquely identified in terms of observational distributions using the information included in the causal graphs.
Identifiability has been defined in [69] as the following:

Definition 3.3.0.1. *Given a causal model $\mathcal{M}$ and any computable quantity $Q(\mathcal{M})$, $Q(\mathcal{M})$ is identifiable if it can be estimated consistently using statistical data without necessarily specifying the structural equations of $\mathcal{M}$.*

For example, in this context $Q(\mathcal{M})$ could be the interventional distribution $P(Y(a))$. Hence, to prove non-identifiability of $Q(\mathcal{M})$ it is sufficient to find two sets of structural equations for the model having different values of Q but resulting in identical distributions of the observed variables of the model $\mathcal{M}$.

As it was mentioned in the previous section, when considering the necessary conditions for the identifiability of interventional distributions, two different approaches can be taken. On one side, assumptions based on causal graphs can be used [66]; this approach has the advantage of providing clear graphical conditions for identifiability. Differently, the second approach consists in using assumptions on potential outcomes. Compared to the previous approach, the latter one replaces graphical conditions with conceptually meaningful assumptions, allowing for a more interpretable approach when estimating the interventional quantities from observed data [82]. In this work, in order to assess the identifiability of causal quantities, both assumptions on causal graphs and on potential outcomes will be considered. A combination of these could indeed potentially help the reader in better understanding the multiple identifiability criterion.

3.3.1. Identification of Interventional Distributions

The goal of this section is to find the identification formula for the interventional distribution $P(Y(a))$ in the different possible causal structures. It will be shown that for Markovian models interventional distributions are always identifiable, while this task is more complex for semi-Markovian models.

In Markovian models it is always possible to identify interventional distributions. The generalized identification formula for different causal structures and sets of variables is the Truncated Factorization Formula [69] [49] [64], otherwise called "Manipulation Theorem" [93] or *g-formula* [77]. According to this mathematical formulation, the interventional distribution $P(V(a))$, where V is the entire set of observed variables, is identified by the following Theorem:

Theorem 3.3.1. ***Truncated Factorization Theorem*** *For any Markovian model, the distribution generated by an intervention $A = a$ on a set V of endogenous variables, such that $A \in V$, is given by the truncated factorization:*

$$P(V(a) = v) = \begin{cases} \prod_{V_i \in V \setminus A} P(V_i = v_i | Pa_i = pa_i) & \textit{if } a \textit{ consistent with } v, \\ 0 & \textit{otherwise.} \end{cases} \tag{3.6}$$

Here $P(V_i = v_i | Pa_i = pa_i)$ are the pre-intervention probabilities conditional on the parents of V_i.

The proof of the g-formula is provided in the Appendix A. The main idea behind the Truncated Factorization formula is the following: the intervened model, so the post-intervention graph $\mathcal{G}_{\bar{a}}$, is a Markovian model itself (Proposition 3.2.3). Thus, the Causal Markovian Condition can be applied to the new graph. Because of the intervened variables, in the proof in the Appendix A it is indeed shown that Equation 3.3 yields to the Equation 3.6.

An example on Figure 3.2 will now be considered. It could be interesting to quantify the interventional probability distribution $P(V(a) = v)$ corresponding to the causal graph in Figure 3.2; in this example we have $V = \{C, Y, M, A\}$ with respective instances $v = \{c, y, m, a\}$. Applying the g-formula 3.6 to this example, the following equation holds:

$$\begin{aligned} P(V(a) = v) &= P(C(a) = c, M(a) = m, Y(a) = y) \\ &= P(Y = y | A = a, C = c, M = m) P(M = m | A = a) P(C = c). \end{aligned}$$

In the g-formula of Equation 3.6, the interventional quantity involves the entire set of observable variables V; nevertheless, the interventional distribution on a selected set of variables could be considered, for instance $P(Y(a) = y)$. Hence, getting back to the original identification problem, the Truncated Factorization formula 3.6 can be used in the following way by applying marginalization [84]:

$$P(Y(a) = y) = \sum_{V \setminus \{A,Y\}} \prod_{V_i \in V \setminus \{A\}} P(V_i = v_i | Pa_i = pa_i), \tag{3.7}$$

where, as before, V is the set of observable variables. Since the intervention is not estimated on the entire set of variables V, the interventional quantity is averaged based on the conditioned probabilities of the variables that are not in Y and A. Specifically, the summation is executed over all the values in the state space of the variables in $V \setminus \{A, Y\}$. In the previous examples, this corresponds to:

$$\begin{aligned} P(Y(a) = y) &= \sum_{V \setminus \{A,Y\}} P(Y = y | A = a, C = c, M = m) P(M = m | A = a) P(C = c) \\ &= \sum_{m \in \mathcal{X}_M, c \in \mathcal{X}_C} P(Y = y | A = a, C = c, M = m) P(M = m | A = a) P(C = c), \end{aligned}$$

where $\mathcal{X}_M$ and $\mathcal{X}_C$ are the state spaces of M and C respectively.

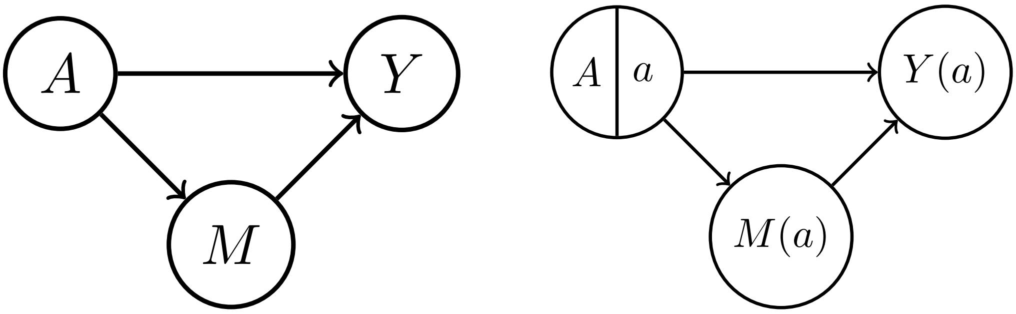
In the next paragraphs, the link between the Truncated Factorization formula (Equation 3.6) and the d-separation applied to potential outcomes will be analysed. Specifically, this content will show the usefulness of SWIGs in order to analyze independence relationships among random variables and potential outcomes. It is important to comprehend these simplified examples in order to understand the formulation of the edge g-formula for complex nested interventional quantities (it will be introduced in Lemma 4.2.1).

The most simplified graphical structure for interventional quantities will now be considered. This involves no observed or unobserved common causes between the treatment variable A and the outcome

Y, for instance the DAG in Figure 3.5a. This condition on graphical structure is equivalent to the *weak ignorability condition* [82]: $Y(a) \perp\!\!\!\perp A$. This means that, after the intervention on the variable A as a is made, any information on the random variable A does not provide information on the outcome Y. This assumption can be immediately noticed analysing the respective SWIG $\mathcal{G}(a)$ in Figure 3.5b corresponding to the intervention on A as a in the causal structure of Figure 3.5a. More formally, because of the d-separation between the nodes A and $Y(a)$ due to the absence of edges connecting A to $Y(a)$ in Figure 3.5b, the weak ignorability assumption can be obtained considering the set S of the d-separation criterion (Criterion 3.2.0.4) as $S = \{\emptyset\}$. As a reminder, Property 3.2.4 of SWIGs justifies the application of the d-separation criterion to these graphical tools.
In this scenario, since $Y(a) \perp\!\!\!\perp A$, we have:

$$P(Y(a) = y) = P(Y = y|A = a), \tag{3.8}$$

meaning that the interventional distribution is equivalent to the conditional probability, regardless of any mediator variables in the causal model.



(a) Causal Graph

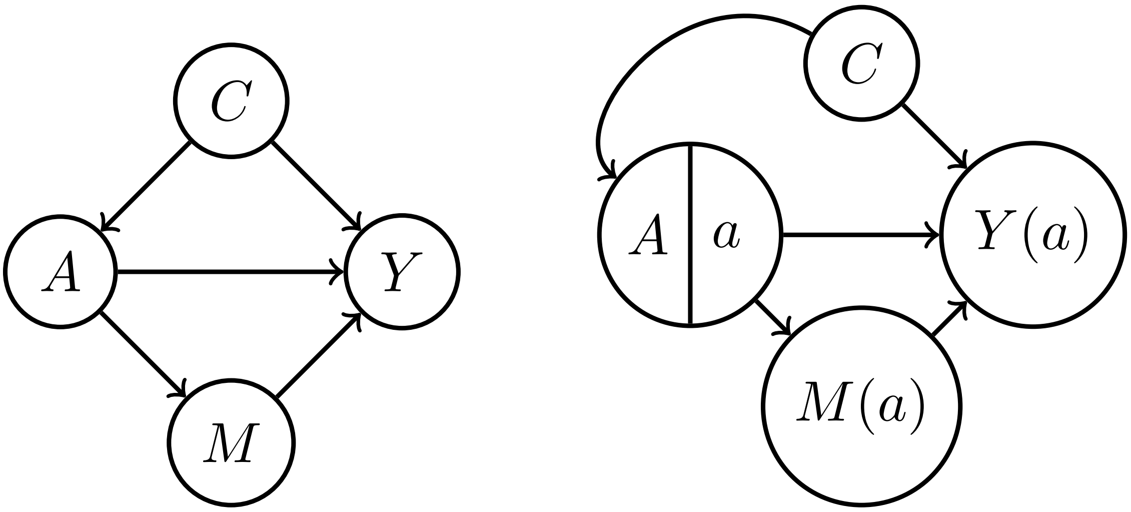
(b) Respective SWIG for Intervention on A as a

Figure 3.5: Example causal graph with its respective SWIG for intervention on A as $a \in \mathcal{X}_A$

Further generalizing the weak ignorability condition, when A and Y have measurable common causes, for example in Figure 3.6a, the respective *conditional ignorability assumption* holds. For instance, if node C is a common cause between A and Y, the conditional ignorability assumption consists in $Y(a) \perp\!\!\!\perp A|C$ [82]. In other words, this assumption means that, when intervening on A, the outcome Y is independent of the information regarding the variable A conditioned on the common cause C. Similarly to before, this assumption can be easily deduced from the SWIG in Figure 3.6b; this follows from the d-separation of A and $Y(a)$ given the variable C.
In this scenario, the interventional distribution $P(Y(a) = y)$ is identifiable under the conditional ignorability assumption $Y(a) \perp\!\!\!\perp A|C$ and is decomposed as:

$$\begin{aligned} P(Y(a) = y) &= \sum_{c \in \mathcal{X}_C} P(Y(a) = y|C = c)P(C = c) & (3.9) \\ &= \sum_{c \in \mathcal{X}_C} P(Y(a) = y|A = a, C = c)P(C = c) & (3.10) \\ &= \sum_{c \in \mathcal{X}_C} P(Y = y|A = a, C = c)P(C = c). & (3.11) \end{aligned}$$

In the above equations, the first one is implied by the Law of Total probability, while the second and the third ones by conditional ignorability ($Y(a) \perp\!\!\!\perp A|C$) and the consistency assumption (Property 3.2.2) respectively. As it can be observed from the final equation, the interventional distribution is computed as the association between the variables A and Y for each value of the confounder $c \in \mathcal{X}_C$, averaging over all the values of C. This is equivalent to the g-formula 3.6 applied to Figure 3.6a.
The above formulation of the interventional distribution is also referred to as the *adjustment formula* or *back-door formula*, which later will be introduced in Definition 3.12. The intuition behind Formula 3.11 consists in controlling or "adjusting" for the possible values the common cause C could assume.



(a) Causal Graph

(b) Respective SWIG for Intervention on A as a

Figure 3.6: Example causal graph Figure 3.2 with its respective SWIG for Intervention on A as a

The above independence relations can also be interpreted graphically by analysing the original causal graph. Regarding the first setting, where no unmeasured variables and no common causes between A and Y are present, the graphical interpretation consists in the absence of any back-door paths from A to Y. By definition, the back-door paths from A to Y are all the paths in the causal diagram that start with an arrow pointing into A [65] (refer to Criterion 3.3.1.2). Differently, analysing the setting in which there is a common cause C between A and Y and all the variables are observable, the assumption $Y(a) \perp\!\!\!\perp A|C$ graphically corresponds to the variable C blocking all back-door paths from A to Y.

In the previous sections, the g-formula was presented as a useful tool in order to identify interventional quantities when all of the variables are measured, hence when there is no need to introduce a latent space (Markovian models). Nevertheless, in general causal structures the g-formula is not always applicable. In Equation 3.11 it was shown that under certain assumptions it is possible to define a set of variables (in the example C) such that it acts as an adjustment term for the identification formula. Covariate adjustment is an useful identification formula; this is equivalent to the g-formula in Markovian models. The adjustment formula can be defined as:

Definition 3.3.1.1. ***Adjustment Formula*** *[79] Given a causal graph $\mathcal{G}$ over a set of variables V, a set Z is called adjustment set for estimating the causal effect of A on Y, if, for any distribution $P(V)$ compatible with $\mathcal{G}$, it holds that*

$$P(Y(a) = y) = \sum_{z \in \mathcal{X}_Z} P(Y = y|A = a, Z = z)P(Z = z), \tag{3.12}$$

where $Y, A, Z \in V$, $a \in \mathcal{X}_A, y \in \mathcal{X}_Y$.

The simplest adjustment set is the set of parent nodes of A, Pa_A. Indeed, the conditional ignorability assumption ($Y(a) \perp\!\!\!\perp A|Pa_A$) always holds since the set Pa_A by definition d-separates $Y(a)$ and A in the respective SWIG $\mathcal{G}(a)$. Nevertheless, in some scenarios it might happen that the parents of A are inaccessible for measurement (i.e. unobserved), leading the set Pa_A not to be a suitable adjustment set for identifying $Y(a)$. In these scenarios the Back-Door Criterion can be useful for defining a suitable adjustment set.

The Back-Door criterion can be used for deriving a broad class of adjustment sets. Presenting the definition provided in [69], we have:

Definition 3.3.1.2. ***Back-Door Criterion*** *A set of variables Z satisfies the back-door criterion relative to a pair of variables (A, Y) in a* DAG *$\mathcal{G}$ if:*

1. *no node in Z is a descendant of A; and*
2. *Z blocks every path between A and Y that contains an arrow into A (i.e. back-door paths)*

The name *Back-Door criterion* comes from the terminology of back-door paths. A back-door path from V_j to V_i in a DAG $\mathcal{G}$ is a path with no emitted edge from V_j $(V_j \leftarrow ...V_i)$. The Back-Door criterion is highly relevant, as if a set of variables Z satisfies it relatively to the couple (A, Y), then the interventional distribution $P(Y(a) = y)$ is identifiable.
As an example, the causal graph in Figure 3.7 will be used for analysing the backdoor criterion relative to (A, Y). In the following graphs U is represented as dashed-edges emitting variable but it could also be graphically drawn as one dashed bi-directional edge between A and M. It can be noticed that the variable M guarantees that the back-door path $A \leftarrow U \rightarrow M \rightarrow Y$ is blocked by M. Nevertheless, M is a descendant of A, meaning that the first condition of the criterion is not satisfied.

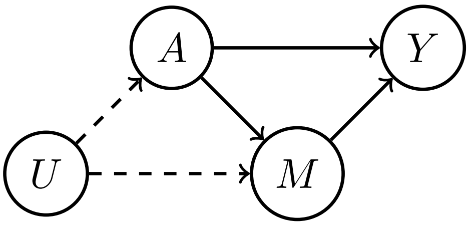


Figure 3.7: Example causal graph where back-door criterion is not satisfied

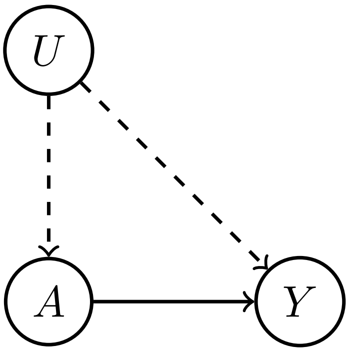


Figure 3.8: Example causal graph with unobserved common cause between A and Y

Like in the previous example, it can be proven that if a set Z satisfied the Back-Door criterion conditions, then the formula 3.11 can be used. More formally, the following theorem from [66] can be analysed:

Theorem 3.3.2. ***Back-Door Adjustment*** *If a set of variables Z satisfies the back-door criterion relative to (A, Y), then the causal effect of A on Y is identifiable and is given by the Adjustment Formula in Equation 3.12.*

Albeit the assumptions needed for the g-formula might be violated, in many scenarios it is possible to select admissible adjustment sets for computing interventional distributions from observed data thanks to the Back-Door criterion. Now the proof of Theorem 3.3.2, following the outline of [66] is given.

Proof. When the Back-Door criterion is satisfied on an admissible set Z, then it is possible to identify the interventional quantity as:

$$P(Y(a) = y) = \sum_{z \in \mathcal{X}_Z} P(Y(a) = y|Z = z)P(Z(a) = z) \tag{3.13}$$

$$= \sum_{z \in \mathcal{X}_Z} P(Y(a) = y|Z = z)P(Z = z) \tag{3.14}$$

$$= \sum_{z \in \mathcal{X}_Z} P(Y = y|A = a, Z = z)P(Z = z). \tag{3.15}$$

The first equality follows from the rule of Total Probability. The second one follows from the first condition of the Back-Door criterion; since no node in Z is a descendant of A, then $Z(a) = Z$. The third equality follows from the third condition of the Back-Door criterion; the back-door paths from A to Y are blocked by Z, so the interventional distribution of Y with respect to A is equivalent to the conditional one when conditioning on the value of Z (blocking the causal path). By expressing this condition with conditional independence relations, $Y(a) \perp\!\!\!\perp A|Z$. The Adjustment Formula follows.

□

The intuition of the Back-Door Criterion follows from the blocking Definition (Definition 3.2.0.3). Looking back at the Back-Door conditions, the selected set Z should not be descendant of A, as otherwise it would be influenced by the intervention on the variable A, hence leading the intervention not to propagate to Y along all the paths from A to Y. Furthermore, in order to ensure that the causal relation between A and Y is fully represented, it is necessary to block the paths that might be source of confounding. In fact, the paths with an arrow entering in A will not propagate the causal effect of A on Y through an intervention, but they might still make these two variables dependent. By conditioning on the set of variables Z that satisfies the back-door criterion, these spurious paths between A and Y actually get blocked.
In many settings, it might be useful to estimate an interventional quantity for a selected set of data values, specifically, by setting a condition on one or more variables that have been observed. For instance, it could be convenient to estimate the Z-specific causal effect, thus estimating the causal effect that A has on Y for a specific value of Z (such as z). These scenarios will be explored in the next sections.

In certain graphical structures, the Back-Door criterion is not applicable. In the presence of unobserved variables, the front-door criterion can also be useful for the identification of interventional quantities. This can be applied to causal graphical structures that do not satisfy the back-door criterion. Indeed, frequently, it might happen that the back-door criterion cannot be used if a set of unobserved variables has as descendants both the sensitive attribute A and the outcome Y, like in Figure 3.8. In these scenarios, the back-door criterion is not applicable because there is a back-door path mediated only by U, which is unobserved. Intuitively, the identifiability problem rises form the impossibility to distinguish between the spurious correlation between A and Y due to the presence of U and the true causal effect that A has on Y. Nevertheless, this would change if there was a mediator between A and Y, such as in Figure 3.9. The intuitive reasoning behind the use of mediator for identifying the causal effect in these scenarios will be explained in the next paragraph.

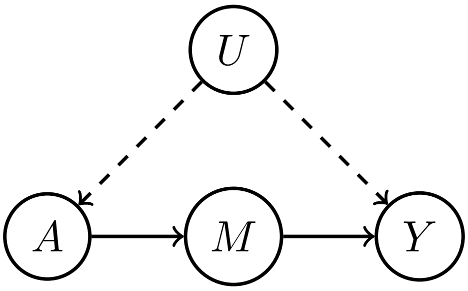


Figure 3.9: Example causal graph front-door criterion

Comparing the front-door criterion and the back-door criterion, the following consideration can be made. While the first condition of the back-door criterion reflects the observations being unaffected from the treatment, the front-door criterion can be useful for analysing causal models where the set Z is affected by the treatment.

Definition 3.3.2.1. ***Front-Door Criterion*** *A set of variables Z satisfies the front-door criterion relative to a pair of variables (A, Y) in a* DAG *$\mathcal{G}$ if:*

1. *Z intercepts all directed paths from A to Y;*
2. *there is no unblocked path from A to Z ; and*
3. *all back-door paths from Z to Y are blocked by A.*

Similarly to before, finding a set of variables Z such that the front-door criterion is satisfied corresponds to confirming the identifiability of the interventional distribution.

Definition 3.3.2.2. ***Front-Door Adjustment*** *If a set of variables Z satisfies the front-door criterion relative to (A, Y), then the causal effect of A on Y is identifiable and is given by the Adjustment Formula:*

$$P(Y(a) = y) = \sum_{z \in \mathcal{X}_Z} P(Z = z | A = a) \sum_{a_0 \in \mathcal{X}_A} P(A = a_0) P(Y = y | Z = z, A = a_0), \tag{3.16}$$

where $a \in \mathcal{X}_A, y \in \mathcal{X}_Y$.

The intuition of this formula consists in applying multiple times the back-door formula to the same causal structure. For instance, referring to Figure 3.9, the interventional quantity in Equation 3.16 can be seen as:

$$P(Y(a) = y) = \sum_{m \in \mathcal{X}_M} P(Y(a) = y | M(a) = m) P(M(a) = m). \tag{3.17}$$

Indeed, given the value a that A is being intervened as, the causal effect it has on Y can be estimated as an interventional distribution through the different values in the domain of M, averaged out on the probability distribution respective the values that M would naturally assume after an intervention on A as a. The two interventional quantities that are used in the right side of Equation 3.17 are identifiable from observable quantities by using the back-door criterion.

In order to prove the identification formula the following steps can be taken:

$$\begin{aligned} P(Y(a) = y) &= \sum_{m \in \mathcal{X}_M} P(Y(a) = y | M(a) = m) P(M(a) = m) \\ &= \sum_{m \in \mathcal{X}_M} P(Y(m) = y) P(M(a) = m | A = a) \\ &= \sum_{m \in \mathcal{X}_M} P(Y(m) = y) P(M = m | A = a) \\ &= \sum_{m \in \mathcal{X}_M, a_0 \in \mathcal{X}_A} P(Y = y | M = m, A = a_0) P(A = a_0) P(M = m | A = a). \end{aligned}$$

Above, the first equality follows from the rule of total probability. The second one follows from the first condition of the front-door criterion, stating that Z intersects all the directed paths from A to Y. The second condition of the front-door criterion, stating that there are no back-door paths from A to M ($M(a) \perp\!\!\!\perp A$) is used in the equality $P(M(a) = m | A = a) = P(M = m | A = a)$. Lastly, the back-door criterion can be used in order to identify $P(Y(m) = y)$, since A blocks all the back-door paths from M to Y; it can be noticed that the condition of the back-door criterion stating that A is not a descendant of M is already enforced in the second condition of the front-door criterion.

The presented Back-Door and Front-Door criterion are extremely useful for assessing the identification of simple graphical structures that are often met in the literature. These criterion are critical for understanding the key aspects of identifiability in Semi-Markovian models and the respective challenges through the use of adjustment sets. These criterion will be useful for better understanding future content of this work.

Assessing identifiability of arbitrary Semi-Markovian graphical structures is relevantly more complicated compared to Markovian Models. Further to the Back-Door and Front-Door criterion, a general identification algorithm for interventional distribution in Semi-Markovian models is presented in multiple works. Specifically, the identification algorithm (ID Algorithm), proposed in [96] and simplified in [83], is proposed. These results have been further explored in [85], by reformulating the ID Algorithm into a one-line formula. Most of the respective proofs are presented in the working paper [75]. As this topic is not the focus of this project, it will not be analysed it in detail. However, interesting insights regarding the identifiability criterion that could potentially benefit the intuition behind the next chapters will be introduced.

The identifiability criterion for semi-Markovian models is based upon the notion of districts, introduced in Definition 3.2.0.7. In [83], districts are referred to as C-components. In this section, an identification criterion for causal effects in semi-Markovian models will be presented. Specifically, it will focus on the identification of $P(Y(a) = y)$ with A and Y as disjoint sets of nodes. The main obstacle towards identification of causal effects in semi-Markovian models is the presence of a hedge graphical structure. Indeed, it will be proven that the above causal effect is identifiable if and only if an hedge structure does not exist. In order to introduce the hedge structure, it is necessary to define C-forests.

Definition 3.3.2.3. ***C-forest*** *[83] Let $\mathcal{G}$ be a semi-Markovian graph, where Y is a set of nodes. Then, $\mathcal{G}$ is a Y-rooted C-forest if all nodes in $\mathcal{G}$ form a district, all observable nodes have at most one child and $An_{\mathcal{G}}(Y) = \mathcal{G}$*

Based on this notion, it is possible to define a hedge structure in the following way:

Definition 3.3.2.4. ***Hedge*** *[83] Let X, Y be disjoint sets of variables in $\mathcal{G}$. Let F, F' be R-rooted C-forests such that $F' \cap A = \emptyset$, $F \cap A \neq \emptyset$, $F' \subseteq F$ and $R \in An_{\mathcal{G}_{\bar{a}}}(Y)$. Then, F and F' form a hedge for $P(Y(a) = y)$ in $\mathcal{G}$.*

For better understanding of hedge structures, it is possible to start by considering a C-forest F' that does not contain A. Then, it is possible to extend F' by growing more branches, while retaining the same root-set, becoming F. By adding new branches, A is intersected by F. The intervention on A has the effect of removing some incoming arrows in $F \backslash F'$, leading to the hedge structure of the graph. Hedges generalize the scenario where Y is a child of A and they are connected through a bi-directed edge, leading $P(Y(a) = y)$ not to be identifiable [96]. Referring to the Back-Door and Front-Door criterion, this scenario was indeed not covered by the different conditions. By generalizing these graphical structures considering both A, Y and the respective mediators in the same district, a hedge structure can be defined. Indeed, it can be proven that:

Theorem 3.3.3. ***Hedge criterion*** *[83] $P(Y(a) = y)$ is identifiable in $\mathcal{G}$ if and only if there does not exist a hedge for $P(Y'(a') = y')$ in $\mathcal{G}$, for any $A' \subseteq A$ and $Y' \subseteq Y$, having $a' \in \mathcal{X}_{A'}, y' \in \mathcal{X}_{Y'}$.*

Proof. The proof of the above Theorem will be provided, following structure in [96]. In particular, it will now be proven that if a hedge structure exists for $P(Y'(a') = y')$, then $P(Y(a) = y)$ is not identifiable. The rest of the proof can be found in the original document.
Two R-rooted C-forests F and F' will be considered. First, it can be proven that $P(R(a) = r)$ is not identifiable in $\mathcal{G}$. It is possible to define two models $\mathcal{M}_1$ and $\mathcal{M}_2$ with the same induced graph $\mathcal{G}$ such that the observed distributions are the same $P^1(R) = P^2(R)$ but such that the interventional distributions are not, proving that indeed $P(R(a) = r)$ is not identifiable in $\mathcal{G}$. Once this has been proven, since $P(Y(a) = y) = \sum_r P(R(a) = r)P(Y = y|R = r)$, it is possible to construct $P(Y = y|R = r)$ such that the mapping between $P(Y(a) = y)$ and $P(R(a) = r)$ is one-to-one. Since $P(R(a) = r)$ is not identifiable in $\mathcal{G}$, the same will hold for $P(Y(a) = y)$.
The authors of [96] choose as $\mathcal{M}_1$ and $\mathcal{M}_2$ the models that assign to every variable the bit parity (sum modulo 2) of its parents, meaning that every variable is binary. These two models differ from the fact that in $\mathcal{M}_2$ the nodes in F' which have parents in F ignore the values of those parents. All the unobserved variables U are fair coins in both of the models. It can be noticed that these two models are observationally indistinguishable, as the bit parity of R is even since the unobserved variables are

counted twice in both of the models. Nevertheless, given the hedge formed by F and F', any intervention in $F \backslash F'$ will lead to two distinct distributions on R. The bit-parity circuit will be broken, which will affect R in the first model but not in the second one, as the variables in F' with parents in F ignored those specific terms in the sum.
Hence, it was proven that if there exists an hedge for $P(Y'(a') = y')$, then $P(R(a) = r)$ is not identifiable in $\mathcal{G}$. This leads to $P(Y(a) = y)$ not being identifiable because of the one-to-one mapping. The reader can refer to [96] for the proof of the other statement of the if-and-only-if condition. □

In this section the different guidelines for deriving the identification formula for interventional distributions in different causal structures have been presented. In particular, it has been explained how the presence of unmeasured confounders can lead to further complications in this task. In general, once the identifiability assumptions have been assessed, it is always possible to provide mathematical formulation for interventional quantities using the observed-data distribution.
As previously mentioned, an entire identification algorithm is provided in [96] to be applied to semi-Markovian models. The main idea of the identification algorithm can be found in the Appendix B. This topic will not be explored further in details because it falls outside the scope of the project.
In the next paragraph, the powerful machinery of *do-calculus* is introduced, along with the relations with the earlier introduced adjustment criterion.

Do-Calculus

The *do-calculus* is characterized by three rules; these rules are equalities that can be used to manipulate interventional distributions. The final scope of this calculus is to turn interventional distributions into observational ones. Even though these rules have not been explicitly introduced in this report yet, many of the stated identification results could have been addressed by using do-calculus, for instance the d-separation, the Back-Door or the Front-Door criterion.
In the below formulation, a causal graph $\mathcal{G}$ containing the disjoint sets of nodes $A, Y, Z, D \in V$ is considered. The do-calculus rules are the following:

- **Rule 1**:
  $$P(Y(a) = y|Z = z, D = d) = P(Y(a) = y|D = d) \text{ if } (Y \perp\!\!\!\perp Z|A, D)_{\mathcal{G}_{\overline{A}}}, \tag{3.18}$$
  where $\mathcal{G}_{\overline{A}}$ is defined as the causal graph $\mathcal{G}$ in which all arrows leading to A have been deleted. This rule is also called *ignoring observations rule*, as it states the equivalence of the interventional distribution $P(Y(a) = y|D = d)$ with or without observing the variable Z under specific graphical conditions.
  It was previously explained that node interventions on A can be graphically represented by cutting the edges entering in the treatment variable A. As a matter of fact, given the causal graph $\mathcal{G}_{\overline{A}}$ where A is set to a constant value a, the probability distributions of the different variables along this causal graph are equivalent to the interventional distributions of the respective variables with A set as a. Hence, intuitively, if $(Y \perp\!\!\!\perp Z|A, D)_{\mathcal{G}_{\overline{A}}}$ holds, then the variable Z is independent of Y when intervening on the variable A and conditioning on D.

- **Rule 2**:
  $$P(Y(a, z) = y|D = d) = P(Y(a) = y|D = d, Z = z) \text{ if } (Y \perp\!\!\!\perp Z|A, D)_{\mathcal{G}_{\overline{A}\underline{Z}}}, \tag{3.19}$$
  where $\mathcal{G}_{\overline{A}\underline{Z}}$ is the graph in which the arrows entering in A and the arrows emitted from Z are deleted. The assumption $(Y \perp\!\!\!\perp Z|A, D)_{\mathcal{G}_{\overline{A}\underline{Z}}}$ means that in the graph $\mathcal{G}_{\overline{A}\underline{Z}}$ the variables $\{A \cup D\}$ block all the back-door paths from Z to Y. Indeed, when deleting the arrows emitted by Z, the only paths remaining in the graph that connect Z and Y are the back-door paths (ending with an entering edge in Z). Hence, since the condition on $\mathcal{G}_{\overline{A}\underline{Z}}$ guarantees that $\{A \cup D\}$ satisfy the Back-Door criterion (refer to Definition 3.3.1.2) in the interventional graph $\mathcal{G}_{\overline{A}}$ (corresponding to the intervention on A). Then, when conditioning on D, an intervention on Z by setting $Z = z$ has the same effect on Y as the conditioning on $Z = z$. Indeed, when all back-door paths are blocked, then the spurious correlations between Z and Y are blocked, leaving the directed paths from Z to Y unperturbed. It is important to notice that in the Back-Door criterion an additional

assumption is required: $\{A \cup D\}$ are not descendants of Z. Nevertheless, when intervening on A and conditioning on D, even if they originally were descendants of Z, this requirement is met. In conclusion, the operation of the intervention is the same as the one of conditioning.

- **Rule 3**:
$$P(Y(z,a) = y|D = d) = P(Y(z) = y|D = d) \text{ if } (Y \perp\!\!\!\perp A|Z, D)_{\mathcal{G}_{\overline{ZA^*}}}. \tag{3.20}$$
Here, we have $A^* = A \backslash An_{\mathcal{G}_{\overline{Z}}}(D)$, which consists in deleting the arrows entering A if A does not precede D. This rule means that it is possible to ignore an intervention on a variable A if this operation does not influence Y through any uncontrolled path. Hence, it is possible to remove the intervention on A if there are no unblocked paths from A to Y.
The most simplified scenario of the Rule 3 of do-calculus consists in transforming an interventional quantity into an observational one. Specifically, by choosing $Z = \{\emptyset\}$ and $D = \{\emptyset\}$, the condition becomes $(Y \perp\!\!\!\perp A)_{\mathcal{G}_{\overline{A}}}$. Hence, we have that, if there are no causal paths between A and Y, $P(Y(a)) = P(Y)$.

All of the adjustment formulas that were formulated in the previous section can be proven and derived using the do-calculus rules. Even though these rules partly fall outside the scope of this work, Rule 1 (Eq. 3.18) and Rule 2 (Eq. 3.19) will be useful in Chapter 6.

3.3.2. Identification of Unit-Level Counterfactuals

Unit-level counterfactuals of a variable $A = a$ on Y were introduced in Definition 3.2.4.1, denoting with $Y_a(u)$ "the value of Y in unit u, had A been a". Based on the definition of SCMs (Definition 3.2.0.1) and of the consistency property (Property 3.2.2), since the structural equations remain invariant among the entire population, the instances of background variables characterize similar individuals, leading to specific values of the observed variables V and of the respective responses to interventions. Given an arbitrary counterfactual, such as $\mathbb{E}[Y_{A=a}|X = x]$ with X being a set of observed covariates, in [68] a three steps process is proposed in order to identify this unit-level intervention.

- **Abduction**: for a given prior U, compute the posterior distribution of U given the evidence $X = x$, having $P(U = u|X = x)$ for any arbitrary $u \in \mathcal{X}_U$;
- **Action**: modify the model by removing the structural equations for the variables in A and replacing them with the appropriate functions $A = a$ with $a \in \mathcal{X}_A$, as previously seen in the definition of interventions;
- **Prediction**: use the modified structural equations in the model and the posterior $P(U|X = x)$ to compute the expectation of Y.

Even though this procedure can be straightforward in datasets with a known generating process, it might be hard in practice with real-world datasets. Indeed, it is necessary to make use of the structural equations, which are usually not known in advance. Otherwise, assumptions on the form of these equations need to be made, specifically for estimating the posterior distribution $P(U|X = x)$. A further assumption in this approach is that the values of the unobserved variables are the same for observed data points and their respective unit-level counterfactuals. It is important to notice that the unit-level counterfactual quantities are identifiable if the involved interventional quantities are identifiable.

As an example of the identification of unit-level counterfactuals, the following Structural Causal Model is given:

$$\begin{aligned} A &= f_A(U_A) \\ M &= f_M(A, U_M) \\ Y &= f_Y(M, A, U_Y). \end{aligned}$$

In order to analyse the counterfactual for a unit with observed values (a_i, m_i, y_i) by intervening on $A = a$, then the following steps need to be taken. Given the functions (f_A, f_M, f_Y), using the value (a_i, m_i, y_i)

it is possible to deterministically derive the values (u_A^i, u_M^i, u_Y^i). Next, the unit-level counterfactuals can be determined by solving the following system:

$$\begin{aligned} A &= a \\ m &= f_M(a, u_M^i) \\ y &= f_Y(m, a, u_Y^i). \end{aligned}$$

Differently, when the structural equations are not known, a deterministic approach is not possible. Nevertheless, following the variational Abduction-Action-Prediction approach, it is possible to use the inferred posterior distribution of the latent space, in this example $P(U|A, M, Y)$, in order to derive the unit-level counterfactuals.

4

Causal Fairness Metrics

In the previous chapters, an overview on fairness and causality was given, motivating the focus on causality-based fairness metrics. This chapter will introduce and compare the most commonly used causality-based fairness metrics. In conclusion, path-specific metrics will be chosen as the most suitable metrics for the research scope of this project.

Analysing causal fairness metrics, discrimination can be defined based on two different frameworks: disparate impact and disparate treatment [9]. The former corresponds to decisions that do not ensure a fair impact on all the classes (defined by the values of A). Meanwhile, the latter reflects decisions that enact a disparate treatment for people who have similar attributes.
It will be shown that the disparate impact framework includes different fairness measures, such as the Total Effect and the Effect of Treatment on the Treated. Differently, fairness notions corresponding to mediation analysis like Direct Effect, Indirect Effect and Path-Specific Effect belong to the disparate treatment framework.
An additional classification on fairness metrics will be made into individual-level and population-level metrics. The advantages and disadvantages of these two frameworks will be analysed.

By using a consistent notation with respect to the rest of the work, in this section the treatment variable (sensitive attribute) that is being intervened upon is A. The state space of this variable is assumed to be $\mathcal{X}_A = \{a_0, a_1\}$. Specifically, $A = a_1$ characterizes the possibly discriminated group of people. The scope of the research on fairness is to evaluate the causal relations between the treatment variable A and the outcome Y.

4.1. Total Effect and Effect of Treatment on the Treated

In this section, two of the most common causal fairness metrics belonging to the disparate impact framework will be presented: the Total Effect and the Effect of Treatment on the Treated. Both of these metrics are commonly used in the research on fairness; nevertheless, the drawbacks of these approaches will be shown, specifically with respect to the purpose of this project.

The *Total Effect* has the purpose of estimating the total causal effect that the treatment variable has on the outcome at a population level [20]. This can be done by comparing the interventional distributions of $Y(a_0)$ and $Y(a_1)$, where $a_0, a_1 \in \mathcal{X}_A$. This effect measures the difference in the distribution of Y once two distinct interventions on A are made (setting the variable to two different values, for example $a_0 = 0$ and $a_1 = 1$). In particular, the literature focuses on evaluating the average Total Effect (also called Average Treatment Effect):

$$TE_{a_0,a_1}(Y) = \mathbb{E}\big[Y(a_0)\big] - \mathbb{E}\big[Y(a_1)\big].$$

The Total Effect is formulated as the difference between the mean outcome between the inactive and active treatment scenarios, in this case the scenarios in which A is set to a_0 and the one where A is set to a_1.

A metric that is closely related to the Total Effect is the *Effect of Treatment on the Treated* [66]. This is a counterfactual metric, measuring the mean response of the treatment variables on the active treatment subjects (for instance with $A = a_1$) [37]. The Effect of Treatment on the Treated is defined as:

$$ETT_{a_0,a_1}(Y) = \mathbb{E}\big[Y(a_0)|A = a_1\big] - \mathbb{E}\big[Y(a_1)|A = a_1\big].$$

The probability distribution of $P(Y(a_1) = y|A = a_1)$ is equivalent to the conditional distribution $P(Y = y|A = a_1)$ by consistency (refer to Property 3.2.2), as it corresponds to intervening on variable A setting it to a_1 in data-points that already had the value of A equal to a_1. Instead, $P(Y(a_0) = y|A = a_1)$ is a counterfactual quantity that can be interpreted as the "the probability that Y would be y had A been set to a_0, given that in the real world $A = a_1$" [108]. The Effect of Treatment on the Treated is commonly used in epidemiology [57] and in social experiments, to analyse their impact on participants [37].

Both the Total Effect and the Effect of Treatment on the Treated are commonly used as fairness metrics in the literature. Applied to the setting of research on fairness of an outcome Y with respect to the sensitive attribute A, the Total Effect evaluates the overall effect that the sensitive attribute has on the outcome. Moreover, it is a common practice in the state of the art to refer the subjects that have not been treated as the ones belonging to the potentially discriminated group. This means that the Effect of Treatment on the Treated can be interpreted as the total effect solely estimate among the potentially discriminated demographic group defined by A.

The Total Effect and the Effect of Treatment on the Treated present different drawbacks as fairness metrics. These metrics focus on evaluating the effect that the sensitive attribute has on the outcome along all the causal paths in the graphical structures. Indeed, when the distinction of the effect among the different pathways from A to Y is not relevant for the scope of the research, these metrics are suitable. Nevertheless, in the fairness setting, not only we are interested in analysing the causal effect of the sensitive attribute on the output of the ML algorithm, but it is also critical to evaluate in which way this effect is propagated. What was now referred to as "ways" of propagation can be graphically interpreted with the different paths from A to Y. The main weakness of these two fairness metrics is thus not capturing the Path-Specific Effects that the sensitive attribute has on the outcome. This concept is critical in the fairness research, as not all the causal paths from the sensitive attribute should always be considered unfair. For these reasons, metrics focusing on the total effect are not considered suitable for this research; a generalization of these metrics to Path-Specific Effects is preferred, allowing for a more flexible interpretation of fairness on causal graphs. This concept will be explored in the next section.

4.2. Path-Specific Effect

In this section, first an introduction on the motivation of Path-Specific Effect (Path-Specific Effect (PSE)) as a fairness metric will be given. Moreover, specific notations regarding edge interventions and the respective identifiability conditions will be explored.

Since the research on causal fairness focuses on evaluating the causal effect that a sensitive variable A has on an outcome Y, mediation analysis [82] [67] is a common approach. The scope of this analysis is to decompose the causal effect of A on Y into the different causal pathways from A to Y in the causal graph. For instance, it could be interesting to analyse the *Direct Effect* that A has on Y, which corresponds to estimating the causal effect solely along the direct edge $(AY)_{\rightarrow}$ in the causal graph. Similarly, the same intuition can be applied to the *Indirect Effect* (along all the indirect paths) and in general to the *Path-Specific Effect* (i.e. effects corresponding to specific selected pathways from A to Y). In fact, in the fairness setting it might happen that only part of the pathways from the sensitive

attribute to the outcome are considered illegitimate.
Generally, in our work and in most of the causal fairness works among the state of the art literature, it will be assumed that the ethical interpretation of the causal pathways in the graphical structure of the data is known. In order to assess which causal pathways can be considered allowed (fair) or not allowed (unfair), researchers often rely on domain experts knowledge.

In order to comprehend the intuition of the Path-Specific Effect as a fairness metric, the following introductory example can be analysed. In Figure 4.1, the causal graph of a construction job hiring procedure model is presented. Different variables can influence the hiring decision (Y) in the model: the name of the applicant (N), the gender of the applicant (A) and his/her physical strength (S), which is a valuable skill for a construction job. By domain expert knowledge, it can be supposed that the direct causal path from the gender variable to the hiring decision can is unfair, as it should not be allowed for the gender to directly causally influence the outcome of the application decision. In order to assess the causal effect that the gender has along this direct path $(AY)_{\rightarrow}$, the Natural Direct Effect can be used. By analysing the other variables in the causal graph too, it can be noticed that the variable N, representing the names of the applicants, can be considered as a *proxy* for the gender variable. By proxy it is meant that relevant information regarding the gender of the applicant can be deducted by his/her name itself. Indeed, from an ethical perspective, the causal path $(ANY)_{\rightarrow}$ can be assumed unfair too. Hence, two distinct paths of the causal graph from the sensitive attribute to the outcome are considered to be unfair in this example. In order to assess the causal effect that A has on Y solely along the unfair paths $\pi : \{(AY)_{\rightarrow}, (ANY)_{\rightarrow}\}$, the Path-Specific Effect can be used. The reason why the path $(ASY)_{\rightarrow}$ is considered a fair path is that, generally, it should be allowed to consider the physical strength of the candidate in the hiring process, as it is a necessary skill for a construction worker.
The above example treated an ideal world, where it is immediate to understand when a causal path can be considered fair or unfair. In general, this procedure is more complicated and consultation with domain expertise is necessary.

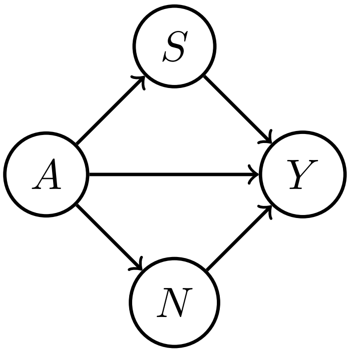


Figure 4.1: Example hiring process

Now that motivation behind path-specific metrics has been introduced, the Path-Specific Effect can be defined as the following:

Definition 4.2.0.1. ***Path-Specific Effect*** *Given a set of paths π part of a causal graph $\mathcal{G}$, two variables A and Y belonging to $\mathcal{G}$ such that $a_0, a_1 \in \mathcal{X}_A$, the Path-Specific Effect of A on Y along π is formulated as:*

$$PSE^{\pi}_{a_0,a_1}(Y) = \mathbb{E}\big[Y(\pi, a_1, a_0)\big] - \mathbb{E}\big[Y(a_0)\big]. \tag{4.1}$$

Here, the probability distribution $P(Y(\pi, a_0, a_1) = y)$ is the probability that $Y = y$ if along the path π, A is intervened upon as a_1, meanwhile, along the other paths not belonging to the set π, A is intervened upon as a_0. The Path-Specific Effect $PSE^{\pi}_{a_0,a_1}(Y)$ has the scope of evaluating the effect that A has on Y along π. By decomposing the Average Total Effect into the Path-Specific Effects [8] of the sensitive attribute on the outcome, an effective and more interpretable fairness assessment can be given. The sum of Path-Specific Effects along all the paths from A to Y is thus equivalent to the Average Total Effect. It can be noticed that this metric is asymmetric with respect to the values $a_0, a_1 \in \mathcal{X}_A$; the formulation of $PSE^{\pi}_{a_0,a_1}(Y)$ follows from the ethical context and common practices in the fairness literature. In this work it is always assumed that $A = a_1$ is the possibly discriminated group.

The Path-Specific Effect is a highly versatile metric. As a matter of fact, the Path-Specific Effect is a general version of the Natural Direct Effect (NDE) and the Natural Indirect Effect (NIE), which instead consist in selecting π as the direct edge and the set of indirect paths respectively.
Until now, this work has focused on node interventions (i.e. interventions corresponding to variables, not edges). Nevertheless, when using mediation analysis, it is necessary to introduce and analyse the concept of *edge interventions*. It is interesting to notice that node interventions are a specific subgroups of edge interventions, in which all the edges emitted by the treatment variable are intervened upon with the same value. In fact, the Total Effect is a special case of the Path-Specific Effect with π is the entire set of paths from A to Y. In the next section, the definition and identifiability conditions of edge interventions will be presented.

4.2.1. Identifiability of Path-Specific Effects

In this Section, first the identification of Path-Specific Effects will be defined with the edge g-formula for arbitrary causal graphs $\mathcal{G}$ and arbitrary edge interventions. This identification formula will further be applied to one of the graphical structures that was analysed in the previous chapters. Moreover, the necessary identifiability assumptions will be given, considering both Markovian models (Definition 3.2.0.2) and semi-Markovian models (Definition 3.2.0.5).

Edge g-formula

In the previous Chapter 3.3.1 on identification of interventional distribution, the focus was set on node interventions. When considering path-specific interventions, it is necessary to introduce the concept of *edge interventions*. Differently from edge interventions, in node interventions the effect of a single intervention is estimated along all the causal pathways emitted from the treatment node itself. Instead, by analysing edge interventions, the causal effect is evaluated along specific edges of the causal graph. In this section, a causal graph $\mathcal{G}$ with vertices $V = \{V_1, ..., V_n\}$ subject to arbitrary edge interventions will be considered. The edge g-formula will be introduced and proven; this formula is used for the identification of Path-Specific Effects and requires specific assumptions that will later be presented.

An overview of the notation used in [86] for defining edge interventions will now be given. The authors of [86] introduced the concept of edge interventions, mapping a set of directed edges in the causal graph $\mathcal{G}$ to values of their source nodes. By notation, given a node V_i, $\mathcal{X}_{V_i}$ defines the state space of the variable V_i, which is the set of all the values that the variable can assume. An edge from node V_i to V_j will be referred to as $(V_iV_j)_{\rightarrow}$. Following the notation used in [86], α is the set of edges in a DAG $\mathcal{G}$ along which the intervention is made. Since each edge is characterized by a source variable, $so(\alpha)$ is defined as the set of source variables of the edges in α. For instance, by intervening on the edge $(AY)_{\rightarrow}$, the value of the source node A is set to an arbitrary value a along that edge. Thus, the state space of the set of edges α is defined as $\mathcal{X}_\alpha = \mathcal{X}_{so(\alpha)}$, which is the Cartesian product of the state spaces of the source variables of all directed edges in α. $\mathcal{X}_\alpha$ corresponds to all the possible interventions that can be done along the selected edges; the interventional values of these edge interventions are denoted as $\mathfrak{a}_\alpha \in \mathcal{X}_\alpha$. If all edges share the same source node, for instance A, an element $\mathfrak{a}_\alpha$ of $\mathcal{X}_\alpha$ could still present different values $a \in A$. As a matter of fact, it is important to take into account that conflicting value assignments of A might be present in $\mathcal{X}_\alpha$, representing conflicting interventions along different edges. This is coherent with how edge interventions were introduced, meaning that the same attribute could be intervened upon as different values along distinct edges.
As an example, the causal graph in Figure 4.2 will be used. The following edge intervention is considered: A is set as a_0 along $(AM)_{\rightarrow}$ and as a_1 along $(AY)_{\rightarrow}$. Given this edge intervention, the respective notation is the following: $\alpha = \{(AM)_{\rightarrow}, (AY)_{\rightarrow}\}$, $\mathcal{X}_A = \{a_0, a_1\}$, $\mathcal{X}_\alpha = \{a_0, a_1\} \times \{a_0, a_1\}$, $\mathfrak{a}_\alpha = [a_1, a_0]$ where $\mathfrak{a}_\alpha \in \mathcal{X}_\alpha$.

Edge interventions are formally defined in [86] in the following way.

Definition 4.2.0.2. ***Response to Edge Interventions*** *For every node variable Y, given a set of edges α in $\mathcal{G}$ with a respective interventional element $\mathfrak{a}_\alpha \in \mathcal{X}_\alpha$, the potential outcome corresponding to this edge intervention $\eta_{\mathfrak{a}_\alpha}$ can be written with the following recursive definition:*

$$Y(\mathfrak{a}_\alpha) = Y\Big(\mathfrak{a}_{\{(*Y)_\rightarrow \in \alpha\}}, \{pa_{\mathcal{G}}^{\hat{\alpha}}(Y)\}(\mathfrak{a}_\alpha)\Big), \tag{4.2}$$

where $pa_{\mathcal{G}}^{\hat{\alpha}}(Y) = \{W \in Pa_Y | (WY)_\rightarrow \notin \alpha\}$.

By notation, $\{pa_{\mathcal{G}}^{\hat{\alpha}}(Y)\}(\mathfrak{a}_\alpha)$ is the interventional distribution of the nodes $pa_{\mathcal{G}}^{\hat{\alpha}}(Y)$ following the interventional element $\mathfrak{a}_\alpha$. $Y(\mathfrak{a}_\alpha)$ is the response of Y to the edge intervention $\eta_{\mathfrak{a}_\alpha}$ that assigns the set of values $\mathfrak{a}_\alpha$ to the respective edges α. The above formulation is defined recursively, based on the lack of directed cycles in DAGs. The edge-specific interventional distribution on the variable (or set of variables) Y can be interpreted as the potential outcome such that the parents that belong to an edge in α are assigned the interventional value in $\mathfrak{a}_\alpha$, while the other parents $pa_{\mathcal{G}}^{\hat{\alpha}}(Y)$ are set the value they would naturally have after the same edge intervention on their respective parents.
Going back to the example in Figure 4.2, the following quantities were defined $\alpha = \{(AM)_\rightarrow, (AY)_\rightarrow\}$, $\mathcal{X}_A = \{a_0, a_1\}$, $\mathcal{X}_\alpha = \{a_0, a_1\} \times \{a_0, a_1\}$, $\mathfrak{a}_\alpha = [a_1, a_0]$. By using the above recursive definition, the potential outcome of the previously introduced edge intervention can be interpreted in the following way: the only parent of Y such that $(*Y)_\rightarrow \in \alpha$ is the source variable A itself, which is set to a_0 along this edge. $M \in pa_{\mathcal{G}}^{\hat{\alpha}}(Y)$ is subject to the intervention on the edge $(AM)_\rightarrow$, along which A is set to a_1. Furthermore, $C \in pa_{\mathcal{G}}^{\hat{\alpha}}(Y)$ is invariant for the intervention since it is not part of the descendants of A. This potential outcome is referred to as $Y(a_0, M(a_1))$.

When considering nested counterfactuals, it is necessary to involve conditional independence relations among interventional quantities characterised by both conflicting and not conflicting interventions. In order to introduce the edge g-formula for arbitrary causal structures, it is important to also consider the assumptions involved defining nested counterfactuals.
The authors of [86] introduced two distinct models, based on the respective assumptions made on the dependence relations of interventional quantities. The former ones can be defined in the following way:

Definition 4.2.0.3. ***Single World Model (SWM) [86]*** *The Single World Model associated with a DAG $\mathcal{G}$ with nodes $V = \{V_1, ..., V_d\}$ is the set of all possible potential outcomes satisfying the assumption that the elements:*

$$\Big\{V_i(v_{pa_{\mathcal{G}}(V_i)}) : V_i \in V\Big\} \tag{4.3}$$

are mutually independent for every $pa_{\mathcal{G}}(V_i) \in \mathcal{X}_{Pa_{V_i}}$.

For instance, for Figure 4.2, $M(a) \perp\!\!\!\perp Y(a, m)$ for the same value of $a \in \mathcal{X}_A$. In [86], by Single World Models the authors refer to the causal model Finest Fully Randomized Causally Interpretable Structured Tree Graphs (FFRCISTG) proposed in [76].
These assumptions can be easily tested by analysing the SWIG corresponding to the specific intervention and causal structure. SWIGs have been introduced in [74] (see Definition 4.2.0.3) as a tool that unifies the graphical and potential outcomes approaches used in causal inference, leading to a graphical interpretation of the independence relations among potential outcomes. In Figure 4.3 the SWIG corresponding to the interventional quantities $M(a)$ and $Y(a, m)$ of Figure 4.2 is presented. As it can be noticed from this example, in order to represent interventional quantities, a “node” operation is made. The node A is split in a random half and a fixed half, where the fixed half inherits the outgoing edges, while the random part the incoming edges. The same holds for the variable $M(a)$, as the interventional quantity $Y(a, m) = Y(a, M(a) = m)$ is the response to two interventions, one on $M(a)$ and one on A.

In the identification of PSE, it is necessary to involve conflicting interventions along distinct edges too, thus involving Multiple Worlds Models. The underlying assumption when involving multiple worlds is presented by [86] in the following way.

Definition 4.2.0.4. ***Multiple Worlds Model (MWM) [86]*** *The Multiple Worlds Model associated with a DAG $\mathcal{G}$ with nodes $V = \{V_1, ..., V_d\}$ is the set of all possible potential outcomes satisfying the* *assumption that the elements:*

$$\Big\{\{V_i(pa_{\mathcal{G}}(v_i)) : pa_{\mathcal{G}}(v_i) \in \mathcal{X}_{Pa_{V_i}}\} : V_i \in V\Big\} \tag{4.4}$$

are mutually independent.

For instance, the above condition imposes that these sets of interventional quantities, for instance for Figure 4.2, $\{M(a) : a \in \{0,1\}\}$, $\{Y(a,m) : a \in \{0,1\}, m \in \{0,1\}\}$, are mutually independent. By Multiple Worlds Model the authors of [86] refer to the Non-Parametric Structural Equation Model with Independent Errors (NPSEM-IE) presented in [66].
Single World and Multiple Worlds Models are characterized by assumptions made on the interventional distributions. It is immediate to notice that the assumption in Formula 4.4 implies the assumption 4.3, hence Multiple World Models are a subgroup of Single World Models. Indeed, Pearl's NPSEM-IE model is a sub-model of Robins' FFRCISTG model.

The edge g-formula makes use of the characteristics of Single World and Multiple Worlds Models. This formula will now be introduced; note that both the Lemma and the Proof are taken from [86].

Lemma 4.2.1. ***Edge g-formula*** *For a DAG $\mathcal{G}$ with vertices $V = \{V_1, ..., V_d\}$, and an edge intervention $\eta_{\mathfrak{a}_\alpha}$ on an edge set α, in Multiple Worlds Model for $\mathcal{G}$ we have that:*

$$P(V(\mathfrak{a}_\alpha) = v) = \prod_{V_i \in V} P(V_i = v_i | v_{Pa^{\hat{\alpha}}_{\mathcal{G}}(V_i)}, \mathfrak{a}_{\{(*Y)_{\to} \in \alpha\}}), \tag{4.5}$$

where $pa^{\hat{\alpha}}_{\mathcal{G}}(Y) = \{W \in Pa_Y | (WY)_{\to} \notin \alpha\}$.

Proof. In this paragraph, the proof of the edge g-formula is presented. Given a causal graph $\mathcal{G}$ with nodes $V = \{V_1, ..., V_d\}$ and topological ordering $\prec_{\mathcal{G}}$, the interventional distribution of the set of nodes V, given the edge intervention $\eta_{\mathfrak{a}_\alpha}$, is the following:

$$\begin{aligned}
P\Big(V(\mathfrak{a}_\alpha) = v\Big) &= \prod_{V_i \in V} P\Big(V_i(\mathfrak{a}_\alpha) = v_i | pre_{\prec_{\mathcal{G}}}(V_i)(\mathfrak{a}_\alpha) = v_{pre_{\prec_{\mathcal{G}}}(V_i)}\Big) \\
&= \prod_{V_i \in V} P\Big(V_i(v_{pa_{\mathcal{G}^{\hat{\alpha}}(V_i)}}, \mathfrak{a}_{\{(*V_i)_{\to} \in \alpha\}}) = v_i | \\
&\{W(v_{pa_{\mathcal{G}^{\hat{\alpha}}(W)}}, \mathfrak{a}_{\{(*W)_{\to} \in \alpha\}}) = w : W \in pre_{\prec_{\mathcal{G}}}(V_i)\}\Big) \\
&= \prod_{V_i \in V} P\Big(V_i(v_{pa_{\mathcal{G}^{\hat{\alpha}}(V_i)}}, \mathfrak{a}_{\{(WY)_{\to} \in \alpha\}}) = v_i : W \in pre_{\prec_{\mathcal{G}}}(V_i)\Big) \\
&= \prod_{V_i \in V} P\Big(V_i = v_i | v_{pa_{\mathcal{G}^{\hat{\alpha}}(V_i)}}, \mathfrak{a}_{\{(WY)_{\to} \in \alpha\}} : W \in pre_{\prec_{\mathcal{G}}}(V_i)\Big),
\end{aligned}$$

where the values $v_i, w, v_{pre_{\prec_{\mathcal{G}}}(V_i)}, v_{pa_{\mathcal{G}^{\hat{\alpha}}(.)}}$ are consistent with v. In the above equations, the first one follows from the Law of Total Probability. The second equality follows from the definition of the responses of outcomes to edge interventions $V(\mathfrak{a}_\alpha)$ in Definition 4.2.0.2, using recursive substitution on the parents of the variables V_i. The third equality, instead, can be justified using the assumptions that characterize Multiple Worlds Models, hence that the quantities in 4.4 are mutually independent. Following this property, the potential outcome corresponding to V_i after the above edge intervention $\eta_{\mathfrak{a}_\alpha}$, is independent from the potential outcomes of the parents $W \in pre_{\prec_{\mathcal{G}}}(V_i)$ induced by $\eta_{\mathfrak{a}_\alpha}$, even though they might involve conflicting interventional distributions. Based on this, the conditioning on the potential outcomes that are parents of V_i can be ignored, as it does not influence the interventional distribution on the different V_i with $V_i \in V$. The last equality is a consequence of the assumption that in Single World Models, the elements in 4.3 are mutually independent, given that this assumption also holds in Multiple Worlds Models.

□

Now, an example of the application of the edge g-formula will be given. The step-by-step identification procedure for this example will be analysed. Moreover, it will be shown that this is equivalent to the edge g-formula itself.

Example

In this example, the causal graph in Figure 4.2 will be analysed. The nodes C, A, M, Y are assumed to be discrete random variables, among which A is binary ($\mathcal{X}_A = \{a_0, a_1\}$). $\mathcal{X}_C$ and $\mathcal{X}_M$ are the domain spaces of the variables M and C. In this causal graph, the Path-Specific Effect of A along $(AY)_\rightarrow$ will be calculated. A is set as a_0 along $(AMY)_\rightarrow$ and as a_1 along $(AY)_\rightarrow$. Given this edge intervention, the respective notation is the following: $\alpha = \{(AM)_\rightarrow, (AY)_\rightarrow\}$, $\mathcal{X}_A = \{a_0, a_1\}$, $\mathcal{X}_\alpha = \{a_0, a_1\} \times \{a_0, a_1\}$, $\mathfrak{a}_\alpha = [a_1, a_0]$ where $\mathfrak{a}_\alpha \in \mathcal{X}_\alpha$. Since in this example the effect is estimated only along the direct edge from A to Y, the Path-Specific Effect can be addressed as the NDE. Based on this, referring to the causal graph in Figure 4.2, by intervening with A as a_1 along $(AY)_\rightarrow$ and with A as a_0 on $(AMY)_\rightarrow$, the respective Path-Specific Effect can be defined in the following way:

$$PSE^{\pi}_{a_0,a_1}(Y) = \mathbb{E}\big[Y(a_0, M(a_1))\big] - \mathbb{E}\big[Y(a_1)\big], \tag{4.6}$$

where π is the set of paths along which the effect is calculated: $\pi = \{(AY)_\rightarrow\}$. Following the notation introduced in the previous chapters, the quantity $Y(a_0, M(a_1))$ represents the outcome Y when A is set to a_0 along the direct path and M, mediator of the indirect path, has the value it would naturally assume when A is set to a_1. It can be noticed that the formulation of the Path-Specific Effect involves nested counterfactuals, such as $Y(a_0, M(a_1))$. These quantities cannot be identified with the previously introduced node intervention methods, as they include conflicting value attributions of the treatment attribute.

In the Path-Specific Effect in Equation 4.6 two interventional quantities need to be identified: $Y(a_1)$ and $Y(a_0, M(a_1))$. The former is a node intervention, meaning that the treatment attribute A is set to a_1 along all the paths from A to Y. This can be identified using the Truncated Factorization formula in 3.6. Differently, the latter $Y(a_0, M(a_1))$ can be identified with the following steps:

$$\begin{aligned}
&P(Y(a_0, M(a_1)) = y)\\
&= \sum_{c\in\mathcal{X}_C, m\in\mathcal{X}_M} P(Y(a_0, M(a_1) = m) = y | M(a_1) = m, C = c) P(M(a_1) = m, C = c)\\
&= \sum_{c\in\mathcal{X}_C, m\in\mathcal{X}_M} P(Y(a_0, M(a_1) = m) = y | M(a_1) = m, C = c) P(M(a_1) = m | C = c) P(C = c)\\
&= \sum_{c\in\mathcal{X}_C, m\in\mathcal{X}_M} P(Y(a_0, m) = y | C = c) P(M(a_1) = m | C = c) P(C = c)\\
&= \sum_{c\in\mathcal{X}_C, m\in\mathcal{X}_M} P(Y = y | A = a_0, M = m, C = c) P(M = m | A = a_1, C = c) P(C = c)
\end{aligned}$$

In the above equations, the first one can be immediately derived from the Law of Total Probability. The second one follows from the Markov Factorization property (see Equation 3.3) applied to the SWIG in Figure 4.3. Since the SWIGs are Directed Acyclic Graphs, the truncated factorization can be applied on its respective nodes, hence $P(M(a_1) = m, C = c) = P(M(a_1) = m | C = c) P(C = c)$ (see Property 3.2.3). The third equality follows from the independence condition $Y(a_0, m) \perp\!\!\!\perp M(a_1) | C$; this condition includes conflicting interventions on the treatment attribute, causing multiple worlds being involved. This assumption is satisfied when considering Multiple Worlds Models. In the last equality, instead, Single World Model assumptions were used. In particular:

$$M(a) \perp\!\!\!\perp A | C \tag{4.7}$$

$$Y(a, m) \perp\!\!\!\perp M(a), A | C \tag{4.8}$$

$\forall a \in \mathcal{X}_A$ and $\forall m \in \mathcal{X}_M$, where the last assumption is equivalent to the following two combined: $Y(a, m) \perp\!\!\!\perp A | C$ and $Y(a, m) \perp\!\!\!\perp M(a) | \{A, C\}$, $\forall a \in \mathcal{X}_A, \forall m \in \mathcal{X}_M$. The last equality can be derived by noticing that, if Assumption 4.7 holds, then $P(M(a_1) = m | C = c) = P(M(a_1) = m | A = a_1, C = c) = P(M = m | A = a_1, C = c)$, since $M(a_1) = M$ on the event $A = a_1$ (as intervening on a variable with an event that has already happened does not have any effect on the outcome), following Property 3.2.1. The same procedure can be applied to $Y(a, m)$ given Assumption 4.8.

Differently from the cross-world assumptions, the single-world ones can be easily interpreted using SWIGs, by leveraging their properties 3.2.3 and 3.2.4. In particular, specific graphical criterion can be used in order to assess the conditioned independence assumptions. In this scenario, using the d-separation criterion, it can be noticed that C d-separates $M(a)$ from A in Figure 4.3, while C d-separates $Y(a,m)$ from the set of nodes $\{A, M(a)\}$. For instance, in the conditional independence relation $M(a) \perp\!\!\!\perp A|C$ can be interpreted as C always containing one emitted arrow in the paths from $M(a)$ to A. Instead, in order to graphically interpret $Y(a,m) \perp\!\!\!\perp M(a), A|C$, it is most immediate to use the d-separation criterion by separately considering $Y(a,m) \perp\!\!\!\perp A|C$ and $Y(a,m) \perp\!\!\!\perp M(a)|\{A, C\}$ in the respective SWIG.

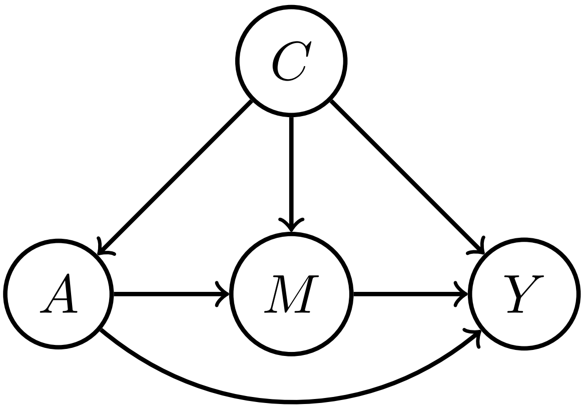


Figure 4.2: Example Causal Graph

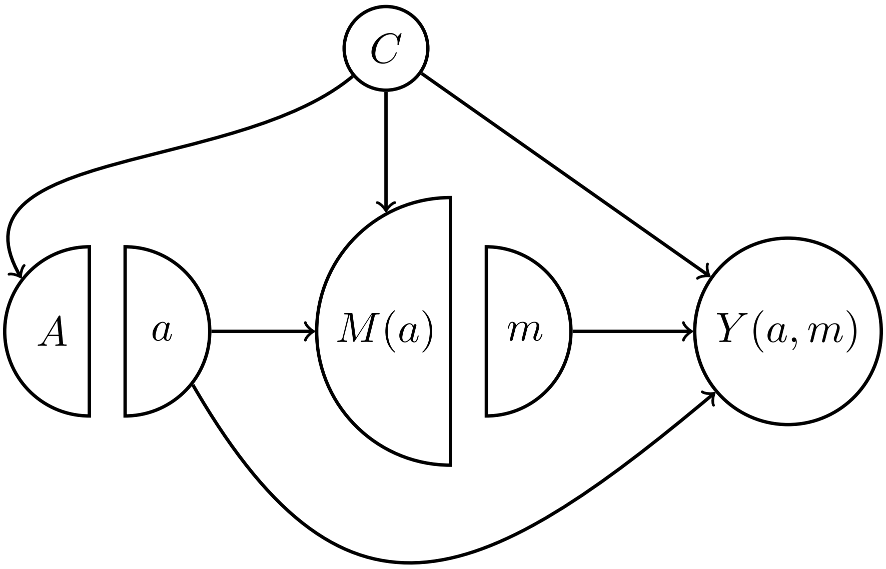


Figure 4.3: Example SWIG of Causal Graph 4.2 for $Y(a,m)$

When estimating the causal effect on a subset of V, for instance $P(V_1(\mathfrak{a}_\alpha))$, it necessary to marginalize over the variables in $V \backslash V_1$, hence $P\big(V_1(\mathfrak{a}_\alpha) = v_1\big) = \sum_{V \backslash V_1} P\big(V(\mathfrak{a}_\alpha) = v\big)$. Getting back to the example of the previous section, by applying the edge g-formula we get:

$$\begin{aligned}
&P(Y(a_0, M(a_1)) = y) \\
&= \sum_{c\in\mathcal{X}_C, a\in\mathcal{X}_A, m\in\mathcal{X}_M} P(Y = y|M = m, A = a_0, C = c)P(M = m|A = a_1, C = c)P(A = a|C = c)P(C = c) \\
&= \sum_{c\in\mathcal{X}_C, m\in\mathcal{X}_M} P(Y = y|M = m, A = a_0, C = c)P(M = m|A = a_1, C = c)P(C = c),
\end{aligned}$$

which is the same formula as the one derived in the previous set of equations.

This example can be also interpreted with the fairness approach. Specifically, in this scenario $\pi = (AY)_{\rightarrow}$ is the unfair path, and the scope is estimating the effect of the sensitive attribute A on the outcome Y along an unfair path.

Identification Assumptions

For single edges interventions, the interventional distributions are always identified under MWN assumptions (Def. 4.2.0.4). Path interventions are a special case of edge interventions, as the treatment is intervened upon not only along the single edges but actually entire paths starting from the treatment attribute itself. Referring to the previous notation, according to path interventions α is not a set of edges but a set of paths. The authors of [8] prove that the identification formula for path interventions is the same as Equation 4.5, nevertheless further assumptions are needed in order to prove identifiability.

The concept of *edge consistency* is key for the formulation of the identifiability conditions. Since the chosen paths α could potentially share certain edges, it is necessary that these path interventions are edge consistent. A nested interventional quantity (like the Path-Specific Effect $V_i(\pi, a_0, a_1)$) is considered to be *edge inconsistent* if it involves interventions that are not compatible. For instance, if paths π_1 and π_2 share the same edge emitted from the source node, the path intervention that intervenes with A as a_0 along π_1 and with A as a_1 along π_2 is edge inconsistent. Intuitively, this means that an outgoing edge from A can only be assigned a single interventional value of A, being edge consistent with respect to the path-specific intervention. Given a set of directed paths π, for every edge $(AW)_{\rightarrow}$ with $W \in V$, a respective path-intervention is edge consistent if the same value of A is assigned to all the paths in π that have $(AW)_{\rightarrow}$ as a prefix. Hence, a path $\{A \rightarrow W \rightarrow ... \rightarrow Y\}$ can be assigned an arbitrary value of A such that the edge $(AW)_{\rightarrow}$ is assigned one unique treatment value of A in the path-intervention. The edge consistency condition will be formalised with the Recanting District Criterion 4.2.1.1.
Edge consistency is a necessary condition for the identifiability of path-specific interventions. It can be interesting to notice that in the context of fairness, edge consistency is satisfied most of the times. Indeed, if a mediator M is considered to be unfair (that should not be allowed) then generally all the paths having as a prefix $(AM)_{\rightarrow}$ will be considered unfair too.

A graphical criterion that can be used to assess identifiability of path-interventions was introduced in [8], proving that in DAGs with independent error terms the violation of assumption of Formula 4.4 occurs in the presence of observed intermediate confounding. More precisely, this is known as the recanting witness criterion, which represents a special case of edge consistency. This criterion is very useful for interpreting cross-world assumptions with a graphical criterion. Two sets of edges π and $\hat{\pi}$ will be considered, along which the sensitive attribute is intervened upon as conflicting values (for instance as a_0 along π and a_1 along $\hat{\pi}$). The criterion used for identifying the potential outcome Y corresponding to this nested intervention is the following [109]:

Definition 4.2.1.1. ***Recanting witness criterion*** *Given a path set π, let Z be a node in $\mathcal{G}$ such that:*

1. *there exists a path from X to Z which is a segment of a path in π;*
2. *there exists a path from Z to Y which is a segment of a path in π;*
3. *there exists another path from Z to Y which is not a segment of any path in π.*

Then, the recanting witness criterion for the π-specific effect is satisfied with Z as a witness.

In order to explain this criterion, an example based on the causal graph in Figure 4.4 will be used. The corresponding structural equations model can be considered as:

$$\begin{aligned} A &= f_A(\epsilon_A) \\ L &= f_L(A, \epsilon_L) \\ M &= f_M(A, L, \epsilon_M) \\ Y &= f_Y(A, L, M, \epsilon_Y), \end{aligned}$$

with $\epsilon_A \perp\!\!\!\perp \epsilon_L \perp\!\!\!\perp \epsilon_M \perp\!\!\!\perp \epsilon_Y$. Given the effect of the path-intervention corresponding to $A = a_0$ along the set of paths $\pi = \{(ALY)_\rightarrow, (AY)_\rightarrow\}$ and $A = a_1$ along the rest of the paths $\hat{\pi} = \{(AMY)_\rightarrow, (ALMY)_\rightarrow\}$, it can be proven that the assumption $Y(a_0, m) \perp\!\!\!\perp M(a_1)$ does not hold because of the intermediate confounder L, satisfying the recanting witness criterion. In fact, by writing:

$$\begin{aligned} L(a_0) &= f_L(a_0, \epsilon_L) \\ L(a_1) &= f_L(a_1, \epsilon_L) \\ M(a_1) &= f_M(a_1, L(a_1), \epsilon_M) \\ Y(a_0, m) &= f_Y(a_0, m, L(a_0), \epsilon_Y), \end{aligned}$$

it can be immediately understood that, since $L(a_0) \not\perp\!\!\!\perp L(a_1)$ due to the common error term, then $Y(a_0, m) \not\perp\!\!\!\perp M(a_1) \quad \forall m$. In this example, the recanting witness criterion was satisfied, leading to the not-identifiability of the respective path-specific interventional distribution.

The above example can also be related to edge consistency. The causal graph in Figure 4.4 and the identification of the counterfactual quantity $Y(a_0, L(a_0), M(a_1))$ corresponding to $Y(a_0, L(a_0), M(a_1, L(a_1)))$ will be considered. As it can be noticed, the identification of $Y(a_0, L(a_0), M(a_1))$ through the edge g-formula is prevented because of the node L, since the edge $\{(AL)_\rightarrow\}$ follows two conflicting interventions (edge inconsistent). In this example, L is a recanting witness. Intuitively, the name *recanting*[1] follows from the idea that the L has to behave as A was an active treatment $(A = a_0)$ along $\{(ALY)_\rightarrow\}$. Subsequently, L needs to “disagree" with the previous statement, since along $\{(ALMY)_\rightarrow\}$ the treatment should not be active $(A = a_1)$. This recantation is impossible, since the two scenarios cannot hold simultaneously.

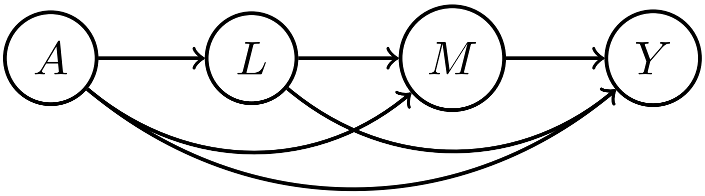


Figure 4.4: Example causal graph with recanting witness node L

In general, in non-parametric structural equation models with independent error terms, the necessary cross-world independence assumptions are violated if and only if the recanting witness criterion is satisfied [3]. Since these assumptions are not intuitive and not empirically testable, special attention should be put in assessing their validity [3].

The same criterion has been generalized to Semi-Markovian models in [82] with the recanting district criterion, which will be presented in the next section.

[1] From the Cambridge Dictionary, *to recant* means: to announce in public that your past beliefs or statements were wrong or not true and that you no longer agree with them

Edge g-formula in Semi-Markovian Models

The recanting witness criterion can be useful for determining identifiability in Markovian models. In order to assess the identifiability of Semi-Markovian models instead, the authors of [82] introduce the *recanting district criterion*, which generalizes the recanting witness criterion [8]. Intuitively, this criterion requires no conflict among the different districts (see Definition 3.2.0.7), instead of the single variables (witnesses) for Markovian models.

Following the same notation and setting as before for the recanting witness, the recanting district criterion is defined by Definition 2 of [82] as:

Definition 4.2.1.2. ***Recanting District*** *Let $\mathcal{G}$ be an acyclic graph, A, Y sets of nodes in $\mathcal{G}$, and π a subset of proper causal paths which start with a node in A and end in a node in Y in $\mathcal{G}$. Let Y^* be the set of nodes not in A which are ancestral of Y via a directed path which does not intersect A. Then a district D in an $\mathcal{G}_{Y^*}$ is called a recanting district for the π-specific effect of A on Y if there exist nodes $z_i, z_j \in D$ (possibly $z_i = z_j$), $a_i \in A$, and y_i, $y_j \in Y$ (possibly $y_i = y_j$) such that there is a proper causal path $a_i \to z_i \to \ldots \to y_i$ in π, and a proper causal path $a_i \to z_i \to \ldots \to y_i$ not in π.*

Based on our notation, $Y^* = An_{\mathcal{G}_{V\setminus A}}(Y)$. The presence of a recanting district prevents the formulation of Path-Specific Effects in terms of observed data. It is possible to express the Path-Specific Effect as a functional of interventional quantities if and only if a recanting district does not exist for this effect. Specifically, the cross-world counterfactual distribution can be written as a functional of interventional distributions in the following way [82]. Given a path intervention of a along π and of a' along $\hat{\pi}$:

$$P(Y(\pi, a_0, a_1) = y) = \sum_{V\setminus Y} \prod_{D \in \mathcal{D}(\mathcal{G})} P\Big(V_D(a_{\{Pa_{\mathcal{G}^\pi}(D)\setminus D\}\cap A}, a'_{\{Pa_{\mathcal{G}^{\hat{\pi}}}(D)\setminus D\}\cap A}, v_{Pa_{\mathcal{G}}(D)\setminus D}) = v_D\Big), \quad (4.9)$$

where V_D is the set of nodes in district D. The causal effect on these nodes is estimated by recursively intervening on the direct parents that are not part of the district itself, taking into account the path they belong to. Indeed, $Pa_{\mathcal{G}^\pi}(D)\setminus D$ is the set of nodes with directed arrows pointing into D along π but which are not in D. The outline of the proof follows the one of [82].

Proof. It will be proven that if a recanting district does not exist, then the Path-Specific Effect is identifiable from interventional distributions. The proof of completeness (if a recanting district exists, the effect is not identifiable) can be referred to in the Supplementary Material of [82]. As it will later be mentioned, the completeness is proven by selecting two models on $\mathcal{G}$ presenting a recanting district that induce the same observational distribution but different path-specific interventional ones.
Given $\{X_1, ..., X_m\} = Y^* = An_{\mathcal{G}_{V\setminus A}}(Y)$ the set of ancestors of Y excluding A and $\{Y_1, ..., Y_d\} = Y$, let the path-specific interventional distribution of Y after the intervention along π on A as a and as a' along the rest of paths be:

$$P(Y(\pi, a, a') = y) = P\Big(Y_1(\pi, a, a') = y_1, ..., Y_d(\pi, a, a') = y_d\Big),$$

where $a, a' \in \mathcal{X}_A$ and $y = [y_1, ..., y_d]$. These single terms can be rewritten using specific interventional terms $x_{Pa_{\mathcal{G}_{Y^*}}(Y_i)}$ that are consistent with the value assignment of the path-specific intervention, meaning that when $Pa_{\mathcal{G}}(X_i)$ intersects A they are equal to a or a' depending on π. Here, $x_{Pa_{\mathcal{G}}(Y_i)}$ are the reference values of the parents of Y_i. It is important to notice that in our notation, the ancestors of a set of variables Y, $An_{\mathcal{G}}(Y)$, is defined as $An_{\mathcal{G}}(Y) \equiv \{V_j | V_j \to \ldots \to Y \text{ in } \mathcal{G}\}$. By convention, $Y \in An_{\mathcal{G}_V}(Y)$, meaning that $m \geq d$ and $\{Y_1, ..., Y_d\} \in \{X_1, ..., X_m\}$. By using this information and by 'unrolling' the previous interventional distribution, the same distribution can be written as:

$$\sum_{x_{Y^*\setminus Y} \in \mathcal{X}_{Y^*\setminus Y}} P\left(X_1(x_{Pa_{\mathcal{G}}(X_1)}) = x_1, ..., X_m(x_{Pa_{\mathcal{G}}(X_m)}) = x_m\right). \quad (4.10)$$

This is the joint distribution among the entire set of variables $Y^* = An_{\mathcal{G}_{V\setminus A}}(Y)$, marginalized by summation with respect to $Y = \{Y_1, ..., Y_d\}$. In the above formula, the values $x_{pa_{\mathcal{G}}(X_i)}$ are consistent with $y_1, ..., y_d$ and with the values $x_{Y^*\setminus Y}$ of the summation. Furthermore, when $A \cap Pa_{\mathcal{G}}(X_i)$ and $X_i \in Y^*$, then the value a or a' is assigned, depending on the set of paths π.

For every V_i, V_j that are not joint by a bi-directed arrow, then the counterfactuals $V_i(v_{Pa(V_i)})$ and $V_j(v_{Pa(V_j)})$ are independent, meaning that the joint distribution can be factorized into marginal ones. The reasoning behind this is that by consistency (Definition 3.2.2), fixing all the observable parent of V_i makes quantity $V_i(v_{Pa(V_i)})$ independent of any counterfactual quantity $V_j(.)$ in the absence of a bi-directed arrow from V_j to V_i. Furthermore, this factorization property can be applied to all the counterfactuals belonging to the same district by using compositionality[2].
Based on this, it is possible to rewrite 4.10 as:

$$\sum_{x_{Y^*\setminus Y}} \prod_{X_1,\ldots X_{k_d} \in D \in \mathcal{D}_{\mathcal{G}_{Y^*}}} P\left(X_1(x_{Pa_{\mathcal{G}}(X_1)}) = x_1, \ldots, X_k(x_{Pa_{\mathcal{G}}(X_{k_d})}) = x_{k_d}\right), \tag{4.11}$$

where $X_1, \ldots, X_{k_d} \in D$ and D is one of the districts in $\mathcal{D}_{\mathcal{G}(Y^*)}$, the set of districts in $\mathcal{G}_{Y^*}$. As the last step, it is possible to use the consistency assumptions (see Property 3.2.2) of the interventional values $x_{Pa_{\mathcal{G}}(X_1)}$ with the values on the summation, the assigned values of the variables in the district, and the assignment to A. Based on this, in conclusion, the expression in 4.10 is equivalent to:

$$\sum_{x_{Y^*\setminus Y}} \prod_{D \in \mathcal{D}_{\mathcal{G}}} P\left(X_D(x_{Pa_{\mathcal{G}}(D)\setminus D}) = x_D\right).$$

In the above equation, $x_{pa_{\mathcal{G}}(D)\setminus D}$ is the value assignment to the parents of the district D that are not part of the district itself. These are consistent with the summation values $x_{Y^*\setminus Y}$ and with the values of the path-specific interventions on A. It is important to notice that the path-specific interventional distribution can be written as a functional of interventional distributions only if the assignments of the treatment attribute are not conflicting. As a matter of fact, if the district has two or more distinct edges leading to it from A with different interventional values a and a', it is not possible to find $x_{Pa_{\mathcal{G}}}$ such that it is consistent with the original intervention.
In [82], the proof is concluded by analysing the path-specific intervention with a respective recanting district. Specifically, two causal models $\mathcal{M}_1$ and $\mathcal{M}_2$ with the same induced graph are proposed such that the observed distributions are the same $P^1(V) = P^2(V)$ but the induced interventional distributions are distinct $P^1(Y(\pi, a, a')) \neq P^2(Y(\pi, a, a'))$. Hence, the interventional distribution is not identifiable in $\mathcal{G}$. □

Equation 4.9 is based on the same concept of the edge g-formula (see Equation 3.6). Nevertheless, since the nodes in the same district share correlated unobserved common causes, the path-specific intervention needs to be evaluated separately on the different districts. In the proof it was shown that the above factorization can be derived from applying the compositionality property to the conditionally independent counterfactual variables, leading to the factorization of the respective joint counterfactual distribution. The above formula can be then seen as an extension of the edge g-formula applied to graphs with possibly mutually correlated hidden variables (semi-Markovian models). Indeed, it can be noticed that, if the disturbance terms are all mutually independent, then the districts actually correspond to the distinct observed covariates themselves, leading Equation 4.9 to be the original edge g-formula. Because of the presence of correlated hidden variables, further exploration needs to be made in order to understand if the single product factors of Equation 4.9 are actually identifiable. This can be done using the identification algorithm (ID Algorithm) proposed in [96] and simplified in [83]; for the general idea of this algorithm, the reader can refer to Section 3.3.1.

4.2.2. Semi-parametric Estimator for Path-Specific Effect

The Path-Specific Effect has been shown to be useful for evaluating the causal effect along a defined set of paths. The respective identification formula is the edge g-formula (see Definition 4.2.1), providing the link between nested interventional quantities and conditional observed distributions. An important step in the application of this metric is defining a suitable estimator for the edge g-formula.

[2] Compositionality applied to counterfactuals states that if $V_k(v_{Pa(V_k)}) \perp\!\!\!\perp V_j(v_{Pa(V_j)})|V_i(v_{Pa(V_i)})$ and $V_w(v_{Pa(V_w)}) \perp\!\!\!\perp V_j(v_{Pa(V_j)})|V_i(v_{Pa(V_i)})$, then $V_i(v_{Pa(V_i)}) \cup V_w(v_{Pa(V_w)}) \perp\!\!\!\perp V_j(v_{Pa(V_j)})|V_i(v_{Pa(V_i)})$

Extensive studies on the performance of multiple semi-parametric estimators for identification formulas have been done in recent years. Graphical models with hidden variables have also been considered in the studies, bridging the gap between identification and estimation of causal effects [11]. These studies have mainly focused on identification formulas for node interventional distributions, such as the standard g-formula 3.6. Differently, due to the previously mentioned obstacles in the identifiability of Path-Specific Effects, defining an estimator for the edge g-formula can be relevantly more complex compared to the g-formula. In the literature, a more relevant focus has been put on the estimation theory of functionals derived by the standard g-formula compared to the edge g-formula [11]. Recent studies have explored the Natural Direct Effect and Natural Indirect Effect but to our knowledge these have not been fully generalized for Path-Specific Effect [95].

The semi-parametric estimator for the edge g-formula used in this project is the plug-in estimator. This consists in parametrically modelling the coefficients of the conditional distributions by maximum likelihood estimator. The marginal distributions of the root nodes are instead empirically estimated from the data.
The choice of the estimator of the edge g-formula is an important research topic. Given a causal graph $\mathcal{G}$ with variables V, a dataset $\mathcal{D}$ induced by a causal model consistent with $\mathcal{G}$ and the parameters β_i corresponding to the conditional distributions $P(V_i = v_i|Pa_i;\beta_i)$ with $v_i \in \mathcal{X}_{V_i}$, then the plug-in estimator $\hat{g}(\mathcal{D};\beta)$ for the $PSE^{\pi}_{a_1,a_0}(V)$ has the form:

$$\begin{aligned}
\hat{g}(\mathcal{D};\beta) = \prod_{V_i \in V\setminus\{A\}} &\Big[P\left(V_i = v_i|a_{1\{(AV_i)_\rightarrow \in\pi\}}, a_{0\{(AV_i)_\rightarrow \notin\pi\}}, Pa_{\mathcal{G}_{V\setminus A}}(V_i);\beta_i\right) - \\
&P\left(V_i = v_i|a_{0\{(AV_i)_\rightarrow \in\pi\}}, Pa_{\mathcal{G}_{V\setminus A}}(V_i);\beta_i\right)\Big] \\
= \prod_{V_i \in V\setminus\{A\cup Ro(\mathcal{G})\}} &\Big[P\left(V_i = v_i|a_{1\{(AV_i)_\rightarrow \in\pi\}}, a_{0\{(AV_i)_\rightarrow \notin\pi\}}, Pa_{\mathcal{G}_{V\setminus A}}(V_i);\beta_i\right)\Big]P_n(Ro(\mathcal{G})) - \\
&\Big[P\left(V_i = v_i|a_{0\{(AV_i)_\rightarrow \in\pi\}}, Pa_{\mathcal{G}_{V\setminus A}}(V_i);\beta_i\right)\Big]P_n(Ro(\mathcal{G})) \\
= \frac{1}{n}\prod_{V_i \in V\setminus\{A\cup Ro(\mathcal{G})\}} &\Big[P\left(V_i = v_i|a_{1\{(AV_i)_\rightarrow \in\pi\}}, a_{0\{(AV_i)_\rightarrow \notin\pi\}}, Pa_{\mathcal{G}_{V\setminus A}}(V_i);\beta_i\right) - \\
&P\left(V_i = v_i|a_{0\{(AV_i)_\rightarrow \in\pi\}}, Pa_{\mathcal{G}_{V\setminus A}}(V_i);\beta_i\right)\Big].
\end{aligned}$$

In the above formulas, the set $Ro(\mathcal{G})$ is the set of root nodes, introduced in 3.1.0.7. Related to this, $P_n(Ro(\mathcal{G}))$ is the empirical distribution of this set. In the third equality, this empirical distribution is estimated as a mass function of $\frac{1}{n}$. The above estimator of the PSE can be applied to any graphical structure, as long as the identifiability conditions are satisfied.
Different semi-parametric estimators for NDE have been proposed in [95] as discussed in Appendix D. In the next chapter, the performance of the plug-in semi-parametric estimator will be analysed by applying it on simulated datasets.

4.3. Counterfactual Fairness

The Counterfactual Fairness metric characterizes a different class of causality-based fairness metrics compared to the previously analysed ones. Fairness metrics could be classified as population-level and individual-level metrics, depending on the target group chosen. On one side, the population-level fairness metrics analyse unfairness among the overall population. For instance, the previously analysed Path-Specific Effect focuses on evaluating unfairness at an aggregate level among the entire population, by evaluating the expected value of the interventional distribution. Nevertheless, Population-based fairness metrics do not guarantee to represent discrimination towards all the single individuals. Differently, individual-level fairness metrics has the scope of quantifying the discrimination for each individual of the population.

The authors of [26] introduced the *non-causal* fairness metric Individual Fairness. The following definition will be useful for better understanding the concept of Counterfactual Fairness. Individual Fairness is defined in [26] as:

Definition 4.3.0.1. ***Individual Fairness*** *Given two individuals i and j having as values of covariates $X^{(i)} = x^{(i)}, A^{(i)} = a^{(i)}$ and $X^{(j)} = x^{(j)}, A^{(j)} = a^{(j)}$ respectively, an outcome Y is individually fair if:*

$$P(\hat{Y}^{(i)} = y | X^{(i)} = x^{(i)}, A^{(i)} = a^{(i)} \approx P(\hat{Y}^{(j)} = y | X^{(j)} = x^{(j)}, A^{(j)} = a^{(j)}) \quad \textit{if} \quad d(i,j) \approx 0, \tag{4.12}$$

where $d(\cdot,\cdot)$ is a metric that describes the similarity of two individuals based on how they should be treated.

This metric is context and task specific; if $d(i,j) \approx 0$ then the individuals i and j should be treated similarly, in accordance to the disparate treatment framework. This is based on a distance measure domain-specific distance measure $d(\cdot,\cdot)$. For instance by using a matching procedure [60], then two very similar individuals with different sensitive attribute can act as the counterfactual of one another. It can be noticed that the definition of this distance metric could lead by itself to possible implicit discrimination, based on the context of the problem.

Many recent works have been focusing on individual fairness, among which the most used one is *Counterfactual Fairness*, introduced in [48]. The authors define a predictor $\hat{Y}$ to be fair towards an individual belonging to the group a_1 if it is the same prediction he/she would have got belonging to a different demographic group a_0. Hence, following the notation introduced in Section 3.2.2, given a Structural Causal Model (U, V, F) with $V = A \cup X$, Counterfactual Fairness is defined in [48] as:

Definition 4.3.0.2. ***Counterfactual Fairness*** *Predictor $\hat{Y}$ is counterfactually fair if under any context $X = x$ and $A = a_0$,*

$$P(\hat{Y}_{A=a_0} = y | X = x, A = a_0) = P(\hat{Y}_{A=a_1} = y | X = x, A = a_0) \tag{4.13}$$

for all $y \in \mathcal{X}_Y$ and for any value $a_1 \in \mathcal{X}_A$.

It is important to notice that the counterfactual quantity $\hat{Y}_{A=a_0}$ conditioned on $A = a_0$ is equivalent to the predictor $\hat{Y}$ itself. This metric is an individual-level fairness metric as the intervened outcomes are conditioned on the entire set of measurable quantities. This fairness definition is a general case of the Individual Fairness definition provided in Formula 4.12, using as distance measure the similarity between two individuals. Specifically, an observed individual i and his/her unit-level counterfactual version j are considered to have $d(i,j) \approx 0$. Counterfactual Fairness can actually be seen as a specific scenario of the Individual Fairness.

In the research community of ethical algorithmic fairness, the use of counterfactuals related to sensitive social attributes (e.g. race or gender) has been criticized due to ontological and epistemological-semantic complications [46]. From an ethical perspective, the counterfactual manipulation of these social categories should not be allowed, as it could actually sharpen the "problematic associations between the sensitive attributes at are the result of social and structural injustice in the first place." [41]. Nevertheless, I believe that, following Haslanger's point of view ([36], [36], [30]) of racial constructivism, the manipulation of of sensitive attributes like race is justifiable if the purpose of the work is actually the fighting and decreasing social falls like racism.

4.3.1. Path-Specific Counterfactual Fairness

Path-Specific Counterfactual Fairness (PSCF) has been first introduced in [15]. It is based on the estimation of path-specific counterfactuals, meaning that the subjects' counterfactuals are intervened upon solely along certain pathways from A to Y. Hence, this metric includes characteristics of both Path-Specific Effect Fairness and Counterfactual Fairness. Differently from the Path-Specific Effect in Formula 4.1, PSCF focuses on detecting unfairness along the not allowed paths at an individual level.

Furthermore, compared to Counterfactual Fairness in Equation 4.13, this metric evaluates the causal effect along the different pathways. Following the approach in [105], the path-specific counterfactual effect can be defined as:

$$PSCE^{\pi}_{a_0,a_1}(\hat{Y}|x,a) = P(\hat{Y}_{(\pi,a_0,a_1)}|X=x, A=a) - P(\hat{Y}_{A=a_0}|X=x, A=a). \tag{4.14}$$

Given the factual observations $X = x$ and $A = a$, PSCE represents the counterfactual effect on Y of the value intervention of A as a_0 along π and with A as a_1 in the rest of causal pathways. The counterfactual notation symbolised with subscripts and not brackets as traditional interventions refers to the notation introduced in Section 3.2.2. Indeed, it was previously mentioned that the identification of unit-level counterfactuals involves inference made on the unobserved variables. This concept is based upon the assumption that counterfactual versions of the same individual obtain the same values obtained by unobserved variables.

4.4. Conclusion

In this section, different causal fairness metrics were evaluated. Two different frameworks were presented, depending on the scope of the metrics. Both the disparate impact and disparate treatment frameworks were analysed, emphasising the advantages that path-specific metrics have. These metrics are indeed capable of evaluating the causal influence of the sensitive attribute along arbitrary causal pathways connecting the sensitive attribute to the outcome. The analysis of the paths as fair or unfair allows for a more flexible and detailed evaluation of this causal effect, leveraging the information along the ‘allowed’ pathways and mediators. Specifically, given a previous domain-based categorization of these paths as allowed or not allowed, path-specific metrics are very suitable for representing fairness. A further classification of the metrics was made, specifically into population-level and individual-level metrics with their respective mathematical formulations. This classification depends on the focus of the metric itself. Among individual-level metrics, Counterfactual Fairness and Path-Specific Counterfactual fairness were presented, underlying the respective strengths and weaknesses. In particular, Path-Specific Counterfactual Fairness was presented as a metric bridging the individual approach of Counterfactual Fairness and the path-specific approach corresponding to path-specific metrics, hence a promising metric in the fairness setting.

5

Existing Methods and Comparison

In this project, multiple causality-based state-of-the-art methods will be analysed. The scope of this chapter is to compare the most promising methods in the literature, analysing both the methodology and the performance on simulated datasets. In Section 5.1 the an overview of the methods will be given, along with their objectives and conceptual differences. In Section 5.2, the methods will also be applied to simulated datasets. First, the application of these methods will be described step-by-step; then, the performance results will be analysed and discussed. The methods will be compared in terms of fairness and accuracy. It will later be explained the relevance of these metrics and their estimation.
Overall, methods based on path-specific metrics will be selected as the most potential ones. In Chapter 4, by comparing the different causality-based metrics, it was motivated that path-specific ones (Path-Specific Effect and Counterfactual Path-Specific Effect) seem the most suitable in a fairness setting. In this Chapter, the empirical comparison of the methods will support this statement, leading this project to focus on the methods *Fair Inference on Outcomes* [61] and *Path-Specific Counterfactual Fairness* [15].

5.1. Selected Methods

In this Section, three state-of-the-art methods will be presented. These methods has two different goals: first to assess unfairness of a Machine Learning algorithm outcome Y with respect to certain predefined metrics, then to predict a new fair, or at least fairer, outcome $\hat{Y}$. The analysed methods are *Fair Inference on Outcomes* [61], *Counterfactual Fairness* [48] and *Path-Specific Counterfactual Fairness* [15], respectively corresponding to the causality-based metrics of Path-Specific Effect, Counterfactual Fairness and Path-Specific Counterfactual Effect, as the titles suggest. These methods were selected as the most promising in terms of results and fairness constraints.

All the selected methods will lead to useful insights regarding the different causality-based used fairness metrics. Since the Counterfactual Fairness [48] and Path-Specific Counterfactual Fairness [15] methods are focused on individual-level fairness, they involve further assumptions on the latent space and generally the necessity of estimating the posterior distribution of unobserved variables. The main difference between these two methods consists in the classification of the different paths from the sensitive attribute to the outcome. On one hand, the authors of [48] consider all these paths as unfair, while in [15] prior information regarding the classification of the paths as allowed or not allowed is taken into account. Differently, Fair Inference on Outcomes method [61] focuses on population-level fairness, making use of Path-Specific Effect as a fairness metric.

5.1.1. Fair Inference On Outcomes [61]

The method Fair Inference on Outcomes (FIO) has the scope of estimating a fair joint distribution of the outcome and the observed variables satisfying constraints on Path-Specific Effect and reflecting as much as possible the observed data distribution.
The joint data distribution of the observed variables $p(Y, X, A)$ induced by a causal model is considered. Following the above notation, Y is the outcome, A is the sensitive attribute $\mathcal{X}_A = \{a_0, a_1\}$ and X is the set of other variables, including the baseline factors C and mediators. It was already shown in Definition 4.2.0.1 that the Path-Specific Effect of A on Y along the unfair set of paths π is defined as:

$$PSE^{\pi}_{a_0,a_1}(Y) = \mathbb{E}\big[Y(\pi, a_0, a_1)\big] - \mathbb{E}\big[Y(a_1)\big]. \tag{5.1}$$

In the first expected value, $Y\,(\pi, a_1, a_0)$ is the output Y such that, along the pathways of interest π variables vary as if the sensitive variable A were set to value a_0, and along other pathways from A to Y, variables vary as if A were set to a_1. Instead, the term $Y\,(a_1)$ is the output value Y if A were set to a_1 along all the paths from A to Y. The intuition behind this metric has been covered in Section 4.2. Furthermore, in Section 4.2 the identifiability of the Path-Specific Effect has been investigated, by defining the necessary conditions that need to be met by the causal model. Assuming that the considered $PSE^{\pi}_{a_0,a_1}(Y)$ is identifiable, the effect can be identified with the edge g-formula in Formula 4.5; for convenience, this identification formula will be referred to as $g^{\pi}_{a_0,a_1}(Y, X, A)$.

The purpose of the Fair Inference on Outcomes method (FIO) of [61] and [62] is transferring the inference problem on the observed joint distribution $p(Y, X, A)$ to a problem on $p^*(Y, X, A)$, a distribution of the *fair world*. The goal is indeed to define a fair distribution $p^*(Y, X, A)$ that is as close as possible from the observed distribution $p(Y, X, A)$ but that satisfies the fairness constraint by limiting the respective Path-Specific Effect to be close to zero. This method can be mathematically formulated with the following constrained optimization problem, defining the fair world distribution as the one minimizing the KL-divergence to $p(Y, X, A)$ while constraining the Path-Specific Effect to lie in the boundary interval ϵ^-, ϵ^+. This is defined as:

$$p^*(Y, X, A) = argmin_q D_{KL}(p||q) \tag{5.2}$$

$$\text{subject to } \epsilon^- \leq g^{\pi}_{a_0,a_1}(Y, X, A) \leq \epsilon^+ \tag{5.3}$$

In a finite sample setting with data $\mathcal{D} = \{(Y_i, X_i, A_i), i = 1, ..., n\}$ drawn from the distributions $p(Y, X, A; \beta)$ parametrized by β, the above problem is equivalent to the following constrained maximum likelihood problem[1]:

$$\hat{\beta} = argmax_{\beta} \mathcal{L}_{Y,X,A}(\mathcal{D}; \beta) \tag{5.4}$$

$$\text{subject to } \epsilon^- \leq \hat{g}(\mathcal{D}; \beta) \leq \epsilon^+, \tag{5.5}$$

where $\hat{g}(\mathcal{D}; \beta)$ is an estimator of the edge g-formula and $\mathcal{L}_{Y,X,A}(\mathcal{D}; \beta)$ is the likelihood function parametrized by β. The parameters $\hat{\beta}$ will then be the estimated parameters corresponding to the fair distribution $p^*(Y, X, A)$, satisfying the constraint on the Path-Specific Effect and resembling at most the observed data distribution. As previously discussed, the baseline estimator of the edge g-formula that will be used in this work is the semi-parametric plug-in estimator $\hat{g}(\mathcal{D}; \beta)$, the formulation of which can be found in Section 4.2.2. The performance of this estimator will be analysed in Section 5.3.
Usually, we have $\epsilon^- = \epsilon^+ = \epsilon$, meaning that the optimization problem can be re-formulated like:

$$\hat{\beta} = argmax_{\beta} \mathcal{L}_{Y,X,A}(\mathcal{D}; \beta)$$
$$\text{subject to } |\hat{g}(\mathcal{D}; \beta)| \leq \epsilon.$$

The FIO method focuses on population-level fairness, since it makes use of the PSE as a fairness metric. A great advantage of this method compared to the two following ones is that it does not involve the

[1] Proof in the Appendix C

estimation of the latent space distribution. Indeed, as long as the PSE identifiability conditions are satisfied, the constraints can be formulated solely in terms of observational distributions. Further details regarding the estimation of the new instances $\hat{Y}$ from the posterior distribution $p(Y|X, A; \hat{\beta})$ are given in the Appendix.

5.1.2. Counterfactual Fairness [48]

The notion of Counterfactual Fairness is introduced in [48]. Referring to Definition 4.3.0.2, this metric is an individual-level fairness definition based on unit-level counterfactuals (Definition 3.2.4.1), as it corresponds to the intuition "what would the output be if the sensitive attribute of a specific individual had been different". Following the intuition of Counterfactual Fairness, two different individuals are compared: an observed individual with an observed value of sensitive attribute and an hypothetical one (called the unit-level counterfactual), corresponding to the set of covariate values the same individual would have with a different sensitive attribute value. As previously mentioned, the unit-level counterfactual approach in causal inference is defined by quantities that can never be observed empirically and characterized by strong assumption that might be difficult to assess [21].

The method introduced in [48] aims at predicting counterfactually fair outcomes. This work is based upon the following lemma, proven by the same authors:

Lemma 5.1.1. *Let $\mathcal{G}$ be the causal graph of the given model (U, V, F). Then Y will be counterfactually fair if it is a function of the non-descendants*[2] *of A.*

Proof. Let W be any non-descendant of A in $\mathcal{G}$. Then, following the notation introduced in Section 4.3, $W_{A=a_0}(D)$ and $W_{A=a_1}(D)$ have the same distribution, where D is the set of non-descendants of A. This means that the distribution of any function Y of the non-descendants of A is invariant with respect to the counterfactual values of A. In conclusion, a function Y of the non-descendants of A and of the unobserved variables is counterfactually fair. □

In order to provide a counterfactually fair outcome $\hat{Y}$, the authors of [48] aim at predicting an output that is independent from the sensitive attribute and all of its descendants. This approach is different from the Fairness through Unawareness method formulated Section 2.3.3. Indeed, in the approach of [48] the descendants of the sensitive attribute that could possibly act as proxies for discrimination, are not used for the prediction either. The condition expressed in Lemma 5.1.1 is actually a stricter condition than Counterfactual Fairness itself. Albeit this condition is not necessary, it is sufficient for guaranteeing Counterfactual Fairness in a predictor $\hat{Y}$.

The strategy of [48] consists in predicting the output $\hat{Y}$ as a function of the non-descendants of A in the causal graph $\mathcal{G}$. Different approaches are defined by the authors, depending on the level of assumptions needed. An overview of these approaches is the following:

- In the *first level*, $\hat{Y}$ is built only using the observed non-descendants of the sensitive variable A, meaning that $\hat{Y} = h_\beta(D)$, where D is the set of nodes $\{D_i : D_i \in V \text{ and } D_i \notin De(A)\}$. Here, the parameter β of the function h minimizes the chosen loss function. The function h can be an arbitrary function, such a regression or a neural network. This level has the benefit of not implying any additional assumption on the unobserved space, but it also leads to some drawbacks. Indeed, if in $\mathcal{G}$ all of the nodes are descendants of the sensitive attribute ($D = \{\emptyset\}$), this approach cannot be used. Albeit this is usually not the case because of the presence of root nodes, a high loss of information regarding the descendants of A is definitely a limit of this approach.

- In the *second level*, the background latent variables are introduced as non-deterministic causes of the observed ones. A unique unobserved variable is assumed to be parent of the observed descendants of A, $De(A)$. This would correspond to an ADMG with all the descendants of A belonging to the same district (hence connected by bi-directed edges). Once the conditional distribution of the latent variable $P(U|X = x)$ has been estimated via variational inference, a predictor $\hat{Y} = h_\beta(U, D)$

[2] The definition of descendants nodes $De(\cdot)$ can be found in Definition 3.1.0.5.

is trained. For every data point $(x^{(i)}, a^{(i)}) \in \mathcal{D}$, $u^{(i)}$ is sampled from the conditioned probability distribution $P(U|X = x^{(i)})$ $(u^{(i)} \sim P(U|X = x^{(i)}))$. Specifically, for every observed data-point i, the dataset is augmented by sampling J instances from $P(U|X = x^{(i)})$. The augmented dataset will thus be composed by $n \times J$ data-points such as $\{u_1^{(i)}, x^{(i)}, a^{(i)}\}, ..., \{u_J^{(i)}, x^{(i)}, a^{(i)}\}$, $\forall i \in \{1, ..., n\}$. The fair prediction $\hat{Y}_i$ for an instance $(x^{(i)}, a^{(i)}) \in \mathcal{D}$ will then be the mean of $\hat{Y}_i^j = h_\beta(u_j^{(i)}, d^{(i)})$ with respect to j varying in $\{1, ..., J\}$, where $d^{(i)} \in \mathcal{X}_D$.
By using the definition of the predictor $\hat{Y} = h_\beta(U, D)$, if $D = \{\emptyset\}$, the fair predictor will only be a function of the latent space, thus $\hat{Y} = h_\beta(U)$. This would lead the coefficients β to be estimated by training the model on $\{u_1^{(i)}, ..., u_J^{(i)}\}$ for $i \in \{1, ..., n\}$.

- The *third level* used by [48] consists in introducing the latent variables ϵ_i as deterministic causes of each of the observed variables. In this scenarios, additive error models are considered. Hence, the distribution of the observed variable V_i given the parents Pa_i is defined as $V_i = f_i(Pa_i) + \epsilon_i$ [70]. Once each ϵ_i has been deterministically estimated from the additive error model, the outcome $\hat{Y} = h_\beta(\epsilon_1, ..., \epsilon_n, D)$ is estimated as a function of ϵ_i and the non-descendants of A.
In this level, a restricted functional class of the Structural Equations is used, characterized by an additive error term. This form is broadly used in causality research to analyze causal graphs [38], but it might not fully represent the original causal structure [31], [71]. Furthermore, in this strategy it is assumed that the error terms ϵ_i are independent from the sensitive variable A, but in practice this is not always true.

It is important to analyse the assumptions that the authors of [48] use. The number of assumptions needed for the method increases accordingly to the level. While in the first level no specific assumptions are needed, the second level assumes that there is a unique set of unobserved variables U that influences all the observed descendants of A. Furthermore, the variable U should be independent from A. The third level is based on stronger assumptions, as the causal generating models are assumed to be additive noise models and that the error terms are independent from the sensitive attribute [43].

Different advantages and disadvantages of this method can be considered. It has the great advantage of not relying on the complicated estimation of counterfactual quantities, since it consists in building a predictor that does not use any variable that is descendant of A. As proven in Lemma 5.1.1, this method satisfies Counterfactual Fairness formulated in Equation 4.13. Hence, if the assumptions on the unobserved variables are met, defining an estimated outcome $\hat{Y}$ that is a function of the non-descendants of A ensures that Counterfactual Fairness is satisfied across all the individuals. Nevertheless, depending on the level of the method, it is necessary to make assumptions and variationally or deterministically estimate the unobserved space.
The main weakness of this method is not involving path-specific fairness, as relevant information is lost in the estimation of the fair outcome. Indeed, albeit some paths might be considered allowed by domain-specific experts, the respective mediators would not considered in the prediction. For instance, in example of Figure 4.2 having as unfair path $(AY)_\rightarrow$, the mediator M would be ignored, even though it is not a proxy for unfairness. Nevertheless, since no path-specific fairness is considered, no further assumptions regarding the classification of the paths as fair and unfair is necessary. Albeit this might seem a simplification of the domain-knowledge based assumptions, consulting expertise from the domain background is still necessary for defining a causal graph, thus the interpretation of the causal paths is a logical step of this process.

5.1.3. Path-Specific Counterfactual Fairness [15]

The aim of the method Path-Specific Counterfactual Fairness (PSCF) introduced in [15] is to build an algorithm that outputs path-specific counterfactually fair predictions, as defined in Formula 4.14. This metric focuses on both individual-level fairness and path-specific fairness. For an overview of the metric itself, the reader can refer to Section 4.3.1.

The method introduced by Chiappa in [15] consists in correcting the outcome prediction at test time by considering the unit-level counterfactual versions along the unfair paths of the different individuals belonging to the potentially discriminated group. As previously defined in Definition 3.2.4.1, the unit-level counterfactual version of an individual is characterized by the set of covariate values it would have assumed if that subject was part of the other sensitive group. Since this method takes into account the fairness classification of the paths, this correction is done solely along the illegitimate paths. For instance, assuming that the sensitive attribute is gender, when predicting the outcome for a female, the individual is considered to be a male along the unfair pathways from A to Y. This means that, for every mediator along an unfair path, it is necessary to estimate its unit-level counterfactual version by using the respective unobserved parents. Unit-level counterfactuals are estimated in [15] by generalising the abduction-action-prediction method introduced in Section 3.3.2.

Now that the intuition of the method has been introduced, an overview of the different steps involved is given. The first part of the method consists in modelling the latent space of the causal model. Indeed, in order to estimate the counterfactual versions of the individuals, it is necessary to estimate the respective unobserved covariates' posterior distributions. This is addressed with a variational approach, by estimating the Gaussian approximations of the posterior distribution $P(U_i|V)$, for every unobserved variable U_i. Here, U_i represents the unobserved variable parent of the mediator V_i from A to Y. Specifically, a Monte-Carlo approximation with a Variational Auto-Encoder (VAE) is used in this task.
Once the posterior distribution of the unobserved space has been inferred, it is possible to estimate the individuals' path-specific counterfactuals. In order to do so, final predictions are estimated with a Monte-Carlo sampling approach. By sampling from the Gaussian approximation of the unobserved distributions $u_i \sim P(U_i|V)$ for every unobserved variable U_i, the values u_i can be used for estimating the counterfactual versions of the mediators along the unfair paths. In particular, by estimating the values of the mediators with the sample values of the unobserved covariates and the estimated values of the mediators' parents, it is possible to define the path-specific counterfactual versions of the mediators and hence of the outcome.
Overall, since the correction happens at test time, the mediators along the fair paths will not vary, while the counterfactual versions of the mediators along unfair paths are estimated thanks to the respective unobserved covariates. Hence, the outcome will be a function of the observed values of the mediators along the fair paths and the counterfactual versions along the unfair ones. More precisely, given an individual with $A = a_0$ and $u_M \sim P(U_M|V)$, for every mediator M lying on an unfair path, the observed value is substituted by its counterfactual version, estimated as $m_{CF} \sim P(M|u_M, a_1, pa_M^{CF})$, hence considering the unobserved variables to remain unchanged in the observed individual and the respective counterfactual version and changing only A as a_1. Here, pa_M^{CF} is the set of observed and counterfactual values of the parents of M, depending on whether they are located along unfair paths. The procedure is repeated for J times; given every sample $u_i^j \sim P(U_i|V)$ with $j \in 1, ..., J$, the final outcome $\hat{Y}$ will be the average of $\hat{Y}^j$ with respect to $j \in 1, ..., J$, following the Monte-Carlo approach. This procedure will be covered in more detail in the application of the method.
It is critical to notice that, differently from the previously introduced methods in which new parameters $\hat{\beta}$ have been estimated such that the fairness constraint was satisfied, this was not the case for PSCF. As a matter of fact, this method uses the same original inferred observed distribution with corrections made at test time in order to estimate the path-specific counterfactual values.

From the analysis of this method, different insights can be derived. By estimating the counterfactual versions of every individual, it is necessary to estimate the respective values of the unobserved covariates. As it can be intuitively understood, it is critical for the method to assess the independence of these unobserved covariates from the sensitive attribute A. Otherwise, relevant (and potentially discriminating) information regarding A could be included in the distribution of unobserved variables. This would mean that the estimation of the counterfactuals would not be reliable, leading the method to fail at ensuring fairness. In order to avoid this issue in the variational approximation of the latent space, the variational auto-encoders enforce the independence of the latent space from A by using a *Maximum Mean Discrepancy* penalization [53], [33]. The Maximum Mean Discrepancy measures the distance between two probabilities, in this scenario the distributions of the Gaussian approximation $P(U_i|V)$ for different values of A. This penalty term is added to the loss function of the VAE with a weighting factor β that determines the independence relation. Even though the method PSCF is highly flexible, the tuning of this additional term in the loss function of the variational auto-encoder might lead to loss of information and instability.

This method has the great advantage of predicting new outcomes that are fair towards all the individuals with respect to the classification of the paths into fair or unfair. This method combines path-specific metrics with individually-fair predictions. Nevertheless, the introduction of the variational approach for estimating the unobserved variables distributions by adding a new regularization term should be regarded at with attention.

5.2. Comparison Methods

In the next sections the state-of-the-art methods will be applied to various simulated datasets. The methods' outlines were described in the previous section, but the details regarding their implementation will now be presented. This should assist the reader in comprehending how these methods work in practice. Furthermore, the different performances will be discussed.

Two examples are taken into consideration, involving two distinct causal structures. The methods are applied to datasets simulated from the proposed causal models, with the scope of assessing the fairness of the model and providing a fair outcome $\hat{Y}$.

Overall, all of the presented methods perform satisfactorily in terms of population-level fairness. Furthermore, since enforcing individual-level fairness is a stronger constraint compared to population-level fairness, the methods of [48] and [15] will be outperformed by the one of [61] in accuracy of the predictions. In addition, since the Fair Inference on Outcomes method [61] does not involve the estimation of the latent space, it is highly more intuitive, also by not involving unit-level counterfactuals estimation and hence not requiring assumptions on the latent space distribution.

5.2.1. Example 1 - Simulated Dataset

In this example, the causal structure of Figure 5.1 is considered. It is assumed that the only unfair path is the direct one $\pi = \{(AY)_{\rightarrow}\}$. Hence, the Path-Specific Effect of this graph is equivalent to the Natural Direct Effect. The variables M and A are assumed to be binary, generated using logistic regression. Differently, Y and C are continuous. The simulated dataset has been generated in the following way:

$$\begin{aligned} C \sim & \mathcal{N}(0,1) \\ logit(P(A=1|C)) \sim & -0.5+0.5C \\ logit(P(M=1|A,C)) \sim & 0.1-1.5C+2A \\ Y = & 1+5C-4A+2M+\epsilon, \end{aligned}$$

with $\epsilon \sim \mathcal{N}(0,1)$. In the next paragraphs, the application of the different methods to this causal structure will be shown.

Fair Inference On Outcomes [61]

In order to apply the Fair Inference on Outcomes method, first the conditional distribution of the dataset are semi-parametrically modelled, consistently with the causal graph. Logistic regressions and linear regressions are respectively used for binary and continuous outcomes. Given the causal graph,

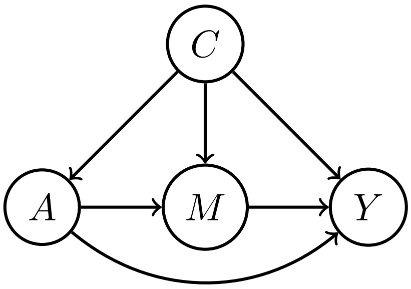


Figure 5.1: Example 1 Causal Graph

the regression model is:

$$\begin{aligned} C &\sim p(C), \\ logit(P(A=1|C))) &\sim \beta_a + \beta_a^c C, \\ logit(P(M=1|A,C))) &\sim \beta_m + \beta_m^a A + \beta_m^c C, \\ Y &= \beta_y + \beta_y^a A + \beta_y^m M + \beta_y^c C + \epsilon. \end{aligned}$$

Before solving the constrained optimization problem in Formula 5.4, it is necessary to definethe semi-parametric plug-in estimator for the Path-Specific Effect. Through the edge g-formula, it is possible to formulate the PSE as a function of observational distributions $PSE^{\pi}_{a_0,a_1}(Y) = g^{\pi}_{a_0,a_1}(Y, M, A, C)$, such that:

$$g^{\pi}_{a_0,a_1}(Y, M, A, C) = \sum_{c \in \mathcal{X}_C, m \in \mathcal{X}_M} \Big(\mathbb{E}\big[Y|A=a_0, M=m, C=c\big] - \mathbb{E}\big[Y|A=a_1, M=m, C=c\big] \Big) \cdot \\ \cdot P(M=m|A=a_1, C=c) P(C=c).$$

The respective semi-parametric plug-in estimator $\hat{g}^{\pi}_{a_0,a_1}(\mathcal{D}, \beta)$ is:

$$\hat{g}^{\pi}_{a_0,a_1}(\mathcal{D}, \beta) = \frac{1}{n} \sum_{i \in \{1,\dots,n\}, m \in \mathcal{X}_M} \Big(\mathbb{E}\big[Y|A=a_0, M=m, C=c_i; \beta_Y\big] - \mathbb{E}\big[Y|A=a_1, M=m, C=c_i; \beta_Y\big] \Big) \cdot \\ \cdot P(M=m|A=a_1, C=c_i; \beta_M),$$

where c_i is the i-th value of C in $\mathcal{D}$ and we have $\beta_M = \{\beta_m, \beta_m^a, \beta_m^c\}$, $\beta_Y = \{\beta_y, \beta_y^a, \beta_y^c, \beta_y^m\}$, $\beta = \{\beta_M, \beta_Y\}$. The marginal distribution $P(C=c)$ with $c \in \mathcal{X}_C$ is empirically estimated from the data by considering $\frac{1}{n}$ mass at every instance c of the data. This was previously shown in Section 4.2.2.
Now that the semi-parametric estimator of the Path-Specific Effect has been defined, the method consists in estimating $\hat{\beta}_M, \hat{\beta}_Y$ that are solutions of the optimization problem 5.4 with $\epsilon^+ = 0.05$ and $\epsilon^- = -0.05$. The new parametric estimator $\hat{\beta}$ is then the optimal set of parameters such that the new joint probability distribution is close to the original one $p(A, M, C, Y)$, having $\epsilon^- \leq \hat{g}^{\pi}_{a_0,a_1}(\mathcal{D}, \beta) \leq \epsilon^+$. Following the original implementation of the work, the optimization algorithm used in the implementation is the Constrained Optimization BY Linear Approximation algorithm (COBYLA). The discussion of different optimization algorithms is left for future work.

Counterfactual Fairness [48]

The implementation of the method presented in [48] can be divided into the three different levels that have been previously introduced. Throughout all of these levels, the predictor function $h_{\beta}(\cdot)$ parametrized by $\hat{\beta}$ is considered to be a linear regression for continuous outcomes and a logistic regression for binary ones.

Starting from the first level, the method consists in creating a predictor that only uses the variables that are non-descendants of A. In this specific example, $\hat{Y} = h_{\hat{\beta}}(C)$, since C is the set of baseline features. As it can be intuitively noticed, albeit Counterfactual Fairness is satisfied, it does not provide trustworthy predictions in terms of accuracy, since most of the information regarding the observed variables is not used. The implementation consists in fitting the following linear regression $Y = \beta_y + \beta_y^c C + \epsilon$.

In the second level, the unobserved variables are modeled as non-deterministic causes of the observable variables descendants of A. This assumption can be graphically represented as Figure 5.2. After estimating the posterior distribution $P(U|X = x^{(i)})$, for every data point (individual) i, J MCMC samples are drawn from this distribution. These samples and the respective values of the baseline features C are used as inputs of the functional $h_{\hat{\beta}}(\cdot)$ used to estimate the fair outcomes $\hat{Y}$. Hence, for every data point $(x^{(i)}, a^{(i)}, y^{(i)})$, d augmented data points $(x^{(i)}, a^{(i)}, y^{(i)}, u_j^{(i)})$ with $j \in \{1, .., J\}$ can be sampled. Then, $\hat{Y}_i = \frac{1}{J} \sum_{j \in \{1,...,J\}} h_{\hat{\beta}}(u_j^{(i)}, c^{(i)})$, following the Monte-Carlo sampling procedure. In the implementation, the posterior distribution is estimated using the probabilistic programming language Stan[3].

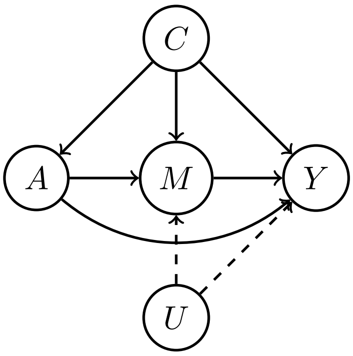


Figure 5.2: Example 1 causal structure for Counterfactual Fairness [48] non-deterministic level

In the third level, a fully deterministic model with latent variables is built, by using additive error terms. This consists in modelling the structural equations as arbitrary functions of the the variables' respective parents plus an error term ϵ_i. For instance, in this scenario, it can be formulated as:

$$\begin{aligned} Y &= \beta_y + \beta_y^a A + \beta_y^m M + \beta_y^c C + \epsilon_Y, \quad \epsilon_Y \sim p(\epsilon_Y) \\ logit(P(M = 1|A, C)) &\sim \beta_m + \beta_m^a A + \beta_m^c C + \epsilon_M, \quad \epsilon_M \sim p(\epsilon_M) \end{aligned}$$

The error terms are estimated as residuals, for instance $\epsilon_M = M - \widehat{M}$. The values $\widehat{M}$ are estimated by fitting a logistic regression on A and C. Similarly to before, the final predictor is a function of the non-descendants of A and of the unobserved variables parents of the descendants of A, except of course for ϵ_Y. For this reason, the additive error terms and C are used as inputs of the predictor $h_{\hat{\beta}}(\cdot)$, where $\hat{\beta}$ are estimated by fitting $Y = h_\beta(\epsilon_M, C) = \beta_y^c C + \beta_y^{\epsilon_M} \epsilon_M$.

Path-Specific Counterfactual Fairness [15]

It has been previously explained that individual-level fairness notions involve certain assumptions on the unobserved space posterior distributions. The first step of the method [15] consists in formulating the causal model by including the potential set of unobserved variables. In this example, the model can be formulated as:

$$\begin{aligned} H_m &\sim p_{\beta_{H_m}}(H_m), \\ logit(P(M = 1|A, C, H_m)) &\sim \beta_m + \beta_m^a A + \beta_m^c C + \beta_m^h H_m, \\ Y &= \beta_y + \beta_y^a A + \beta_y^m M + \beta_y^c C + \epsilon, \end{aligned}$$

[3] Stan is a probabilistic programming language for statistical inference written in C++.

with $\epsilon \sim \mathcal{N}(0,1)$ and where H_m with marginal distribution $p_{\beta_{H_m}}(H_m)$ is the set of unobserved variables parents of M, as represented in Figure 5.3.

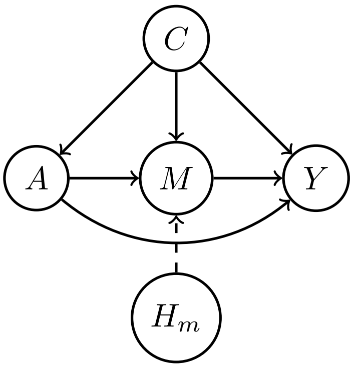


Figure 5.3: Example 1 causal structure for Path-Specific Counterfactual Fairness [15] with latent variable H_m

In order to achieve Path-Specific Counterfactual Fairness with the approach of [15], it is necessary to estimate the posterior probability distributions of the latent variables for the mediators between A and Y given all the observable variables, so in this case $p_{\beta_{H_m}}(H_m|A,C,M)$. This is done by training Variational Auto-Encoders composed by encoders and decoders. The encoder has the purpose of estimating the Gaussian approximation of the probability distribution of the latent space, by outputting the mean and standard deviation of the distribution; the decoder aims at reconstructing the original observed distribution, in this case of M given A, C, H_m. As a matter of fact, the Variational Auto-Encoder formulates the Gaussian approximation $q_{\hat{\mu},\hat{\sigma}}(H_m|A,C,M,L)$ by providing the estimated mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of the distribution.

Once the VAE has been trained, it is possible to estimate the path-specific counterfactually fair outcome $\hat{Y}$ by using a Monte-Carlo approach. As previously mentioned, given an individual, the goal of the method proposed by Chiappa is to estimate the output by 'correcting' the unfair output setting on the unfair paths the values as the ones of the counterfactual version of the selected individual. In this specific example, since there are no mediators on unfair paths, this corresponds to predicting the outcome with the correction made on the sensitive attribute along the direct path from A to Y only. In particular, this is done for the individuals belonging to the 'disadvantaged' group ($A = a_1$). The reasoning behind this method has already been presented in the previous section.

After estimating the Gaussian approximation of posterior distribution of H_m, $q_{\hat{\mu},\hat{\sigma}}(H_m|A,C,M)$, the implementation of Monte-Carlo sampling approach will now be explained for a subject $\{a^{(i)} = a_1, c^{(i)}, m^{(i)}\}$. First, after drawing J MCMC samples $h_{mj}^{(i)} \sim q_{\hat{\mu},\hat{\sigma}}(H_m|a^{(i)}, c^{(i)}, m^{(i)})$ with $j \in \{1, ..., J\}$, then the path-specific counterfactual outcome $\hat{Y}_{PSCF}^{(i)}$ is predicted as:

$$\begin{aligned} m_{PSCFj}^{(i)} &\sim p_{\beta_M}(M|a^{(i)}, c^{(i)}, h_{mj}^{(i)}) \\ \hat{Y}_{PSCF}^{(i)} &= \frac{1}{J}\sum_{i=1}^{J} \mathbb{E}\big[Y|a_0, c^{(i)}, m_{PSCFj}^{(i)}; \beta_Y\big]. \end{aligned}$$

It can be noticed how the fair outcome for an individual i ($\hat{Y}_{PSCF}^{(i)}$) is estimated for the counterfactual value of sensitive attribute a_0. If there had been additional mediators on the unfair paths, it would have been more intuitive to understand why it was necessary to estimate the posterior probability of the latent variables. Indeed, since the latent space is assumed to be unchanged among the real and counterfactual worlds, this is a necessary step to estimate the counterfactual values of the mediators descendants of A. In the next example, a more suitable causal structure will be considered.

5.2.2. Example 2 - Simulated Dataset

In this example, the causal structure represented in Figure 5.4 will be used. The only path that is assumed to be allowed is the one $(ALY)_{\rightarrow}$, while the others are considered to be unfair.
The performance of the different methods will be applied to the following simulated dataset:

$$\begin{aligned} C \sim& Binom(n, 0.7) \\ logit(p(A=1|C)) \sim& -0.5+0.5C \\ logit(p(M=1|A,C)) \sim& -0.5-1C+4A \\ logit(p(L=1|A,C,M)) \sim& -0.5-0.4C-2A+3M \\ Y =& 1+C+5A-4M+3L+\epsilon, \end{aligned}$$

with $\epsilon \sim \mathcal{N}(0,1)$. Now, the different methods will be applied to this simulated dataset.

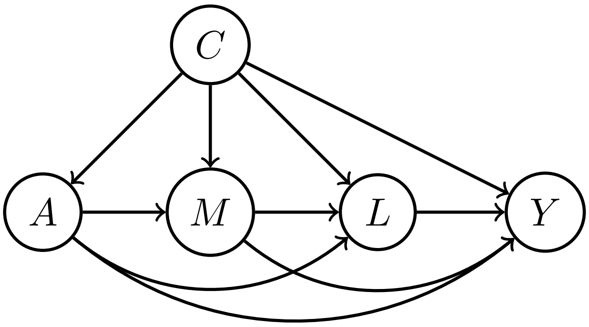


Figure 5.4: Example 2 Causal Structure

Fair Inference On Outcomes [61]

The application of the Fair Inference On Outcomes method to this causal structure is nearly equivalent to the previously presented example.
After parametrically modelling the distributions by making use of logistic regressions for binary variables and linear regressions for continuous ones, the coefficients of the respective fair distribution are estimated as solutions of the constrained optimization problem. Specifically, the set of estimated parameters $\hat{\beta} = \{\hat{\beta}_Y, \hat{\beta}_L, \hat{\beta}_M\}$ will be such that the estimated Path-Specific Effect is bounded. In this causal structure the Path-Specific Effect can be estimated with the plug-in estimator in the following:

$$\begin{aligned} \hat{g}^{\pi}_{a_0,a_1}(\mathcal{D},\hat{\beta}) = \frac{1}{n} \sum_{i\in\{1,\dots,n\}, m\in\mathcal{X}_M} & \Big(\mathbb{E}\big[Y|A=a_1, M=m, C=c_i; \hat{\beta}_Y\big] P(M=m|A=a_1, C=c_i; \hat{\beta}_M) \\ & - \hat{\mathbb{E}}\big[Y|A=a_0, M=m, C=c_i; \hat{\beta}_Y\big] P(M=m|A=a_0, C=c_i; \hat{\beta}_M)\Big) \cdot \\ & \cdot P(L=l|M=m, A=a_0, C=c_i; \hat{\beta}_M), \end{aligned}$$

where $\hat{\beta}_M = \{\hat{\beta}_m, \hat{\beta}^a_m, \hat{\beta}^c_m\}$, $\hat{\beta}_Y = \{\hat{\beta}_y, \hat{\beta}^a_y, \hat{\beta}^c_y, \hat{\beta}^m_y, \hat{\beta}^l_y\}$ and $\hat{\beta}_L = \{\hat{\beta}_l, \hat{\beta}^a_l, \hat{\beta}^c_l, \hat{\beta}^m_l\}$. Hence, the fair distribution defined by $\hat{\beta}$ will be such that the Path-Specific Effect is bounded between ϵ^+ and ϵ^-, while guaranteeing the likelihood function to be maximised.

Counterfactual Fairness [48]

The application of the method [48] to this causal structure is similar to the previously shown example. The main differences in this scenario are the following. In the non-deterministic method, the unobserved latent variable is modeled as a parent of both Y, L, M, as these variables are the descendants of A. Differently, in the deterministic scenarios both the additive error terms ϵ_L, ϵ_M will be estimated, leading the predictor to have the form $h_{\hat{\beta}}(C, \epsilon_L, \epsilon_M) = \hat{\beta}_y^c C + \hat{\beta}_y^{\epsilon_M}\epsilon_M + \hat{\beta}_y^{\epsilon_L}\epsilon_L$.

Path-Specific Counterfactual Fairness [15]

Differently from the other two methods, modifications to the previous implementation need to be made with respect to this causal structure. Specifically, in Example 1 the PSE is equivalent to the NDE, while in this causal structure unfair indirect paths are present.
For the previously explained reasons, the variables' distributions are modelled according to:

$$\begin{aligned}
H_m &\sim p_{\beta_{H_m}}(H_m), \\
H_l &\sim p_{\beta_{H_l}}(H_l) \\
logit(P(M=1|A,C,H_m)) &\sim \beta_m + \beta_m^a A + \beta_m^c C + \beta_m^h H_m, \\
logit(P(L=1|A,C,M,H_l)) &\sim \beta_l + \beta_l^a A + \beta_l^c C + \beta_l^m M + \beta_l^h H_l, \\
Y &= \beta_y + \beta_y^a A + \beta_y^m M + \beta_y^c C + \beta_y^m M + \epsilon,
\end{aligned}$$

where H_m and H_l are the unobserved variables respectively parents of M and L, as it can be noticed in Figure 5.5 and $\epsilon \sim \mathcal{N}(0,1)$.
Variational Auto-Encoders are used for estimating the Gaussian approximations of $p_{\beta_{H_m}}(H_m|A,C,M,L)$ and $p_{\beta_{H_l}}(H_l|A,C,M,L)$, referred to as $q_{\hat{\mu}_m,\hat{\sigma}_m}(H_m|A,C,M,L)$ and $q_{\hat{\mu}_l,\hat{\sigma}_l}(H_l|A,C,M,L)$. Thereafter, the Monte-Carlo approach that was previously introduced is used for estimating the path-specific counterfactually fair outcomes of the individuals belonging to the potentially discriminated group ($A = a_1$). Specifically, given an individual i with $\{a^n = a_1, c^{(i)}, m^{(i)}, l^{(i)}\}$, the aim of the method is to predict $\hat{Y}^{(i)}_{PSCF}$ as the outcome the subject would have obtained by having the sensitive attribute equal to the complementary value a_0 along the unfair paths, hence every path except for $(ALY)_{\rightarrow}$. Like in the previous example, after drawing J samples $h^{(i)}_{mj} \sim q_{\hat{\mu}_m,\hat{\sigma}_m}(H_m|a^n, c^n, m^n, l^n)$ and $h^{(i)}_{lj} \sim q_{\hat{\mu}_l,\hat{\sigma}_l}(H_l|a^n, c^n, m^n, l^n)$ with $j \in \{1, ..., J\}$, then the output is estimated for each of them in the following way:

$$\begin{aligned}
m^{(i)}_{PSCFj} &\sim p_{\beta_M}(M|a_0, c^{(i)}, h^{(i)}_{mj}) \\
l^{(i)}_{PSCFj} &\sim p_{\beta_L}(L|a^{(i)}, c^{(i)}, m^{(i)}_{PSCFj}, h^{(i)}_{lj}) \\
\hat{Y}^{(i)}_{PSCF} &= \frac{1}{J}\sum_{i=1}^{J}\mathbb{E}[Y|a_0, c^{(i)}, l^{(i)}_{PSCFj}, m^{(i)}_{PSCFj}; \beta_Y].
\end{aligned}$$

It can be noticed that, since M and Y are connected to A through unfair paths, $m^{(i)}_{PSCFj}$ and $\hat{Y}^{(i)}_{PSCF}$ are estimated by inputting the different value of the sensitive attribute (a_0). Differently, since L is along a fair path, the input value corresponding to the sensitive attribute remains unchanged, without corrections made on the conditioning values. It is now more intuitive to understand the necessity of estimating the posterior distribution of the latent variables. Indeed, the counterfactual versions of the mediators and the observed one share the same values of unobserved variables. This means that, independently on the fairness or unfairness of the paths connecting a mediator, the sampled values of the unobserved variables will remain unchanged.

5.3. Results

In this section, the results of the application of the state-of-the-art methods to the datasets simulated from the graphical structures of Figure 5.1 and Figure 5.4 (Example 1 and Example 2 respectively) will be presented. The performances of the methods will be compared in terms of fairness and accuracy. Regarding the dataset, a sample of 6000 data points was simulated, among which 80% was used for

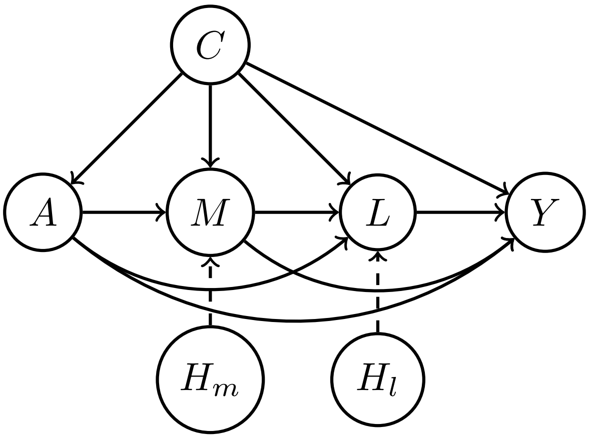


Figure 5.5: Example 2 causal structure for Path-Specific Counterfactual Fairness [15] with latent variables H_m and H_l

training and the rest for testing and validating. The results show the performance of the methods on the test set, after being trained on the training set.
The methods are compared according to three different metrics. In prediction tasks, with continuous outcomes, the Mean-Squared Error (MSE) is used for measuring accuracy. The MSE is estimated as $\frac{1}{n}\sum_i (Y_i - \hat{Y}_i)^2$ with $i \in \{1, ..., n\}$ and measures the accuracy of the prediction with the average squared difference between the the observed values Y_i and the outputs of the methods $\hat{Y}_i$. PSE and Average Treatment Effect (ATE) are used for measuring the population-level fairness of the predictions. It has been previously motivated why the PSE is relevant for this work; evaluating the ATE of the predictions could lead to further interesting insights regarding the different methods' characteristics.
Albeit both accuracy and fairness of predictions are measured, they do not have the same relevance in the evaluation of the methods. As mentioned in Chapter 2, in the considered setting, the focus of this work is set on defining a method that returns a fair output.Even though the fairness of the predictions represents the key measure to consider in this work, it is also important to evaluate how the method performs in the prediction task itself with accuracy metrics.

The compared methods are Fair Inference on Outcomes (FIO [61]), Path-Specific Counterfactual Fairness (PSCF [15]), Counterfactual Fairness ([48]). Furthermore, other baseline methods will be applied to the same dataset, as the comparison with their performance will support the evaluation of the methods. These baseline methods are the Unconstrained Method and the Fairness Through Unawareness Method; the former aims at predicting the output with maximum accuracy, without any constraints on fairness set, while the latter simply 'ignores' the sensitive attribute in the prediction/classification task (refer to Section 2.3.3). The performance of the Fairness through Unawareness method is included in the comparison, confirming that this non-causality based metric is not be suitable for detecting unfairness in causal models presenting mediators that are proxies for discrimination.

In the next sections, the results of the application of the methods on Example 1 and Example 2 simulated datasets will be presented. Furthermore, the performance of the plug-in estimator will be discussed.

5.3.1. Results Example 1

First of all, the results of the methods applied to the dataset simulated from the causal model of Example 1 will be analysed. The application of the different methods is straightforward in this scenario, since it was shown that the PSE matches the NDE.

In Table 5.1, the results corresponding to PSCF, FIO, Counterfactual Fairness, Unconstrained and Unaware methods are presented. By analysing these results, it can be immediately noticed that the method with highest accuracy is the Unconstrained one (lowest MSE of 0.8). As expected, this leads to high (in absolute value) PSE and ATE. Indeed, no fairness constraints are enforced.

The results of the Fairness Through Unawareness method (ignoring sensitive attribute) present a PSE value close to 0. This might seem counter-intuitive, as in the previous sections it was argued that this method is not suitable for guaranteeing unfairness since some mediators in the dataset could act as proxies for discrimination. Nevertheless, since in this dataset the only unfair path is the direct one from A to Y, all the mediators M lie on an allowed path, meaning that they should be rightfully considered in the prediction. However, this does not hold for the ATE; indeed, by using M in the prediction with no fairness constraints, the Average Treatment Effect on all the paths is not null.

Both the methods FIO and PSCF succeed at ensuring nearly null Path-Specific Effect, meaning that these methods achieve the path-specific fairness constraint. The higher Average Treatment Effect encountered in FIO can be positively interpreted as the consequence of enforcing a high causal effect through the allowed paths. Indeed, if the effect along the unfair paths is almost zero while the one along the entire set of paths is high, it can be deducted that the allowed paths are being sufficiently leveraged in order to predict $\hat{Y}$ with high accuracy. Since PSCF is focused on individual-level fairness, due to the increased strength of the fairness constraints, the accuracy of the predictions is lower (higher MSE compared to FIO). This is the consequence of a high accuracy for data-points not belonging to the sensitive group and low accuracy for the potentially discriminated group. As a matter of fact, as mentioned earlier in the method introduction, the outcomes of the data-points i that are not part of the sensitive group ($A_i = 0$) are equivalent to the observed ones, meaning that $Y_i = \hat{Y}_i$. Differently, the predicted value $\hat{Y}_i$ of potentially discriminated data-points i with $A_i = 1$ is corrected by using the respective counterfactual values on the unfair paths. In this example, the Maximum Mean Discrepancy (MMD) penalty term hyper-parameter is set to $\beta = 0$; since no mediators belong to unfair paths, the choice of this value is not relevant for the results.

The Counterfactual Fairness method aims at ensuring individual-level fairness, without taking into account the fairness/unfairness of the different paths. As previously discussed, one of the main disadvantages of this approach is discarding relevant information that is transmitted through the fair paths and that could improve the predictions' accuracy. Differently, when analysing the fairness metrics, the method's performance is successful. It has been proven in Section 4.3.0.2 that the condition enforced in the Counterfactual Fairness method in [48] is a stricter constraint on the predicted outcome $\hat{Y}$ compared to Counterfactual Fairness itself. Hence, by ensuring counterfactual fairness on every individual, also population-level metrics such as PSE and ATE are met, leading these metrics to be nearly null. The non-deterministic level and the level 1 achieve fairness with the drawback of the relevantly low accuracy. Overall, since the assumptions made in the different levels of the Counterfactual Fairness method are met in this graphical structure, as it was simulated with an additive error model with independent noise terms, the performance of the method in terms of fairness is successful.

Method	MSE	PSE	ATE
Unconstrained Method	0.8	-11	-9.3
Unaware Method	4.8	0.0001	-0.9
Constrained FIO [61]	9.0	-0.02	-1.8
PSCF [15]	11.1	0.0001	1.1
Counterfactual Fairness [48] - deterministic level	18.8	0.001	0.002
Counterfactual Fairness [48] - non deterministic level	21.1	0.001	0.0003
Counterfactual Fairness [48] - level 1	25.1	0.0001	0.00001

Table 5.1: Results Example 1 causal structure

In general, it can be useful to compare the performance of the method on different causal structures, specifically by adding unmeasured variables to the model. The same graphical structure of Figure 4.2 is used, but with an additional unobserved variable with edges towards A and Y. In the literature, this is referred as a common unobserved confounder between A and Y. The simulation procedure is presented in the Appendix J.1.5. In this causal structure, the path-specific effect is still identifiable with the edge g-formula, so the previously introduced metrics can be compared again. It can indeed interesting to analyse the performance if one or more assumptions of a model are violated.

In Table 5.2, the performance per each method is presented. Analysing the results, it can be noticed that both the Unconstrained and Unaware methods fail at ensuring fairness of the outcomes $\hat{Y}$, as expected. Furthermore, the additional unobserved confounding leads to a loss in accuracy compared to the previous example. Analysing the Constrained FIO method, this outperforms the other state-of-the-art methods in terms of accuracy, still satisfying the constraint on the PSE. This is also the case for the method PSCF, since the graphical structure satisfies the underlying assumptions of the method; indeed, the unobserved variable is not a parent of the mediator M. Nevertheless, the accuracy is relevantly lower compared to the other methods, due to a high MSE in the groups with $A = 1$. This loss in accuracy is not due to the additional unobserved confounding, but only depends on the simulations procedure.
It can be noticed that the non-deterministic level of Counterfactual Fairness performs worse compared to the example with no unobserved variables. Indeed, the assumptions required by the non-deterministic level of Counterfactual Fairness are not satisfied in this example. The unique unobserved variable modelled by the method (represented in Figure 5.2) is assumed to be independent from A, which is not the case in this graphical structure, as the unobserved variable parent of Y is also parent of A. Nevertheless, the performance is promising in terms of fairness and accuracy.

Method	MSE	PSE	ATE
Unconstrained Method	3.3	-9.3	-7.5
Unaware Method	20.4	0	-0.4
Constrained FIO [61]	8.9	-0.01	-1.8
PSCF [15]	42.2	0.0001	1.1
Counterfactual Fairness [48] - deterministic method	11.8	0.001	0.002
Counterfactual Fairness [48] - non deterministic method	15.9	-0.07	-0.08

Table 5.2: Results Example 1 causal structure with additional unobserved variable parent of A and Y

5.3.2. Results Example 2

In Table 5.3, the results from the application of the methods to the datasets generated by the causal model in Example 2 are represented. In this scenario, the PSCF method is proposed for different values of the MMD penalty term hyper-parameter β, with both $\beta = 0$ and $\beta = 100$.
It can be noticed that in this Table the results reflect similar patterns as the previous Table 5.1. Differently from the previous causal structure, the PSE here is not equivalent to the NDE, meaning that the Fairness Through Unawareness method in this example does not ensure path-specific fairness, as expected. Similarly to before, the Constrained FIO method is the best performing one in terms of balance between accuracy and path-specific fairness. Indeed, while having PSE $= -0.07$ on the test set, the MSE is also relevantly lower compared to the other state-of-the-art methods. In this example, due to the absence of additional unobserved variables in the causal model, not relevant change can be noticed among the performance of PSCF with different values of β. Analysing the Counterfactual Fairness method instead, the accuracy of the method increases by varying the level of the method, hence by increasing the needed assumptions. The reason behind the better performance in terms of accuracy of the deterministic model compared to PSCF is that the stricter assumptions are met, specifically in this example the additive noise terms were indeed independent among each others. Again, the fairness metrics for this method are nearly zero.

In the next paragraph, a simulation procedure with additional unobserved variables will be considered. The same graphical structure of Figure 5.4 is used, with the addition of different unobserved variables. Specifically, in Table 5.4, the results corresponding to the causal structure with an unobserved variable parent of both M and Y is considered. Differently, in Table 5.5, two unobserved variables are considered: a parent of M and Y and a parent of A and L. The details regarding the simulation procedure

Method	MSE	PSE	ATE
Unconstrained Method	1.0	-3.4	-3.1
Unaware Method	14.4	-0.2	-0.6
Constrained FIO [61]	1.7	-0.07	0.7
$PSCF_{\beta=0}$ [15]	8.7	0.06	0.05
$PSCF_{\beta=100}$ [15]	8.8	0.06	0.06
Counterfactual Fairness [48] - deterministic	3.6	0.01	0.001
Counterfactual Fairness [48] - non deterministic	12.8	0.004	-0.001
Counterfactual Fairness [48] - level 1	14.9	0.00001	0.00001

Table 5.3: Results Example 2 Causal Structure

can be found in the Appendix J.2.6. In both of these scenarios, unobserved confounding of the couples of variables (M, Y) or (L, A) is involved. It is important to notice that the Path-Specific Effect is identifiable in both of these graphical structures, according to the recanting district criterion.

In Table 5.4 no relevant difference can be found compared to the previous Table corresponding to the same graphical structure without the additional unobserved variables. This probably happens because the additional unobserved variable is not a parent of A too. Nevertheless, the pattern changes when considering the results in Table 5.5. First of all, it can be noticed that the PSCF method improves in terms of fairness when considering $\beta = 100$ compared to $\beta = 0$, since the former penalizes the dependency between the unobserved variables used for the predictions and A. Similarly, the Counterfactual Fairness method performs less fairly compared to the previous scenarios, even though path-specific fairness is still achieved. Similarly to before, this probably happens due to the violation of the assumptions made in the deterministic level of the method.

Method	MSE	PSE	ATE
Unconstrained Method	4.2	-1.2	-1.3
Unaware Method	4.7	0.02	-0.06
Constrained FIO [61]	4.4	-0.04	-0.2
$PSCF_{\beta=0}$ [15]	4.8	-0.002	-0.02
$PSCF_{\beta=100}$ [15]	4.8	-0.003	-0.02
Counterfactual Fairness [48] - deterministic	5.7	0.0001	0.001

Table 5.4: Results Example 2 Causal Structure Unobserved MY

Method	MSE	PSE	ATE
Unconstrained Method	4.2	-1.3	-1.4
Unaware Method	5.4	0.02	-0.3
Constrained FIO [61]	4.5	-0.04	0.4
$PSCF_{\beta=0}$ [15]	4.9	0.09	0.02
$PSCF_{\beta=100}$ [15]	5.0	0.008	0.02
Counterfactual Fairness [48] - deterministic	5.4	0.01	0.02

Table 5.5: Results Example 2 Causal Structure Unobserved MY, LA

5.3.3. Performance Plug-in Estimator

In the implementation of the FIO method, the semi-parametric plug-in estimator for the edge g-formula was used. In the next paragraphs, the performance of this estimator will be evaluated, empirically showing that it is suitable for the goal of this project. In order to analyse the performance of the plug-in estimator, an empirical study has been done. The plug-in estimator of the edge g-formula will be applied to different causal structures; the next paragraph will focus on evaluating the characteristics of the estimator.

The simulation procedure corresponding to the graphical structure of Example 1 is considered, specifically the simulation procedure introduced in J.1.3. The performance of the estimator on different causal models can be found in the Appendix E. Before moving into the analysis of the estimator, the steps for deriving its true sampling distribution is explained. Given the simulation procedure, an arbitrary number d of datasets composed by n data-points are simulated. By using the plug-in estimator for the PSE quantity of each dataset, the true sampling distribution of this estimate can be derived. The standard deviation of the sampling distribution (standard error) can be useful for quantifying the variability of the estimator. It can hence be used as a measure of precision, describing the variability of the estimator among the experiments.

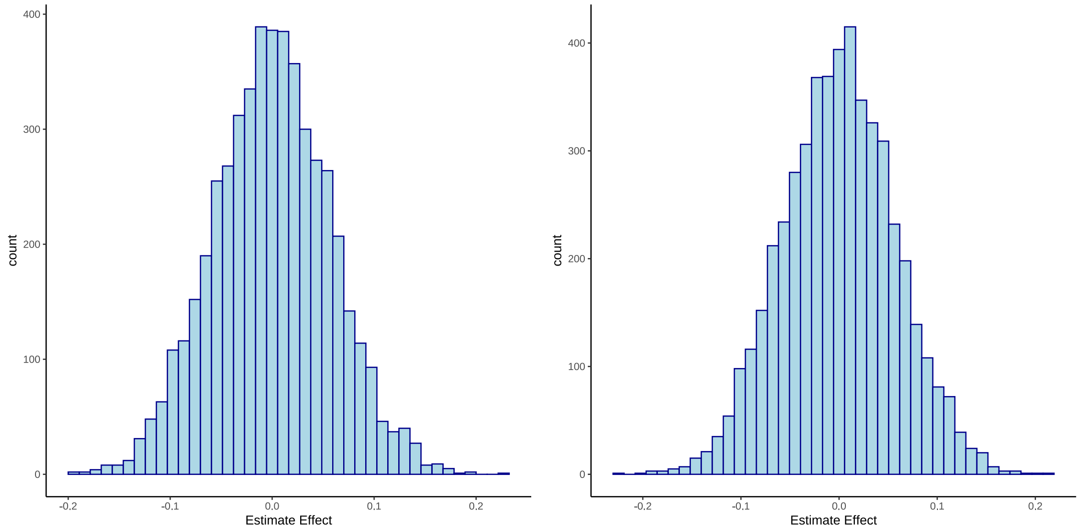
As a first step, the consistency of the estimator will be analysed. The true sampling distribution is derived for $d = 1000$ datasets of different sizes n. By varying the sample size, it is indeed possible to check the consistency of the estimator in a simulation procedure. Indeed, in Table 5.6 the Standard Deviation of the true sampling distribution for each of these simulation procedures by sample size n is presented. It can be noticed that the Standard Deviation varies as expected, given by the square root of n in the denominator of its mathematical formulation. Furthermore, the consistency of the estimator is supported by the convergence to zero of the standard error for the simulation procedure increasing n.

n	Standard Deviation
100	0.42
1000	0.13
4000	0.06
8000	0.04
100000	0.01
1000000	0.004

Table 5.6: Plug-in Estimator Performance on datasets with n data-points

In order to evaluate the performance of the estimator, a further analysis on the empirical distribution of the estimator can be made. The empirical distribution is derived from the application of bootstrap sampling with replacement to a given dataset $\mathcal{D}_n$. The plug-in estimator is applied to each bootstrap sample, hence composing the bootstrap distribution. The bootstrap approximates the shape of the true sampling distribution by simulating replicate experiments based on the observed data $\mathcal{D}_n$. By comparing the variability of this distribution with the true sampling distribution, relevant information regarding the performance of the estimator can be deduced. In order to simplify the comparison among the two distributions, the analysis is made on the shifted ones, centering the distributions in 0. Since the bootstrap distribution will be centered around the estimate and not the true value, unlike the true sampling distribution, these distributions are shifted to wipe out the initial bias.

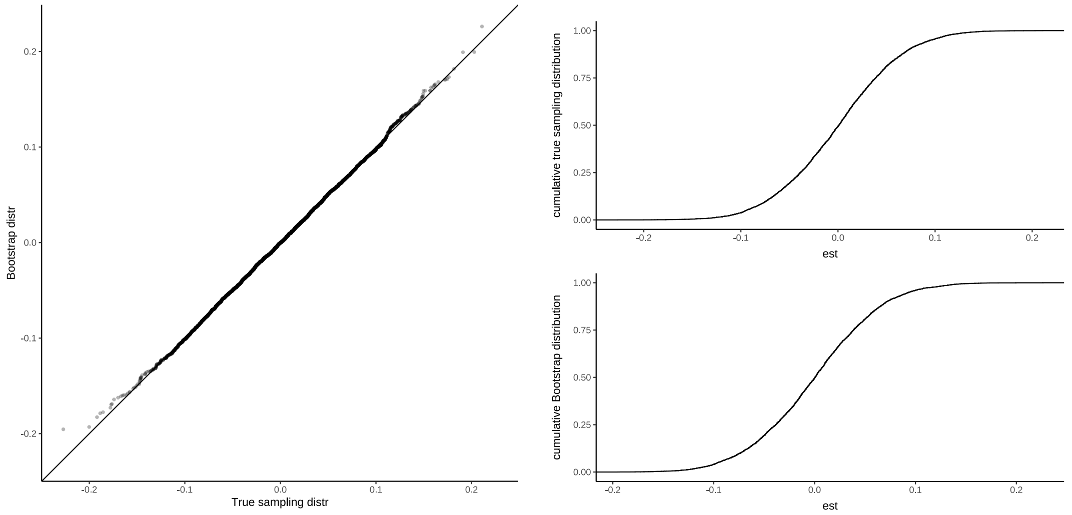
In Figures 5.6a, 5.6b the two shifted distributions are represented. The variances of the two distributions are compared. The standard deviation of the bootstrap and true sampling distributions are highly comparable, respectively 0.05726 and 0.05737. Furthermore, it might be interesting to compare the two shifted distributions in terms of quantiles and cumulative distributions in Figures 5.7a, 5.7b.
By the similar variability of the two distributions, it can be deduced that the inherent variance of the estimator is not high. Indeed, the noise in the bootstrap distribution is comparable to the inherent variance of the estimator without bootstrapping.
Overall, the semi-parametric plug-in estimator seems suitable for the tasks of this project. In Appendix E these results are also supported by analysing the estimator on another causal structure.



(a) Shifted true sampling distribution

(b) Shifted bootstrap distribution

Figure 5.6: Original and Shifted Distribution



(a) Quantile-Quantile plot Bootstrap distribution and True Sampling Distribution

(b) Cumulative bootstrap and true sampling distribution

5.4. Conclusions

Overall, the experiments were really useful for evaluating the performance of the selected state-of-the-art methods and for comparing them among each others. Different metrics have been used, both portraying the fairness and accuracy of the prediction. The methods performed as expected, as the strengths and weaknesses underlined in the methodology section were also portrayed in the results. Furthermore, the peculiar accuracy and fairness trends of the methods were encountered in all the graphical structures, specifically analysing the *trade-off* between accuracy and fairness. Particularly interesting was the variation of the graphical structures with the addition of unobserved variables, as the strength of the assumptions used in the different methods could be emphasized.

Analysing all of the state-of-the-art methods, different conclusion can be made regarding their performances. These experiments gave an additional confirmation that the Fairness through Unawareness method is not suitable for detecting and correcting unfairness in scenarios with mediators along unfair paths. The Counterfactual Fairness method and Path-Specific Counterfactual Fairness method both involve the estimation of the latent space of unobserved variables, since they focus on individual-level fairness. This has shown to possibly lead to instability due to the assumptions made on the unobserved variables. In both of the methods it is indeed necessary to ensure that the estimated unobserved variables are independent from the sensitive attribute. Differently, the population-focused FIO method does not involve assumptions on the unobserved variables. For this reason, this method is actually more interpretable and intuitive. Furthermore, the PSE is bounded in a small interval and the accuracy is higher compared to the other methods. Albeit its inability of portraying individual-level fairness, this method seems highly promising. Similarly, the path-specific individual-level fairness achieved by Path-Specific Counterfactual Fairness method has a lot of potential in the fairness setting in combining path-specific fairness and individual-level fairness.

At the end of the chapter, the performance of the semi-parametric plug-in estimator used in the implementation of the FIO method for estimating the PSE was analysed. An empirical study of the properties of this estimator shows that it is indeed suitable for the tasks of this method.

6

Fairness through conditional Path-Specific Effect

In this chapter, the newly proposed method will be presented. With this method, the authors aim at proposing an alternative version of Fair Inference on Outcomes method (FIO) [61] by attempting to bridge the gap with individual-level fairness, represented by Path-Specific Counterfactual Fairness method (PSCF) [15]. Both of these methods have been presented in Chapter 5. It will be shown that, compared to PSCF, this novel method has the advantage of leveraging the identifiability conditions of Path-Specific Effect via the edge g-formula (Definition 4.2.1), providing nearly individually-fair predictions by avoiding assumptions on the posterior distributions of the unobserved variables. This newly proposed approach is based upon conditional Path-Specific Effects; it will therefore be referred to as **Conditional Path-Specific Fairness (CPSF)**.

This chapter will be structured in the following way. First, conditional Path-Specific Effects will be introduced. In particular, the motivation regarding the focus on individual-level fairness will be given, along with the intuition, definition and identifiability conditions of conditional Path-Specific Effect. Moreover, the novel method will be introduced and compared to the original versions of FIO [61] and PSCF [15].

6.1. Conditional Path-Specific Effect

The conditional Path-Specific Effect can be a useful tool for estimating and evaluating the Path-Specific Effect on distinct subgroups of individuals. Albeit the concept of conditional Path-Specific Effect is closely related to the overall Path-Specific Effect, relevant focus should be put on the identifiability conditions of such effects. These topics will be extensively covered in the next two sections.

6.1.1. Motivation of Conditional Interventional Distributions

Comparing population-level and individual-level fairness metrics, different strengths and weaknesses can be assessed in both. It was previously explained that effects such as the Path-Specific Effect are population-level fairness metrics, as they consist in defining the effects among the entire population. By using the Path-Specific Effect as a fairness metric, the effect along certain unfair paths is estimated among the entire population.Intuitively, when addressing the fairness topic, a constraint on the PSE consists in limiting the significance of this effect when considering as treatment a sensitive attribute and as the outcome a decision that might lead to potential discrimination.
Population-level metrics have been questioned for not addressing individual-level [48]. As a matter of fact, even though a model is fair at a population level with respect to a sensitive attribute, it could

still happen that among subjects with the same sensitive attribute, half of the individuals is negatively discriminated and the other half is positively discriminated [48].
Population-level fairness metrics have relevant advantages, based on the overall identification theory that has been explored in the past years. If the respective identifiability conditions are met, it is then possible to apply the edge g-formula and to identify this effect without considering further assumptions on the latent space. It was shown that this is instead a necessary step when considering individual-level fairness.
In this chapter, conditional Path-Specific Effect will be introduced as a metric that bridges individual-level and population-level fairness metrics.

The identification of conditional Path-Specific Effect is useful for estimating the PSE on certain subgroups of the population [54]. Intuitively, if a given dataset is fair towards the different subgroups of the population, then it will also be fair towards the entire population. Nevertheless, it is important to notice a fair model towards the entire population does not imply fairness among the different demographic subgroups. These subgroups can be in general defined by post-treatment variables or pre-treatment variables, corresponding to descendants or not descendants of the treatment attribute respectively. In the context of fairness, the focus is set on the baseline features, usually representing the demographic attributes characterized by specific values of the pre-treatment covariate C. In Appendix F, an example describing the difference between the use of a certain covariate as a conditioning attribute and as a treatment attribute will be given.

The concept of conditional effect can be seen as closely related to the individual-level fairness approach. In most of the real-world scenarios, by analysing population-level fairness metrics separately among the different demographic groups, it is possible to have a clearer idea on how to interpret and evaluate fairness at an individual level. Indeed, it can often be assumed that the different individuals can be satisfactorily represented by the specific demographic subgroup they are part of.
Overall, ensuring that the dataset is fair towards the subgroups of the population can be seen as a step towards individual-level fairness, without involving the complexity that characterizes the estimation of individuals' counterfactuals.

In this paragraph, an example regarding the relevance of evaluating fairness on different demographic subgroups is given. A hiring process with the gender A as the sensitive attribute can be considered. In this scenario, the purpose of the research would be to evaluate the Path-Specific Effect of A on the hiring decision Y given certain unfair paths. Albeit the estimated effect among the entire population is assumed to be low, it might still be interesting to separately evaluate the effect that A has on the hiring decision in younger generations and older generations. When assessing the fairness of the dataset using population-level metrics, the goal is usually to prove that this effect is on average very low. Nevertheless, it might happen that the younger generations are actually positively discriminated, while the older generations negatively discriminated. In other words, this means that the effect of the gender on the hiring decision among younger generations leads to more beneficial results for women rather than for men. Differently, when analysing older generations applicants, it means that the causal effect can be interpreted as a negative discrimination (lack of incentive) on female applicants. In this context, this phenomenon could be interpreted by analysing an hiring scenario in which younger females, usually applying for entry-level positions, are more encouraged to apply compared to older females, typically looking for higher-level positions involving more responsibilities. This is actually realistic, as nowadays companies often need to ensure a specific yearly percentage of overall hired women (e.g. female quotas). Unfortunately, it often happens that companies tend to advantage younger female applicants compared to older ones because of the range of responsibilities involved with the requested jobs.

The choice of the conditioning subgroups is a widely-discussed question in the literature, as the estimation of causal effects on subgroups of the population is a relevant topic in many applications ([5], [91]). Among the proposed methods, most of them make use of the Propensity Score ([27], [58], [4], [7], [47]). The Propensity Score (PS) is defined as the probability of being assigned to a specific treatment value conditional on all the other relevant covariates. Different methods can be implemented based upon the propensity score estimation, such as PS matching, PS stratification, Inverse Probability Weighting using PS [27]. The propensity score is particularly useful in applications such as personalised medicine,

in which it is relevant to evaluate the probability of treatment assignment depending on the other covariates values. In the setting of the research of fairness of Machine Learning, approaches based upon the propensity score are not necessarily useful, as the treatment variable is actually a demographic attribute, such as gender or race; hence, this means that in most of these scenarios the estimation of the propensity score is not relevant for the scope of the research. Furthermore, in other methods in the literature, the subgroups of the population are chosen by defining different scores that summarize relevant information regarding the individuals [90]. In various applied contexts, it can also be useful to analyse the relevance of the different covariates on the outcome and to choose the best conditioning set by using covariate selection methods ([103], [73], [51]).
The choice of the conditioning set is not part of the focus of this work. This aspect of the conditional PSE is left for future work.

The identification of conditional interventional distributions is a common identifiability question in the causality research. This has been treated in the past years by many authors, such as Pearl in [69]. A complete algorithm based on graphical conditions aiming at identifying conditional node-interventional distributions has been developed in [84], referred to as Compete Identification Algorithm for Conditional Effects (IDC). Following this work, the authors of [54] developed a framework based on potential outcomes with the goal of applying these results to nested counterfactuals, such as the ones involved in path-specific interventions; this last algorithm is called Compete Identification Algorithm for Conditional Path-Specific Effects (PS-IDC).
In Section 6.1.2 the algorithm for identifying conditional node-interventional distributions proposed in [84] (IDC) will be presented. In particular, it will be shown how under certain graphical conditions it is possible to use the Rule 2 of do-calculus in order to identify conditional interventional distributions. Next, in Section 6.1.3, the application of this method to conditional path-specific interventional distributions will be described, following the approach of [54] (PS-IDC). These will be generalized in order to analyse the identification of conditional path-specific interventional distributions on any arbitrary graphical structure. Furthermore, in both of these sections, the application to the research on fairness of Machine Learning will be taken into consideration.

6.1.2. Do-Calculus for the Identifiability of Conditional Interventional Distribution

The method introduced in [84] has the goal of estimating the conditional interventional probability $P(Y(a) = y|C = c)$, where A, C, Y are disjoint and arbitrary sets of nodes in a causal graph $\mathcal{G}$ and $a \in \mathcal{X}_A, y \in \mathcal{X}_Y, c \in \mathcal{X}_C$. This algorithm is applicable to any graphical causal structure. Nevertheless, as previously mentioned, the interest of this project is focused conditional path-specific interventional distributions. Thus, only the results that will be useful for introducing the algorithm for the identification of conditional path-specific interventional distributions (PS-IDC) will be presented.
The algorithm IDC returns the identification formula of the needed conditional interventional distribution in terms of observational quantities. The algorithm is able to succeed in the identification of these distributions, as long as the identifiability conditions are met.

In this section, the algorithm IDC will not be presented in detail. Indeed, the formalization of the identification of the conditional path-specific effect will already be covered in detail in Section 6.1.3. As shown in the previous chapters, node-interventional distributions are a specific case of path-specific ones. Hence, the identification algorithm PS-IDC will also cover the scenarios encountered in IDC. The scope of this section is instead showing that the Rule 2 of do-calculus (Formula 3.19) can be leveraged for identifying conditional interventional distributions, as shown in the algorithm IDC [84]. For this reason, in this section the algorithm will be simplified by taking into account causal structures that completely satisfy the condition of Rule 2 of do-calculus, which are also the typical causal structures of the datasets analysed so far in the report. This is a strong assumption that will be alleviated in the next section on conditional path-specific interventional distributions, by considering arbitrary graphical structures.

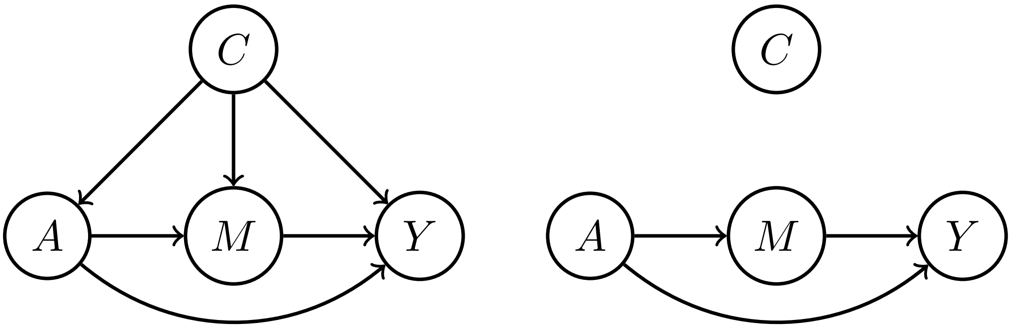
Typically, in the fairness research setting, the demographic conditioning set C is a set of root nodes

(Definition 3.1.0.7). In the previous chapters it has been explained that the scope of this project is partly set on analysing the effect of the sensitive attribute A on the outcome Y. The motivation for focusing on conditional interventional distribution rose from the interest of estimating causal effects on specific subgroups of the population. These subgroups are defined by the covariate C, including certain demographic information. Generally, the covariate C is a non-descendant of A, specifically a set of root nodes. Indeed, if the demographic attribute is a mediator between A and Y, then it is important to inspect whether this mediator is considered to be along a fair or unfair path; this means that descendants of A are generally not suitable conditioning sets, as previously discussed.
Even though it might seem that the necessary condition corresponding to Rule 2 is a strong assumption, this holds in the previously analysed graphical structures. It will be explained why this condition holds in these scenarios. It is important to notice that these graphical structures are very specific and simplified, with the purpose of showing how to apply the Rule 2 for the identification of conditional interventional distributions. Nevertheless, the algorithm IDC can be applied to any arbitrary causal structure.

The Rule 2 of do-calculus can be leveraged for identifying conditional interventional distributions, as long as the respective graphical condition is satisfied. Referring to the notation used in Rule 2 in Formula 3.19 ($P(Y(a,z) = y|D = d) = P(Y(a) = y|D = d, Z = z)$ if $(Y \perp\!\!\!\perp Z|A, D)_{\mathcal{G}_{\overline{A}\underline{Z}}}$), the following is taken into consideration. D can be defined as the null set $D = \{\emptyset\}$ and Z as the conditioning variable C. Following these assignments, the Rule 2 of do-calculus corresponds to

$$P(Y(a,c) = y) = P(Y(a) = y|C = c) \text{ if } (Y \perp\!\!\!\perp C|A)_{\mathcal{G}_{\overline{A}\underline{C}}}.$$

In the above condition, the causal graph $\mathcal{G}_{\overline{A}\underline{C}}$ corresponds to the original causal graph with the arrows emitted by C and the arrows entering A deleted. In the mentioned graphical structures, the set of demographic covariates C is a root node set. Following the previously mentioned graphical assumptions, since the demographic variables set C does not have any edge entering from other observed covariates, the condition $(Y \perp\!\!\!\perp C|A)_{\mathcal{G}_{\overline{A}\underline{C}}}$ holds, assuming that the disturbance terms are mutually independent. In general, it will then be possible to identify $P(Y(a) = y|C = C)$ as $P(Y(a,c) = y)$, according to Rule 2. As an example, Figure 6.1b can be referred to, corresponding to the original causal graph in Figure 6.1a. Hence, in this causal graph, $P(Y(a) = y|C = c) = P(Y(a,c) = y) = \sum_{m \in \mathcal{X}_M} P(Y = y|A = a, M = m, C = c)P(M = m|A = a, C = c)$.



(a) Causal Graph $\mathcal{G}$

(b) Example Corresponding Causal Graph $\mathcal{G}_{\overline{A}\underline{C}}$

Figure 6.1: Graphical structure $\mathcal{G}$ with respective representation of $\mathcal{G}_{\overline{A}\underline{C}}$

In conclusion, when analysing causal graphs with mutually independent error terms and fulfilling certain structural requirements, it is possible to use Rule 2 of do-calculus to simplify the conditional interventional distribution into a joint interventional distribution.
In the above paragraphs, a big assumption was made regarding the causal structure of the variables. Even though this structure is encountered in the commonly analysed datasets in the context of fairness of ML (COMPAS, UCI Adult, ...), it is important to generalise the identifiability of interventional distribution to any causal structure. This will further be explored in the next section analysing conditional path-specific interventions. All of the results presented in the next section can be equivalently proven also for simple node interventions, thus for IDC.

6.1.3. Conditional Path-Specific Interventional Distribution

In the previous section, it was shown how do-calculus can be used for identifying conditional interventional distributions. This section will focus instead on the more generic conditional path-specific interventional quantities.
Through IDC, in [84] it was shown that it is possible to identify conditioned interventional distributions by making use of the Rule 2 of do-calculus, simplifying it into a joint interventional distribution. The authors of [54] leverage these results in order to build an algorithm (PS-IDC) that identifies conditional path-specific interventional distributions.

Notation and Definitions

The relevant notation and definitions that will be useful in this chapter will now be presented. The concept of potential outcomes has already been introduced in the previous chapters; in this paragraph, these will be defined following the notation in [54].
A causal graph $\mathcal{G}$ with nodes $V = \{V_1, ..., V_n\}$ is given. The parents of each V_i are denoted as Pa_i. It is possible to define the set of potential outcomes $\mathbb{V}$ in the following way:

$$\mathbb{V} \equiv \{V_i(pa_i)|i \in \{1, ..., k\}, pa_i \in \mathcal{X}_{Pa_i}\}, \tag{6.1}$$

where $\mathcal{X}_{Pa_i}$ is the state space of the parents of V_i, Pa_i. In the above notation, the single potential outcomes $V_i(a)$ can be defined by using a recursive substitution, similarly to the one used in [83] for edge interventions:

$$V_i(a) \equiv V_i(a_{A \cap Pa_i}, \{V_j(a)|V_j \in Pa_i \backslash A\}). \tag{6.2}$$

Here, $V_i(a)$ is the response V_i after the intervention on the attribute A as $a \in \mathcal{X}_A$. Based on recursive substitution, the parents of V_i that belong to A are set as a, meanwhile all the parents of V_i that are not in A are recursively defined as the values they would have if A was set to a. It will be useful to notice that the notation $a_{A \cap Pa_i}$ consists indeed in setting the set $A \cap Pa_i$ as a. For instance, in Figure 4.2, $Y(a) = Y(a, M(a), C(a))$. In general, when Y is a set of nodes, $Y(a)$ is the set of $\{Y_i(a)|Y_i \in Y\}$. By using the above notation, the authors of [54] allow for the use of potential outcomes such as $A(a)$. If A is composed by a single variable, then by the Equivalence 6.2 $A(a)$ is the random variable A itself, not the interventional value a. This can also be generalized for set of variables.

For the next steps, it is useful to introduce the consistency property. In Property 3.2.2, the consistency property was defined as:

$$B = b \implies V_i(b) = V_i, \tag{6.3}$$

where $B \in V$ and $b \in \mathcal{X}_B$. This property can be interpreted as $P(V_i(b)|B = b) = P(V_i|B = b)$. Equivalently, the consistency property can be defined in terms of node interventions on multiple treatment variables, for instance $V_i(a, b)$. In this scenario, the consistency property can still be applied. Specifically, following the recursive definition of potential outcomes in the Equivalence 6.2, given two disjoint sets $A, B \in V$, $V_i \in V \backslash (A \cup B), a \in \mathcal{X}_A, b \in \mathcal{X}_B$:

$$B(a) = b \implies V_i(a, b) = V_i(a). \tag{6.4}$$

This can be proven by applying the Equivalence 6.2 to both $V_i(a, b)$ and $V_i(b)$. Indeed,

$$V_i(a, b) = V_i(a_{A \cap Pa_i}, b_{B \cap Pa_i}, \{V_j(a, b)|V_j \in Pa_i \backslash \{A \cup B\}\})$$

meanwhile for $V_i(a)$:

$$V_i(a) = V_i(a_{A \cap Pa_i}, \{V_j(a)|V_j \in Pa_i \backslash A\}) = V_i(a_{A \cap Pa_i}, \{V_j(a)|V_j \in Pa_i \backslash \{A \cup B\}\}, b_{B \cap Pa_i}\}).$$

Now that the useful notation introduced in [54] has been presented, the next paragraphs will focus

on the use of potential outcomes calculus in order to identify conditional path-specific interventional distributions.

In [54] a new framework called *po-calculus* is introduced; this is based upon do-calculus and potential outcomes. The po-calculus enables the generalization of do-calculus to nested counterfactuals. In fact, while the do-calculus is not suitable for treating path-specific interventions, as they involve multiple worlds, po-calculus can overcome these obstacles.
Single World Interventional Graphs (refer to Definition 4.2.0.3) have been introduced in the previous chapters as an useful tool for unifying the approaches of identifiability based on potential outcomes and on graphical conditions. Following the notation in [54], given a DAG $\mathcal{G}$ and $x \in \mathcal{X}_X, X \in V$, $\mathcal{G}(x)$ is the SWIG corresponding to $\mathcal{G}$ in which the nodes belonging to X are split into two nodes, with one half assigned the arbitrary value x. The notation $\mathcal{G}(x, z)$ corresponds to the node splitting operation applied to both X and Z with respective interventional values x and z.
It was previously shown how the Rule 2 of do-calculus can be useful for simplifying conditioned interventional distributions. The authors of [54] use the same approach through potential outcomes, rephrasing the do-calculus Rules in terms of conditional independencies implied by SWIGs. Following this approach, the Rule 1 and Rule 2 of po-calculus are presented. Given a causal graph $\mathcal{G}$ containing the disjoint set of nodes $A, Y, Z, D \in V$ the following relations on the potential outcomes and the respective SWIGs $\mathcal{G}(a)$, $\mathcal{G}(a, z)$ hold.

- The Rule 1 of po-calculus states that:

$$P(Y(a) = y|D(a) = d) = P(Y(a) = y|D(a) = d, Z(a) = z) \tag{6.5}$$

$$\text{if } \Big(Y(a) \perp\!\!\!\perp Z(a)|D(a)\Big)_{\mathcal{G}(a)}, \tag{6.6}$$

- Rule 2 of po-calculus states that:

$$P(Y(a, z) = y|D(a, z) = d) = P(Y(a) = y|D(a) = d, Z(a) = z) \tag{6.7}$$

$$\text{if } \Big(Y(a, z) \perp\!\!\!\perp Z(a, z)|D(a, z)\Big)_{\mathcal{G}(a,z)}. \tag{6.8}$$

An important result regarding these po-calculus rules is presented in [54]: the Rule 1 and Rule 2 of do-calculus respectively hold if and only if the Rule 1 and Rule 2 of po-calculus hold. This follows from the definition of the two couples of graphs $\mathcal{G}(a, z), \mathcal{G}_{\overline{A}\underline{Z}}$ and $\mathcal{G}(a), \mathcal{G}_{\overline{A}}$ along with the definition of d-separation (refer to Definition 3.2.0.4).
The soundness of Rule 1 of po-calculus can be easily proven by using the d-separation criterion, which is complete for SWIGs (see Property 3.2.4). If the random variable $D(a)$ d-separates $Y(a)$ and $Z(a)$ in $\mathcal{G}(a)$, then the variable $Y(a)$ is independent of $Z(a)$ given $D(a)$. From this, Rule 1 of po-calculus follows.
The proof of soundness of Rule 2 of po-calculus can be derived:

$$\begin{aligned} P(Y(a, z) = y|D(a, z) = d) &= P(Y(a, z) = y|D(a, z) = d, Z(a, z) = z) \\ &= P(Y(a, z) = y|D(a, z) = d, Z(a) = z) \\ &= P(Y(a) = y|D(a) = d, Z(a) = z). \end{aligned}$$

The first equation follows from the application of Rule 1 to the assumption $\Big(Y(a, z) \perp\!\!\!\perp Z(a, z)|D(a, z)\Big)_{\mathcal{G}(a,z)}$.
The second equation follows from the equality $Z(a, z) = Z(a)$, given by minimal labelling of Equivalence 6.2 (it was previously shown that by $Z(z)$ is meant the random variable itself). Furthermore, the consistency property can be used to derive the third equation. According to 6.4, given the event $Z(a) = z$, then by consistency $D(a, z) = D(a)$ and $Y(a, z) = Y(a)$. Intuitively, the intervention set with an event that has already occurred can be ignored.

Before diving into the identifiability conditions presented in [54], it is important to understand how the nested counterfactuals in Path-Specific Effects can be analysed by using graphical representations based on SWIGs. Since the focus of this research is to focus on conditional Path-Specific Effects, it is necessary to graphically represent path-specific interventions with a potential outcome framework, so that the Rule 2 of po-calculus can actually be applied to nested counterfactual quantities. The authors of [78] introduce the concept of *extended graph* $\mathcal{G}^e$, a graphical tool that enables the use of po-calculus to nested potential outcomes. Thanks to this framework, the results of [84] can be applied to nested interventions. Thanks to extended graphs, it is possible to represent the path-specific intervention $V(\pi, a_0, a_1)$, with $V \in \mathcal{G}$ and $a_0, a_1 \in \mathcal{X}_A$, in terms of potential outcomes on $\mathcal{G}^e$. Hence, the concept of intervening on the same attribute with different values depending on the selected paths in $\mathcal{G}$ can be interpreted as multiple node interventions in $\mathcal{G}^e$.

Given a DAG $\mathcal{G}$ with nodes V and edges $\mathcal{E}$, it is possible to define the extended graph $\mathcal{G}^e$ [54] as follows. In this project, the only treatment variable taken into consideration is the variable A. For this reason, a specific version $\mathcal{G}^e_A$ of the extended causal graph will be considered. The graphical structure that will now be presented can be easily generalised by selecting different treatment covariates, referring to [54]. In fact, in the general version of $\mathcal{G}^e$ all the attributes can be intervened upon with path-specific interventions.

A formal definition of the expanded graph $\mathcal{G}^e_A$ is now provided. The variable A is defined such that $|\mathcal{X}_A| > 1$, where $\mathcal{X}_A$ is the set of values that A can admit (state space of A). $Ch(A)$ is the set of children nodes such that there exists an edge $\{A \to Ch(A)\}$ in $\mathcal{G}$. In the extended graph $\mathcal{G}^e_A$, for each children node of A, $V_j \in Ch(A)$, a new variable is introduced; this leads to the creation of a new set of variables $A^{Ch} := \{A_j | V_j \in Ch(A)\}$. Analysing the newly created variables, each A_j can assume the same values as A, meaning that $\mathcal{X}_A = \mathcal{X}_{A_j}$. The extended graph corresponding to the original graph $\mathcal{G}$ with treatment variable A is then a causal graph with nodes $\{V \cup A^{Ch}\}$, having the same set of edges $\mathcal{E}$ with some modifications on $(AV_j)_{\to}$ with $V_j \in Ch(A)$. The edges going from A to its children $Ch(A)$ are deleted and replaced with the intermediate variables A_j and the respective edges $(AA_j)_{\to}$ and $(A_jV_j)_{\to}$. In other words, the edges going from the sensitive attribute to its respective children are modified by inserting specific mediators A^{Ch} having the same state space of A. In order for $\mathcal{G}^e$ to be consistent with $\mathcal{G}$, the edges $(AA_j)_{\to}$ are deterministic equality relations, such as $A_j(a_j) = a_j$, with $a_j \in \mathcal{X}_A$. Specifically, each $A_j \in A^{Ch}$ has a single parent A such that $A_j(a_j)$ is a constant random variable corresponding to a point-mass at a_j. It can be noticed that if A has d children, then there exist $|\mathcal{X}_A|^d$ different graphs $\mathcal{G}^e_A$, based upon the deterministic relations embodied by the edges $(AA_j)_{\to}$. In order to formalise this complicated graphical structure, the following definition is considered:

Definition 6.1.0.1. ***Extended Graph*** *Given the DAG $\mathcal{G}$ with nodes V such that $A \in V$, the extended graph $\mathcal{G}^e_A$ is defined by inserting the deterministic arrows from A to A_j, such that $V_j \in Ch(A)$. The deterministic relations $A_j = a_j$, $a_j \in \mathcal{X}_A$, are consistent with the path-specific interventions applied to $\mathcal{G}$. Then, the interventional distribution on the values $v \in \mathcal{X}_{V \setminus A}$, induced by the extended graph $\mathcal{G}^e_A$, is identified as:*

$$P_{\mathcal{G}^e_A}(V \setminus A = v) = \prod_{j: V_j \notin Ch(A)} P(V_j = v_j | Pa_j = pa_j) \cdot \prod_{j: V_j \in Ch(A)} P(V_j = v_j | Pa_j \setminus A = pa_j \setminus a_j, A = a_j). \tag{6.9}$$

If the identifiability assumptions of the edge g-formula are met, the interventional distribution in Equation 6.9 is equivalent to the standard edge g-formula with path-specific interventions that are consistent with the graphical representation $\mathcal{G}^e_A$.

As an example, the extended graph corresponding to Figure 4.2 with A as the sensitive attribute is presented in Figure 6.2. In this plot the children of A are M and Y, meaning that $A^{Ch} = \{A_M, A_Y\}$. It can be noticed that the red edges indicate deterministic relations. In this example, the path-specific intervention of $A = a_0$ along $(AM)_{\to}$ and $A = a_1$ along $(AY)_{\to}$ in $\mathcal{G}$ consists in having $A_M = a_0$ and $A_Y = a_1$ in the extended graph $\mathcal{G}^e_A$ of Figure 6.2.

The extended graph $\mathcal{G}^e$ is a graphical tool that can be used to represent path-specific interventions. Based on the definition of $\mathcal{G}^e_A$, following the notation in [54], $\mathcal{G}^e(a^\pi)$ represents the interventional graph corresponding to setting $A_j = a_1$ if in the original graph we have that $(AV_j)_{\to} \in \pi$ and setting $A_j = a_0$ if $(AV_j)_{\to} \notin \pi$. Furthermore, in [54] it is proven that, given a set of nodes Y, $P(Y(\pi, a_1, a_0))$ is identified in $\mathcal{G}$ if and only if $P(Y(a^\pi))$ is identified in the respective extended graph $\mathcal{G}^e(a^\pi)$. These two distributions are indeed identified by the same functional, as it was shown in Definition 6.1.0.1. The

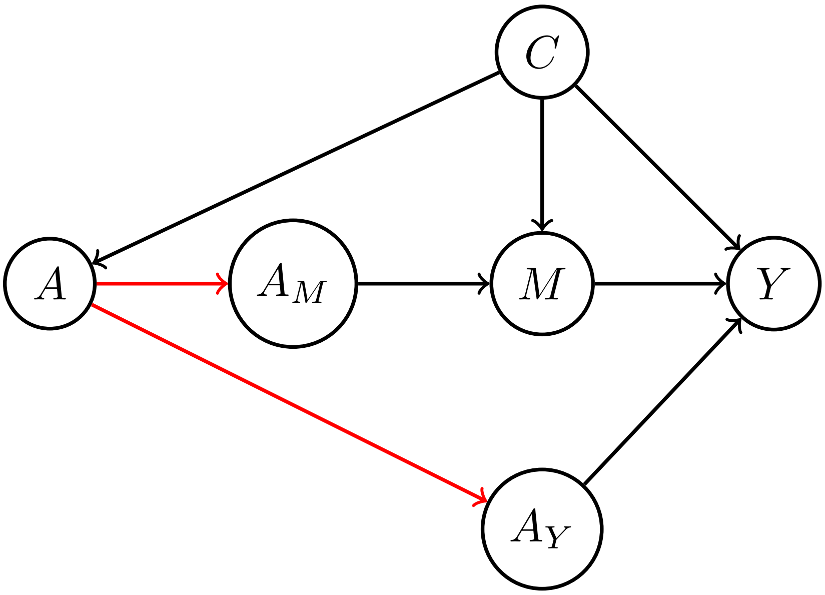


Figure 6.2: Example Extended Causal Graph $\mathcal{G}^e$

respective proposition and proof can be found in the Proposition G.0.1 of the Appendix.
Moreover, following the notation of [54], the graph $\mathcal{G}^e(a^\pi, z)$ is constructed using the usual node-splitting operation on the nodes Z; moreover, in $\mathcal{G}^e(a^\pi, z)$ the newly introduced nodes A_j are set to their fixed values, where A_j is a fixed copy of a_1 if $(AV_j)_\rightarrow \in \pi$ and and of a_0 if $(AV_j)_\rightarrow \notin \pi$. According to this notation $\mathcal{G}^e(a_0)$ corresponds to the SWIG $\mathcal{G}(a_0)$.

In the above section the notation used in [54] is introduced, along with new graphical tools. It was shown how path-specific interventions on A in $\mathcal{G}$ can be interpreted as multiple node interventions in $\mathcal{G}^e_A$. It is now possible to extend the Rule 2 of po-calculus to path-specific interventional distributions.

Identification of Conditional Path-Specific Effect

The general idea of the algorithm defined in [54] will now be presented. Similarly to the IDC algorithm, the goal of [54] is to leverage Rule 2 of po-calculus in order to transform the conditional Path-Specific Effect into a joint interventional distribution. First, the results of [54] applied to arbitrary causal structures will be presented. Subsequently, the same criterion will be simplified for the previously analysed specific causal structures often encountered in the fairness research.

The identification of conditional path-specific interventional distribution will now be introduced for generic causal structures. Given an extended graph $\mathcal{G}^e_A$ and the values $y \in \mathcal{X}_Y, c \in \mathcal{X}_C$, in order to identify $P(Y(a^\pi) = y|C(a^\pi, c) = c)$, similarly to conditional node-interventional distributions, Rule 2 can be used for simplifying this distribution as a joint intervention on A and C. When the causal structure satisfies the assumption of Rule 2 $(Y(a^\pi, c) \perp\!\!\!\perp C(a^\pi, c))_{\mathcal{G}^e(a^\pi, c)}$, then the identification procedure is immediate and follows directly from the application of Rule 2. In generic graphical structures, this condition is not always satisfied. Nevertheless, this does not necessarily mean that the Rule 2 of po-calculus cannot be leveraged in order to simplify the identification of the interventional distribution. Albeit in some scenarios it is not possible to apply Rule 2 to the entire conditioning set C, it could still be useful to apply it to a subset of the conditioning set. The following Corollary can indeed be used in order to guarantee the existence of a unique maximal subset Z of C such that Rule 2 can be applied.

Corollary 6.1.0.1 ([54]). *For any $\mathcal{G}^e(a)$ and any conditional distribution $P(Y(a) = y|C(a) = c)$, there exists a unique maximal set $Z(a) = \{Z_i(a) \in C(a) : P(Y(a) = y|C(a) = c) = P(Y(a, z_i) = y|\{C(a, z_i)\backslash Z_i(a, z_i)\} = c\backslash z_i)\}$ such that Rule 2 applies for $Z(a, z)$ in $\mathcal{G}^e(a, z)$ for $P(Y(a, z) = y|C(a, z) =$*

c).

In the above corollary, $a \in \mathcal{X}_A, c \in \mathcal{X}_C, y \in \mathcal{X}_Y, z_i \in \mathcal{X}_{Z_i}, z \in \mathcal{X}_Z$. The proof of [54] can be found in the Corollary G.0.3.1 in the Appendix. Given this result, the authors of [54] introduce an important theorem regarding the identification of conditional path-specific interventional distributions:

Theorem 6.1.1 ([54])**.** *Let $P(Y(\pi, a_1, a_0) = y | C(\pi, a_1, a_0) = c)$ be a conditional path-specific distribution in the causal model for $\mathcal{G}$, and let $P(Y(a^\pi) = y | C(a^\pi) = c)$ be the corresponding distribution in the extended causal model for $\mathcal{G}^e_A$. Let Z be the maximal subset of C such that $P(Y(a^\pi) = y | C(a^\pi) = c) = P(Y(a^\pi, z) = y | C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$. Then, $P(Y(a^\pi) = y | C(a^\pi) = c)$ is identifiable in $\mathcal{G}^e_A$ if and only if $P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is identifiable in $\mathcal{G}^e_A$.*

In the above, $a_1, a_0 \in \mathcal{X}_A, c \in \mathcal{X}_C, y \in \mathcal{X}_Y, z \in \mathcal{X}_Z, c \backslash z \in \mathcal{X}_{C \backslash Z}$.

Proof. The proof can be found in the Supplement material of [54]. An overview of the proof is presented here, a more detailed version can be found in the original paper.
If $P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is identified in $\mathcal{G}^e_A$, then also $P(Y(a^\pi) = y | C(a^\pi) = c)$ is. This can be easily proven by using the chain rule. Indeed, it is possible to write:

$$\begin{aligned} P(Y(a^\pi) = y | C(a^\pi) = c) &= P(Y(a^\pi, z) = y | C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z) \\ &= \frac{P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)}{P(C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)}. \end{aligned}$$

It follows immediately that if $P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is identified in $\mathcal{G}^e_A$, also $P(Y(a^\pi) = y | C(a^\pi) = c)$ is identified in $\mathcal{G}^e$ with the above formulation. Notice that by definition if $P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is identified in $\mathcal{G}^e_A$, then also $P(C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is. Differently, if $P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is not identified, there are two possible scenarios. The interventional distribution $P(C(a^\pi, z) = c)$ can be either identified or not. If $P(C(a^\pi, z) = c)$ is identified, following the above equations, $P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is identified if and only if $P(Y(a^\pi, z) = y | C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is. Since it was assumed that $P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is not identified in $\mathcal{G}^e_A$, the interventional distribution $P(Y(a^\pi) = y | C(a^\pi) = c)$ is not either. If $P(C(a^\pi, z) = c)$ is not identified, proving that $P(Y(a^\pi) = y | C(a^\pi) = c)$ is not identified is more complicated. Based on the theoretical insights that were presented in the previous chapters, the cause of unidentifiability of $P(C(a^\pi, z) = c)$ can be due to either the presence of a hedge structure, or the presence of a recanting district. In [54], it is proven that in both of these scenario the conditional interventional distribution $P(Y(a^\pi) = y | C(a^\pi) = c)$ is not identified in $\mathcal{G}^e_A$. □

Summarizing, two scenarios are possible in the identification of conditional path-specific interventional distribution $P(Y(a^\pi) = y | C(a^\pi) = c)$. If the condition of Rule 2 of po-calculus holds for the entire set C, then it is possible to identify the conditional interventional distribution as $P(Y(a^\pi, c) = y)$. If the condition does not hold for the entire set C, it is still possible to graphically find the unique maximal set that satisfies it in order to simplify the identification procedure by leveraging the application of Rule 2. It was proven in Corollary 6.1.0.1 that a unique maximal set $Z \subseteq C$ (possibly empty) such that $P(Y(a^\pi) = y | C(a^\pi) = c) = P(Y(a^\pi, z) = y | C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ always exists. In order to prove the identifiability of $P(Y(a^\pi, z) = y | C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$, it is then necessary to prove that the joint interventional distribution $P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$ is identifiable. If identifiability holds for this quantity, then the chain rule can be used to identify $P(Y(a^\pi, z) = y | C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)$, similarly to the proof of Theorem 6.1.1, in the following way:

$$\begin{aligned} P(Y(a^\pi) = y | C(a^\pi) = c) &= P(Y(a^\pi, z) = y | C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z) \\ &= \frac{P(Y(a^\pi, z) = y, C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)}{P(C(a^\pi, z) \backslash Z(a^\pi, z) = c \backslash z)}. \end{aligned}$$

This expression allows us to generalize the identification of conditional path-specific interventional distributions, as long as the identifiability conditions of the joint interventional distribution are met.

Now, the scenario in which the entire conditioning set C satisfies the condition of the Rule 2 of po-calculus will be analysed. Referring to Formula 6.7, in this scenario the conditioning variable D is the empty set and the variable Z is the set of demographic covariates C, graphically represented as root nodes (see Definition 3.1.0.7). Hence, Rule 2 can be rewritten as:

$$\begin{aligned} &P(Y(a^{\pi}, c) = y) = P(Y(a^{\pi}) = y | C(a^{\pi}, c) = c) \\ &\text{if } (Y(a^{\pi}, c) \perp\!\!\!\perp C(a^{\pi}, c))_{\mathcal{G}^{e}(a^{\pi}, c)}. \end{aligned}$$

In the previously described dataset structure, C is not a descendant of A, meaning that the path-specific intervention with A as a treatment attribute has no impact on the random variable C. Furthermore, following the notation of [54], the intervened variable $C(c)$ is equivalent to the random variables itself. Hence, we have $C(a^{\pi}, c) = C$, leading to the graphical condition $(Y(a^{\pi}, c) \perp\!\!\!\perp C)_{\mathcal{G}^{e}(a^{\pi}, c)}$, which holds in these graphical structures (refer to 6.1.2). As a matter of fact, by analysing for instance the graph in Figure 6.1b, since C only presents emitting edges in the original graph, when using the node-splitting operation on C, the random variable itself is not connected to $Y(a^{\pi}, c)$ anymore. It is thus possible to simplify Rule 2 as:

$$\begin{aligned} &P(Y(a^{\pi}, c) = y) = P(Y(a^{\pi}) = y | C = c) \\ &\text{if } (Y(a^{\pi}, c) \perp\!\!\!\perp C)_{\mathcal{G}^{e}(a^{\pi}, c)}. \end{aligned}$$

In conclusion, in order to prove the identifiability of $P(Y(a^{\pi}) = y | C = c)$ with the above presented graphical conditions holding, it is sufficient to prove the identifiability of $P(Y(a^{\pi}, c) = y)$. If the identifiability conditions are met, this can be done with the edge g-formula. For instance, for Figure 6.1a:

$$P(Y(a^{\pi}, c) = y) = P(Y(a_1, M(a_0), c) = y) = \sum_{m \in \mathcal{X}_M} P(Y = y | A = a_1, C = c, M = m) P(M = m | A = a_0, C = c). \tag{6.10}$$

There are different causal structure in which the conditional path-specific interventional distribution is not identifiable due to the unidentifiability of the joint interventional distribution. In general, identifiability conditions can be assessed with the Recanting District Criterion (refer to Definition 4.2.1.2). In the next paragraphs, two examples will be given. In the first one, the conditional path-specific interventional distribution is not identifiable, while in the second one it is.

In Figure 6.3, the interventional path-specific distribution $P(Y(a_1, M(a_0)) = y)$ can be identified, while this is not possible for $P(Y(a_1, M(a_0)) = y | C = c)$ because $P(Y(a_1, M(a_0)) = y, C = c)$ is not identifiable in $\mathcal{G}$. In the previous chapters it was shown that the identification of Path-Specific Effects is linked to the condition of absence of recanting districts. Indeed, referring to $P(Y(a_1, M(a_0)) = y)$ in $\mathcal{G}$, no recanting district exists. This can be noticed by analysing the set of nodes that are present in $Y^* = An_{\mathcal{G}_{\setminus A}}(Y)$. In the graph $\mathcal{G}_{\setminus A}$, the set of ancestors of the outcome is $\{M, Y\}$. In $\mathcal{G}_{Y^*}$ these two nodes correspond to two separate districts of the model, leading to the absence of recanting districts in $\mathcal{G}_{Y^*}$. Hence, $P(Y(a_1, M(a_0)) = y)$ can be identified. Differently, when considering $P(Y(a_1, M(a_0)) = y | C = c)$, it is first necessary to prove that $P(Y(a_1, M(a_0)) = y, C = c)$ can be identified. Nevertheless, it will now be shown that this is not possible due to the presence of a recanting district. In this scenario, the target set is $\{C, Y\}$, meaning that $\{C, Y\}^* = An_{\mathcal{G}_{\setminus A}}(\{C, Y\})$ is the set $\{C, M, Y\}$. Differently from before, $\mathcal{D}(\mathcal{G}_{\{C,Y\}^*}) = \{C, M, Y\}$, as all nodes are connected by bi-directed edges in $\mathcal{G}_{\{C,Y\}^*}$. This means that $\{C, M, Y\}$ is actually a recanting district, since there are two edges $(AM)_{\rightarrow}$ and $(AY)_{\rightarrow}$ with respectively different assigned values of A (a_1 and a_0). This result proves that $P(Y(a_1, M(a_0)) = y | C = c)$ for any $c \in \mathcal{X}_C, a_0, a_1 \in \mathcal{X}_A, y \in \mathcal{X}_Y$ is not identifiable in the causal graph of Figure 6.3.

Differently, in Figure 6.4 $P(Y(a_1, M(a_0)) = y | C = c)$ is identified in $\mathcal{G}$. There, $\{C, Y\}^* = An_{\mathcal{G}_{\setminus A}}(\{C, Y\})$ is the set of nodes $\{C, Y, M\}$. The respective districts are $\mathcal{D}(\mathcal{G}_{\{C,Y\}^*}) = \{\{C, Y\}, \{M\}\}$. For the previously explained reasons, $\mathcal{D}\left(\mathcal{G}_{\{C,Y\}^*}\right)$ does not contain any recanting district, since M and Y are not in the same district.

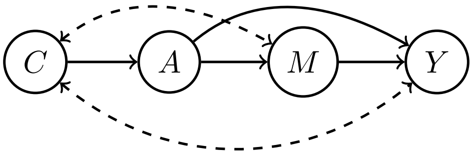


Figure 6.3: Example Directed Acyclic Graph where $P(Y(a_1, M(a_0)) = y|C = c)$ is not identified because $P(Y(a_1, M(a_0)) = y, C = c)$ is not identified

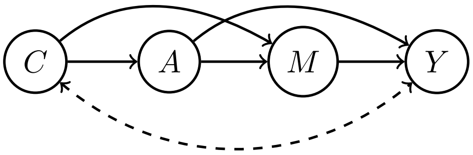


Figure 6.4: Example Directed Acyclic Graph where $P(Y(a_1, M(a_0)) = y, C = c)$ is identified

Conclusion

The motivation behind this section on Conditional Path-Specific Interventional distributions is that, as previously mentioned, it might be interesting to estimate the effect that the sensitive attribute has on a subgroup of the population, for instance people from a specific age group, country of origin, family background. The focus of this research on causal fairness of Machine Learning is to analyse the causal effect that the sensitive attribute has on the final decision. Nevertheless, while the Path-Specific Effect focuses on the effect that A has on Y on average among the entire population, the conditional Path-Specific Effect could help in potentially ensuring that possible fairness requirements on this effect are met among different subgroups of the population.
In the above section, the identification formula for path-specific conditional interventional distributions was presented, along with the respective conditions. Specifically, closer attention was given to the graphical structures that are used in this research, simplifying the identification of conditional path-specific interventional distribution with the Rule 2 of po-calculus.

In the next section, the novel method proposed in this project will be presented along with its respective motivation, analysing the comparison with the Fair Inference on Outcomes [61] and Path-Specific Counterfactual Fairness [15] methods.

6.2. Novel Method

In this problem definition, we are given a dataset consisting of a sensitive attribute A, an unfair output Y and the rest of observed covariates X, along with a DAG representing the causal relations among the variables and a classification of the paths from the sensitive attribute to the outcome as fair/unfair. The set of baseline features is part of X, and it includes the demographic attributes C, generally represented in the graph as root nodes. The main characteristic of this set is not being part of the descendants of A. This graphical assumption has been justified in the previous section, by motivating that conditioning the Path-Specific Effect on a set of mediators would miss the scope of the method. Indeed, by definition of natural effects, the mediators assume the values they would naturally have subsequent to an intervention on the sensitive attribute. Conditioning on a mediator would hence be in contrast with the definition of fairness with Path-Specific Effect.

The method presented in this work consists in estimating a fair world distribution on the observed variables $p^*(A, X, Y)$ that is not only fair among the whole population with respect to the sensitive attribute A and the set of unfair paths, but it is also fair towards specific subgroups of individuals represented in the dataset. This is achieved by enforcing a constraint on the conditional Path-Specific Effect of these subgroups, instead of the overall Path-Specific Effect. The demographic attribute defining these subgroups of individuals will be denoted as the set $C_1 \in C$, possibly empty. C_1 could include a set of multiple covariates, up to include the entire set of demographic covariates $C_1 = C$. As previously mentioned, in this project the choice of the conditioning variables C_1 will not be analysed in details. Nevertheless, in the next examples, it will be shown how the choice of C_1 could influence the performance of the method in terms of accuracy and balance between individual- and population-level fairness. It will be shown how this novel method can be useful for bridging the individual-level and population-level fairness approaches analysed so far.

The overview of the method CPSF (Conditional Path-Specific Fairness) will now be presented. In this section the same notation of Fair Inference on Outcomes [61] will be used. In the next paragraphs, it will be assumed that the demographic covariates are discrete, as it will make the next steps more intuitive. This assumption can easily be refrained for continuous variables, by either constraining the Path-Specific Effect conditioned on disjoint intervals of C_1 and or on the observed instances of C_1.
Given the discrete state space $\mathcal{X}_{C_1}$ of the conditioning covariates, the formulation of the constrained optimization problem is the following:

$$\hat{\beta} = argmax_{\beta} \mathcal{L}_{Y,X,A}(\mathcal{D}; \beta) \tag{6.11}$$

$$\text{subject to } \epsilon^{-} \leq \hat{g}_{(C_1=c_1)}(\mathcal{D}; \beta) \leq \epsilon^{+} \quad \forall c_1 \in \mathcal{X}_{C_1}. \tag{6.12}$$

In the above constrained optimization problem, $\hat{g}_{(C_1=c_1)}(\mathcal{D}; \beta)$ is a semi-parametric estimator for the Path-Specific Effect conditional on the value $C_1 = c_1$. When C is the only observed root node of the DAG, it can be noticed that if C_1 is the entire set of baseline features, then the estimator $\hat{g}_{(C_1=c_1)}(\mathcal{D}; \beta)$ will actually not depend on the data $\mathcal{D}$, as it will be solely a function of the parameters β. Differently, in other scenarios the marginal distribution of $C \backslash C_1$ is empirically estimated from $\mathcal{D}$. Similarly to Chapter 5, the semi-parametric plug-in estimator applied to the edge g-formula will be used. Indeed, in Section 4.2.2 the performance of this semi-parametric estimator has been analysed, concluding that this estimator is suitable for the tasks of this project.
The constrained optimization problem is composed by an objective function that needs to be maximized and by multiple constraints, depending on the number of variables included in C_1 and on the respective state spaces $\mathcal{X}_{C_1}$. It should now be clear how the fair distribution $p(A, X, Y; \hat{\beta})$ will achieve the scope of the method: finding a joint distribution that is as accurate as possible compared to the original one, and such that the outcome is fair towards all the subgroups defined by C_1 with respect to A, by using the Path-Specific Effect as a fairness metric.
Instead of choosing the conditioning set C_1 as the entire set of demographic attributes C, it could be interesting to apply a variable selection or dimensionality reduction in order to improve the performance of the method. It would thus decrease the number of constraints, leading to less computational time needed. Since this task falls outside the scope of this project, it is left for future work.

As mentioned in the previous section, the reason for choosing the conditional Path-Specific Effect on the different demographic subgroups is mainly related to the contrast between the assumptions needed in individual- and population-level fairness approaches. If the goal of a method is solely ensuring individual fairness, the reader might think that it would be more intuitive to directly make use of one of the counterfactuals-based methods [15] [48] analysed in Chapter 5. The choice of combining the method of Fair Inference on Outcomes [61] with conditional Path-Specific Effect is a direct consequence of the conclusions achieved by analysing and implementing selected state-of-the-art methods ([15] [48] [61]) focusing on population- and individual-level fairness. Indeed, it has been shown that individual-level methods necessarily involve approximation methods for inferring the distribution of the unobserved variables, necessary for accurately estimating counterfactuals. Differently, the Path-Specific Effect can be formulated solely in terms of observational distributions, thanks to the edge g-formula. This leads Path-Specific Effect based methods to be more intuitive and inherently more understandable. As a matter of fact, the reader could probably agree that methods based on unit-level counterfactuals, involving multiple worlds and multiple versions of the same individual, can be complex and counter-intuitive to the eyes of an inexpert (but also expert) user. Furthermore, it was shown that further assumptions need to be considered regarding independence conditions among the unobserved variables, specifically with respect to the sensitive attribute; this can lead to potential instability of the method (refer to Chapter 5).

This proposed novel method CPSF aims at combining the advantages of the methods Fair Inference on Outcomes [61] and Path-Specific Counterfactual Fairness [15]. Specifically, the strength of this method relies in focusing on fairness constraints involving a more individual-level perspective compared to FIO [61], but avoiding the variational approach in estimating the latent space and the use of counterfactuals in PSCF [15]. In the section below, a comparison of this novel method with FIO and PSCF will be presented, focusing on how the bridge between individual- and population-level fairness was achieved. The advantages and disadvantaged of the new method CPSF will be further underlined.

6.2.1. Comparison With Other Methods

As anticipated, a methodological comparison of the new method CPSF (Conditional Path-Specific Fairness) with Fair Inference of Outcomes [61] and Path-Specific Counterfactual Fairness [15] is proposed. Subsequently, in Chapter 7, the performance of these methods in terms of fairness and accuracy will be compared on simulated and real datasets. Eventually, further experiments on the choice of the conditioning set and in-depth study on the optimization process will be done.

The CPSF method is based upon the FIO method [61]. It was previously explained that the proposed method is an extension of FIO [61], by increasing the number of constraints on distinct subgroups of the population. This means that, overall, the two methods are identical, with the exception of the formulation of the constraints. Specifically, the new method CPSF is a more general version of the original one FIO, by adding more flexibility regarding which demographic group the fairness constraint is focused on. As a matter of fact, if the conditioning set $C_1 = \{\emptyset\}$, then the new method and the original one are equivalent. There are also further scenarios in which the two methods are equivalent, specifically when the identification formula of the Path-Specific Effect is equivalent to the one of Conditional Path-Specific Effect. An example of these scenarios is covered in Appendix H. In general, we expect the new method CPSF to increase the level of individual fairness compared to FIO; in the next paragraphs close focus will be put in defining a suitable metric for evaluating individual fairness.

By comparing the new method CPSF and PSCF [15] various methodological differences can be analysed. First of all, the PSCF method does not impose any correction on the prediction of the non-disadvantaged group. As a matter of fact, while the new CPSF method predicts the outcomes of new instances based on the new fair joint distribution $p(A, X, Y; \hat{\beta})$ for people in both the originally disadvantaged and not groups, the PSCF method corrects the prediction at test time only for individuals belonging to the initially disadvantaged group. This means that in PSCF the $\hat{Y}$ will be equivalent to the observed outcome Y for individuals belonging to the originally non-disadvantaged group. Furthermore, as it has been previously explained, the new method CPSF does not involve any assumption or estimation regarding the latent space, since the edge g-formula is used as a fairness metric and constraint.
Even though CPSF and PSCF are hardly comparable at a population level, it is interesting to investigate if the newly proposed method actually converges to the same level of individual fairness for certain conditioning attributes C_1. Specifically, we would expect the new method to compute a more individually fair prediction the more 'detailed' C_1 is. Indeed, if the set of values of C_1 is capable of uniquely defining every individual, a different constraint is enforced on every subject, leading to a more individually fair prediction.

Because of the conceptual differences of individual- and population-level fairness, comparing these two fairness approaches can be complex. In order to compare the results of the new method with hte ones of FIO [61] and PSCF [15], it is necessary to make use of fairness metrics capable of portraying both individually fair predictions without estimating the latent space. It was shown in Chapter 5 that Path-Specific Effect can be applied to the different methods in order to evaluate population-level fairness, by making only use of the original graphical structure and observed distributions. Differently, it has previously been discussed that, in order to accurately estimate counterfactuals, it is generally necessary to make assumptions and to estimate the values of the unobserved variables. Since the proposed method does not involve the estimation of unobserved values, it is necessary to find a different metric in order to assess individual-level fairness.

Generally, individual-level fairness consists in evaluating the difference of outcomes of the observed individual and his/her counterfactual version. The estimation of the counterfactuals without modelling the unobserved variables. It has already been explained that estimating the unobserved variables is very important for identifying counterfactuals; it is assumed that in the multiple worlds involved with counterfactuals, the same individual will have the same values of the unobserved variables and the non-descendants of A. The values of the unobserved variables will then help uniquely defining each individual of the population. Differently, since our metric does not involve unobserved variables, we will assume that the non-descendants of A (baseline features) are sufficient for defining the different individuals. This is an approximation of the real world, since two distinct individuals might have the same set of values $c \in \mathcal{X}_{C_1}$. Nevertheless, in the scenario in which C_1 is unique for every individual, the approximation can be highly accurate. The metric applied to the newly proposed method is the following. Given an individual i, the metric is the following:

$$CF(i, \hat{\beta}) = (\hat{Y}_i^{CF} - \hat{Y}_i)\mathbb{I}_{A_i=a_1} + (\hat{Y}_i - \hat{Y}_i^{CF})\mathbb{I}_{A_i=a_0}, \tag{6.13}$$

where $\hat{Y}_i$ is the outcome of the model given the solutions of the constrained optimization problem $\hat{\beta}$, while $\hat{Y}_i^{CF}$ represents the outcome of the counterfactual version of individual i. As a reminder, in this report we have referred as the counterfactual version of an individual the set of values that his/her covariates would have attained if along the unfair paths of the causal graph the sensitive attribute was set to the opposite value of the observed one. These counterfactuals are estimated making use of the edge g-formula; specifically, given an individual i with $A_i = a_1$, it is done by estimating the value $\mathbb{E}\big[\hat{Y}(a_0, a_1, \pi)|C = c_i\big]$ with Formula 4.5. Here, $\hat{Y}(a_0, a_1, \pi)$ is the outcome $\hat{Y}$ by intervening $A = a_0$ along π and $A = a_1$ on the other paths. Instead, as a reminder, the estimation of the counterfactuals in the metric 6.13 for PSCF method [15] consists in estimating $\mathbb{E}\big[\hat{Y}(a_0, a_1, \pi)|C = c_i, U = u_i\big]$. Equivalently, our metric will also be applied to individuals with $A_i = a_0$.

This metric can be interpreted as a quantification of individual path-specific discrimination. The lower the absolute values are for all the individuals in the population, the more individually-fair the prediction $\hat{Y}$ is.

From the application of the method, we would expect the method PSCF to perform with a very low metric $CF(i, \hat{\beta})$ (in absolute value) for individuals i belonging to the potentially discriminated group, while higher for the others. Differently, we expect the new method to outperform FIO in terms of $CF(i, \hat{\beta})$, by returning overall lower values. Compared to PSCF, it will instead be outperformed when considering individuals in the disadvantaged group and vice versa for the rest of the population.

The reader might have doubts regarding the use of counterfactuals in this method with the CF metric. In fact, in the previous chapter it was mentioned that these can be counter-intuitive, as they involve multiple worlds models. Nevertheless, it is important to notice that the newly proposed method does not make use of counterfactuals, but they are only leveraged to build a suitable metric. This has been done with the only purpose of comparing the new method to FIO and PSCF from an individual-level perspective. This will be eventually shown to be a useful approach for comparing the different methods and for evaluating their advantages and disadvantages.

7 Experiments

The goal of this section is to analyse the performance of the novel CPSF method in comparison with FIO and PSCF. These methods will be applied on multiple datasets with different causal structures and data generating processes.
The experiments will be structured in the following way. First, simulated datasets will be considered. In particular, different generating processes corresponding to the causal structures of Example 1 and Example 2 in Section 5.2 will be used. The different simulated datasets vary by the task of the algorithm (prediction/classification), the state space of the conditioning attribute C and the presence of unobserved variables. Furthermore, the methods will be applied to real-world datasets (COMPAS Dataset and UCI Adult Dataset).

The results will be compared in terms of accuracy and fairness. Specifically, among the fairness metrics, both the PSE and the counterfactual fairness metric $CF(i, \hat{\beta})$ for $i \in \{1, ..., n\}$ are taken into consideration. This last metric will be referred to as the CF metric. Both the plots of $CF(i, \hat{\beta})$ for $i \in \{1, ..., n\}$ and statistical measures of its distributions will be analysed. The terms $Mean_{CF}$ and Var_{CF} respectively refer to the mean and variance of the $CF(i, \hat{\beta})$ values among all the data-points.
Overall, four different metrics will be used. All of these will be evaluated on the test set. The accuracy is measured with the MSE for prediction tasks and with percentage of correct outcomes for the classification task. Population-level path-specific fairness is measured with the PSE, while individual-level fairness with the mean and the variance of $CF(i, \hat{\beta})$ among all the individuals i. Furthermore, plots representing the conditional PSE for the different demographic subgroups and the distribution of the CF metric also lead to useful deductions on the methods.

7.1. Simulated Datasets

In this section, the performances of the different methods will be compared on simulated datasets. This allows for a more immediate interpretation of the results, since the data-generating process is known. Both Example 1 and Example 2 causal structures (refer to Section 5.2) will be used. The causal model of Example 2 is most suitable for this task, as the PSE is not equivalent to the NDE. For this reason, the results corresponding to this causal structure will be covered in more detail.
The simulated datasets are composed by 5000 data points, among which 20% is used as a test set. The results of the metrics and the plots refer to the test set of the datasets.

7.1.1. Example 1 Structure

In this section, the methods will be applied to simulated datasets consistent with the causal graph in Figure 4.2 (Example 1). In this causal graph, the only unfair path is the direct one $(AY)_{\rightarrow}$.

As a first example, the linear setting will be analysed. By linear setting, it is meant the data-generating process in which the mediators and the outcome are modelled via linear regression. In Appendix H, an example of a linear setting in which the FIO and CPSF method are equivalent is given.

The first simulated dataset to be taken into consideration is generated from the following linear model:

$$P(C=c) = \begin{cases} \frac{1}{3} & \text{if} \quad c=0 \\ \frac{1}{3} & \text{if} \quad c=1 \\ \frac{2}{9} & \text{if} \quad c=2 \\ \frac{1}{9} & \text{if} \quad c=3 \end{cases}$$

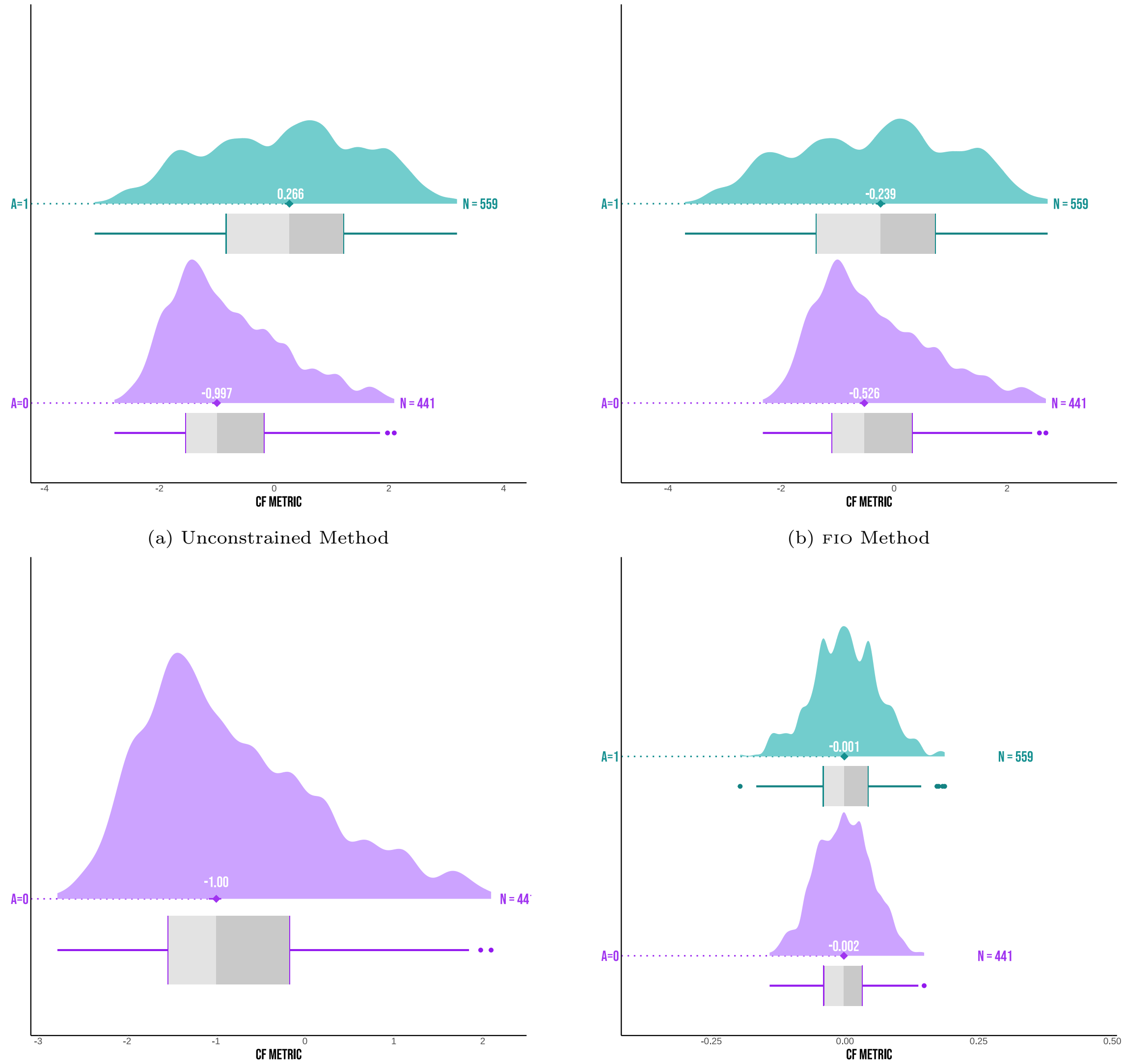
$$\begin{aligned} logit(P(A=1|C)) \sim & -0.5 + C; \\ M = & 0.1 + 2C - A + 3.5AC + \epsilon_M; \\ Y = & 1 - C + 2A - M - 3AC - 0.5AM - 3CM + \epsilon_Y, \end{aligned}$$

where $\epsilon_M, \epsilon_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. In Table 7.1, the results of the different methods applied to the simulated dataset are shown. By analysing the accuracy of the methods, measured with the MSE, the PSCF and CPSF methods are the less accurate, as expected due to the stricter constraints. Furthermore, Path-Specific Fairness is achieved by all of the methods except of course for the unconstrained one. Hence, by solely comparing the MSE and PSE metrics, the preferred method would be the original FIO method. Nevertheless, the scope of this project is to take into account the individual-level fairness too. The mean and variance of the CF values can be analysed, along with the plots found in Figure 7.1. These plots present the density of the metric CF for each method, grouped by the value of the sensitive attribute. In the plot corresponding to the PSCF method, it can be noticed that only the results for the data points with $A = 0$ are represented, since for the rest of the data-points the CF metric is null. Even though PSCF method focuses on individual-level fairness, the variance of CF is still higher than the CPSF method. This happens because for individuals in the discriminated group ($A_i = 1$) by methodology we have that $CF(i, \hat{\beta}) = 0$, while for the rest of individuals $\hat{Y_i} = Y_i$, meaning that the respective $CF(i, \hat{\beta})$ is high in absolute value, specifically nearly the same as in the Unconstrained method.
As expected, the novel method CPSF succeeds at achieving more individually fair predictions compared to the original method FIO. This can be inferred by analysing the distribution of the CF metric along with the respective mean and variance.

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	0.9	-0.6	-0.2	1.7
FIO Method	1.04	-0.05	-0.3	1.6
PSCF Method	1.4	-0.0002	-0.005	1.1
CPSF Method (new)	1.3	-0.01	0.004	0.04

Table 7.1: Results Example 1 Causal Structure - linear discrete C

The CF metric is an indicator for individual-level fairness. In contexts with a discrete conditioning attribute C, like in this simulated dataset, it can also be interesting to analyse the values of the conditional Path-Specific Effect, comparing the original FIO method with the newly introduced CPSF. In Figure 7.2 the different values of conditional PSE are presented, grouped by the different values of the state space of the conditioning covariate C. The reader should focus on analysing and comparing the difference in the x-axis scale of the two plots. As expected, the conditional PSE of the FIO method highly exceeds the threshold $[-0.05, 0.05]$, albeit having an overall PSE of -0.054. Differently, in the



(a) Unconstrained Method

(b) FIO Method

(c) PSCF method

(d) CPSF Method

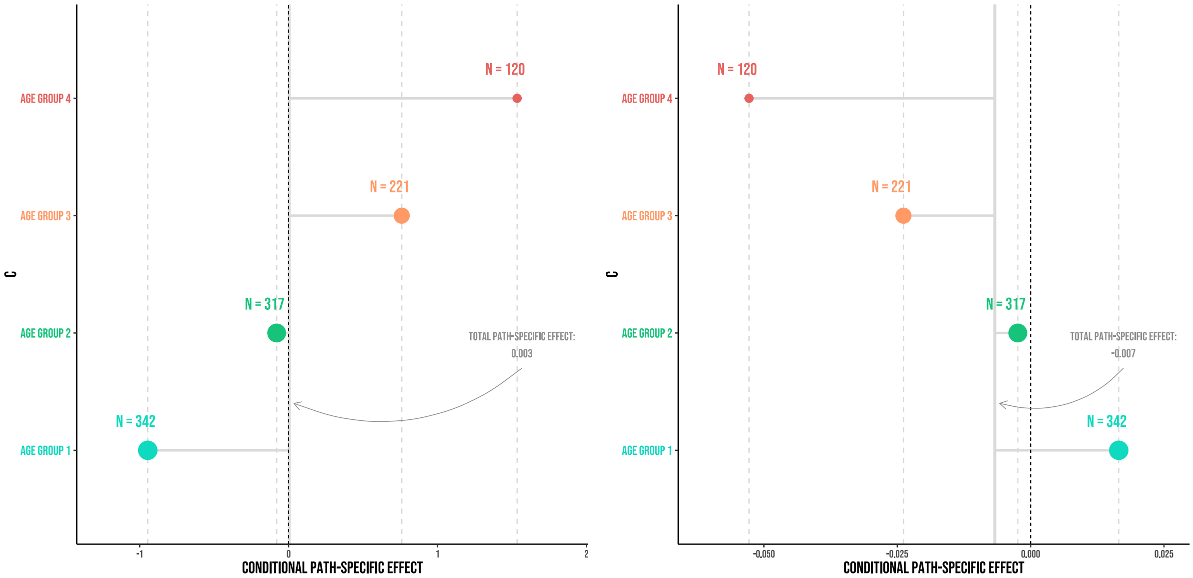
Figure 7.1: Plot of the Counterfactual Fairness metric

newly introduced method the conditional PSE are nearly zero, for any value of $c \in \mathcal{X}_C$.

Comparison by Conditioning Set

In this section, the same graphical structure and simulation procedure will be considered, by varying only the conditioning variable C. The scope of this section is indeed to evaluate the variation of individual-level fairness with different conditioning sets, varying from discrete to continuous. The main idea behind this comparison is that we would expect the measure of individual fairness to improve by considering a more “detailed” set of conditioning variable. As a matter of fact, the more representative of the different data-points the conditioning attribute is, the more individually-fair the CPSF method should be. Similarly, as the number of constraints increases, we would expect the accuracy to decrease.

In Table 7.2, the results corresponding to the previous simulation process with $C \sim \mathcal{N}(0, 1)$ instead are presented. The simulation process can be found in Appendix J.1.2. Analysing the CPSF method, the trend and values of the fairness and accuracy metrics do not seem too far apart compared to the



(a) Original method FIO

(b) CPSF Method (new)

Figure 7.2: Plot of the conditional Path-Specific Effect for different values of the conditioning attribute C

ones presented in Table 7.1 referring to the simulation procedure with discrete C. Nevertheless, the comparison with the original FIO method is interesting. Indeed, compared to Table 7.1, the variance of the CF metric for CPSF is higher. Nevertheless, the variance gap between the newly proposed method and the original FIO is greater compared to the discrete C setting. These results reflect the intuitive expectations of the method, also supported by a drop in accuracy of the CPSF method. Overall, the CPSF method still outperforms the FIO method in terms of individual fairness, further improving the performance gap compared to the discrete C simulation setting.

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	1.0	-2.0	-0.9	5.1
FIO Method	1.6	-0.04	-0.8	2.8
PSCF Method	1.9	-0.0003	2.7	1.3
CPSF Method (new)	1.9	-0.05	-0.6	0.7

Table 7.2: Results Example 1 Causal Structure - linear continuous C

Classification Tasks

In this Section, the classification task will be analysed. The same graphical structure and simulation procedure of the previous examples is kept, except for the outcome and mediator that are modelled via logistic regression, leading to a classification setting. The modified setting is:

$$\begin{aligned} logit(P(M=1|A,C)) &\sim 0.05 - 1.5C + 2A - 0.05AC; \\ logit(P(Y=1|M,A,C)) &\sim -1 + 5C - 4A + 2M - 7AC - 3AM + 3CM. \end{aligned}$$

The entire simulation process can be found in Appendix J.1.4. In classification tasks, the accuracy is not measured anymore by MSE, but by the percentage of correct classifications in the test set.

In Tables 7.3 and the results of the simulation procedure with binary conditioning attribute C can be found. Overall, the previously observed trends in the fairness metrics are met in this example too.

Indeed, the CPSF method outperforms the FIO one in terms of individual-level fairness.
It is also interesting to notice that the CPSF method achieves higher accuracy compared to the FIO one. Since logistic regressions are less interpretable compared to linear regression, given the multiplicative and not additive weights, the decrease of accuracy due to additional constraints on the weights is probably not as straightforward.

Method	Accuracy	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	94%	-0.8	0.0003	0.9
FIO Method	52%	-0.05	-0.2	0.5
CPSF Method (new)	62%	-0.05	-0.2	0.3

Table 7.3: Results Example 1 Causal Structure - Classification binary C

Comparison by Unobserved Variables

In this section, the performance of the method in causal graphs presenting unobserved variables will be evaluated. Specifically, the dataset is simulated by considering an additional unobserved variable parent of both the variable A and Y (simulation procedure in Appendix J.1.5). By the recanting district criterion, the PSE of this graphical structure is identifiable. We are interested in analysing the performance of CPSF in semi-Markovian models. If the PSE is identifiable, we would expect CPSF to still outperform FIO in terms of individual fairness.

In Table 7.4, the same trends observed in the previous examples are presented. Indeed, the novel method CPSF presents a relevantly lower variance of the CF metric compared to the FIO method. Furthermore, CPSF predictions are less accurate compared to the Unconstrained one, while almost as equally accurate as FIO. In the Chapter 5.2 it was already explained that the performance of the PSCF method is not relevantly affected by the additional unobserved confounder, both in terms of accuracy and fairness because of the absence of any mediator on unfair paths in this causal model is the main reason.
In the next section regarding Example 2 causal structure, further experiments on causal graphs with unobserved variables will be considered. In particular, interesting methods' performances are observed.

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	38.1	2.4	0.26	10.3
FIO Method	64.1	0.08	-0.09	10.7
PSCF Method	36.5	0.00001	-3.7	23.3
CPSF Method (new)	65.6	0.02	-0.5	5.0

Table 7.4: Results Example 1 Causal Structure - prediction binary C with unobserved variable

7.1.2. Example 2 Structure

Now, an examples of a linear regression model following the causal graph in Figure 5.4 analysed in Example 2 of Chapter 5 will be analysed. As mentioned in Appendix H, in the presence of mixed terms in the linear regression model, the Path-Specific Effect differs from the conditional Path-Specific Effect.

A simulated dataset from the following regression model will be considered:

$$\begin{aligned} C \sim & \mathcal{N}(0,1) \\ logit(P(A=1|C)) \sim & -0.5-0.5C; \\ logit(P(M=1|A,C)) \sim & -0.5-C+4A-3AC; \\ logit(P(L=1|M,A,C)) \sim & -0.5-0.4C-2A+3M-0.6AC+2AM+0.5ACM; \\ Y = & 1+C+5A-4M+3L-2AC+AM+AL-1.5AML+\epsilon_Y, \end{aligned}$$

where $\epsilon_Y \sim \mathcal{N}(0,1)$. Similarly to the previous example, the different methods are implemented on the simulated dataset. In this model, all of the variables are discrete, except for the conditioning attribute C and the outcome Y, which are continuous.

In Table 7.6, the results are presented. Starting from the accuracy, the same trend of the previous causal structure is encountered, characterized by a lowest accuracy for PSCF due to the individual focus; the CPSF method is second-lowest, presenting a relevantly higher MSE compared to the FIO method and to, of course, the unconstrained one. Due to the continuity of the conditioning variable C, a high number of constraints is set in CPSF, theoretically equal to n. Albeit in the optimization problem CPSF aims at maximising the likelihood of the data, the large number of constraints does not allow for high accuracy of predictions. Moreover, as expected, the population-level path-specific fairness (measured by the PSE) is achieved by all the methods, except of course for the unconstrained one. Differently, a steep improvement of individual-level fairness of the CPSF can be observed compared to FIO, especially supported by a relevantly lower variance of the CF metric. This is also shown in the plots of Figure 7.3, where the difference between the FIO and CPSF performances is emphasized. Indeed, by analysing the domain of the CF difference metric, the CPSF method outperforms the others in terms of individual fairness. As it was previously explained before, the subplot in Figure 7.3d corresponding to PSCF method should be analysed by taking into account that for the potentially discriminated group ($A = 1$), the method is completely individually fair, with a null CF metric.

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	1.0	-4.9	2.5	63
FIO Method	6.6	-0.07	4.3	43.6
PSCF Method	16.15	0.002	-3.5	10.1
CPSF Method (new)	14.4	-0.04	0.3	8.6

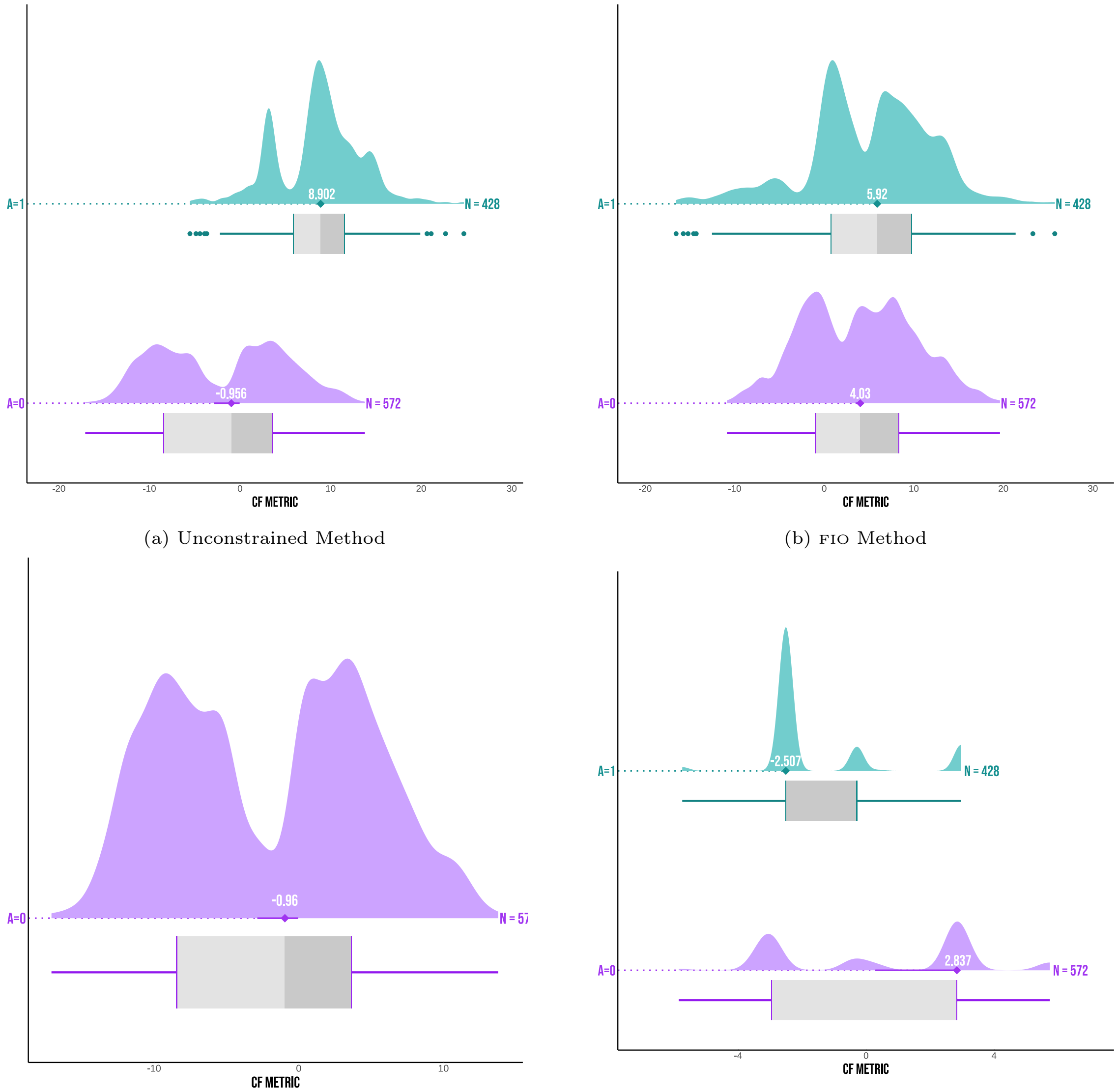
Table 7.5: Results Example 2 Causal Structure - prediction continuous C

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	1.0	-4.9	2.5	63
FIO Method	6.6	-0.07	4.3	43.6
PSCF Method	16.15	0.002	-3.5	10.1
CPSF Method (new)	14.4	-0.04	0.3	8.6

Table 7.6: Results Example 2 Causal Structure - prediction continuous C

Comparison by conditioning set

In this paragraph, the same simulation procedure of the previous example was considered, but with a binary conditioning set $C \sim Bin(0.7)$. By varying the distribution of the conditioning set, we would expect the performance of the CPSF method to improve the more ‘detailed’ the conditioning attribute is. By considering a binary conditioning set, we hence expect the method to have a less individually



(a) Unconstrained Method

(b) FIO Method

(c) PSCF method

(d) CPSF Method

Figure 7.3: Plot of the Counterfactual Fairness metric

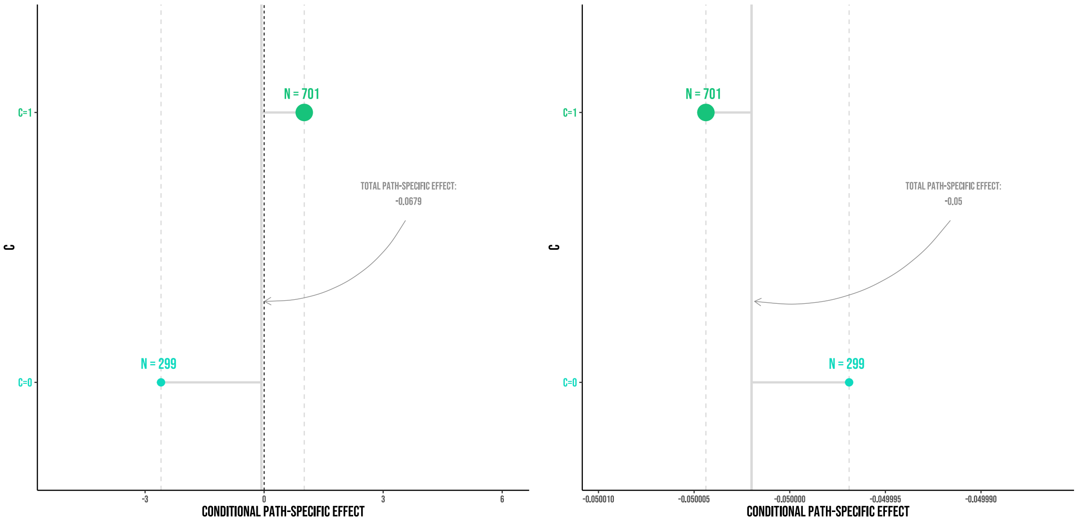
fair performance compared to the previous example in Table 7.6. The complete simulation process can be found in Appendix J.2.3.

The results of the application of these methods to this simulated dataset can be found in Table 7.7. It is interesting to observe that the newly proposed method does not highly outperform the original FIO one in terms of individual fairness, as the variance of the CF metric is higher. Nevertheless, by comparing the values of the conditional PSE for the binary values of C in Figure 7.4, it can be noticed that the CPSF method satisfies path-specific fairness among all the population groups determined by the values of C.

To support the intuition that the performance of CPSF improves the larger the dimension of the state space $\mathcal{X}_C$ is, another simulation procedure is considered. The setting is the same as the one in the

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	0.9	-3.4	1.0	25.6
FIO Method	1.7	-0.07	-2.6	30
PSCF Method	4.7	-0.001	-2.3	3.9
CPSF Method (new)	3.6	-0.05	-2.3	36

Table 7.7: Results Example 2 Causal Structure - prediction binary C



(a) Original method FIO

(b) CPSF Method (new)

Figure 7.4: Plot of the conditional Path-Specific Effect for the binary values of the conditioning attribute C

previous two examples, but the variable C has the following distribution:

$$P(C=c)=\begin{cases}\frac{1}{3} & \text{if} \quad c=0\\ \frac{1}{3} & \text{if} \quad c=1\\ \frac{2}{9} & \text{if} \quad c=2\\ \frac{1}{9} & \text{if} \quad c=3.\end{cases}$$

Based on the theoretical considerations of the method, we would expect CPSF to perform better than in the previous example with the binary conditioning variable C, but not as good as the one with continuous variable C, the results of which are presented in Tables 7.7 and 7.6 respectively. The simulation process can be found in Appendix J.2.2. In Table 7.8 the results of the application of the methods to the new simulation procedure are presented. Differently from the previous example in Table 7.7, the method CPSF outperforms FIO in terms of individual fairness. Nevertheless, it can be noticed that compared to Table 7.6 the gap between the CF metrics of CPSF and FIO is not as wide. Hence, by comparing the CPSF method with FIO, the novel method is best performing when applied to the simulated dataset with continuous C and generally an increase of performance is observed by increasing the value $|\mathcal{X}_C|$.

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	1.0	-3.4	-0.6	35.7
FIO Method	2.0	-0.05	-1.2	25.9
PSCF Method	4.5	-0.0001	-2.6	3.8
CPSF Method (new)	3.7	-0.01	0.6	6.9

Table 7.8: Results Example 2 Causal Structure - prediction discrete C

Classification tasks

Similarly to the example of the previous causal structure, in this section the performance of FIO and CPSF will be analysed for classification tasks. The simulated dataset has the same simulation procedure as the previous examples, but the outcome is generated from a logistic regression function. In the simulation procedure, we have

$$logit(P(Y=1|L,M,A,C)) \sim -1-C-5A+4M-3L+2AC-AM-AL+1.5AML.$$

The entire simulation procedures can be found in Appendix J.2.5 and Appendix J.2.4. In Tables 7.9 and 7.10 the results corresponding to the simulated dataset with continuous and binary conditioning variable C respectively are shown. Overall, as in the classification task of the previous graphical structure, the performance of the CPSF method has a less steep improvement in terms of fairness compared to the FIO method, differently from the results of the prediction tasks. In Table 7.9, the expected trends of results in accuracy and fairness are observed. Indeed, even though the accuracy of the CPSF drops compared to the other two methods, the CF metric has mean and variance closer to zero, probably consequence of a better performance as individually-fair classifications.

It is also interesting to compare the results of Tables 7.9 and 7.10 among each others. As it was observed in the previous example, in a setting with a binary conditioning set, the CPSF method does not outperform the original FIO one. Nevertheless, the drop of accuracy observed in Table 7.9 is higher compared to the one in Table 7.10. As previously mentioned, this behaviour is due to the fact that a binary C is poorly representative of the individuals in the population.

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	89%	0.3	-0.2	0.3
FIO Method	78%	0.05	-0.03	0.3
CPSF Method (new)	65%	0.03	0.008	0.01

Table 7.9: Results Example 2 Causal Structure - classification continuous C

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	93%	0.1	0.05	0.02
FIO Method	91%	0.05	0.03	0.01
CPSF Method (new)	88%	0.05	0.05	0.01

Table 7.10: Results Example 2 Causal Structure - classification binary C

Comparison by Unobserved Variables

In this section the comparison with datasets simulated from semi-Markovian models are presented. In the first simulated dataset, a causal graph with an unobserved parent of both M and Y is considered. In the second dataset, an additional unobserved variable that is parent of both A and L is considered. The simulation procedure can be found in the Appendix J.2.6.

As a reminder, the $PSCF_{\beta=0}$ Method considered in the tables below has the hyper-parameter β of the Maximum Mean Discrepancy set to 0. Differently, this values is equal to 100 for $PSCF_{\beta=100}$. For this reason, we expect the PSCF method not to fully correct unfairness in the application to the second simulated dataset. Indeed, PSCF with $\beta = 0$ assumes that the unobserved parents of the mediators are independent from the sensitive attribute. Nevertheless, in the second simulated dataset L shares unobserved parent with A, leading this assumption to fail. In the previous Chapter 5.2, it was shown that selecting a higher value of β, like $\beta = 100$ improves the performance of PSCF in terms of PSE.

In both of the two considered graphical structures the PSE and conditional PSE are identifiable, according to the recanting district criterion.

In Tables 7.11 and 7.12 the results of these applications are presented, respectively for the first and second simulated datasets. In both of the simulated datasets, the method CPSF presents a lower variance of CF compared to FIO. Furthermore, as expected, it can be noticed that the PSCF is outperformed in terms of population-level fairness (PSE) by both FIO and CPSF in Table 7.12. Compared to the previous simulated datasets, the PSE achieved by the predictions of PSCF method was generally relevantly lower compared to the other methods. This results is indeed probably consequence of the additionally unobserved confounding between L and A.

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	4.2	-1.2	0.3	2.0
FIO Method	4.5	-0.05	1.1	2.3
$PSCF_{\beta=0}$ Method	4.8	-0.02	-0.9	0.4
$PSCF_{\beta=100}$ Method	4.7	-0.02	-0.9	0.4
CPSF Method (new)	4.9	0.02	0.9	0.3

Table 7.11: Results Example 2 Causal Structure - prediction continuous C with unobserved variable

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	4.4	-1.3	0.2	1.8
FIO Method	4.7	-0.06	0.4	2.8
$PSCF_{\beta=0}$ Method	5.4	0.09	-0.6	0.4
$PSCF_{\beta=100}$ Method	5.4	0.008	-0.6	0.4
CPSF Method (new)	5.3	-0.05	0.9	2.0

Table 7.12: Results Example 2 Causal Structure - prediction continuous C with unobserved variables

7.2. Real-World Datasets

In this section, the newly proposed method will be applied to the real datasets COMPAS and UCI Adult. This will give us interesting results in terms of robustness of the method in model misspecification scenarios. The pre-processing procedure of the real-world datasets can be found in Appendix I.
Overall, the results corresponding to the COMPAS dataset are coherent with the expectation and with the previous results. Differently, no conclusions can be derived from the results of the pre-processed UCI Adult dataset because all of the methods (even the unconstrained one) present bounded conditional PSE in the different demographic groups.

7.2.1. COMPAS Dataset

The COMPAS Dataset has been introduced in the previous Section 2.2 and in the Appendix I. In this section, the CPSF, FIO and PSCF methods will be applied to the pre-processed real dataset. The scope of this domain-specific example is to predict whether an inmate will reoffend in the future, by making a recidivism classification. The outcome Y represents the recidivism decision, while the mediator M the number of prior convictions. The sensitive attribute A stands for the race of the defendants and is a binary variable, respectively for Caucasian an non-Caucasian subjects. In this example, the only unfair path is the direct path from A to Y, meaning that the PSE coincides with the NDE.
Differently from the previously presented examples, in this dataset the demographic attribute is a set of two distinct variables, representing the age (discrete) and gender (binary) of the defendant. In particular, the age of the defendants is divided into four main age groups.

In Table 7.13, the results of the different methods are presented. Three different versions of the CPSF method are applied, differentiated by the chosen conditioning attribute. In the first one the conditioning covariate only includes the gender, while in the second version only the age. At last, the CPSF is applied with both the age and gender variables as conditioning sets. The results are coherent with the previous performance on simulated datasets. Indeed, the constraint on PSE is achieved by both FIO and CPSF. Furthermore, similar to the previous classification tasks, the CPSF actually presents a higher accuracy compared to FIO. Analysing individual fairness instead, the CPSF (age and gender) relevantly outperforms the FIO method. Instead, the other two CPSF versions are slightly better at detecting individual fairness than FIO but not as well performing as CPSF (age and gender). By comparing the three proposed versions of CPSF, it can be noticed that the accuracy decreases in CPSF (age and gender), coherently with the additional number of constraints. This is indeed also supported by an increase in individual-level fairness performance. In Figure 7.5, the conditional PSE by age group is represented. The only method in which the conditional PSE falls outside the threshold area is the CPSF (age) in Figure 7.5b, as expected.

Overall, it is interesting to see that the trends encountered in the simulation setting are also observed in real-world datasets, corresponding to possible model misspecification scenarios.

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	69%	-0.7	-0.01	0.01
FIO Method	58%	-0.06	-0.01	0.04
CPSF Method (gender)	65%	-0.06	-0.01	0.004
CPSF Method (age)	65%	0.01	0.01	0.003
CPSF Method (age and gender)	63%	-0.004	0.003	0.001

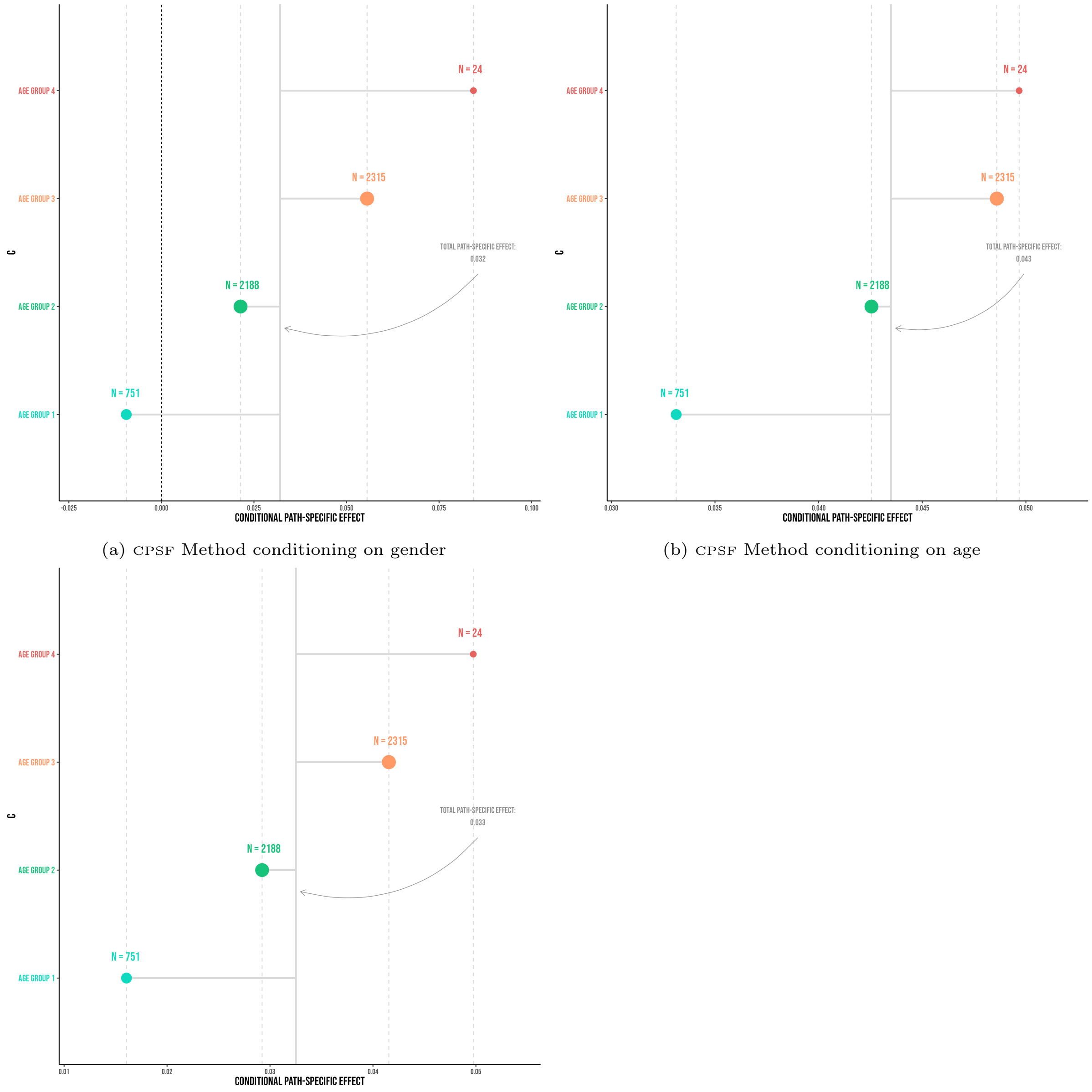
Table 7.13: COMPAS Real-World Dataset Results

7.2.2. UCI Adult Dataset

The results of the different methods applied to the UCI Adult pre-processed dataset can be found in Table 7.14. The results are not quite relevant, as the unconstrained method already achieved the requirements in terms of individual-level and population-level fairness, with a low conditional PSE among all the demographic variables. For this reason, the application of the methods to these datasets will not be analysed in detail. As previously mentioned, these peculiar results are probably due to the specific pre-processing of the data that has been done.

Method	MSE	PSE	$Mean_{CF}$	Var_{CF}
Unconstrained Method	82%	0.002	-0.001	0.006
FIO Method	81%	0.004	-0.01	0.01
CPSF Method (new)	81%	0.003	0.001	0.01

Table 7.14: UCI Adult Real-World Dataset Results



(a) CPSF Method conditioning on gender

(b) CPSF Method conditioning on age

(c) CPSF Method conditioning on age and gender

Figure 7.5: Plot of the conditional Path-Specific Effect for the age groups

8 Discussion and Conclusion

Overall, analysing the results of the experiments in the chapter above, the CPSF method is a competitive alternative to FIO and PSCF methods. This method has indeed the goal of bridging the population-level approach of FIO with the individual-level one of PSCF. The performance of CPSF on simulated and real-world datasets shows high flexibility and interpretability. In fact, by comparing the performance of CPSF with the ones of FIO and PSCF, different conclusions can be made. Starting from the comparison with the former, the CPSF method outperforms FIO in terms of individual-level fairness when the conditioning set is sufficiently representative of the data-points. It was shown that by choosing a binary conditioning set $\mathcal{X}_C = \{0, 1\}$ the performance of CPSF is very similar to the one of FIO; by increasing the state space, specifically from $|\mathcal{X}_C| \geq 4$ a steep improvement of individual-level fairness is observed, coherently with the more representative the conditioning set. Furthermore, due to the higher number of constraints, a drop in accuracy of the CPSF method is observed, especially in prediction tasks. Similarly, the decrease in accuracy is relevant in scenarios with a high number of constraints (high $|\mathcal{X}_C|$).
It has previously been discussed that the CPSF method is methodologically different to PSCF, but the two methods can still be compared in terms of accuracy, Path-Specific Effect and individual-level fairness. Overall, PSCF outperforms CPSF in individual-level fairness for the potentially discriminated group, as the respective CF metric is null for all the corresponding data-points. For the rest of the data-points, the CF metric of CPSF is generally lower in absolute value. The accuracy and interpretability of CPSF emphasize its potential as an alternative of the original PSCF method, showing promising improvements. Additional experiments on semi-Markovian models and real-datasets confirm the high flexibility and stability of the CPSF method, showing similar trends as the previously mentioned ones in comparison to FIO and PSCF methods.

The scope of this project was to analyse the field of fairness of Machine Learning and to possibly improve the state-of-the-art methods. In this thesis, an overview on fairness of Machine Learning was given; in particular, the motivation behind the choice of using a causal approach was explored. The state-of-the-art causality-based fairness metrics and methods were compared, both methodologically and experimentally. Their application on simulated datasets emphasised their advantages and drawbacks. Because of their high flexibility and their scalability to arbitrary causal structures, path-specific causality-based metrics were chosen as the most suitable for portraying the fairness of a model. Furthermore, by comparing individual- and population-level fairness methods, a significant gap was encountered, given by the lack of individual focus in population-level fairness metrics and by the low interpretability and complexity of individual-level fairness metrics. Because of these differences, a newly proposed method CPSF was defined with the scope of bridging the two approaches, leveraging the strengths of both. Overall, the performance the new method on simulated and real-datasets confirms its competitiveness both in terms of fairness and accuracy. The method CPSF presents two main characteristics. First, the outcomes can be considered to be more individually fair compared to state-of-the-art population-level causality-based methods. Further, the strong assumptions on the latent space that characterize individual-level methods are avoided, allowing for a more interpretable approach.

In conclusion, this thesis proposed a method CPSF with a high potential and flexibility that should be taken into account for future developments on the topic of Fairness of Machine Learning. Furthermore, an useful overview on the methodological and conceptual differences between individual-level and population-level fairness approaches was given, which had not been done in the literature yet with this perspective. Hopefully, this work will further lead to constructive debates regarding these two approaches in fairness applications.

8.0.1. Limitations and Future Work

In the previous sections, different strengths of the CPSF method has been analysed. In this section, its limitations will be presented, along with possible future works. First of all, it was previously mentioned that one of the main strengths of this method is the absence of assumptions on the latent space distribution. Nevertheless, there are certain conditions that need to be met with respect to the identifiability of the conditional PSE. These were presented with the edge g-formula in Section 4.2.1 and in Section 6.1.3. Hence, when applying this method to a new causal structure, it is critically important to assess the validity of these conditions on the causal graph. If these conditions are not met, the CPSF method is then not applicable. The problem of unidentifiability has been widely discussed in the causality literature, further research on the identification of PSE could be very useful in the fairness research. It is important to point out that the original FIO method was also affected by similar identifiability limitations. Furthermore, in the application of CPSF it was noticed that increasing the constraints in the optimization problem often leads to an longer computational time. Future works should focus on the choice of the most suitable optimization algorithms used in the constrained optimization problem. Moreover, in the literature different semi-parametric estimators have been proposed for the edge g-formula [84]. This work focused on the plug-in estimator, but further options should be considered in order to increase the level of robustness of the estimator. Additionally, this work on conditional PSE did not focus on the criterion for selecting the conditioning set, part of the baseline features. It has previously been discussed in Section 6 that it might be interesting to identify a strategy for choosing the suitable conditioning subset when the set of baseline features C includes a high number of covariates. Lastly, it would be interesting to expand this method not only to prediction and classification tasks, but also to decision scenarios with missing labels, as explained in the Chapter 2.

I

Appendix

A

Proof G-Formula

The proof of the Truncated Factorization formula proposed in this project will follow the approach used in [71]. An additional proof based on Single World Intervention Graphs (SWIGs) has been recently proposed in [74].

As mentioned in Section 3.2.2, interventions can also be interpreted graphically in a causal model. Indeed, intervening on a set of variables V_k corresponds to deleting the edges entering the variables in V_k, while the omitted edges are fixed to the arbitrary interventional values. In the next steps, the original causal graph will be referred to as $\mathcal{G}$, meanwhile the one corresponding to the intervention $V_k = v_k^{int}$ as $\mathcal{G}_{\overline{v_k^{int}}}$. In this proof, the *do* notation will be different from the rest of the project; in order to simplify the possible formulation redundancies in the proof, the *do*() notation of Pearl will be used. Specifically, $P(V_1 = v_1, ..., V_k = v_k, ..., V_d = v_d | do(v_k^{int}))$ will correspond to the previously introduced $P(V_1(v_k^{int}) = v_1, ..., V_k(v_k^{int}) = v_k, ..., V_d(v_k^{int}) = v_d)$.

Theorem A.0.1. ***Truncated Factorization Formula*** *For any Markovian model, the distribution generated by an intervention $V_k = v_k$ on a set V of endogenous variables, such that $V_k \in V$, is given by the truncated factorization:*

$$P(V(v_k) = v) = \begin{cases} \prod_{V_i \in V \setminus V_k} P(V_i = v_i | Pa_i = pa_i) & \textit{if } v_k \textit{ consistent with } v, \\ 0 & \textit{otherwise.} \end{cases} \tag{A.1}$$

Here $P(V_i = v_i | Pa_i = pa_{V_i})$ are the pre-intervention conditional probabilities.

Proof. In order to begin the proof of the Truncated Factorization formula, the property of *Modularity* can be useful. This was formally introduced in Property 3.2.1. This property states that when intervening on a variable or a set of variables V_k, the mechanisms corresponding to the other variables remain invariant. In mathematical notation, given an intervention on a set of variables V_k that is disjoint from V_j, this property can be formulated as:

$$P^{\mathcal{G}}(V_j = v_j | Pa_j = pa_j) = P^{\mathcal{G}_{\overline{v_k^{int}}}}(V_j = v_j | Pa_j = pa_j), \tag{A.2}$$

where $\mathcal{G}_{\overline{v_k^{int}}}$ is causal graph $\mathcal{G}$ subsequent to the intervention on V_k as v_k^{int}. Using the above property, the joint distribution of the set of observed variables $V_1, ..., V_d$ after the intervention on V_k as v_k^{int}, assuming that the set V is discrete, can be written as:

$$
\begin{aligned}
P(V_1 = v_1, ..., V_k = v_k, ..., V_d = v_d | do(v_k^{int})) &= P^{\mathcal{G}_{\overline{v_k^{int}}}}(V_1 = v_1, ..., V_k = v_k, ..., V_d = v_d) \\
&= \prod_{j \in \{1,...,d\}} P^{\mathcal{G}_{\overline{v_k^{int}}}}(V_j = v_j | Pa_j = pa_j) \\
&= \prod_{j \in \{1,...,k-1,k+1,...,d\}} P^{\mathcal{G}_{\overline{v_k^{int}}}}(V_j = v_j | Pa_j = pa_j) \mathbb{1}_{(v_k = v_k^{int})} \\
&= \prod_{j \in \{1,...,k-1,k+1,...,d\}} P(V_j = v_j | Pa_j = pa_j) \mathbb{1}_{(v_k = v_k^{int})}.
\end{aligned}
$$

In the above equalities, different properties have been used. The first equality follows the graphical definition of intervention, indeed analysing an intervention on the variable $V_k = v_k^{int}$ is equivalent to analysing the causal graph $\mathcal{G}_{\overline{v_k^{int}}}$. The second equality follows from Equation 3.3, as $\mathcal{G}_{\overline{v_k^{int}}}$ is also a Markovian model by definition. The third equality is a consequence of the definition of $\mathcal{G}_{\overline{v_k^{int}}}$, as the probability distribution of V_k consistent with this graph is non-null only if $V_k = v_k^{int}$. Lastly, the Modularity property of Equation A.2 guarantees the third equality; indeed, $P^{\mathcal{G}_{\overline{v_k^{int}}}}(V_j = v_j | Pa_j = pa_j)$ is equivalent to $P(V_j = v_j | Pa_j = pa_j)$, as it represents the conditioned distribution corresponding to the causal graph $\mathcal{G}_{\overline{v_k^{int}}}$.

□

B

Identification Formula - ID Algorithm

Once the districts $\mathcal{D}(\mathcal{G})$ of the graphical structure have been detected, it is possible to simplify the identification of interventional distributions leveraging the structure of the unobserved variables. Based on this simplification, interventional distributions can be factorized into the multiple interventional distributions on the single districts. This is presented in Lemma 4 of [83], which can be found here:

Lemma B.0.1. *Let $\mathcal{M}$ be a causal model with graph $\mathcal{G}$ with set of observed nodes V. Let y, a be value assignments respectively belonging to $\mathcal{X}_Y$ and $\mathcal{X}_A$. Let $\mathcal{D}(\mathcal{G}_{\setminus A}) = \{D_1, ..., D_k\}$ with $0 \leq k \leq n$. Then,*

$$P(Y(a) = y) = \sum_{V\setminus(Y\cup A)} \prod_{i\in\{1,..,k\}} P(D_i(V\setminus D_i = v\setminus d_i) = d_i).$$

Here, the values of v and d_i are consistent with the intervention $A = a$.

The proof will follow the one in [83]:

Proof. Assume first that $A = \{\emptyset\}$ and that Pa_{D_i} is the set of parents of the district D_i excluding D_i itself. Then, we have:

$$\begin{aligned}
\prod_i P(D_i(V\setminus D_i = v\setminus d_i) = d_i) &= \prod_i P(D_i(Pa_{D_i} = pa_{D_i}) = d_i) \\
&= \prod_i \prod_{V_j\in D_i} P(V_j(Pa_{D_i} = pa_{D_i}) = v_j | pre_{\prec_\mathcal{G}}(V_j)\setminus Pa_{D_i}) \\
&= \prod_i \prod_{V_j\in D_i} P(V_j = v_j | pre_{\prec_\mathcal{G}}(V_j)) \\
&= \prod_i P(V_i = v_i | pre_{\prec_\mathcal{G}}(V_i)) \\
&= P(V = v)
\end{aligned}$$

By marginalization, $P(Y = y) = \sum_{V\setminus Y} \prod_i P(D_i(V\setminus D_i) = d_i)$. In the above equalities, the first one follows from Rule 3 of do-calculus (see Formula 3.20), since the elements of D_i are independent from the variables $V\setminus\{Pa_{D_i} \cup D_i\}$ when intervening on Pa_{D_i}. The second equality follows from the chain rule of probability, referring to the topological ordering of observable nodes in $\mathcal{G}$ (Definition 3.1.0.6). Regarding the third equality, for every term corresponding to the variable V_j in D_i, the nodes $pre_{\prec_\mathcal{G}}(V_j)$ can either intersect or not with the nodes Pa_{D_i}. In either of these scenarios, the equality is satisfied by using either Rule 2 or Rule 3 of do-calculus (Formulas 3.19, 3.20). The last two equalities follow from grouping up the terms and the chain rule.
The same factorization can be applied to the quantity $P(V(a) = v)$, by considering the causal graph $\mathcal{G}_{\bar{a}}$, corresponding to the intervention $do(A = a)$ on the entire set of nodes V. □

This iterative composition of the effect into districts is what characterizes the ID algorithm. The algorithm itself will not bw explained in detail, as it falls outside the scope of the project. Nevertheless, Lemma B.0.1 will be helpful for understanding the identification of Path-Specific Interventions 4.2.0.1. The reader can explore further into the ID Algorithm by referring to the sources [96] and [83].

C
Objective Function Formulation

The objective functions of the optimization problems in (5.2) and (5.4) are equivalent. Indeed,

$$\begin{aligned} D_{KL}\Big[P(X;\beta^*)||P(X;\beta)\Big] &= \mathbb{E}_{X\sim P(X;\beta^*)}\left[log\frac{P(X;\beta^*)}{P(X;\beta)}\right] \\ &= \mathbb{E}_{X\sim P(X;\beta^*)}\Big[logP(X;\beta^*) - logP(X;\beta)\Big] \\ &= \mathbb{E}_{X\sim P(X;\beta^*)}\big[logP(X;\beta^*)\big] - \mathbb{E}_{X\sim P(X;\beta^*)}\big[logP(X;\beta)\big] \end{aligned}$$

Hence, we have:

$$\begin{aligned} \hat{\beta} &= argmin_\beta D_{KL}\Big[P(X;\beta^*)|P(X;\beta)\Big] \\ &= argmin_\beta\Big\{\mathbb{E}_{X\sim P(X;\beta^*)}\big[logP(X;\beta^*)\big] - \mathbb{E}_{X\sim P(X;\beta^*)}\big[logP(X;\beta)\big]\Big\} \\ &= argmin_\beta\Big\{-\mathbb{E}_{X\sim P(X;\beta^*)}\big[logP(X;\beta)\big]\Big\} \\ &= argmin_\beta\left\{-\frac{1}{N}\sum_i^N logP(X_i;\beta)\right\} \\ &= argmin_\beta\Big\{c\cdot NLL\Big\} \\ &= argmax_\beta\mathcal{L}(D;\beta) \end{aligned}$$

Where c is a constant and NLL the negative log-likelihood. All the previous equalities can be proved using probability rules. Among these, in the fourth equality, N of $X_i \overset{i.i.d.}{\sim} P(X;\beta^*)$ were sampled and the Law of Large Number for N that goes to infinity was used.

D

Fair Inference on Outcomes - Detailed Overview

In this Appendix, further details regarding the method Fair Inference on Outcomes method introduced in [61] and further developed in [62] will be given. Different aspects of the method will be discussed, such as the prediction procedure on the test set, the choice of the optimization algorithm, the estimator of the edge g-formula and the algorithm training procedure.

D.1. Prediction at Test Level

The prediction step of the FIO method has been widely discussed in different works, emphasizing its importance [61], [62],[15]. An original idea of the prediction procedure for new instances at test time was introduced in [61]; given the limitations of this approach, this method was further developed by the same authors in [62]. In the implementation of this project, the version of [62] was used.

In section 5.1.1, an overview of FIO method was given. In particular, it has been motivated the choice of transferring the inference problem from the observed data distribution $P(Y, X, A)$ to the fair distribution $P^*(Y, X, A)$. Based on this introduction, we would expect the outcome y corresponding to new instances $\{a, x\}$ to be predicted as $y = \mathbb{E}^*(Y|X = x, A = a)$, where $\mathbb{E}^*$ is the expected value induced by the fair conditional probability distribution $P^*(Y = y|X = x, A = a)$. Nevertheless, further aspects of the optimization problem need to be taken into account.
In the formulation of the optimization problem, a set of variables $W \in V$ can be defined such that $P^*(W = w|Pa(W)) = P(W = w|Pa(W))$, for any value $w \in \mathcal{X}_W$. These are thus variables that are shared between the observed and the fair distributions. These distributions depend on the solution of the constrained optimization problem; for instance, if the coefficients corresponding to the posterior distribution of W are not part of the input parameters of the constrained optimization problem β, then it will hold that $P^*(W = w|Pa(W)) = P(W = w|Pa(W))$. Generally, by using the plug-in estimator for edge g-formula, both the baseline features (root nodes) and the sensitive attribute are not used in the optimization algorithm, meaning that in the example given by Figure 4.2 $W = \{A, C\}$.
When predicting a new outcome using the fair distribution $P^*(Y = y|X = x, A = a)$ on new instances $\{a, x\}$, these should be drawn from the fair distribution itself. Nevertheless, the new instances are actually drawn from the observed data distribution $P(X, A)$, which does not ensure that unfairness is removed. Hence, it is important to find a respective fair version of the new instance $\{a, x\}$. The values of $\{a, x\}$ over the set of variables W can be seen as allowed inputs, since the observed and fair distribution on these variables are equivalent, leading the set W to be potentially drawn from P^* itself.
In order to solve this problem, in [61], the new instances are averaged over the possible values of $V \setminus \{W \cup Y\}$, weighted by their fair probability P^*. In fact, since it is not possible to estimate the

values of the variables $V \setminus \{W \cup Y\}$ that the new 'fair instance' would obtain, they can be estimated by how likely they are to be drawn from P^*. Thus, in [61] the outcome Y is predicted by using $\mathbb{E}^*[Y|W = w]$. In the example of Figure 4.2 by using the plug-in estimator, given a new instance $\{a_i, m_i, c_i\}$ with $w_i = [a_i, c_i]$, we have that the newly predicted outcome $\hat{Y}_i$ is the following:

$$\hat{Y}_i = \mathbb{E}^*[Y|W = w_i] = \mathbb{E}^*[Y|A = a_i, C = c_i] = \sum_{m \in \mathcal{X}_M} \mathbb{E}^*[Y|A = a_i, C = c_i, M = m]P^*(M = m|A = a_i, C = c_i).$$

Analysing the approach of predicting new instances by averaging over the mediators values, it can be noticed that a high lack of accuracy would be involved, as not all the information regrading the new instances is considered. For instance, with this averaging operation, the outcomes of any two data points with the same values of W are the same, independently of the other variables. In [62], the same authors find an alternative training procedure in order to avoid the complication of predicting the outcomes of new instances.

In the original FIO method, the constrained optimization problem is applied only to the training set. Thus, at test time the averaging operation is used to apply the new 'fair' distribution to new instances of the test time. Nevertheless, in [62] it is shown that training the model in a batch setting could relevantly improve the predictions' accuracy. It is hence assumed that the missing labels corresponding to the test set are 'missing at random' labels. In order to take this into account, a new variable R is introduced, where $R = 0$ denotes the missingness of the label Y_i, $R = 1$ the opposite. All the historical data points (training data) will have the value of R set to 1, meanwhile the rest of the data points (test data) to 0. Based on this, the plug-in estimator will only be estimated on the training data points, while the test data-points will be used for the cost function of the constrained optimization problem. Indeed, given the number of training data points n_1 and number of test data points n_2, the likelihood function for the dataset $\mathcal{D}$ composed by the data points $\{x_i, a_i, y_i, r_i\}$ with $i \in \{1, ..., n_1 + n_2\}$ will be:

$$\prod_{i=1}^{n=n_1+n_2} P(X = x_i, A = a_i)P(Y = y_i|X = x_i, A = a_i)^{r_i},$$

where $r_i = 0$ if the data point is part of the test data and $r_i = 1$ otherwise.

In conclusion, by modifying the training procedure it is possible to predict the outcomes of new instances without the averaging operation.

D.2. Algorithm Improvements and Choice of Estimator

In the new work [62], further improvements on the methodology and the choice of the estimator were made. These involve a reparametrization of the likelihood function and the PSE, the choice of the inputs of the optimization algorithm and the choice of the semi-parametric estimator for the edge g-formula.

Because of the computational challenges of the constrained optimization problem, a reparametrization of the problem is made in order to express the PSE as causal parameters and the observed data likelihood in terms of these parameters. In this way, the computational challenges are improved because the constraints of the optimization problem will be box constraints, leading the resolution of the problem to be more efficient.

Further modification have been proposed in [62] for improving the performance of the method. In the original FIO method, only part of the likelihood is constrained. As a matter of fact, given a set of baseline features C, the respective marginal density is not constrained, but estimated by using $\frac{1}{n}$ mass over all the observed points belonging to $\mathcal{X}_C$. The improvement of [62] consists in using a hybrid/semi-parametric empirical likelihood method in order to estimate the marginal distribution of C non-parametrically.

In the following theorem it is shown that constraining a larger part of the joint distribution will potentially lead to a KL-closer fair distribution as a solution:

Theorem D.2.1. *Let $p(V)$ denote the observed data distribution, $M_1 = \{p_1^*(V) = argmin_{q(V)} D_{KL}(p||q)$, s.t. $\epsilon_l \leq g(q(V)) \leq \epsilon_u$, and $q(V_1) = p(V_1)\}$, and $M_2 = \{p_2^*(V) = argmin_{q(V)} D_{KL}(p||q)$, s.t. $\epsilon_l \leq g(q(V)) \leq \epsilon_u$, and $q(V_2) = p(V_2)$. If $V_2 \subseteq V_1 \subseteq V$, then $D_{KL}(p||p_2^*) \leq D_{KL}(p||p_1^*)$.*

From the above theorem, it can be proven that constraining an additional part of the joint distribution, for instance the distribution of C, can lead to a higher accuracy of the model's predictions. Indeed, the estimated probability distribution will be closer to the original one.

In both of the works of [61] and [62], different semi-parametric estimators for the edge g-formula are considered. In [95] different consistent estimators for the edge g-formula are introduced: the plug-in estimator, the inverse probability weighting estimator (IPW), Mixed estimator and the augmented inverse probability weighting (AIPW). In this work only the plug-in estimator was used, nevertheless for future work it is suggested to implement the same methods varying the used estimator in order to improve its robustness. In the previous works, these estimators have mainly been applied to NDE identification formulas, but they could potentially be generalized to arbitrary PSE in the future.

E

Estimator Performance

In this Appendix, further results regarding the performance of the semi-parametric plug-in estimator on another causal structure are presented.
The simulation procedure of Appendix J.2.4 is considered, referring to the graphical structure of Figure 6.4.
In the below figures, the shifted bootstrap and true sampling distributions are presented, along with Table E.1 useful for deriving the consistency of the estimator. The plots of the quantiles and the cumulative distribution of the shifted versions show the similarity between their variability. The standard deviation of the true sampling distribution and the bootstrap one are respectively 0.104 and 0.0097, very comparable. The conclusions encountered in Chapter 5 can also be similarly inferred from this example. The only aspect that differs in the two examples is the QQplot in Figure E.2, showing a slight difference in the quantiles distributions. This is however a minor detail, that can indeed be ignored for the scope of the project.

n	Standard Deviation
1000	0.0165
2000	0.0111
4000	0.0076
8000	0.0054
...	...
100000	10^{-5}

Table E.1: Plug-in Estimator Performance on datasets with n data-points

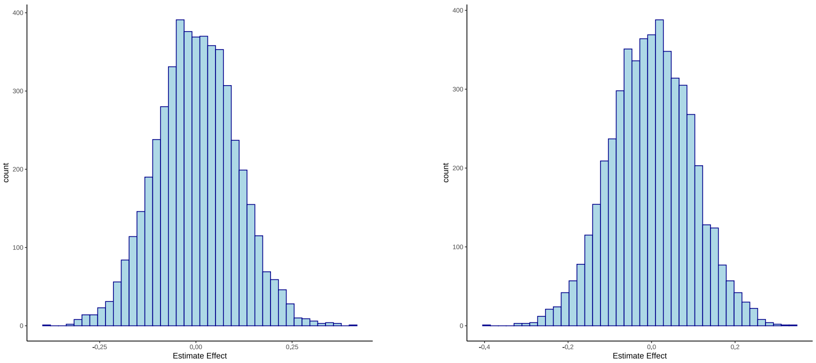


Figure E.1: True sampling distribution (left) and bootstrap distribution (right)

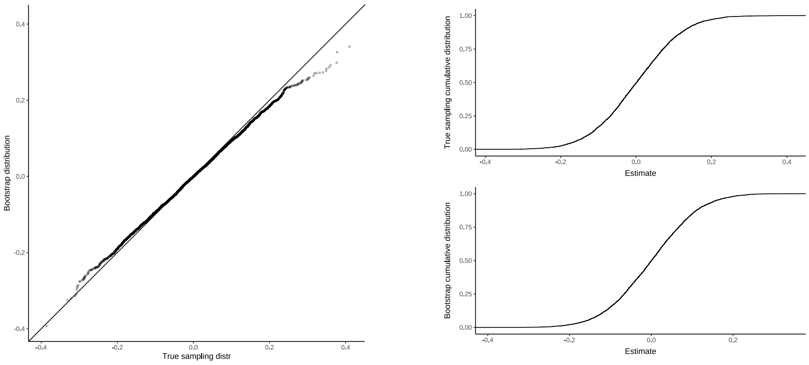


Figure E.2: Quantile-Quantile plot (left) and cumulative distributions (right)

F

Conditioning or Intervening?

In this paragraph, the difference between conditioning and intervening will be explored. Specifically, referring to a causal structure $\mathcal{G}$, given an interventional distribution of a variable Y with respect to a treatment variable A and an additional set of variables C, we are interested in analysing the difference between using C as a further treatment attribute or as a conditioning one. Hence, the comparison between $P(Y(a) = y|C = c)$ and $P(Y(a, c) = y)$ with $c \in \mathcal{X}_C, a \in \mathcal{X}_A, y \in \mathcal{X}_Y$ will be given. The methodological difference has been covered in the Chapter 6.1.3, in this section the intuitive difference will be analysed, in order to help the reader better understanding the motivation of the newly proposed method CPSF based upon conditional interventional distributions.

An example illustrating the difference between $P(Y(a) = y|C = c)$ and $P(Y(a, c) = y)$ will now be presented. When interpreting interventions, it is more immediate to use practical examples in which the treatment variable can be controlled in experiments. For instance, applications of interventions in the field of personalised medicine can help in the intuition of the concepts. A research on sleeping pills consumption is considered. The treatment variable in this example is the consumption of sleeping pills ($A = \{0, 1\}$ active treatment / non-active treatment). In this example, the graphical structure is presented in Figure F.1, where Y is the approximated hours of sleep per night ($\mathcal{X}_Y = \{0, 1, ..., 24\}$), C is the time of going to bed ($\mathcal{X}_C = \{c \in \mathbb{R} : 0 \leq c \leq 24\}$), W is the working employment ($\mathcal{X}_W = \{employed, unemployed, student, retired, ...\}$).
The scope of the example is to analyse the causal effect that the sleeping pills have on the sleeping behaviour of the individuals, specifically with respect to a threshold of 7 hours/night. In this example, considering the interventional distribution $P(Y(A = 1) > 7)$ answers the question: what is the effect that the treatment (sleeping pills) has on the hours of sleep? Differently, by involving the age of the applicants C, $P(Y(A = 1, C = 21) > 7)$ corresponds to the question: what is the effect of the treatment if the patients assume the sleeping pills and go to bed at $21h$? Moreover by instead conditioning on the time of going to bed, $P(Y(A = 1) > 7|C = 21)$ answers the question: what is the effect of the treatment on the people who normally go to bed at $21h$?
The main conceptual difference between the interventional distributions $P(Y(A = 1, C = 21) > 7)$ and $P(Y(A = 1) > 7|C = 21)$ is that by selecting individuals that go to bed at $21h$, then additional information is taken into consideration regarding the subjects, for instance their working state. Indeed, since C is a child of A, the individuals going to bed at $21h$ are the ones with a certain working status, subjects to specific stress that might influence the quality of sleep. Hence in $P(Y(A = 1) > 7|C = 21)$, only individuals in that specific condition are considered. Differently, $P(Y(A = 1, C = 21) > 7)$ analyses the sleeping behaviour of all the individuals, independently of their working status, if they happened to go to bed at $21h$. This concept is supported by the graphical representation of interventions, according to which in this example the edges entering C would be eliminated.

This example shows that the difference between the two interventional distributions corresponds to the scope of the intervention itself and to the question that we want to answer.

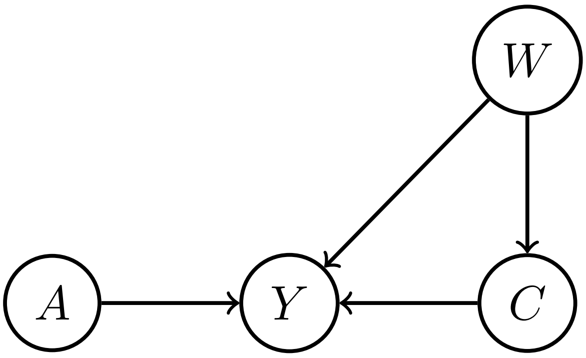


Figure F.1: Example Personalized Medicine

G

Conditional Path-Specific Interventions Proofs

The following Propositions and proofs are introduced in [54].

Proposition G.0.1. *Given a* DAG $\mathcal{G}$ *with distribution on the potential outcomes* $P(\mathbb{V})$. *Consider the corresponding element* $P_{\mathcal{G}^e_A}(V^e)$ *in the extended causal model associated with* $\mathcal{G}^e_A(V \cup A^{Ch})$. *Then* $P(V(\pi, a_1, a_0)) = P_{\mathcal{G}^e_A}(V(a^\pi))$.

Proof. By definition of the model $\mathcal{G}$ and structural equations:

$$P(V(\pi, a_1, a_0) = v) = \sum_{\epsilon_i : f_i(a_{0 Pa_i^\pi}, a_{1 Pa_i^{\hat{\pi}}}, v_{Pa_i \setminus A}) = v_i} P(\epsilon_1, ..., \epsilon_k),$$

where ϵ_i are the disturbance terms of the structural equations. Where for each V_i, Pa_i^π is the subset of $Pa_i \cap A$ with an edge from Pa_i to V_i in π, and $Pa_i^{\hat{\pi}}$ is the subset of $Pa_i \cap A$ with an edge from Pa_i to V_i not in π. Similarly, for $\mathcal{G}^e(V \cup A^{Ch})$,

$$P_{\mathcal{G}^e_A}(V(a^\pi) = v) = \sum_{\epsilon_i : f_i(a^\pi_{Pa_i \cap A_i}, v_{Pa_i \setminus A}) = v_i} P(\epsilon_1, ..., \epsilon_k),$$

where by $a^\pi_{Pa_i \cap A_i}$ is meant the interventional values consistent with the newly added nodes A^{Ch}. Since the two equations are equivalent, the argument follows. □

Proposition G.0.2. *For any* $V \subseteq Y$, $p(Y(\pi, a, a'))$ *is identified in the ADMG* $\mathcal{G}_V$ *if and only if* $p(Y(a^\pi))$ *is identified in the ADMG* $\mathcal{G}^e(V, A^{Ch})$. *Moreover if* $p(Y(a^\pi))$ *is identified, it is by the same functional as* $p(Y(\pi, a, a'))$.

Proof. The proof follows from the proof of the previous proposition. □

Proposition G.0.3. *If* $\Big(Y(x,z) \perp\!\!\!\perp Z(x,z) | W(x,z)\Big)_{\mathcal{G}^e(x,z)}$ *and* $T \subseteq W$ *then* $\Big(Y(x,t) \perp\!\!\!\perp T(x,t) | Z(x,t), W_1(x,t)\Big)_{\mathcal{G}^e(x,t)}$ *if and only if* $\Big(Y(x,z,t) \perp\!\!\!\perp T(x,z,t) | Z(x,z,t), W_1(x,z,t)\Big)_{\mathcal{G}^e(x,z,t)}$, *where* $W_1 = W \backslash T$.

Proof. By comparing the d-connecting paths between the pairs $(Y(x,z,t), T(x,z,t))$ in $\mathcal{G}^e(x,z,t)$ and $(Y(x,t), T(x,t))$ in $\mathcal{G}^e(x,t)$, it can be noticed that the possible connecting paths from $Y(x,z,t)$ to $T(x,z,t)$ in $\mathcal{G}^e(x,z,t)$ is a subset of the paths from $Y(x,t)$ to $T(x,t)$ in $\mathcal{G}^e(x,t)$. Then, if there exists a set $W_1(x,t)$ in $\mathcal{G}^e(x,t)$ that blocks a path, then $W_1(x,z,t)$ also blocks the ones in $\mathcal{G}^e(x,z,t)$. Further, by construction of $\mathcal{G}^e(x,z,t)$, if $Z(x,t)$ blocks a path in $\mathcal{G}^e(x,t)$, then this is also blocked in $\mathcal{G}^e(x,z,t)$. If instead a path is clocked by descendants of $Z(x,t), W_1(x,t)$, then the same will happen in $\mathcal{G}^e(x,z,t)$. Hence, if $(Y(x,t) \perp\!\!\!\perp T(x,t) | Z(x,t), W1(x,t))_{\mathcal{G}^e(x,t)}$, then $(Y(x,z,t) \perp\!\!\!\perp T(x,z,t) | W1(x,z,t))_{\mathcal{G}^e(x,z,t)}$.

We will now assume for contradiction that $(Y(x,z,t) \perp\!\!\!\perp T(x,z,t)|W1(x,z,t))_{\mathcal{G}^e(x,z,t)}$ but that $(Y(x,t) \perp\!\!\!\perp T(x,t)|Z(x,t), W1(x,t))_{\mathcal{G}^e(x,t)}$ does not hold. If there is a connecting path in $\mathcal{G}^e(x,t)$ but not in $\mathcal{G}^e(x,z,t)$, then it must contain a non-collider (introduced in Section 6.1.3) through an element of Z. If this is still a possible connecting path in $\mathcal{G}^e(x,z,t)$, then there must exist a collider (with no descendants in $W_1(x,z,t)$) that blocks the path in $\mathcal{G}^e(x,z,t)$.
Hence, there must exist a path in $\mathcal{G}^e(x,t)$ from (an element of) $Y(x,t)$ to (an element of) $Z(x,t)$ given $W_1(x,t)$. Furthermore, since in $\mathcal{G}^e(x,t)$ the node-splitting variable $T(x,t)$ will not block this path by construction, the previous consideration can be made for the entire set $W(x,t)$. For this reason, the assumption made $\Big(Y(x,z) \perp\!\!\!\perp Z(x,z)|W(x,z)\Big)_{\mathcal{G}^e(x,z)}$ is contradicted, leading to the if and only if proposition. □

Corollary G.0.3.1. *For any $\mathcal{G}^e(a)$ and any conditional distribution $p(Y(x)|W(x))$, there exists a unique maximal set $Z(x) = \{Z_i(x) \in W(x)|p(Y(x)|W(x)) = p(Y(x,z_i)|W(x,z_i)\backslash\{Z_i(x,z_i)\})\}$ such that Rule 2 applies for $Z(x,z)$ in $\mathcal{G}^e(a)$ for $p(Y(x,z)|W(x,z))$.*

Proof. This proof is based on the proposition G.0.3. Getting back to the Corollary, if two sets $Z_1(x)$ and $Z_2(x)$ are fixed such that Rule 2 is satisfied with respect to $\mathcal{G}^e(x,z)$ for $p(Y(x,z)|W(x,z))$. If $Z_1(x) \neq Z_2(x)$, specifically $Z_2(x) \subseteq Z_1(x)$, then it is possible to define $T(x) = Z_1(x)\backslash Z_2(x)$. Based on this, the hypothesis of $Z_2(x)$ being the maximal set is actually contradicted because, following the proposition that was mentioned above, $Z_2(x) \cup T(x)$ satisfies the Rule 2 condition. □

H

Example Equivalence Path-Specific Effect and Conditional Path-Specific Effect

First, a model in which the application of CPSF is equivalent to FIO will be considered. This could indeed help the reader in better understanding the similarity between FIO and CPSF. Then, different data-generating processes will be considered.

The Figure 5.1 will be considered, where C is the set of conditioning covariates. The first analysed linear setting is the most simplified one, as it does not include mixed terms in the regressions. This is a special scenario, as the conditional PSE on C is equivalent to the PSE, meaning that the original FIO method and the newly proposed CPSF are equivalent.
The considered linear model is the following:

$$\begin{aligned}
C \sim& \mathcal{N}(0,1) \\
logit(P(A=1|C)) \sim& \beta_A + \beta_A^C C; \\
M =& \beta_M + \beta_M^C C + \beta_M^A A + \epsilon_M; \\
Y =& \beta_Y + \beta_Y^C C + \beta_Y^A A + \beta_Y^M M + \epsilon_Y,
\end{aligned}$$

where $\epsilon_M, \epsilon_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. In this scenario, it is immediate to re-write the Path-Specific Effect via the edge g-formula in terms of the linear coefficients β. By remembering that the fair path from A to Y is considered to be $(AMY)_{\rightarrow}$, then the Path-Specific Effect is:

$$\begin{aligned}
&\mathbb{E}[Y(a_0, M(a_1))] - \mathbb{E}[Y(a_1, M(a_1))] = \\
=&\sum_{c \in \mathcal{X}_c} \left(\beta_Y + \beta_Y^C c + \beta_Y^A a_0 + \beta_Y^M (\beta_M + \beta_M^C c + \beta_M^A a_1)\right) P(C=c) \\
&- \sum_{c \in \mathcal{X}_c} \left(\beta_Y + \beta_Y^C c + \beta_Y^A a_1 + \beta_Y^M (\beta_M + \beta_M^C c + \beta_M^A a_1)\right) P(C=c) \\
=&\sum_{c \in \mathcal{X}_c} \left(\beta_Y^A a_0\right) P(C=c) - \sum_{c \in \mathcal{X}_c} \left(\beta_Y^A a_1)\right) P(C=c) \\
=&\sum_{c \in \mathcal{X}_c} \left(\beta_Y^A (a_0 - a_1)\right) P(C=c) \\
=&\beta_Y^A (a_0 - a_1).
\end{aligned}$$

In this simplified setting, it can be noticed that the Path-Specific Effect is independent of C, meaning that the conditional PSE and the PSE are equivalent. In this specific example, in order to ensure that the PSE is equal to 0 for $a_1 \neq a_0$, it is necessary to estimate $\hat{\beta}_Y^A = 0$.

The above example represents a simplified version of a causal model with linear structural equations. Indeed, linear regressions including mixed coefficient terms could be considered. For instance, referring to causal graph in 4.2, the following linear setting will be used:

$$\begin{aligned}
C \sim& \mathcal{N}(0,1) \\
logit(P(A=1|C)) \sim& \beta_A + \beta_A^C C; \\
M =& \beta_M + \beta_M^C C + \beta_M^A A + \beta_M^{AC} AC + \epsilon_M; \\
Y =& \beta_Y + \beta_Y^C C + \beta_Y^A A + \beta_Y^M M + \beta_Y^{AC} AC + \beta_Y^{CM} CM + \beta_Y^{AM} AM + \epsilon_Y,
\end{aligned}$$

where $\epsilon_M, \epsilon_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. In this causal model, following the same procedure are previously described, it can be immediately proved that the Path-Specific Effect is:

$$\begin{aligned}
&\mathbb{E}(a_0, M(a_1)) - \mathbb{E}(a_1, M(a_1)) = \\
&= \sum_{c \in \mathcal{X}_c} \left(\beta_Y^A (a_0 - a_1) + \beta_Y^{AC} (a_0 - a_1) c + \beta_Y^{AM} (a_0 - a_1)(\beta_M + \beta_M^C c + \beta_M^A a_1 + \beta_M^{AC} a_1 c)\right) P(C = c) \\
&= \sum_{c \in \mathcal{X}_c} (a_0 - a_1)(\beta_Y^A + \beta_Y^{AC} c + \beta_Y^{AM} (\beta_M + \beta_M^C c + \beta_M^A a_1 + \beta_M^{AC} a_1 c)) P(C = c).
\end{aligned}$$

As it can be noticed, the formulation of the Path-Specific Effect is dependent on the covariate C, trivially leading to distinct values of PSE and conditional PSE.

I

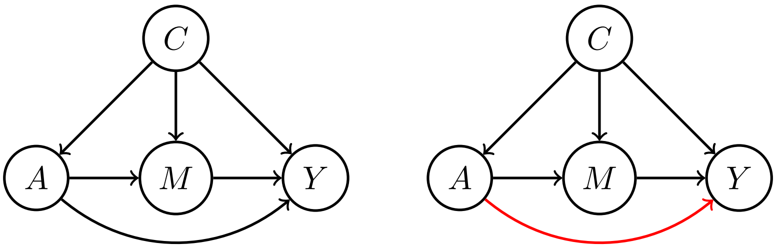
Causal Graphs of Real-World Examples

In this section, the causal graphs of the real-world examples introduced in Section 2.2 will be presented. In particular, the causal graphs of both the COMPAS Dataset and the UCI Adult dataset will be analysed.

I.1. COMPAS Dataset - Parole Prediction

One of the most exemplifying causal structures of the fairness research is the one corresponding to the COMPAS dataset model, introduced in Section 2.2. The causal graph corresponding to this dataset is represented in Figure I.1a. The node C represents the set of demographic features gender and age, the sensitive attribute A is the race, the mediator M is the number of prior convictions and Y is the recidivism decision. The pre-processing of the dataset involved grouping the age covariate into four main subgroups. The same approach of [61] was used.

From the causal structure, it can be noticed that the race has a possible direct causal influence on both the number of prior convictions and the recidivism decision. Based on domain knowledge and ethical intuition, the causal paths between the sensitive attribute A and the recidivism decision Y can be classified as ethically plausible or not. Specifically, in the literature it is common to assume that it is fair to consider the number of prior convictions in order to make the recidivism decisions; nevertheless, this does not hold for the race. Hence, following the terminology of this work, the set of unfair paths is $\pi = \{(AY)_{\rightarrow}\}$, referred to the color red in the Figure I.1b.



(a) Causal Structure

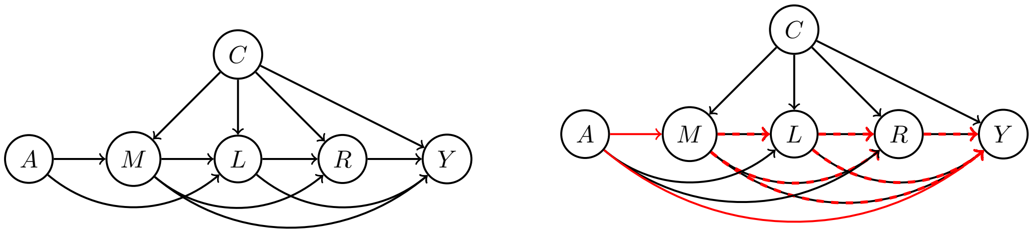
(b) Causal Structure with Unfair Paths in Red

Figure I.1: COMPAS Dataset

I.2. UCI Adult Dataset - Loan Application

The second example introduced in Section 2.2 regards the loan application setting. It was mentioned that the most used dataset in this application is the UCI Adult dataset, the causal structure of which is represented in Figure I.2a. The dataset includes job related variables for 48842 individuals. In this dataset, the sensitive attribute A represents the gender of the applicants and the binary outcome Y is the salary prediction of the candidate, classifying it as higher or lower than $\$50k/year$. The baseline features C include the age and nationality, while the mediator M represents the marital status. Furthermore, L is the number of years of education and R the set of working-related features such as the number of hours worked and the kind of occupation. The original dataset includes more variables (14 in total); however, a similar approach to [110] was used for the pre-processing. Nevertheless, only part of the working-related variables have been used; these are the number of hours spent working per week and a binary variable stating whether the occupation is a private one or not. The same causal structure presented in the literature is used in Figure I.2a. Nevertheless, we do not necessarily agree with this causal structure, especially regarding the Marital Status. Indeed, the causal edge from the feature gender A and the marital status M (binary variables for married or non-married status) is not justified in a straightforward way in this context. By analysing the conditional distribution of the marital status in Figure I.3, it can be noticed that among the Non-Married individuals (considered as Divorced, Separated, Single, Widowed) a relevantly higher number of women joined the survey. This could happen because in married couples it is more likely that men apply for a loan. Nevertheless, based on our common knowledge a more suitable graph would include the Married status not as a mediator but as a baseline feature.

The causal structure of the UCI Adult Dataset is represented in Figure I.2a. The respective classification of causal paths from A to Y into fair and unfair paths is represented in Figure I.2b. The red edges are considered unfair, and the dashed one are unfair as a consequence of the red ones. Indeed, for instance, $(LR)_{\rightarrow}$ is unfair along the path $A \rightarrow M \rightarrow L \rightarrow R \rightarrow Y$, but fair along $A \rightarrow L \rightarrow R \rightarrow Y$. In the literature, it is assumed that the unfair paths are $\pi = \{A \rightarrow Y, A \rightarrow M \rightarrow ... \rightarrow Y\}$. As earlier mentioned, we do not agree with this causal structure and classification of the causal pathways as fair or unfair, nevertheless a similar approach to the state-of-the-art works was used.



(a) Causal Structure

(b) Causal Structure with Unfair Paths in Red

Figure I.2: UCI Adult Dataset

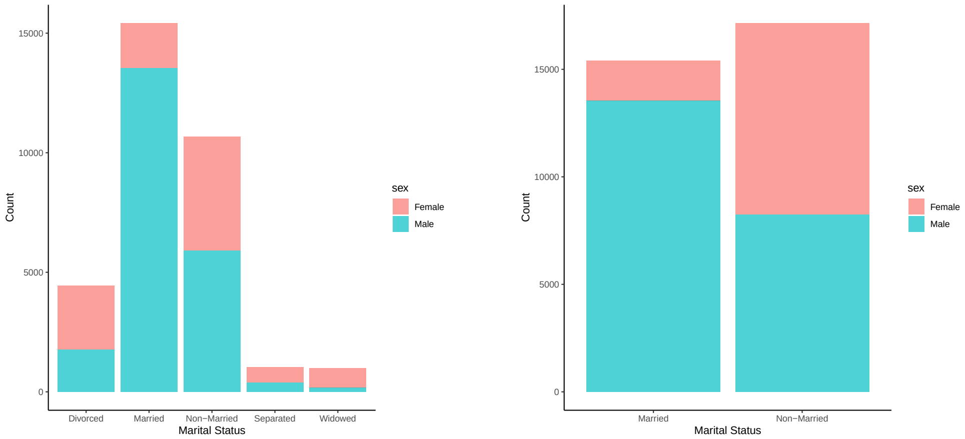


Figure I.3: Count Plot of Marital-Status

J

Data Simulation

In this chapter, the different simulation dataset procedures will be presented, referring to the experiments of Chapter 5.2 and Chapter 7.

J.1. Data Structure 1

The causal graph of Example 1 is presented in Figure J.1. In the next sections the different data simulation procedures will be concisely presented.

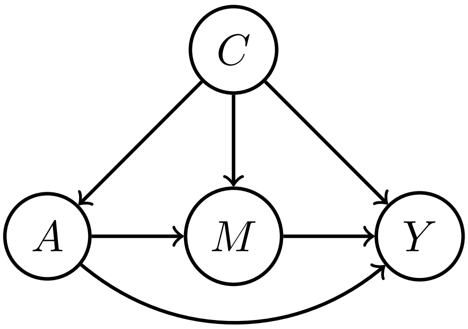


Figure J.1: Data Structure 1 Causal Graph

J.1.1. Linear Prediction Setting Discrete C

$$P(C=c)=\begin{cases}\frac{1}{3} & \text{if} \quad c=0\\ \frac{1}{3} & \text{if} \quad c=1\\ \frac{2}{9} & \text{if} \quad c=2\\ \frac{1}{9} & \text{if} \quad c=3\end{cases}$$

$$\begin{aligned} logit(P(A=1|C)) &\sim 0.5-C;\\ M &= 0.1+2C-A+3.5AC+\epsilon_M;\\ Y &= 1-C+2A-M-3AC-0.5AM-3CM+\epsilon_Y.\end{aligned}$$

Here, $\epsilon_M, \epsilon_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

J.1.2. Linear Prediction Setting Continuous C

$$\begin{aligned} C \sim & \mathcal{N}(0,1); \\ logit(P(A=1|C)) \sim & 0.5 - C; \\ M = & 0.1 + 2C - A + 3.5AC + \epsilon_M; \\ Y = & 1 - C + 2A - M - 3AC - 0.5AM - 3CM + \epsilon_Y. \end{aligned}$$

Here, $\epsilon_M, \epsilon_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

J.1.3. Linear Prediction Setting Binary C

$$\begin{aligned} C \sim & Bin(0.7); \\ logit(P(A=1|C)) \sim & 0.5 - C; \\ M = & 0.1 + 2C - A + 3.5AC + \epsilon_M; \\ Y = & 1 - C + 2A - M - 3AC - 0.5AM - 3CM + \epsilon_Y. \end{aligned}$$

Here, $\epsilon_M, \epsilon_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

J.1.4. Classification Setting Binary C

$$\begin{aligned} C \sim & Bin(0.7); \\ logit(P(A=1|C)) \sim & -0.5 + 0.5C; \\ logit(P(M=1|A,C)) \sim & 0.05 - 1.5C + 2A - 0.05AC; \\ logit(P(Y=1|M,A,C)) \sim & -1 + 5C - 4A + 2M - 7AC - 3AM + 3CM. \end{aligned}$$

J.1.5. Linear Setting with Unobserved Variables

$$\begin{aligned} C \sim & Bin(0.7); \\ U \sim & \mathcal{N}(1,1); \\ logit(P(A=1|C)) \sim & -0.5 + 0.5C; \\ M = & -0.1 + 1.5C - 2A + 0.05AC + U + \epsilon_M; \\ Y = & 1 - 5C + 4A - 2M + 7AC + 3AM - 3CM + 2U + \epsilon_Y. \end{aligned}$$

Here, $\epsilon_M, \epsilon_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

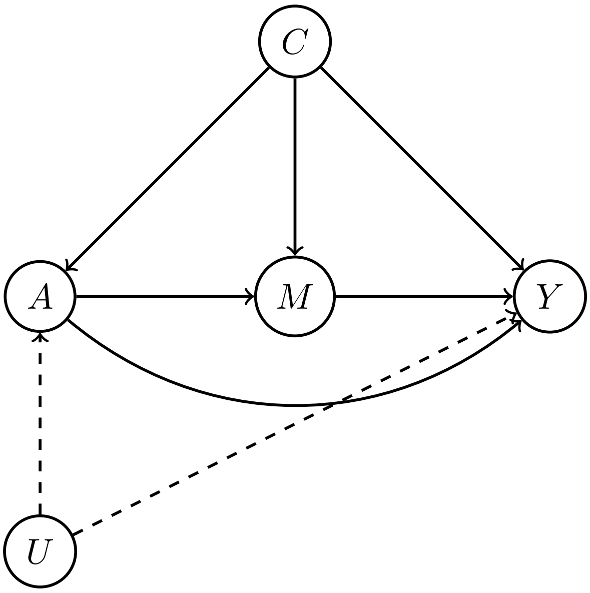


Figure J.2: Data Structure 1 with additional unobserved variable

J.2. Data Structure 2

The causal graph of Example 2 is presented in Figure J.3. Similarly to the previous graphical structure, the different data simulation procedures will now follow.

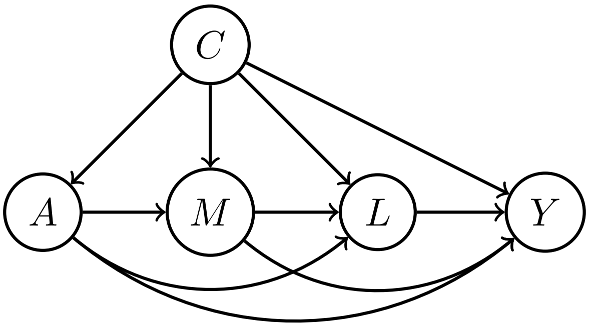


Figure J.3: Data Structure 2 Causal Graph

J.2.1. Prediction Setting Continuous C

$$\begin{aligned}
C \sim & \mathcal{N}(0,1); \\
logit(P(A=1|C)) \sim & -0.5 - 0.5C; \\
logit(P(M=1|A,C)) \sim & -0.5 - C + 4A - 3AC; \\
logit(P(L=1|M,A,C)) \sim & -0.5 - 0.4C - 2A + 3M - 0.6AC + 2AM + 0.5ACM; \\
Y = & 1 + C + 5A - 4M + 3L - 2AC + AM + AL - 1.5AML + \epsilon_Y,
\end{aligned}$$

where $\epsilon_Y \sim \mathcal{N}(0,1)$.

J.2.2. Prediction Setting Discrete C

$$
\begin{aligned}
P(C=c) =& \begin{cases} \frac{1}{3} & \text{if} \quad c=0 \\ \frac{1}{3} & \text{if} \quad c=1 \\ \frac{2}{9} & \text{if} \quad c=2 \\ \frac{1}{9} & \text{if} \quad c=3 \end{cases} \\
logit(P(A=1|C)) \sim& -0.5-0.5C; \\
logit(P(M=1|A,C)) \sim& -0.5-C+4A-3AC; \\
logit(P(L=1|M,A,C)) \sim& -0.5-0.4C-2A+3M-0.6AC+2AM+0.5ACM; \\
Y =& 1+C+5A-4M+3L-2AC+AM+AL-1.5AML+\epsilon_Y,
\end{aligned}
$$

where $\epsilon_Y \sim \mathcal{N}(0,1)$.

J.2.3. Prediction Setting Binary C

$$
\begin{aligned}
C \sim& Bin(0.7); \\
logit(P(A=1|C)) \sim& -0.5-0.5C; \\
logit(P(M=1|A,C)) \sim& -0.5-C+4A-3AC; \\
logit(P(L=1|M,A,C)) \sim& -0.5-0.4C-2A+3M-0.6AC+2AM+0.5ACM; \\
Y =& 1+C+5A-4M+3L-2AC+AM+AL-1.5AML+\epsilon_Y,
\end{aligned}
$$

where $\epsilon_Y \sim \mathcal{N}(0,1)$.

J.2.4. Classification Setting Binary C

$$
\begin{aligned}
C \sim& Bin(0.7) \\
logit(P(A=1|C)) \sim& -0.5-0.5C; \\
logit(P(M=1|A,C)) \sim& 0.5+C-4A+3AC; \\
logit(P(L=1|M,A,C)) \sim& -0.5-0.4C-2A+3M-0.6AC+2AM-0.5ACM; \\
logit(P(Y=1|L,M,A,C)) \sim& -1-C-5A+4M-3L+2AC-AM-AL+1.5AML.
\end{aligned}
$$

J.2.5. Classification Setting Continuous C

$$
\begin{aligned}
C \sim& \mathcal{N}(0,1); \\
logit(P(A=1|C)) \sim& -0.5-0.5C; \\
logit(P(M=1|A,C)) \sim& 0.5+C-4A+3AC; \\
logit(P(L=1|M,A,C)) \sim& -0.5-0.4C-2A+3M-0.6AC+2AM-0.5ACM; \\
logit(P(Y=1|L,M,A,C)) \sim& -1-C-5A+4M-3L+2AC-AM-AL+1.5AML.
\end{aligned}
$$

J.2.6. Prediction Setting with Unobserved Variables

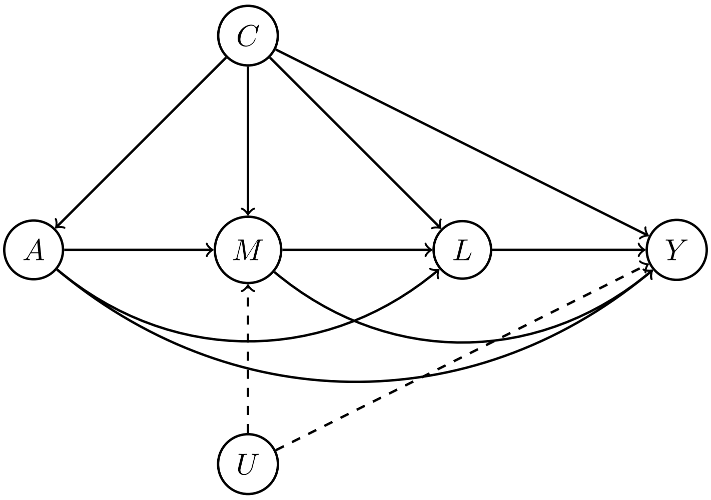


Figure J.4: Data Structure 2 Causal Graph with additional unobserved variable

$$\begin{aligned}
C \sim & Bin(0.7); \\
U \sim & \mathcal{N}(-1,1); \\
logit(P(A=1|C)) \sim & -0.5-0.5C; \\
logit(P(M=1|A,C)) \sim & -0.5-C+4A-3AC+U; \\
logit(P(L=1|M,A,C)) \sim & -0.5-0.4C-2A+3M-0.6AC+2AM+0.5ACM; \\
Y = & 1+C+5A-4M+3L-2AC+AM+AL-1.5AML+2U+\epsilon_Y,
\end{aligned}$$

where $\epsilon_Y \sim \mathcal{N}(0,1)$.

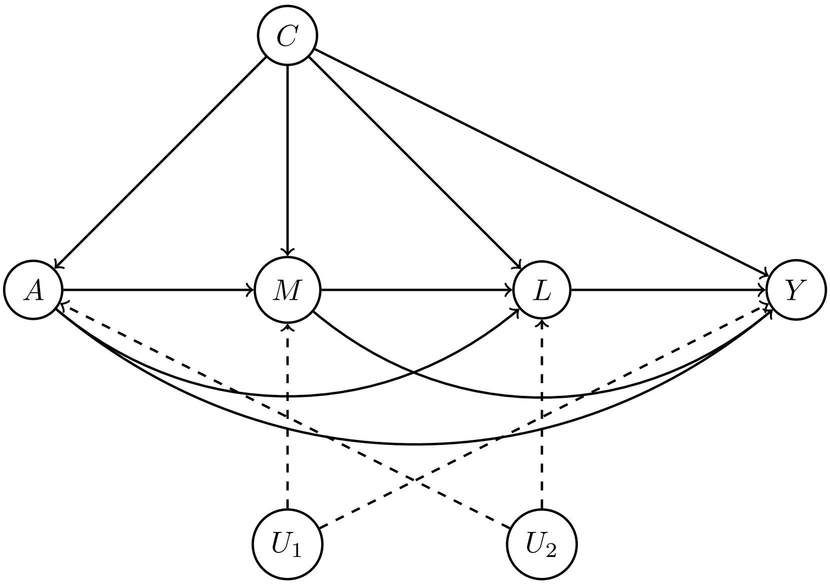


Figure J.5: Data Structure 2 Causal Graph with additional unobserved variables

$$\begin{aligned}
C \sim & Bin(0.7); \\
U_1 \sim & \mathcal{N}(-1,1); \\
U_2 \sim & \mathcal{N}(-1,1); \\
logit(P(A=1|C)) \sim & -0.5-0.5C+U_2; \\
logit(P(M=1|A,C)) \sim & -0.5-C+4A-3AC+U_1; \\
logit(P(L=1|M,A,C)) \sim & -0.5-0.4C-2A+3M-0.6AC+2AM+0.5ACM+U_2; \\
Y = & 1+C+5A-4M+3L-2AC+AM+AL-1.5AML+2U_1+\epsilon_Y,
\end{aligned}$$

where $\epsilon_Y \sim \mathcal{N}(0,1)$.

Bibliography

[1] Detecting a new generation of biases in ai-related recruitment platforms. *PennLaw*, April 2021.

[2] Alfred V. Aho, Michael R Garey, and Jeffrey D. Ullman. The transitive reduction of a directed graph. *SIAM Journal on Computing*, 1(2):131–137, 1972.

[3] Ryan M Andrews and Vanessa Didelez. Insights into the" cross-world" independence assumption of causal mediation analysis. *arXiv preprint arXiv:2003.10341*, 2020.

[4] Joshua Angrist and Jinyong Hahn. When to control for covariates? panel asymptotics for estimates of treatment effects. *Review of Economics and statistics*, 86(1):58–72, 2004.

[5] Joshua Angrist, David Autor, and Amanda Pallais. Marginal effects of merit aid for low-income students. Technical report, National Bureau of Economic Research, 2020.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23(2016):139–159, 2016.

[7] Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

[8] Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. 2005.

[9] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[10] Jeremy Bentham. The collected works of jeremy bentham: Deontology. together with a table of the springs of action and the article on utilitarianism. 1983.

[11] Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

[12] Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.

[13] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[14] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[15] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[16] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[17] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[18] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[19] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[20] National Research Council et al. *Measuring racial discrimination.* National Academies Press, 2004.

[21] A Philip Dawid. Causal inference without counterfactuals. *Journal of the American statistical Association*, 95(450):407–424, 2000.

[22] A Philip Dawid. Beware of the dag! In *Causality: objectives and assessment*, pages 59–86. PMLR, 2010.

[23] Jared N Day. Credit, capital and community: informal banking in immigrant communities in the united states, 1880-1924. *Financial History Review*, 9(1):65, 2002.

[24] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. *arXiv preprint arXiv:1807.00905*, 2018.

[25] Simon DeDeo. Wrong side of the tracks: Big data and protected categories. *arXiv preprint arXiv:1412.4643*, 2014.

[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[27] Markus C Elze, John Gregson, Usman Baber, Elizabeth Williamson, Samantha Sartori, Roxana Mehran, Melissa Nichols, Gregg W Stone, and Stuart J Pocock. Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology*, 69(3):345–357, 2017.

[28] Pan Mohamad Faiz. Teori keadilan john rawls (john rawls' theory of justice). *Jurnal Konstitusi*, 6(1):135–149, 2009.

[29] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, (2020-1), 2020.

[30] Joshua Glasgow, Sally Haslanger, Chike Jeffers, and Quayshawn Spencer. *What is Race?: Four Philosophical Views.* Oxford University Press, 2019.

[31] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[32] Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*, 2018.

[33] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[34] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016.

[35] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

[36] Sally Haslanger. Gender and race:(what) are they?(what) do we want them to be? *Noûs*, 34(1): 31–55, 2000.

[37] James J Heckman. Randomization and social policy evaluation. *Evaluating welfare and training programs*, 1:201–30, 1992.

[38] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21:689–696, 2008.

[39] Daniel Kahneman, Jack L Knetsch, and Richard Thaler. Fairness as a constraint on profit seeking: Entitlements in the market. *The American economic review*, pages 728–741, 1986.

[40] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

[41] Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236, 2021.

[42] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

[43] Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in Artificial Intelligence*, pages 616–626. PMLR, 2020.

[44] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[45] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

[46] Issa Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163, 2018.

[47] Tobias Kurth, Alexander M Walker, Robert J Glynn, K Arnold Chan, J Michael Gaziano, Klaus Berger, and James M Robins. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American journal of epidemiology*, 163(3):262–270, 2006.

[48] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[49] Steffen L Lauritzen. Causal inference from graphical models. *Complex stochastic systems*, pages 63–107, 2001.

[50] Moshe Lichman et al. Uci machine learning repository, 2013.

[51] Ilya Lipkovich, Alex Dmitrienko, and Ralph B D'Agostino Sr. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in medicine*, 36(1):136–196, 2017.

[52] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[53] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[54] Daniel Malinsky, Ilya Shpitser, and Thomas Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088. PMLR, 2019.

[55] Stijn Meganck, Sam Maes, Philippe Leray, Bernard Manderick, INSA Rouen, and France St Etienne du Rouvray. Learning semi-markovian causal models using experiments. In *Probabilistic Graphical Models*, pages 195–206. Citeseer, 2006.

[56] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[57] Olli S Miettinen. Proportion of disease caused or prevented by a given exposure, trait or intervention. *American journal of epidemiology*, 99(5):325–332, 1974.

[58] Gouri Shankar Mishra, Regina R Clewlow, Patricia L Mokhtarian, and Keith F Widaman. The effect of carsharing on vehicle holdings and travel behavior: A propensity score and causal mediation analysis of the san francisco bay area. *Research in Transportation Economics*, 52:46–55, 2015.

[59] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

[60] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[61] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[62] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Optimal training of fair predictive models. *arXiv preprint arXiv:1910.04109*, 2019.

[63] Osonde A Osoba and William Welser IV. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation, 2017.

[64] J Pearl and TS Verma. A theory of inferred causation. 1991. In *2-nd Conference on the Principles of Knowledge Representation and Reasoning, Cambridge, MA*, 1991.

[65] Judea Pearl. [bayesian analysis in expert systems]: Comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

[66] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge university press, 2009.

[67] Judea Pearl. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention science*, 13(4):426–436, 2012.

[68] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[69] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[70] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

[71] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[72] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

[73] Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

[74] Thomas Richardson and J Robins. Single world intervention graphs (swigs). Technical report, Technical report, University of Washington, 2013.

[75] Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.

[76] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

[77] James Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of chronic diseases*, 40:139S–161S, 1987.

[78] James M Robins and Thomas S Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, pages 103–158, 2010.

[79] Mojdeh Saadati and Jin Tian. Adjustment criteria for recovering causal effects from missing data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 561–577. Springer, 2019.

[80] Michael J Sandel. *Justice: What's the right thing to do?* Macmillan, 2010.

[81] Kailash Karthik Saravanakumar. The impossibility theorem of machine fairness–a causal perspective. *arXiv preprint arXiv:2007.06024*, 2020.

[82] Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035, 2013.

[83] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[84] Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. *arXiv preprint arXiv:1206.6876*, 2012.

[85] Ilya Shpitser and Eli Sherman. Identification of personalized effects associated with causal pathways. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2018. NIH Public Access, 2018.

[86] Ilya Shpitser and Eric Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of statistics*, 44(6):2433, 2016.

[87] Ilya Shpitser, Thomas S Richardson, and James M Robins. An efficient algorithm for computing interventional distributions in latent variable causal models. *arXiv preprint arXiv:1202.3763*, 2012.

[88] Herbert A Simon. Causal ordering and identifiability. In *Models of Discovery*, pages 53–80. Springer, 1977.

[89] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[90] Chathura Siriwardhana, KB Kulasekera, and Somnath Datta. Personalized treatment selection using data from crossover designs with carry-over effects. *Statistics in medicine*, 38(28):5391–5412, 2019.

[91] Andrew J Spieker, Joseph AC Delaney, and Robyn L McClelland. A method to account for covariate-specific treatment effects when estimating biomarker associations in the presence of endogenous medication use. *Statistical methods in medical research*, 27(8):2279–2293, 2018.

[92] Peter Spirtes. Building causal graphs from statistical data in the presence of latent variables. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 813–829. Elsevier, 1995.

[93] Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, prediction, and search, volume 81 of. *Lecture notes in statistics*, 1993.

[94] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

[95] Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816, 2012.

[96] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

[97] Jin Tian and Ilya Shpitser. On the identification of causal effects. 2003.

[98] Jin Tian and Ilya Shpitser. On identifying causal effects. *Heuristics, Probability and Causality: A Tribute to Judea Pearl (R. Dechter, H. Geffner and J. Halpern, eds.). College Publications, UK*, pages 415–444, 2010.

[99] Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. *arXiv preprint arXiv:1805.10318*, 2018.

[100] Tyler J VanderWeele and Miguel A Hernán. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *American journal of epidemiology*, 175(12):1303–1310, 2012.

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

[102] Thomas Verma, Judea Pearl, et al. Equivalence and synthesis of causal models. 1991.

[103] Robin Anno Wester, Julian Rubel, and Axel Mayer. Covariate selection for estimating individual treatment effects in psychotherapy research: A simulation study and empirical example. 2021.

[104] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. 2019.

[105] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. *arXiv preprint arXiv:1910.12586*, 2019.

[106] H Peyton Young. *Equity.* Princeton University Press, 2020.

[107] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

[108] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[109] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*, 2016.

[110] Haojun Zhu. Predicting earning potential using the adult dataset. $https://rpubs.com/H_Zhu/235617$, 2016.