TECHNICAL REPORT
R-90
First version: Jul 2022
Last version: Dec 2023

Foundations and Trends® in Machine Learning

Causal Fairness Analysis

A Causal Toolkit for Fair Machine Learning

Suggested Citation: Drago Plečko and Elias Bareinboim (2024), "Causal Fairness Analysis", Foundations and Trends® in Machine Learning: Vol. 17, No. 3, pp 1–238. DOI: 10.1561/2200000106.

Drago Plečko
Seminar für Statistik, ETH Zürich
drago.plecko@stat.math.ethz.ch

Elias Bareinboim
Department of Computer Science, Columbia University
eb@cs.columbia.edu

now
the essence of knowledge
Boston — Delft

Contents

Causal Fairness Analysis

Drago Plečko[1] and Elias Bareinboim[2]

[1] *Seminar für Statistik, ETH Zürich; drago.plecko@stat.math.ethz.ch*
[2] *Department of Computer Science, Columbia University; eb@cs.columbia.edu*

ABSTRACT

Decision-making systems based on AI and machine learning have been used throughout a wide range of real-world scenarios, including healthcare, law enforcement, education, and finance. It is no longer far-fetched to envision a future where autonomous systems will drive entire business decisions and, more broadly, support large-scale decision-making infrastructure to solve society's most challenging problems. Issues of unfairness and discrimination are pervasive when decisions are being made by humans, and remain (or are potentially amplified) when decisions are made using machines with little transparency, accountability, and fairness. In this paper, we introduce a framework for *causal fairness analysis* with the intent of filling in this gap, i.e., understanding, modeling, and possibly solving issues of fairness in decision-making settings. The main insight of our approach will be to link the quantification of the disparities present in the observed data with the underlying, often unobserved, collection of causal mechanisms that generate the disparity in the first place, a challenge we call the Fundamental Problem of Causal Fairness Analysis (FPCFA). In order to solve the FPCFA, we study the problem of decomposing variations and empirical measures of fairness that attribute such variations to structural mechanisms and different units of the population. Our effort culminates in the Fairness Map, the first systematic attempt to organize and explain the relationship between various criteria found in the literature. Finally, we study which causal assumptions are minimally needed for performing causal fairness analysis and propose the Fairness Cookbook, which allows one to assess the existence of disparate impact and disparate treatment.

1

Introduction

As society transitions to an AI-based economy, an increasing number of decisions that were once made by humans are now delegated to automated systems, and this trend will likely accelerate in the coming years. Automated systems may exhibit discrimination based on gender, race, religion, or other sensitive attributes, so considerations about fairness in AI are an emergent discussion across the globe. The European Union, for instance, recently passed sweeping regulations putting substantial constraints over automated decision-making and AI systems (Commission, 2021). While we believe it is evident that a novel legal framework is needed to organize and regulate this new, emerging economy, it is less clear, however, that the proper scientific understanding and tools for designing such regulations are currently available. Even though one may surmise that issues of unfairness in AI are a recent development, the problem's origins can be traced to long before the advent of AI and the prominence these systems have reached in the last years. This is perhaps best witnessed by the civil rights movements of the twentieth century. Interestingly, Martin Luther King Jr. spoke of having a dream that his children "will one day live in a nation where they will not be judged by the color of their skin, but by the content of their character." So little could he have anticipated that machine algorithms would one day use race for making decisions, and that the issues of unfairness in AI would be legislated under Title VII of the Civil Rights Act of 1964 (Act, 1964), which he advocated and fought for (Oppenheimer, 1994; Kotz, 2005).

The critical challenge underlying fairness in AI systems lies in the fact that biases in decision-making exist in the real world from which various datasets are collected. Perhaps not surprisingly, a dataset collected from a biased reality will contain aspects of this bias as an imprint. In this context, algorithms are tools that may replicate or potentially even amplify the biases that exist in reality in the first place. As automated systems are a priori oblivious to ethical considerations, deploying and using them blindly could lead to the perpetuation of unfairness in the future.

More pessimistic analysts take this observation as a prelude to doomsday, which, in their opinion, suggests that we should be extremely wary and defensive against any AI. We believe a degree of caution is necessary, of course, but take a more positive perspective and consider this transition to a more AI-based society as a unique opportunity to improve the current state of affairs. While human decision-makers are hard to change, even when aware of their own biases, AI systems may be less brittle and more flexible. Still, one of the requirements to realize the AI's potential is a new mathematical framework that allows the description and assessment of legal notions of discrimination in a formal way. This situation is somewhat unique in the context of AI because a new definition of "ground truth" is required. The decision-making system cannot rely purely on learning from the data, which is contaminated with unwanted bias. It is currently unclear how to formulate the ideal inferential target[1] that could help bring about a fair world when deployed. This degree of flexibility in deciding the new ground truth also emphasizes the importance of normative work in this context.[2]

In this paper, we build on two legal doctrines applied to large bodies of cases throughout the US and the EU known as *disparate treatment* and *disparate impact* (Barocas and Selbst, 2016). One of our goals will be to develop a framework for causal fairness analysis grounded in these doctrines and translate them into exact mathematical language amenable to AI optimization. The disparate treatment doctrine enforces the equality of treatment of different groups, prohibiting the use of the protected attribute (e.g., race) in the decision process. One of the legal formulations for proving disparate treatment is that "a similarly situated person who is not a member of the protected class would

[1] We believe this explains the vast number of fairness criteria described in the literature, which we will detail later on in the paper.

[2] One way of seeing this point a bit more formally goes as follows. We first consider the current version of the world, say π, and note that it generates a probability distribution $\mathcal{P}$. Training the machine learning algorithm with data from this distribution ($\mathcal{D} \sim \mathcal{P}$) is replicating patterns from this reality, π. We would want an alternative, counterfactual reality π', which induces a different distribution $\mathcal{P}'$ without the past biases. The challenge here is that thinking about and defining $\mathcal{P}'$ relies on going beyond $\mathcal{P}$, or the corresponding dataset, which is non-trivial, and yet one of our main goals.

not have suffered the same fate" (Barocas and Selbst, 2016)[3]. On the other hand, the disparate impact doctrine focuses on *outcome fairness*[4], namely, the equality of outcomes among protected groups. Disparate impact discrimination occurs if a facially neutral practice has an adverse impact on members of the protected group. Under this doctrine most commonly fall cases where discrimination is unintended or implicit. The analysis can become somewhat intricate when variables are correlated with the protected attribute and may act as a proxy. The law may not necessarily prohibit their usage due to their relevance to the business itself, legally known as "business necessity" or "job-relatedness". Taking business necessity into account is the essence of disparate impact (Barocas and Selbst, 2016).

In fact, as we demonstrate intuitively and formally later in the text, the disparate treatment and disparate impact doctrines can be seen as spanning a spectrum of fairness notions (see Fig. 1.1). On the one end, the disparate treatment doctrine ensures that there is *no direct effect* of the protected attribute on the outcome, which can be seen as the minimal fairness requirement. On the other end, the disparate impact doctrine (in the extreme case), ensures that the protected attribute has *no effect* on the outcome. In practice, however, business necessity considerations determine where on this spectrum the appropriate fairness notion is, given the requirements and specific details of the application in question.

The connection of fairness with causal inference might be seen as natural for two reasons. Firstly, business necessity considerations are inherently causal, as they require attributing the observed disparity to the underlying causal mechanism. Our framework will therefore allow the data scientist to quantify the disparity explained by mechanisms that do not fall under business necessity and are considered discriminatory, thereby accommodating application-specific requirements. Secondly, the legal frameworks of

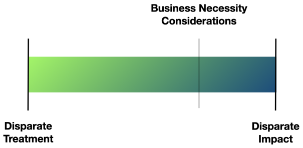


Figure 1.1: The spectrum of fairness notions spanned by the disparate treatment and disparate impact doctrines.

[3] This formulation is related to a condition known as *ceteris paribus*, which represents the effect of the protected attribute on the outcome of interest while keeping everything else constant. From a causal perspective, this suggests that the disparate treatment doctrine is concerned with direct discrimination, a connection we draw formally later on in the manuscript.

[4] Interestingly, both of the above-discussed doctrines are usually considered under the rubric of outcome fairness, that is, focusing on the disparity in the outcome itself. An important complementary notion to outcome fairness is *process fairness*, which is instead focused on how the decision process is carried out, and not specifically on the outcomes themselves (Grgic-Hlaca *et al.*, 2016). In this context, the causal approach offers a key strength, discussed in detail in Appendix E.

anti-discrimination laws (for example, Title VII in the US) often require that to establish a *prima facie* case of discrimination the plaintiff must demonstrate "a strong causal connection" between the alleged discriminatory practice and the observed statistical disparity (e.g., Texas Dept. of Housing and Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. 519 (2015)). Therefore, as discussed in subsequent sections, another requirement of our framework will be the ability to distinguish between notions of discrimination that would otherwise be statistically indistinguishable.

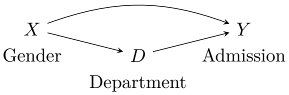


Figure 1.2: A partial causal model for the Berkeley Admission example.

Consider the Berkeley Admission example, in which admission results of students applying to UC Berkeley were collected and analyzed (Bickel *et al.*, 1975). The analysis showed that male students are 14% more likely to be admitted than their female counterparts, which raised concerns about the possibility of gender discrimination. The discussion of this example is often less focused on the accuracy and appropriateness of the used statistical measures and more on the plausible justification of disparity based on the mechanism underlying this disparity. A visual representation of the dynamics in this setting is shown in Fig. 1.2. In words, each student chooses a department of application (D). The department's choice and the student's gender (X) might, in turn, influence the admission decision (Y). In this example, there is a clear need to determine how much of the observed statistical disparity can be attributed to the direct causal path from gender to admission decision vs. the indirect mechanism[5] going through the department choice variable. Looking directly at gender for determining university admission would indeed be disallowed, whereas using department choice, which may be influenced by gender, might be deemed acceptable. [6] The need to explain an observed statistical disparity, say in this case the 14% difference in admission rates, through the underlying causal mechanisms – direct and indirect – is a recurring theme when assessing discrimination, even though it is sometimes considered only implicitly.

When AI tools are deployed in the real world, a similar pattern of questions emerges. Examples include (but are not limited to) the debate over the origins and interpretation of discrimination in the criminal justice system (COMPAS, Angwin *et al.*, 2016), the contribution of data vs. algorithms in the observed bias in face detection (e.g., Harwell, 2019; Buolamwini and Gebru, 2018),

[5] As discussed later on, even among indirect paths, one may need to distinguish between mediated causal paths and confounded non-causal paths, or, more generally, among a specific subset of these paths.

[6] Society may be "guilty" of creating the wrong incentives, and perhaps fewer female applicants are considering certain departments, but the university itself may not be deemed discriminatory.

and the business necessity vs. risk of digital redlining in targeted advertising (Detrixhe and Merrill, 2019). Intuitively, through these questions, society wants to draw a line between what is seen as discriminatory on the one hand and what is seen as acceptable or justified by economic principles on the other. Put differently, such discussions aim to determine where on the fairness spectrum in Fig. 1.1 the appropriate notion of fairness lies.

A practitioner interested in implementing a fair AI system will need to detect and quantify undesired discrimination based on society's current ethical standards, and then design learning methods capable of removing such unfairness from future predictions and decisions. In doing so, the practitioner will face two challenges. The first stems from the fact that the current literature is abundant with different fairness measures, some of which are mutually incompatible (Corbett-Davies and Goel, 2018), and choosing among these measures, even for the system designer, is usually a non-trivial task. This challenge is compounded with the second challenge, which arises from the statistical nature of such fairness measures. As we will show both formally and empirically later in the text, statistical measures alone cannot distinguish between different causal mechanisms that transmit change and generate disparity in the real world, even if an unlimited amount of data is available. Despite this apparent shortcoming of purely statistical measures, much of the literature focuses on casting fair prediction as an optimization problem subject to fairness constraints based on such measures (Pedreschi *et al.*, 2008; Pedreschi *et al.*, 2009; Luong *et al.*, 2011; Ruggieri *et al.*, 2011; Hajian and Domingo-Ferrer, 2012; Kamiran and Calders, 2009; Calders and Verwer, 2010; Kamiran *et al.*, 2010; Zliobaite *et al.*, 2011; Kamiran and Calders, 2012; Kamiran *et al.*, 2012; Zemel *et al.*, 2013; Mancuhan and Clifton, 2014; Romei and Ruggieri, 2014; Dwork *et al.*, 2012; Friedler *et al.*, 2016; Chouldechova, 2017; Pleiss *et al.*, 2017), to cite a few. In fact, these methods may be insufficient for removing bias and perhaps even lead to unintended consequences and bias amplification, as it will become clear later on.

As outlined briefly in previous paragraphs, the behavior of AI/ML-based decision-making systems is an emergent property following a complex combination of past (possibly biased) data and interactions with the environment. Predicting or explaining this behavior and its impact on the real world can be difficult, even for the system designer who knows how the system is built. Ensuring fairness of such decision-making systems, therefore, critically relies on contributions from two groups, namely:

a. the AI and ML engineers who develop methods to detect bias and ensure adherence of ML systems to fairness measures, and

b. the domain experts, policymakers, economists, social scientists, and legal experts who study the origins of these biases and can provide the societal interpretations of fairness measures and their expectations in terms of norms and standards.

Currently, these groups do not share a common starting point. It is challenging for them to understand each other and work together towards developing a fair specification of such complex systems, aligned with the many stakeholders involved in the process.

In this paper, we argue that the language of structural causality can provide a unique perspective on the issues of fairness and facilitate the discussion and exchange of ideas, goals, and expectations between these groups. Issues of unfairness are fundamentally linked to considerations of responsibility and blame, and thus a causal analysis of the problem is mandated from legal, logical and philosophical standpoints (Moore, 2019; Halpern, 2016)[7]. A causal analysis, as will be discussed in detail, is contingent on obtaining rich enough causal models of unobserved or partially observed reality, which may be non-trivial in practice, yet it is crucial in the context of fair ML as it allows one to relate observed disparities to existing causal mechanisms. Causal models must be built using inputs from domain experts, social scientists, and policy-makers, and a formal language is needed to express and scrutinize their assumptions. In this work, we lay down the foundations for interpreting legal doctrines of discrimination through causal reasoning, which we view as an essential step towards the development of a new generation of more ethical and transparent AI systems.

Paper's Roadmap & Contributions

We develop a general and coherent framework of Causal Fairness Analysis to overcome the challenges described above. This framework provides a common language to connect computer scientists, statisticians, and data scientists on the one hand and legal, social, and ethical experts on the other, to tackle challenges of fairness in automated decision-making. Further, this new framework grounds the legal doctrines of disparate impact and disparate treatment through the semantics of structural causal models. The critical elements of our proposal are shown in Fig. 1.3, which also serves as a roadmap of how this paper is organized and how causal fairness analysis should be conducted. Specifically, in Sec. 2, we cover the basic notions of causal inference needed to build our framework, including structural causal models, causal diagrams, and data collection. In Sec. 3, we introduce the essential elements of our theoretical framework. In particular, we define the notions of structural fairness that will serve as a baseline, ground truth for determining the presence or absence of discrimination under disparate impact and disparate treatment doctrines. In

[7] We remark that the causal perspective on fairness is not the only viewpoint, and a number of important works have been developed entirely outside this rubric.

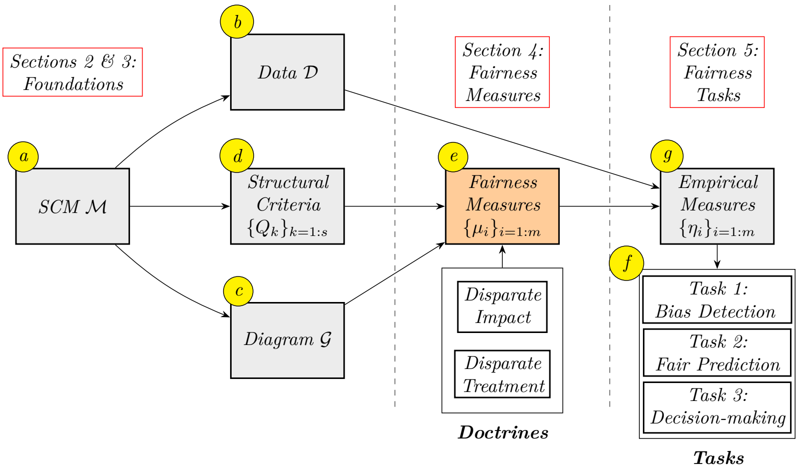

Sections 2 & 3: Foundations
Data $\mathcal{D}$
Section 4: Fairness Measures
Section 5: Fairness Tasks
a
b
c
d
e
f
g
SCM $\mathcal{M}$
Structural Criteria $\{Q_k\}_{k=1:s}$
Fairness Measures $\{\mu_i\}_{i=1:m}$
Empirical Measures $\{\eta_i\}_{i=1:m}$
Diagram $\mathcal{G}$
Disparate Impact
Disparate Treatment
Doctrines
Task 1: Bias Detection
Task 2: Fair Prediction
Task 3: Decision-making
Tasks

Figure 1.3: A mental map of the Causal Fairness Analysis framework.

Sec. 4, we introduce causal measures of fairness that can be computed from data in practice. We further draw the connection between such measures and the aforementioned legal doctrines. In Sec. 5, we introduce the tasks of Causal Fairness Analysis – (1) bias detection and quantification, (2) fair prediction, and (3) fair decision-making – and show how they can be solved by building on the tools developed earlier. In Sec. 6 we develop tools for decomposing indirect and spurious variations on a variable-specific level, which leads to a general approach for evaluating fairness under arbitrary business necessity sets. More specifically, our contributions are as follows:

1. We study the problem of decomposing variations between the protected attribute X and the outcome variable Y, using the technique of factual and counterfactual contrasts (Def. 3.7). We prove the structural basis expansion formula for such contrasts, which highlights the fundamental difference between causal and non-causal variations (Thm. 3.1). Furthermore, this result allows us to show how the total variation (TV)[8] can be decomposed based on different causal mechanisms and across different groups of units. These developments lead to the construction of the *explainability plane* (Fig. 3.3).

2. We introduce the Fundamental Problem of Causal Fairness Analysis

[8] What we refer to in this manuscript as the total variation (TV) measure is also known in the literature as the *parity gap*, or simply the difference in conditional expectations, $\mathbb{E}[Y \mid x_1] - \mathbb{E}[Y \mid x_0]$, where x_0, x_1 are the two values of the protected attribute X, and Y is the outcome of interest.

(FPCFA, Def. 3.6), which formalizes the key properties that empirical measures of fairness should exhibit, including admissibility and decomposability. Subsequently, we develop increasingly refined solutions to the FPCFA, proved in Thms. 4.2, 4.3, 4.4, and 4.5.

3. We design the first version of the *Fairness Map* (Thm. 4.8 and Fig. 4.5), putting many well-known fairness measures under the same theoretical umbrella and uncovering the structure that connects them. In particular, the Map connects all the measures in the so-called TV family (Tab. 4.1). We provide a detailed analysis of the causal properties of well-known measures found in the literature, including counterfactual fairness, individual fairness, and predictive parity (Sec. 4.4).

4. We propose a simplified type of (clustered) graphical model called the *Standard Fairness Model* (SFM, Def. 2.7), which requires fewer modeling assumptions than typically used causal diagrams. We show that the SFM strikes a balance between simplicity of construction and informativeness for causal analysis (Thm. 4.11), allowing us to perform causal inference even when detailed knowledge about the underlying decision-making process is scarce.

5. We develop the first non-parametric decomposition of the predictive parity measure in terms of the underlying causal mechanisms. Building on this, we define causal predictive parity (Def. 4.14), and show how this new notion is complementary to statistical parity, thereby addressing a well-known impossibility result from the literature (Thms. 4.12, 4.13).

6. By putting all the above results together, we develop a practical procedure called the Fairness Cookbook (Alg. 5.1) that allows data scientists to assess the presence of disparate treatment and disparate impact and quantify their degree. Furthermore, we provide an R-package called `faircause` for performing this task.

7. We study the implications of Causal Fairness Analysis on the fair prediction problem. In particular, we prove the Fair Prediction Theorem (Thm. 5.1) that shows that making TV equal to zero during the training stage is almost never sufficient to ensure that causal measures of fairness are well-behaved. We further propose solutions that can provide causal guarantees for the constructed predictors (Thms. 5.2, 5.3).

8. Based on the implications of the Fair Prediction Theorem to decision-making (Cor. 5.5), we develop new procedures for achieving fairness in particular single-step decision-making (Algs. 5.3 and 5.5).

9. We prove the first non-parametric decomposition for spurious effects in Semi-Markovian models (Thms. 6.1, 6.3). We further show results that establish what is the most fine-grained path-specific analysis that can be performed in practice (Thm. 6.9, Cor. 6.10), and develop an algorithm for testing arbitrary business necessity requirements (Alg. 6.4).

Readers familiar with causal inference may want to move straight to Sec. 3, even though the examples provided in the next section are used to motivate the problem of fairness discussed throughout the manuscript.

2

Foundations of Causal Inference

In this section, we introduce three fundamental building blocks that will allow us to formalize the challenges of fairness described above through a causal lens. First, we will define in Sec. 2.1 a general class of data-generating models known as *structural causal models* (shown in Fig. 1.3a). The key observation here is that the collection of mechanisms underpinning any decision-making scenario are causal, and therefore should be modeled through proper and formal causal semantics. Second, we will discuss in Sec. 2.2 qualitatively different probability distributions that are induced by the causal generative process, and which will lead to the observed data and counterfactuals (Fig. 1.3b). Third, we will introduce in Sec. 2.3 an object known as a *causal diagram* (Fig. 1.3c), which will allow the data scientist to articulate non-parametric assumptions over the space of generative models. These assumptions can be shown as necessary for the analysis, in a broader sense. Finally, we will define the *standard fairness model* (SFM), which is a special class of diagrams that act as a template, allowing one to generically express entire classes of structural models. The SFM class, in particular, requires fewer modeling assumptions than the more commonly used causal diagrams.

2.1 Structural Causal Models

The basic semantical framework of our analysis rests on the notion of structural causal models (SCM, for short), which is one of the most flexible class of generative models known to date and that allows modeling various tasks (Pearl, 2000; Bareinboim and Pearl, 2016). The section will follow the presentation in (Bareinboim *et al.*, 2022), which contains more detailed discussions and proofs. First, we introduce and exemplify SCMs through the following definition:

Definition 2.1 (Structural Causal Model (SCM) (Pearl, 2000))**.** A structural causal model (SCM) is a 4-tuple $\langle V, U, \mathcal{F}, P(u)\rangle$, where

1. U is a set of exogenous variables, also called background variables, that are determined by factors outside the model;

2. $V = \{V_1, ..., V_n\}$ is a set of endogenous (observed) variables, that are determined by variables in the model (i.e. by the variables in $U \cup V$);

3. $\mathcal{F} = \{f_1, ..., f_n\}$ is the set of structural functions determining V, $v_i \leftarrow f_i(\mathrm{pa}(v_i), u_i)$, where $\mathrm{pa}(V_i) \subseteq V \setminus V_i$ and $U_i \subseteq U$ are the functional arguments of f_i;

4. $P(u)$ is a distribution over the exogenous variables U.

In words, each structural causal model can be seen as partitioning the variables involved in the phenomenon into sets of exogenous (unobserved) and endogenous (observed) variables, respectively, U and V. The exogenous variables are determined "outside" of the model and their associated probability distribution, $P(u)$, represents a summary of the world external to the phenomenon that is under investigation. In our setting, these variables will represent the units involved in the phenomenon, which correspond to elements of the population under study, for instance, patients, students, customers. Naturally, their randomness (encoded in $P(u)$) induces variations in the endogenous set V.

Inside the model, the value of each endogenous variable V_i is determined by a causal process, $V_i \leftarrow f_i(\mathrm{pa}(v_i), u_i)$, that maps the exogenous factors U_i and a set of endogenous variables pa_i (so-called parents) to V_i. These causal processes – or mechanisms – are assumed to be invariant unless explicitly intervened on (as defined later in the section). Together with the background factors, they represent the data-generating process according to which the values of the endogenous variables are determined. For concreteness and grounding of the definition, we revisit the Berkeley admission example through the lens of SCMs.

Example 2.1 (Berkeley Admission (Bickel *et al.*, 1975))**.** During the application process for admissions to UC Berkeley, potential students choose a department to which they apply, labeled as D (binary with $D = 0$ for arts & humanities, $D = 1$ for sciences). The admission decision is labeled as Y (y_1 accepted, y_0

rejected) and the student's gender is labeled as X (x_0 female, x_1 male)[1].

The SCM $\mathcal{M}$ is the 4-tuple $\langle V = \{X, D, Y\}, U = \{U_X, U_D, U_Y\}, \mathcal{F}, P(U)\rangle$, where U_X, U_Y, U_D represent the exogenous variables, outside of the model, that affect X, Y, D, respectively. Also, the causal mechanisms $\mathcal{F}$ are given as follows [2]:

$$X \leftarrow \mathbb{1}(U_X < 0.5) \tag{2.4}$$
$$D \leftarrow \mathbb{1}(U_D < 0.5 + \lambda X) \tag{2.5}$$
$$Y \leftarrow \mathbb{1}(U_Y < 0.1 + \alpha X + \beta D), \tag{2.6}$$

and $P(U_X, U_D, U_Y)$ is such that U_X, U_D, U_Y are independent Unif$[0, 1]$ random variables.

In words, the population is partitioned into males and females with equal probability (the exogenous U_X represents the population's biological randomness). Each applicant chooses a department D, and this decision depends on U_D and gender X. The exogenous variable U_D represents the individual's natural inclination towards studying science. Whenever $\lambda > 0$ in Eq. 2.5, the threshold for applying to a science department is higher for female individuals, which is a result of various societal pressures. Finally, the admission decision Y possibly depends on gender (if $\alpha \neq 0$ in Eq. 2.6) and/or department of choice (if $\beta \neq 0$ in Eq. 2.6). In this case, the exogenous variable U_Y represents the impression the applicant left during an admission interview. Notice that female students and arts & humanities students may need to leave a better interview impression in order to be admitted (depending on Eq. 2.6). □

Another important notion for our discussion is that of a submodel, which is defined next:

Definition 2.2 (Submodel (Pearl, 2000))**.** Let $\mathcal{M}$ be a structural causal model, X a set of variables in V, and x a particular value of X. A submodel $\mathcal{M}_x$ (of $\mathcal{M}$) is a 4-tuple:

$$\mathcal{M}_x = \langle V, U, \mathcal{F}_x, P(u)\rangle \tag{2.7}$$

[1] In the manuscript, gender is discussed as a binary variable, which is a simplification of reality, used to keep the presentation of the concepts simple. In general, one might be interested in analyses of gender discrimination with gender taking non-binary values.

[2] The given SCM can also be written as

$$X \leftarrow \text{Bernoulli}(0.5) \tag{2.1}$$
$$D \leftarrow \text{Bernoulli}(0.5 + \lambda X) \tag{2.2}$$
$$Y \leftarrow \text{Bernoulli}(0.1 + \alpha X + \beta D). \tag{2.3}$$

where

$$\mathcal{F}_x = \{f_i : V_i \notin X\} \cup \{X \leftarrow x\}, \tag{2.8}$$

and all other components are preserved from $\mathcal{M}$.

In words, the SCM $\mathcal{M}_x$ is obtained from $\mathcal{M}$ by replacing all equations in $\mathcal{F}$ related to variables X by equations that set X to a specific value x. In the context of Causal Fairness Analysis, we might be interested in submodels in which the protected attribute X is set to a fixed value x. Building on submodels, we introduce next the notion of a potential outcome:

Definition 2.3 (Potential Outcome / Response (Rubin, 1974; Pearl, 2000))**.** Let X and Y be two sets of variables in V and $u \in \mathcal{U}$ be a unit. The potential outcome/response $Y_x(u)$ is defined as the solution for Y of the set of equations $\mathcal{F}_x$ evaluated with $U = u$. That is, $Y_x(u)$ denotes the solution of Y in the submodel $\mathcal{M}_x$ of $\mathcal{M}$.

In words, $Y_x(u)$ is the value variable Y would take if (possibly contrary to observed facts) X is set to x, for a specific unit u. In the Admission example, $Y_x(u)$ would denote the admission outcome for the specific unit u, had their gender X been set to value x by intervention (e.g., possibly contrary to their actual gender).

Notation in the Potential Outcomes (PO) Literature. For readers familiar with the PO framework (Rubin, 1974; Rubin, 2005), we mention how our notation translates the standard PO notation. The potential outcome under intervention $X = x$ is usually denoted by $Y(x)$, corresponding to Y_x in our notation. When indicating a specific unit, in PO framework one may write $Y(x, u)$, corresponding to $Y_x(u)$ in our notation. That is, the argument of $Y(\cdot)$ always indicates the unit in this manuscript, and the subscript indicates the (possibly counterfactual) intervention.

2.2 Observational & Counterfactual Distributions

Each SCM $\mathcal{M}$ induces different types of probability distributions, which represent different data collection modes and will play a key role in fairness analysis. We start with the observational distribution that represents a state of the underlying decision-making system from which the fairness analyst just passively collects data, without interfering in any decision-making processes:

Definition 2.4 (Observational Distribution (Bareinboim *et al.*, 2022))**.** An SCM $\mathcal{M} = \langle V, U, \mathcal{F}, P(u)\rangle$ induces a joint probability distribution $P(V)$ such that for each $Y \subseteq V$,

$$P^{\mathcal{M}}(y) = \sum_u \mathbb{1}\Big(Y(u) = y\Big)P(u), \tag{2.9}$$

where $Y(u)$ is the solution for Y after evaluating $\mathcal{F}$ with $U = u$.

In words, the procedure can be described as follows:

1. for each unit $U = u$, the structural functions $\mathcal{F}$ are evaluated following a valid topological order, and
2. the probability mass $P(U = u)$ is accumulated for each instantiation $U = u$ consistent with the event $Y = y$.

Throughout this manuscript, all the sums should be replaced by the corresponding integrals whenever suitable, e.g., when underlying densities exist[3]. To ground the discussion about this definition, we continue with the example above and see how the corresponding observational distribution is induced.

Example 2.2 (College Admissions Observational Distribution)**.** Consider the SCM $\mathcal{M}$ in Eq. 2.4-2.6. The total variation (TV for short; also called parity gap) generated by $\mathcal{M}$ depends on the structural mechanisms $\mathcal{F}$ and the distribution of exogenous variables $P(U_X, U_D, U_Y)$. The total variation can be written as:

$$P(y \mid x_1) - P(y \mid x_0) = \frac{P(y, x_1)}{P(x_1)} - \frac{P(y, x_0)}{P(x_0)}. \tag{2.10}$$

Therefore, we compute the terms $P(y, x_1), P(x_1), P(y, x_0), P(x_0)$ based on the true, underlying SCM. Using Def. 2.4 and Eq. 2.4, we can see that:

$$P(x_1) = P(U_X < 0.5) = \frac{1}{2} = P(U_X > 0.5) = P(x_0). \tag{2.11}$$

Using the fact that U_X, U_D, and U_Y are independent in the SCM, $P(y, x_1)$ can be computed in the following way (Def. 2.4):

$$\begin{aligned}
P(y, x_1) &= \sum_u \mathbb{1}(Y(u) = 1, X(u) = 1)P(u) && (2.12)\\
&= P(U_X < 0.5)\big[P(U_D > 0.5 + \lambda)P(U_Y < 0.1 + \alpha) + && (2.13)\\
&\qquad\qquad P(U_D < 0.5 + \lambda)P(U_Y < 0.1 + \alpha + \beta)\big] \\
&= \frac{1}{2}[(\frac{1}{2} - \lambda)(0.1 + \alpha) + (\frac{1}{2} + \lambda)(0.1 + \alpha + \beta)] && (2.14)\\
&= \frac{1}{2}(0.1 + \alpha + (\frac{1}{2} + \lambda)\beta). && (2.15)
\end{aligned}$$

[3] In the continuous case, the existence of densities will be sufficient to write down many of the definitions and results found in the manuscript. However, in such a setting the estimation of target quantities from finite samples may be more complicated, and may require further regularity conditions such smoothness and positivity. These challenges are not discussed in the manuscript.

The computation above can be described as follows. Firstly, $X(u) = 1$ is equivalent with $U_X < 0.5$ (Eq. 2.4). Secondly, when $X(u) = 1$, there are two possibilities for the variable D based on U_D (see Eq. 2.5). Whenever $U_D > 0.5 + \lambda$, then $D(u) = 0$, and to have $Y(u) = 1$, we need $U_Y < 0.1 + \alpha$ (see Eq. 2.6). If $U_D < 0.5 + \lambda$, then $D(u) = 1$, and to have $Y(u) = 1$, we need $U_Y < 0.1 + \alpha + \beta$ (see Eq. 2.6). An analogous computation yields that:

$$\begin{aligned} P(y, x_0) &= \sum_u \mathbb{1}(Y(u) = 1, X(u) = 0)P(u) && (2.16)\\ &= \frac{1}{2}\Big[\frac{1}{2} \cdot 0.1 + \frac{1}{2} \cdot (0.1 + \beta)\Big] = \frac{1}{2}(0.1 + \frac{\beta}{2}). && (2.17) \end{aligned}$$

Putting the results together in Eq. 2.10, the TV equals

$$\begin{aligned} P(y \mid x_1) - P(y \mid x_0) &= \frac{\frac{1}{2}(0.1 + \alpha + (\frac{1}{2} + \lambda)\beta)}{\frac{1}{2}} - \frac{\frac{1}{2}(0.1 + \frac{\beta}{2})}{\frac{1}{2}} && (2.18)\\ &= \alpha + \lambda\beta. && (2.19) \end{aligned}$$

In fact, after analyzing the admission dataset from UC Berkeley, a data scientist computes the observed disparity to be[4]

$$P(y \mid x_1) - P(y \mid x_0) = 14\%. \tag{2.20}$$

In words, male candidates are 14% more likely to be admitted than female candidates. The data scientist (who does not have access to the SCM $\mathcal{M}$ described above) might wonder if this disparity (14%) means that female applicants are discriminated against. Also, she/he might wonder how the observed disparity relates to the SCM $\mathcal{M}$ given in Eq. 2.4-2.6. Our goal in this manuscript is to address these questions from first principles. □

Next, we define another important family of distributions, over possible counterfactual outcomes, which will be used throughout this manuscript:

Definition 2.5 (Counterfactual Distributions (Bareinboim *et al.*, 2022)). Let $\mathcal{M} = \langle V, U, \mathcal{F}, P(u)\rangle$ be an SCM, and let $Y_1, \ldots, Y_k \subset V$, and $X_1, \ldots, X_k \subset V$ be subsets of the observables, and let $x_1, \ldots, x_k$ be specific values of $X_1, \ldots, X_k$. Denote by $(Y_i)_{x_i}$ the potential response of variables Y_i when setting $X_i = x_i$. The SCM $\mathcal{M}$ induces a family of joint distributions over counterfactual events $(Y_1)_{x_1}, \ldots, (Y_k)_{x_k}$:

$$P^{\mathcal{M}}((y_1)_{x_1}, \ldots, (y_k)_{x_k}) = \sum_u \mathbb{1}\Big(\bigwedge_{i=1}^{k} (Y_i)_{x_i}(u) = y_i\Big)P(u). \tag{2.21}$$

[4] The number below was evaluated from the actual real dataset, which is compatible with structural coefficients $\alpha = 0, \beta = \frac{7}{10}$, and $\lambda = \frac{2}{10}$.

The l.h.s. in Eq. 2.21 contains variables with different subscripts, which syntactically represent different potential responses (Def. 2.3), or counterfactual worlds. In words, the equation can be interpreted as follows:

1. For each set of subscripts and variables ($X_1, \dots, X_k$ and $Y_1, \dots, Y_k$), replace the corresponding mechanism with appropriate constants to generate $\mathcal{F}_{x_1}, \dots, \mathcal{F}_{x_k}$ and create submodels $\mathcal{M}_{x_1}, \dots, \mathcal{M}_{x_k}$,

2. For each unit $U = u$, evaluate the modified mechanisms $\mathcal{F}_{x_1}, ..., \mathcal{F}_{x_k}$ to obtain the potential response of the observables,

3. The probability mass $P(U = u)$ is accumulated for each instance $U = u$ that is consistent with the events over the counterfactual variables, that is $(Y_1)_{x_1} = y_1, \dots, (Y_k)_{x_k} = y_k$, that is, $Y_1 = y_1$ in $\mathcal{M}_{x_1}$, …, $Y_k = y_k$ in $\mathcal{M}_{x_k}$.

Example 2.3 (College Admission Counterfactual Distribution)**.** Consider the SCM in Eq. 2.4-2.6 and the following joint counterfactual distribution:

$$P(y_{x_1}, y_{x_0}). \tag{2.22}$$

In the submodel $\mathcal{M}_{x_0}$ (where $X = 0$ is set by intervention), we have that $D_{x_0}(u) = 1$ is equivalent with $U_D < 0.5$. When $D_{x_0}(u) = 1$, $Y_{x_0}(u) = 1$ if and only if $U_Y < 0.1 + \beta$. Similarly, when $D_{x_0}(u) = 0$, $Y_{x_0}(u) = 1$ if and only if $U_Y < 0.1$. Therefore, we have that

$$\begin{aligned} Y_{x_0}(u) = 1 \iff &((U_D < 0.5) \wedge (U_Y < 0.1 + \beta)) \vee \\ &((U_D > 0.5) \wedge (U_Y < 0.1)). \end{aligned} \tag{2.23}$$

In the submodel $\mathcal{M}_{x_1}$, we have

$$\begin{aligned} Y_{x_1}(u) = 1 \iff &((U_D < 0.5 + \lambda) \wedge (U_Y < 0.1 + \alpha + \beta)) \vee \\ &((U_D > 0.5 + \lambda) \wedge (U_Y < 0.1 + \alpha)). \end{aligned} \tag{2.24}$$

Based on this, the expression in Eq. 2.22 can be evaluated using Def. 2.5, which leads to

$$P(y_{x_1}, y_{x_0}) = \sum_u \mathbb{1}(Y_{x_1}(u) = 1, Y_{x_0}(u) = 1)P(u) \tag{2.25}$$

$$= P(U_D < 0.5)P(U_Y < 0.1 + \beta) + P(U_D > 0.5)P(U_Y < 0.1)$$

$$= 0.1 + \frac{\beta}{2}. \tag{2.26}$$

Interestingly, this distribution is never attainable from observational data, since it involves both potential responses Y_{x_0}, Y_{x_1}, which can never be observed simultaneously. □

In most fairness analysis settings, the data scientist will only have data $\mathcal{D}$ in the form of samples collected from the observational distribution. One significant result in this context is known as the *causal hierarchy theorem* (CHT, for short), which says that it is almost never possible (in an information-theoretic sense) to recover the counterfactual distribution from the observational distribution alone (Bareinboim *et al.*, 2022, Thm. 1). Given this impossibility result and the unavailability of the SCM in most settings, the data scientist needs to resort to some assumptions in order to possibly make claims about these underlying mechanisms, which is discussed in the next section.

2.3 Encoding Structural assumptions through Causal Diagrams

Even though SCMs are well defined and provide the semantics to different families of probability distributions, and are essential for fairness analysis, one critical observation is that they are usually not observable by the data scientist. A common way of encoding assumptions about the underlying SCM is through an object called a causal diagram. We describe below the constructive procedure that allows one to articulate a diagram from a coarse understanding of the SCM.

Definition 2.6 (Causal Diagram (Pearl, 2000; Bareinboim *et al.*, 2022))**.** Let $\mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle$ be an SCM. A graph $\mathcal{G}$ is said to be a *causal diagram* (of $\mathcal{M}$) if:

1. there is a vertex for every endogenous variable $V_i \in V$,
2. there is an edge $V_i \to V_j$ if V_i appears as an argument of $f_j \in \mathcal{F}$,
3. there is a bidirected edge $V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding $U_i, U_j \subset U$ are correlated or the corresponding functions f_i, f_j share some $U_{ij} \in U$ as an argument.

In words, there is an edge from an endogenous variable V_i to V_j whenever V_j "listens to" V_i for determining its value[5]. Similarly, the existence of a bidirected edge between V_i and V_j indicates there is some shared, unobserved information affecting how both V_i and V_j obtain their values. Note that while the SCM contains explicit information about all structural mechanisms ($\mathcal{F}$) and exogenous variables ($P(u)$), the causal diagram, on the other hand, encodes information only about which functional arguments were possibly used as

[5] This construction lies at the heart of the type of knowledge causal models represent, as suggested in (Pearl and Mackenzie, 2018, pp. 129): "This listening metaphor encapsulates the entire knowledge that a causal network conveys; the rest can be derived, sometimes by leveraging data."

inputs to the functions in $\mathcal{F}$. That is, the diagram abstracts out the specifics of the functions $\mathcal{F}$ and retains information about their possible arguments.

Furthermore, the existence of a directed arrow, e.g., $V_i \to V_j$, encodes the *possibility* of the mechanism of V_j to listen to variable V_i, but not the necessity. In this sense, the edges are non-committal; for instance, f_j may decide not to take the value of V_i into account. On the other hand, the assumptions are not encoded in the arrows present in the diagram but in the missing arrows; each missing arrow ascertains that one variable is *certainly* not the argument of the other. The data scientist, in general, should try to specify as much knowledge as possible of this type. For concreteness, consider the following example.

Example 2.4 (Admissions Causal Diagram)**.** Consider again the SCM $\mathcal{M}$ in Ex. 2.1, which is unknown to the data scientist trying to analyze the existence of discrimination in the admission process. To apply the graphical construction dictated by Def. 2.6, the data scientist starts the modeling process by examining each of the endogenous variables and the potential arguments of their corresponding mechanisms. For example, the mechanism

$$D \leftarrow f_D(X, U_D) \tag{2.27}$$

suggests that each applicant's department choice (D) is, possibly, a function of their gender X, regardless of the specific form of how this happens in reality. If that is the case, so the causal diagram $\mathcal{G}$ will contain the arrow $X \to D$. Again, an arrow in $\mathcal{G}$ does not commit to how the variables X and D interact, which is significantly less informative than the true mechanism given by Eq. 2.5. Continuing the causal modeling process, the data scientist may think about the admission process, and consider that

$$Y \leftarrow f_Y(X, D, U_Y), \tag{2.28}$$

which represents that admission decisions may be influenced by gender and department choice. If that is the case, the causal diagram $\mathcal{G}$ will also contain the arrows $X \to Y$ and $D \to Y$, respectively. Again, this contrasts sharply with how detailed the knowledge avaiable in the true SCM $\mathcal{M}$ is, as delineated by Eq. 2.6. Interestingly enough, an entirely different functional form than that in Eq. 2.6, say

$$Y \leftarrow \mathbb{1}\big(U_Y < 0.1 + \beta X D\big), \tag{2.29}$$

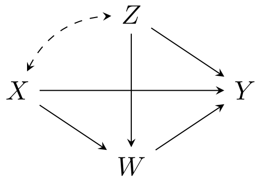
is also compatible with the causal diagram in Fig. 1.2.

Lastly, if the coefficient α is equal to 0 in the mechanism described by Eq. 2.6 (i.e., $Y \leftarrow \mathbb{1}(U_Y < 0.1 + \alpha X + \beta D)$), this would still be compatible with the causal diagram $\mathcal{G}$. Again, the arrow allows for the possibility of functional dependence but does not necessitate it. □

2.3.1 Standard Fairness Model

Specifying the relationship among all pairs of variables, as required by the definition of a causal diagram, is possibly non-trivial in many practical settings. In this section, we will introduce the *Standard Fairness Model*, which is a template-like model that represents a collection of causal diagrams and aims to alleviate the modeling requirements.

Definition 2.7 (Standard Fairness Model (SFM)). The standard fairness model (SFM) is the causal diagram $\mathcal{G}_{\text{SFM}}$ over endogenous variables $\{X, Z, W, Y\}$ and given by



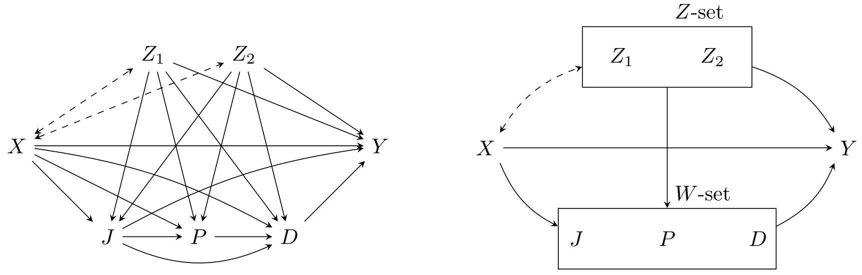
where the nodes represent:

- the *protected attribute*, labeled X (e.g., gender, race, religion),
- the set of *confounding* variables Z, which are not causally influenced by the attribute X (e.g., demographic information, zip code),
- the set of *mediator* variables W that are possibly causally influenced by the attribute (e.g., educational level or other job-related information),
- the *outcome* variable Y (e.g., admissions, hiring, salary).

Nodes Z and W are possibly multi-dimensional or empty. Furthermore, for a causal diagram $\mathcal{G}$, the projection of $\mathcal{G}$ onto the SFM is defined as the mapping of the endogenous variables V appearing in $\mathcal{G}$ into four groups X, Z, W, Y, as described above. The projection is denoted by $\Pi_{\text{SFM}}(\mathcal{G})$ and is constructed by choosing the protected attribute, the outcome of interest, and grouping the confounders Z and mediators W.

When X is a singleton (i.e., we have a single protected attribute), constructing the groups X, Z, W, and Y will be possible for most practical applications. The key assumptions of the SFM template are encoded in the absence of bidirected edges other than $X \dashleftarrow\dashrightarrow Z$. When there are multiple protected attributes, the causal structure that arises may be more complex. In Appendix D we discuss some possibilities for handling multiple protected attributes.

For simplicity, we assume X to be binary (whereas Z, W, and Y could be either discrete or continuous). The adaptation of the framework to the setting of multi-valued or continuous X is discussed in Appendix D, but readers are



(a) Causal diagram of COMPAS dataset.

(b) Causal diagram projected onto the SFM.

Figure 2.1: Causal diagram of COMPAS dataset and its projection onto the SFM.

encouraged to consider the binary case in the main text first, for grounding the key concepts. Furthermore, in Appendix D.1 we discuss the conceptual underpinnings of causal manipulations of protected attributes and explain how one may think about hypothesized manipulations of race, gender or religion. In this appendix, we also address some considerations from the previous works of Kohler-Hausmann, 2018; Hu and Kohler-Hausmann, 2020.

We next ground the notion of the SFM through examples. For instance, by setting $Z = \emptyset$ and $W = \{D\}$, the causal diagram of the Admissions example can be represented by $\mathcal{G}_{\text{SFM}}$. To ground the definition further, consider the following well-known example.

Example 2.5 (COMPAS (Larson *et al.*, 2016))**.** Courts in Broward County, Florida use machine learning to predict whether individuals released on parole are at high risk of re-offending within 2 years (Y). The algorithm is based on the demographic information Z (Z_1 for gender, Z_2 for age), race X (x_0 denoting White, x_1 Non-White), juvenile offense counts J, prior offense count P, and degree of charge D. The causal diagram for this setting is shown in Fig. 2.1a. The bidirected arrows between X and Z_1, Z_2 indicate that the exogenous variable U_X possibly shares information with exogenous variables U_{Z_1}, U_{Z_2}. This diagram can be standardized (projected on the SFM) by grouping the mediators $W = \{J, P, D\}$ and confounders $Z = \{Z_1, Z_2\}$. Formally, the SFM projection can be written as

$$\Pi_{\text{SFM}}(\mathcal{G}) = \langle X = \{X\}, Z = \{Z_1, Z_2\}, W = \{J, P, D\}, Y = \{Y\}\rangle. \tag{2.30}$$

The projection is shown in Fig. 2.1b. Notice that the complete diagram $\mathcal{G}$ is not needed for determining the SFM projection. The data scientist only needs to group the confounders and mediators, and determine whether there is latent confounding between any of the groups.

Going back to Florida, after a period of using the algorithm, it is observed that Non-White individuals are 9% more likely to be classified as high-risk, i.e.,

$$P(y \mid x_1) - P(y \mid x_0) = 9\%. \tag{2.31}$$

The reader may wonder if the disparity of 9% means that racial minorities are discriminated by the legal justice system in Broward County. An important consideration here is how much of the disparity is explained by (i) the spurious association of race with age or gender (which potentially influences the recidivism prediction); (ii) the effect of race on the prediction mediated by juvenile and prior offense counts; or (iii) the direct effect of race on the prediction. □

As noted in the example, the SFM does not explicitly assume the causal structure within the possibly multi-dimensional sets Z, W. In causal language, the SFM can be seen as an equivalence class of causal diagrams[6]. For instance, under the SFM, if $Z = \{Z_1, Z_2\}$, the relationship between Z_1 and Z_2 is not fully specified, and it may be the case that $Z_1 \to Z_2$, $Z_2 \to Z_1$, or the relationship may be of another type. Secondly, the SFM encodes assumptions about the lack of hidden confounding, which is reflected through the absence of bidirected arrows between variable groups. We discuss in Appendix B how the lack of confounding assumptions can be relaxed.

[6] A more detailed study on the properties of clustered diagrams can be found in (Anand *et al.*, 2021).

3

Foundations of Causal Fairness Analysis

In this section, we will introduce two main results that will allow us to understand and possibly solve the problem of fairness using causal tools. First, we will introduce in Sec. 3.1 a structural definition of fairness, which leads to a natural way of expressing legal requirements based on the doctrines of disparate treatment and impact. In particular, we will define the notion of *fairness measure* and two key properties called *admissibility* and *decomposability.* Armed with these new notions, we will then be able to formally state the fundamental problem of causal fairness analysis. In words, these results suggest that reasoning about fairness requires an understanding of how to explain variations, in particular, how the outcome variable Y can be explained in terms of the structural measures following variations of the protected attribute X. In Sec. 3.2, we formalize the notion of a contrast, which allows us to understand the aforementioned variations from a factual-counterfactual perspective. We then prove how to decompose contrasts and re-express them in terms of the structural basis, which lead to the explainability plane and the decomposition of arbitrary types of contrast. The discussion is somewhat theoretical and we will provide examples to ground and make the main points more concrete.

Example 3.1 (College Admissions, inspired by (Bickel *et al.*, 1975))**.** During the process of application to undergraduate studies, prospective students choose a department to which they want to join (D), report their gender X (x_0 female, x_1 male), and after a certain period they receive the admission decisions Y (y_1 accepted, y_0 rejected).

In reality, how applicants pick their department (f_D) and how the university decides on who to admit (f_Y) is represented by the SCM $\mathcal{M}^* = \langle V = \{X, D, Y\}, U = \{U_X, U_D, U_Y\},\ \mathcal{F}^*, P^*(U)\rangle$, where the pair $\langle \mathcal{F}^*, P^*(U)\rangle$ is such that

$$\mathcal{F}^*, P^*(U) : \begin{cases} X & \leftarrow \text{Bernoulli } (0.5) & (3.1) \\ D & \leftarrow \text{Bernoulli } (0.5 + \frac{2}{10} \text{ X}) & (3.2) \\ Y & \leftarrow \text{Bernoulli } (0.1 + 0 \cdot X + \frac{7}{10} \text{ D}). & (3.3) \end{cases}$$

Based on data that it made available from the previous admissions cycle, the school is sued by a group of applicants who allege gender discrimination. In particular, they share with the court the following statistics:

$$P(y \mid x_1) - P(y \mid x_0) = 14\%, \tag{3.4}$$

which seems a devastating piece of evidence against the university. In words, it seems that male candidates are 14% more likely to be admitted than their female counterparts. The natural question that arises is what could explain such a disparity in the observed data? Would this be a textbook case of direct, gender-discrimination?

Despite the fact that the court does not have access to the true $\mathcal{M}^*$, in reality, there is no direct discrimination at all since f_Y (Eq. 3.3) does not take gender into account (note the zero coefficient multiplying X). In fact, female applicants are more likely to apply to arts & humanities departments, which have lower admission rates, in turn causing a disparity in the overall admission rates.

The plaintiffs hire a team of (evil) data scientists that conduct their own study. After some time, the team comes back and claims to have understood the university decision-making process after a series of interviews and research, which is given by the SCM $\mathcal{M}' = \langle V = \{X, D, Y\}, U = \{U_X, U_D, U_Y\}, \mathcal{F}', P'(U)\rangle$, where $\langle \mathcal{F}', P'(U)\rangle$ are such that

$$\mathcal{F}', P'(U) : \begin{cases} X & \leftarrow \text{Bernoulli } (0.5) & (3.5) \\ D & \leftarrow \text{Bernoulli}(0.5 + \frac{2}{10} X) & (3.6) \\ Y & \leftarrow \text{Bernoulli}(0.1 + \frac{14}{100} \cdot X + 0 \cdot D). & (3.7) \end{cases}$$

The only difference between $\mathcal{M}^*$ (the true set of mechanisms) and $\mathcal{M}'$ (the hypothesized one) is f_Y. Interestingly enough, the hypothesized f_Y (Eq. 3.7) takes gender (X) into account while discarding any information about applicants' department choices (D). Clearly, if this was indeed the true decision-making process by which the university selects students, the jury should condemn the university, since that would be a blatant case of direct discrimination. □

Interestingly, both SCMs $\mathcal{M}^*$ and $\mathcal{M}'$ generate the same total variation of 14%. Still, $\mathcal{M}^*$, which is the true generating model, doesn't suggest any type of gender discrimination, while $\mathcal{M}'$, which is false, suggests that the university's admissions decisions are purely based on gender. In summary, SCMs $\mathcal{M}^*$ and $\mathcal{M}'$ are qualitatively different (in the sense that the disparity is transmitted along different causal mechanisms), but they are indistinguishable based on TV measure. We next formalize this issue in more generality.

3.1 Structural Fairness Criteria

To understand the issue discussed in the previous section, we start by noting that qualitative distinctions – such as differentiating direct and indirect discrimination – lie at the heart of some of the most important legal doctrines on discrimination. In particular, the doctrine of *disparate treatment* asks the question on whether a different decision would have been reached for an individual, had she/he been of a different race or gender, while keeping all other attributes the same (Barocas and Selbst, 2016). In causal terminology, the question is about disparities transmitted along the *direct causal mechanism* between the attribute X and the outcome Y. On the other hand, the doctrine of *disparate impact* considers situations in which a facially neutral policy (that does not use race or gender explicitly) results in very different outcomes for racial or gender groups (Rutherglen, 1987). In this case, the concern is also with disparities transmitted along *indirect and spurious* causal mechanisms. Motivated by these legal doctrines, we can mathematically define qualitative assessments about discrimination based on an SCM:

Definition 3.1 (Structural Fairness Criterion)**.** Let Ω be a space of SCMs. A structural criterion Q is a binary operator on the space Ω, that is a map $Q : \Omega \to \{0, 1\}$ that determines whether a set of causal mechanisms between X and Y exist or not, in a given SCM $\mathcal{M} \in \Omega$.

For most of the manuscript, we wish to focus on structural criteria that capture direct, indirect, and spurious discrimination. We consider these criteria as elementary. More refined and detailed structural notions are discussed in Sec. 6. We now formally define the three elementary structural fairness criteria, based on the functional relationships between X and Y encoded in an SCM:

Definition 3.2 (Elementary Structural Fairness Criteria)**.** Let $\text{pa}(V_i)$ and $\text{an}(V_i)$ be the observed parents and ancestors of V_i in the causal diagram $\mathcal{G}$, respectively. Let $\overline{\text{an}}(V_i)$ denote the extended set of ancestors of V_i that also includes the unobserved, exogenous ancestors of V_i. Let $\mathcal{G}_{\underline{X}}$ denote the causal diagram $\mathcal{G}$ with the outgoing edges from X removed. For an SCM $\mathcal{M}$, define the following three structural criteria:

(i) Structural direct criterion:

$$\text{Str-DE}_X(Y) = \mathbb{1}(X \in \text{pa}(Y)).$$

(ii) Structural indirect criterion:

$$\text{Str-IE}_X(Y) = \mathbb{1}(X \in \text{an}(\text{pa}(Y))).$$

(iii) Structural spurious criterion:

$$\text{Str-SE}_X(Y) = \mathbb{1}\Big(\overline{\text{an}}(X) \cap \overline{\text{an}}_{\mathcal{G}_{\underline{X}}}(Y) \neq \emptyset\Big).$$

For $\text{Str-DE}_X(Y) = 0$, $\text{Str-IE}_X(Y) = 0$, and $\text{Str-SE}_X(Y) = 0$, we write $\text{DE-fair}_X(Y)$, $\text{IE-fair}_X(Y)$, and $\text{SE-fair}_X(Y)$, respectively.

In words, the structural direct criterion verifies whether the attribute X is a function of the mechanism f_Y, that is, if Y is a function of X. The structural indirect criterion verifies whether there exist mediating variables, which are affected by X, that in turn influence Y. These two criteria are defined in terms of the functional relationships within $\mathcal{M}$, or $\mathcal{F}$. This means that they convey causal information about the relationship among endogenous variables. Finally, the structural spurious criterion verifies whether there exist variables (either observed or unobserved) that both causally affect the attribute X and the outcome Y, sometimes also referred to as back-door confounding. Different than the previous ones, this criterion also relies on the relationships among the exogenous variables U, which relates to the confounding relation among the observables. We revisit the Admissions example to ground these notions:

Example 3.2 (Admissions – continued)**.** In the SCM $\mathcal{M}$ defined in Eq. 2.1-2.3, the structural direct and indirect effects can be analyzed as follows:

(i) Y is fair w.r.t. X in terms of direct effect if and only if:

$$\alpha = 0 \text{ in } \{Y \leftarrow \text{Bernoulli}(0.1 + \alpha X + \beta D)\}. \tag{3.8}$$

(ii) Y is fair w.r.t. X in terms of indirect effect if and only if:

$$\begin{aligned} &\lambda = 0 \text{ in } \{D \leftarrow \text{Bernoulli}(0.5 + \lambda X)\}, \text{ or} \\ &\beta = 0 \text{ in } \{Y \leftarrow \text{Bernoulli}(0.1 + \alpha X + \beta D))\}. \end{aligned} \tag{3.9}$$

For the SCM $\mathcal{M}^*$ in Eq. 3.1-3.3, we can see that direct discrimination does not exist, since $\alpha = 0$, and therefore $X \notin \text{pa}(Y)$ (see Def. 3.2(i)). However, indirect discrimination is present, since $\lambda = \frac{2}{10}$ and $\beta = \frac{7}{10}$, and therefore $X \in \text{an}(\text{pa}(Y))$ (see Def. 3.2(ii)). In contrast to this, for the SCM $\mathcal{M}'$ in Eq. 3.5-3.7, direct discrimination is present, since $\alpha = \frac{1}{7}$ and thus $X \in \text{pa}(Y)$, but indirect discrimination is not, since $\beta = 0$ and thus $X \notin \text{an}(\text{pa}(Y))$. □

Other meaningful structural fairness criteria could be defined using different logical combinations of these three elementary criteria. For instance, Y can be called *totally fair* with respect to X ($\text{Fair}_X(Y)$) if and only if direct, indirect, and spurious fairness are simultaneously true (i.e., $\text{Fair}_X(Y) = \text{DE-fair}_X(Y) \land \text{IE-fair}_X(Y) \land \text{SE-fair}_X(Y)$). Alternatively, causal fairness could be defined as $\text{Causal-fair}_X(Y) = \text{DE-fair}_X(Y) \land \text{IE-fair}_X(Y)$, which encodes the non-existence of active causal influence from X to Y (neither direct nor mediated).

Structural definitions of fairness represent idealized and intuitive criteria that can be evaluated whenever the true underlying mechanisms are known, i.e., the fully specified SCM $\mathcal{M}$. The importance of these measures, encoded through the structural mechanisms (Def. 3.2), stems from the fact that they underpin existing legal and societal notions of fairness. Therefore, they will be used as a benchmark to understand under what conditions, and how close other measures, which might be estimable from data, approximate these idealized and intuitive notions.

One central question is whether there exist quantitative measures of discrimination that can help us assess whether a structural criterion is satisfied or not. Firstly, we define a general fairness measure that can be computed from the SCM:

Definition 3.3 (Fairness Measure). Let Ω be a space of SCMs. A fairness measure μ is a functional on the space Ω, that is a map $\mu : \Omega \to \mathbb{R}$, which quantifies the association of X and Y through any subset of causal mechanisms, in a given SCM $\mathcal{M} \in \Omega$.

Here, the definition of a fairness measure μ is kept as quite general. In Sec. 3.2, we will restrict our attention to a specific class of measures μ and explain their importance in the context of Causal Fairness Analysis. In the sequel, we introduce a notion that represents when a fairness measure μ is suitable for assessing a structural criterion Q:

Definition 3.4 (Admissibility). Let Ω be a class of SCMs on which a structural criterion Q and a measure μ are defined. A measure μ is said to be admissible w.r.t. the structural criterion Q within the class of models Ω, or (Q, Ω)-admissible, if:

$$\forall \mathcal{M} \in \Omega : Q(\mathcal{M}) = 0 \implies \mu(\mathcal{M}) = 0. \tag{3.10}$$

For simplicity, we will use admissibility instead of (Q, Ω)-admissibility whenever the context is clear. The importance of having an admissible measure μ stems from the contrapositive of Eq. 3.10, namely, if $\mu(\mathcal{M})$ can be measured or evaluated and $\mu(\mathcal{M}) \neq 0$, this means that the structural measure must be true, i.e., $Q(\mathcal{M}) = 1$. In other words, the measure μ will act as a link between the well-defined but unobservable structural measure and the observable and estimable world. For concreteness, consider the following result that formalizes the issue found in Ex. 3.1:

Proposition 3.1 (TV is not admissible w.r.t. Str-DE, IE, SE). Let Ω be the space of Semi-Markovian SCMs which contain variables X and Y. Let μ be the total variation measure $\text{TV}_{x_0,x_1}(y)$. Then μ is not admissible with respect to structural direct, indirect, or spurious criteria. That is,

$$(\text{Str-DE}(\mathcal{M}) = 0) \;\not\Rightarrow\; (\text{TV}_{x_0,x_1}(y) = 0), \tag{3.11}$$
$$(\text{Str-IE}(\mathcal{M}) = 0) \;\not\Rightarrow\; (\text{TV}_{x_0,x_1}(y) = 0), \tag{3.12}$$
$$(\text{Str-SE}(\mathcal{M}) = 0) \;\not\Rightarrow\; (\text{TV}_{x_0,x_1}(y) = 0). \tag{3.13}$$

In fact, the reason why the TV measure is not admissible with respect to structural direct, indirect, and spurious criteria is because it captures the three types of variations together.

To formalize this idea, we introduce the notion of *decomposability* of a measure μ, i.e.:

Definition 3.5 (Decomposability). Let Ω be a class of SCMs and μ be a measure defined over it. μ is said to be Ω-decomposable if there exist measures

$$\mu_1, \dots, \mu_k \text{ such that } \mu = f(\mu_1, \dots, \mu_k), \tag{3.14}$$

and where f is a non-trivial function vanishing at the origin, $f(0, \dots, 0) = 0$.

In words, decomposability states that a measure μ can be written as a function of measures $(\mu_i)_{i=1}^k$, and that if all measures $(\mu_i)_{i=1}^k$ are equal to 0 for an SCM $\mathcal{M}$, then the measure μ must be 0 as well. For concreteness, consider the following example.

Example 3.3 (Covariance decomposition, after (Zhang and Bareinboim, 2018c)). Let μ be the covariance measure between random variables X and Y,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \tag{3.15}$$

which plays a role somewhat analogous to TV (and, more broadly, the observational distribution) whenever the system $\mathcal{F}$ and $P(u)$ are linear and Gaussian. Further, let the causal covariance be defined as

$$\text{Cov}^c_x(X, Y) = \text{Cov}(X, Y - Y_x). \tag{3.16}$$

Furthermore, let the spurious covariance be defined as

$$\text{Cov}^s_x(X, Y) = \text{Cov}(X, Y_x). \tag{3.17}$$

Then, we can write

$$\text{Cov}(X, Y) = f\big(\text{Cov}^c_x(X, Y), \text{Cov}^s_x(X, Y)\big), \tag{3.18}$$

with the function $f(a, b) = a + b$, which satisfies $f(0, 0) = 0$. □

Armed with the definitions of admissibility and decomposability, we are ready to formally define the first version of the problem studied here.

Definition 3.6 (Fundamental Problem of Causal Fairness Analysis (preliminary))**.** Consider a class of SCMs Ω, and let

- $Q_1, \dots, Q_k$ be a collection of structural fairness criteria, and
- μ be a measure,

both defined over Ω. The Fundamental Problem of Causal Fairness Analysis is to find a collection of measures $\mu_1, \dots, \mu_k$ such that the following properties are satisfied:

(1) μ is decomposable w.r.t. $\mu_1, \dots, \mu_k$;

(2) $\mu_1, \dots, \mu_k$ are admissible w.r.t. the structural fairness criteria $Q_1, \dots, Q_k$.

In other words, find measures

$$\mu_1, \dots, \mu_k \text{ that are admissible w.r.t. } Q_1, \dots, Q_k, \tag{3.19}$$

respectively, and such that

$$\mu = f(\mu_1, \dots, \mu_k), \tag{3.20}$$

where f is a non-trivial function vanishing at the origin, $f(0, \dots, 0) = 0$.

For grounding this discussion, we will consider that the measure μ is given by the TV[1] and the structural measures will be Str-{DE,IE,SE}. We refer to this problem as FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$). Fig. 3.1 provides a visual summary of the FPCFA where TV is shown on the top and the structural measures Str-{DE,IE,SE} on the bottom. As we have just seen in Prop. 3.1, TV is not admissible relative to each of these structural measures.

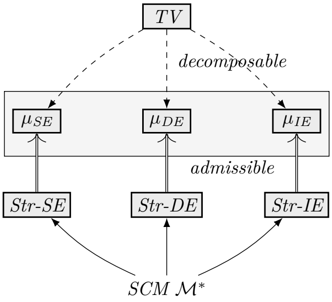


Figure 3.1: Fundamental Problem of Causal Fairness Analysis (TV version).

The FPCFA asks for the existence of a set of measures $(\mu_{DE}, \mu_{IE}, \mu_{SE})$ that could act as a bridge between TV and the more meaningful, albeit unobservable

[1] Naturally, other types of contrasts can be used as measures instead of TV, such as the covariance (Zhang and Bareinboim, 2018c) or equality of odds (Hardt *et al.*, 2016; Zhang and Bareinboim, 2018a).

structural measures Str-{DE,IE,SE}. In fact, the FPCFA is solved whenever TV can be expressed in terms of $(\mu_{DE}, \mu_{IE}, \mu_{SE})$, and each of these measures is admissible w.r.t. to the corresponding structural measures. If that is the case, the measures $(\mu_{DE}, \mu_{IE}, \mu_{SE})$ could be seen as explaining the variations of TV in terms of the most elementary, structural components. Interestingly, this is both a quantitative and a qualitative exercise. From TV's perspective, $(\mu_i)_{i=1}^k$ should account for all its variations, which is naturally a quantitative exercise. From the structural measures perspective, we would like to enforce soundness, namely, discrimination is indeed readable from the corresponding $(\mu_i)_{i=1}^k$, which is a qualitative exercise.

3.2 Explaining Factual & Counterfactual Variations

In this section, the main task is studying how the variations in outcome Y can be explained by changes of the protected attribute X. The result of this study is what we call the *population-mechanism* plane, which we also refer to as the *explainability plane* (Fig. 3.3). The methodology introduced by the plane will allow us to re-express different measures of fairness in a unified manner, which will facilitate their comparison in terms of admissibility, decomposability, and possibly other desirable properties.

We start by introducing a quite general type of measure encoding the idea of contrast.

Definition 3.7 (Contrast). Given a SCM $\mathcal{M}$, a contrast $\mathcal{C}$ is any quantity of the form

$$\mathcal{C}(C_0, C_1, E_0, E_1) = \mathbb{E}[y_{C_1} \mid E_1] - \mathbb{E}[y_{C_0} \mid E_0], \tag{3.21}$$

where E_0, E_1 are observed (factual) clauses and C_0, C_1 are counterfactual clauses to which the outcome Y responds. Furthermore, whenever

(a) $E_0 = E_1$, the contrast $\mathcal{C}$ is said to be counterfactual;

(b) $C_0 = C_1$, the contrast $\mathcal{C}$ is said to be factual.

For simplicity[2], we will focus on the binary case, in which a contrast can be written as

$$P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0). \tag{3.22}$$

The purpose of a contrast is to compare the outcome of individuals who coincide with the observed event E_1 in the factual world and whose values were intervened on (possibly counterfactually) as defined by C_1, against individuals who coincide with the observed event E_0 in the factual world and whose values were intervened on (possibly counterfactually) as defined by C_0. The

[2] The results in this section hold for any real-valued random variable Y.

definition also distinguishes two special cases of contrasts. A counterfactual contrast captures only the difference in outcome induced by the difference in interventions C_0, C_1 (since $E_0 = E_1$). Complementary to this, a factual contrast captures only the difference induced by the observed events E_0, E_1 (since $C_0 = C_1$). We now show why contrasts are useful for explaining variations:

Theorem 3.1 (Contrast's Decomposition & Structural Basis Expansion). Let $\mathcal{M}$ be an SCM and let $\mathcal{C}$ be a contrast $P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0)$. $\mathcal{C}$ can be decomposed into its counterfactual and factual variations, namely:

$$\underbrace{P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_1)}_{\text{counterfactual contrast}} + \underbrace{P(y_{C_0} \mid E_1) - P(y_{C_0} \mid E_0)}_{\text{factual contrast}}. \tag{3.23}$$

Furthermore, the corresponding counterfactual and factual contrasts admit the following structural basis expansions, respectively:

(a) Counterfactual contrast ($\mathcal{C}_{\text{ctf}}$), where $E_0 = E_1 = E$, can be expanded as

$$P(y_{C_1} \mid E) - P(y_{C_0} \mid E) = \sum_u \Big(\underbrace{\mathbb{1}(Y_{C_1}(u) = y) - \mathbb{1}(Y_{C_0}(u) = y)}_{\text{unit-level difference}}\Big) \times \underbrace{P(u \mid E)}_{\text{posterior}}, \tag{3.24}$$

(b) Factual contrast ($\mathcal{C}_{\text{factual}}$), where $C_0 = C_1 = C$, can be expanded as

$$P(y_C \mid E_1) - P(y_C \mid E_0) = \sum_u \underbrace{\mathbb{1}(Y_C(u) = y)}_{\text{unit outcome}} \Big(\underbrace{P(u \mid E_1) - P(u \mid E_0)}_{\text{posterior difference}}\Big). \tag{3.25}$$

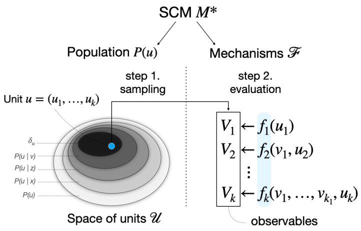


Figure 3.2: Two-step generative process includes sampling a unit from the population (left), and evaluating it against corresponding structural mechanisms (right).

The decomposition and structural basis expansion of contrasts presented in this theorem entail a fundamental connection of causal fairness measures with structural causal models. In particular, the decomposition given in Eq. 3.23 allows us to disentangle factual and counterfactual variations within any contrast.

We note that Eqs. 3.24 and 3.25 reexpresses the variations within the target quantity in terms of the underlying units and activated mechanisms, as referenced by the SCM. We would like to understand these qualitatively different types of variations separately. First, we will take a generative interpretation over how the targeted variations are realized in terms of the SCM $\mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle$. Fig. 3.2 illustrates the two-step generative process that goes as follows:

(1) **Sampling:** A unit $U = u$ is sampled from the population distributed according to $P(U)$;

(2) **Evaluation:** This unit u passes through the sequence of mechanisms $\mathcal{F}$, in causal order, until the values of the endogenous variables V are realized.

The l.h.s. of the figure shows the sampling process while the r.h.s. represents the evaluation process. As discussed in Sec. 2.2, if the system is not submitted to an intervention, this leads to the observational distribution. On the other hand, if the values of certain variables are fixed through intervention, this leads to the corresponding counterfactual distribution.

Considering this two-step generative process, we re-examine the variations encoded in the structural basis expansion of Thm. 3.1. For convenience, we reproduce the equation relative to the counterfactual variations in the sequel (Eq. 3.24), but for simplicity we restrict our attention to the positive outcome $Y = 1$, and replace $\mathbb{1}(Y_C(u) = y)$ terms in Eq. 3.24 with the shorthand $y_C(u)$:

$$P(y_{C_1} \mid E) - P(y_{C_0} \mid E) = \sum_u \big(\underbrace{y_{C_1}(u) - y_{C_0}(u)}_{\text{unit-level difference}}\big) \underbrace{P(u \mid E)}_{\text{posterior}}.$$

First, we consider the second factor in the r.h.s. of the expression. Note that $P(u \mid E = e)$ represents the first step in the generative process in which units who naturally arise to value $E = e$ are drawn from the population. In fact, depending on the granularity of the evidence E, a different fraction of the population (or types of individuals) will be selected. For instance, if $E = \{\}$, the (posterior) distribution $P(u)$ is somewhat uninformative, and represents an average when units are drawn at random from the underlying population, regardless of their predispositions and characteristics. On the other hand, if $E = \{X = x\}$, the posterior distribution $P(u \mid x)$ would be more informative since it now includes units that naturally would have $X = x$. This is less informative compared to more specific events such as $E = \{X = x, Z = z\}$ or $E = \{X = x, Z = z, W = w, Y = y\}$. In fact, the l.h.s. of the figure illustrates this increasingly more refined and informative set of events E, i.e., starting from picking individuals at random from the general population, $P(u)$, to a single individual δ_u, where δ_u is the Dirac delta function. Second, we note that once the unit $U = u$ is selected, all randomness vanishes, and the unit will go through the set of mechanisms $\mathcal{F}$. The first factor of the expression, $y_{C_1}(u) - y_{C_0}(u)$, describes the difference in response y between conditions C_1 and C_0 for a fixed realization of exogenous variables u. As realizations of exogenous variables U determine the identity of different units in the population, the quantity $y_{C_1}(u) - y_{C_0}(u)$ will be an unit-level quantity.

In the context of fairness discussed here, consider the case when $C_1 = x_1$ and $C_0 = x_0$, which could represent the protected attribute, for instance, males and females, or White and African-American. The quantity $y_{x_1}(u) - y_{x_0}(u)$ measures what the change in outcome Y would be when changing the attribute X from x_0 to x_1, for a specific unit u. For this particular choice of C_0, C_1, the quantity captures what is known as the total causal effect of X on Y, that is it includes all the variations from X to Y translated across causal pathways.

In summary, any counterfactual contrast $\mathcal{C}_{\text{ctf}}$ can be decomposed into two parts:

1. A unit-level difference comparing the counterfactual worlds C_1 vs. C_0 for a specific unit $U = u$. This quantity is determined by the causal mechanisms $\mathcal{F}$ of the SCM, and does not depend on the distribution $P(u)$.

2. A posterior distribution $P(u \mid E = e)$ that indicates the probability mass assigned to unit u whenever the event $E = e$. By changing the granularity of the event E, the space of included units is restricted, making the measure more specific to a subpopulation (see Fig. 3.2 (l.h.s.)).

Given that the selection of units is fixed (second factor), and the only thing that varies is the selection of the mechanisms (first factor) through the choices of the counterfactual conditions C_1 and C_0, this will generate downstream variations that are inherently "causal". In fact, the specific instantiation of C_1 and C_0 and $E = \{\}$ (i.e., $P(u)$) matches the very definition of average causal effect, $P(y \mid do(x_1)) - P(y \mid do(x_0))$.

We now re-examine the factual variations encoded in the structural basis expansion of Thm. 3.1. For convenience, we reproduce the corresponding equation (Eq. 3.25):

$$P(y_C \mid E_1) - P(y_C \mid E_0) = \sum_u \underbrace{y_C(u)}_{\text{unit outcome}} \big(\underbrace{P(u \mid E_1) - P(u \mid E_0)}_{\text{posterior difference}}\big)$$

In words, a factual contrast can be expanded as a sum of differences in the posteria $P(u \mid E_1) - P(u \mid E_0)$, weighted by unit-level outcomes $y_C(u)$. We note that the difference in posteria represents the first step in the generative process in which two sets of units that naturally arise to values E_1 and E_0 are drawn from the population, respectively. Similarly to the previous discussion, different sub-populations will be selected depending on the granularity of the evidence E_1, E_0. The scope of these events is the same but their instantiations are different.

This can be seen as complementary when compared to the counterfactual contrasts. Given that the mechanisms are fixed (first factor), the component that generates variations is relative to the choice of units based on the factual conditions E_0 and E_1. We suggest this will generate upstream variations, which will be "non-causal" (also called spurious), as described in more detail later on in the manuscript. For factual contrasts, we are mostly interested in setting $C_1 = C_0 = x$, so that $X = x$ along all causal pathways. The contrast will then capture the difference in probability mass assigned to u in events E_1 and E_0. By definition, spurious effects are generated by variations that *causally precede* X, so these cannot be captured by intervening on X. For this reason, we need to compare events E_1 and E_0, which have resulted in a different instantiation of the value of X. This factorization also suggests mathematically how causal and spurious effects are inherently different from each other.

Explainability plane. By decomposing variations via factual and counterfactual contrasts, and expanding them using the structural basis, we can give the essential structure of the measures used in Causal Fairness Analysis. The approach used for decomposing the total variation is shown in Fig. 3.3, which we call the explainability plane. As the figure illustrates, there are two separate axes of the decomposition. On the mechanism axis, we are decomposing the TV into its direct, indirect, and spurious variations. On the population axis, we are considering increasingly precise subsets of the space of units $\mathcal{U}$, which correspond to different posterior distributions. As we will see later, moving along the population axis will correspond to constructing increasingly more powerful fairness measures that are better suited for detecting discrimination.

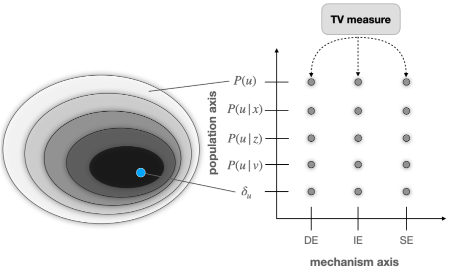


Figure 3.3: In the population axis, contrasts are restricted to smaller subsets of units u in the domain $\mathcal{U}$. At the same time, along the mechanism axis, we distinguish between direct, indirect, and spurious variations.

4

Total Variation Family

In this section, we introduce a family of measures that populate the explainability plane in Fig. 3.3. Since all the measures describe variations included within the TV measure, we refer to them as the *TV-family* (part e of Fig. 1.3). In particular, this section aims to solve the FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) discussed in Sec. 3. The measures in the TV-family are introduced in order. We start with measures that quantify discrimination in the entire population of units u (corresponding to the posterior $P(u)$), and reach measures that quantify discrimination for a single unit u (corresponding to the posterior δ_u, where δ is the Dirac delta function).

4.1 Population-level Contrasts - $P(u)$

We first recall that the TV measure itself is not admissible with respect to structural criteria Str-{DE,IE,SE}, as shown in Prop. 3.1. Specifically, the reason for this is that the TV captures variations between groups generated by any mechanism of association, both causal and non-causal, and does not distinguish between them. Our first step is therefore to disentangle these variations – the causal and non-causal (or spurious) – within the TV.

Definition 4.1 (Total and Spurious Effects). Let the total effect and experimental spurious effect be defined as follows:

$$\text{TE}_{x_0,x_1}(y) = P(y_{x_1}) - P(y_{x_0}) \tag{4.1}$$

$$\text{Exp-SE}_x(y) = P(y \mid x) - P(y_x) \tag{4.2}$$

Further, we write TE-fair$_X(Y)$ whenever $\text{TE}_{x_0,x_1}(y) = 0$, or simply TE-fair when X and Y are clear from the context. Exp-SE-fair is defined analogously.

In words, TE measures the difference in the outcome Y when setting $X = x_1$, compared to setting $X = x_0$. The measure can be visualized graphically as shown in Fig. 4.1a. In this case, Y responds to the change in X from x_0 to x_1 through two mechanisms: (i) the direct link, $X \to Y$, and (ii) the indirect link via W, $X \to W \to Y$. In the context of the COMPAS dataset in Ex. 2.5, the total effect would be the average difference in recidivism prediction had an individual's race been set to White (by intervention) compared to had it been set to Non-White. Since the covariates Z vary naturally in both counterfactual worlds (both sides of the expression), those are canceled out and Y variations can be explained in terms of the downstream variations in response to the change in X[1].

In a complementary manner, the experimental spurious effect measures the average difference in outcome Y when simply observing that $X = x$, compared to setting $X = x$ by intervention, as shown graphically in Fig. 4.1b. Since from Y's perspective X has the same value x in both factors, the Y variations can be explained in terms of the upstream effect in response to how X naturally affected Z versus how Z varies free from the influence of X. In the COMPAS dataset, this would mean the average difference in recidivism prediction for individuals for whom the race is set to White by intervention, compared to simply observing the race to be White.

Syntactically, following the discussion in Sec. 3.2, we can write these quantities in terms of contrasts (Def. 3.7), namely:

$$\text{TE}_{x_0,x_1}(y) = \mathcal{C}(x_0, x_1, \emptyset, \emptyset) \tag{4.3}$$

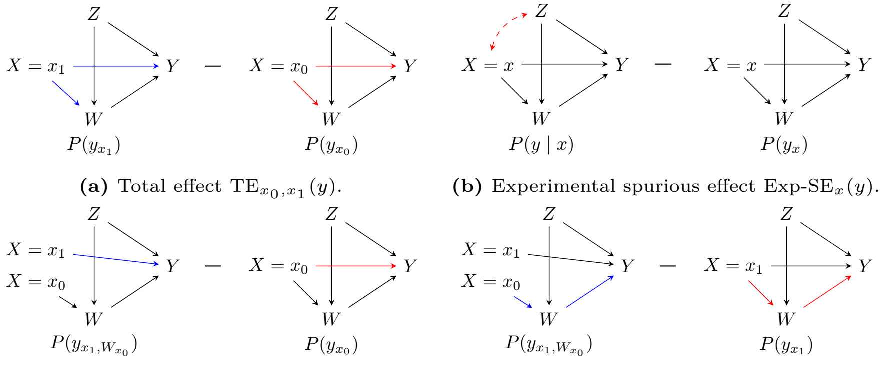
$$\text{Exp-SE}_x(y) = \mathcal{C}(x, \emptyset, \emptyset, x) \tag{4.4}$$

Based on these two notions, the TV can be decomposed into two distinct sources of variation, which correspond precisely to its causal and non-causal mechanisms:

Lemma 4.1 (TV Decomposition I). The total variation measure can be decomposed as

$$\text{TV}_{x_0,x_1}(y) = \text{TE}_{x_0,x_1}(y) + (\text{Exp-SE}_{x_1}(y) - \text{Exp-SE}_{x_0}(y)). \tag{4.5}$$

[1] The TE measure is also called causal effect and sometimes written in do-notation, $P(y \mid do(x_1)) - P(y \mid do(x_0))$. Obviously, this quantity has well-defined semantics given an SCM, despite the fact that no one intends or believes to set any of the protected attributes literally by intervention. Still, through the formal language of causality, one can contemplate these distinct counterfactual realities. In particular, one can disentangle and explain the sources of Y variations in response to changes in X, including the ones through the causal pathways versus the non-causal ones, along the spurious paths.



(a) Total effect $\mathrm{TE}_{x_0,x_1}(y)$.

(b) Experimental spurious effect $\text{Exp-SE}_x(y)$.

(c) Natural direct effect $\mathrm{NDE}_{x_0,x_1}(y)$.

(d) Natural indirect effect $\mathrm{NIE}_{x_1,x_0}(y)$.

Figure 4.1: Graphical representations of measures used in TV decomposition.

Lem. 4.1 shows that the TV measure equals to the total effect on Y when X transitions from x_0 to x_1 plus the difference between the experimental spurious effect of $X = x_1$ and $X = x_0$ [2]. In other words, TV accounts for the sum of the directed (causal) and confounding paths from X to Y. More formally, the lemma shows that the TV is decomposable with respect to TE and Exp-SE (recall Def. 3.5).

Interestingly, the TE itself is still not admissible w.r.t. Str-{DE,IE}, as it captures all causal influences of X on Y, including the direct (through the direct link $X \to Y$) and indirect ones (i.e., paths via W).

Proposition 4.1 (TE Inadmissibility). The total effect measure $\mathrm{TE}_{x_0,x_1}(y)$ is not admissible with respect to structural criteria Str-DE and Str-IE.

To solve FPCFA(Str-{DE,IE,SE}, $\mathrm{TV}_{x_0,x_1}(y)$), therefore, we will further need to disentangle the relationships within TE. In particular, we will need to determine the Y variations that are a direct consequence of the protected attribute, and the ones that are mediated by other variables. In the literature, the total effect was shown to be decomposable into the measures known as the *natural direct and indirect effects* (Pearl, 2001).

[2] An alternative way of interpreting this relation is by flipping TV and TE in the equation, namely:

$$\mathrm{TE}_{x_0,x_1}(y) = \mathrm{TV}_{x_0,x_1}(y) - (\text{Exp-SE}_{x_1}(y) - \text{Exp-SE}_{x_0}(y)). \tag{4.6}$$

This means that the total effect of transitioning X from x_0 to x_1 on Y is equal to the corresponding total variation of Y minus the difference in spurious effects of $X = x_1$ versus the baseline $X = x_0$.

Definition 4.2 (Natural Direct and Indirect Effects)**.** The natural direct and indirect effects are defined, respectively, as follows:

$$\mathrm{NDE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0}) \tag{4.7}$$

$$\mathrm{NIE}_{x_1,x_0}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_1}). \tag{4.8}$$

Further, we write NDE-fair$_X(Y)$ for $\mathrm{NDE}_{x_0,x_1}(y) = 0$, or simply NDE-fair when the attribute/outcome are clear from the context. The condition NIE-fair is defined analogously.

There are several important observations about these definitions. First in terms of semantics, the NDE (Eq. 4.7) captures how the outcome Y changes when setting $X = x_1$, but keeping the mediators W at whatever value they would have taken had X been x_0, compared to setting $X = x_0$ by intervention. This counterfactual statement is shown graphically in Fig. 4.1c. Note that Y "perceives" X through the direct link (marked in blue) as if it is equal to x_1, written in counterfactual language as y_{x_1}, while W perceives X as if it is x_0, formally, W_{x_0}. Putting these two together leads to the first factor in Eq. 4.7, i.e., $y_{x_1,W_{x_0}}$. The second factor in the contrast is y_{x_0}, which can be written equivalently as $y_{x_0,W_{x_0}}$, due to the consistency axiom. It represents the fact that both Y and W perceive X at the same level, x_0[3]. Whenever we subtract one from the other, in some sense, the variations coming from X to Y through W are the same (since it perceives X at the baseline level x_0), and what remains are the variations transmitted through the direct arrows, so the name direct effect. The qualification natural is because W attains its value naturally, depending on the value of X, and not by intervention.

Second, in the context of our COMPAS example, the NDE would measure how much the predicted probability of recidivism would change for an individual whose race was set by intervention to White, but their juvenile and prior offense counts took a value they would have attained naturally (that is, a value naturally attained by Non-White individuals), compared to the race being set to Non-White. The contrast represented by the NDE (in Eq. 4.7) is known as a *nested counterfactual*, since X takes distinct values when influencing different variables. Albeit not realizable in the real world, it encodes significant types of variations that can be evaluated from a collection of mechanisms and fully specified SCM, and which is sometimes computable from data, as discussed in more details in Sec. 4.3.

Third, the definition of NIE follows a similar logic while flipping the sources of variations, as illustrated in Eq. 4.8 and Fig. 4.1d. More specifically, the outcome Y responds to X as being x_1 through the direct link in both factors of the contrast (y_{x_1}), which means that no direct influence from X to Y

[3] For further discussion on counterfactuals, see (Pearl, 2000, Sec. 7.2) and (Bareinboim *et al.*, 2022).

is "active". On the other hand, W responds to X when varying from levels $X = x_1$ to x_0, formally written as W_{x_1} versus W_{x_0}; this, in turn, affects Y, which formally is written as counterfactuals $y_{x_1,W_{x_1}}$ versus $y_{x_1,W_{x_0}}$. [4] The NIE is also a nested counterfactual. For the COMPAS example, the NIE would measure how much the predicted probability of recidivism would change for an individual whose race was set to White, had their race been Non-White along the indirect causal pathway influencing the values of juvenile and prior offense counts, compared against an individual whose race was set to White.

Syntactically, and following the discussion in Sec. 3.2, we can put these observations together and write the NDE and NIE as counterfactual contrasts (Eq. 3.24), namely: [5]

$$\text{NDE}_{x_0,x_1}(y) = \mathcal{C}(x_0, \{x_1, W_{x_0}\}, \emptyset, \emptyset) \tag{4.11}$$

$$\text{NIE}_{x_1,x_0}(y) = \mathcal{C}(x_1, \{x_1, W_{x_0}\}, \emptyset, \emptyset). \tag{4.12}$$

The notions of NDE and NIE, together with Exp-SE, in fact provide the first solution to the FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$), as shown in the next result.

Theorem 4.2 (FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) Solution (preliminary))**.** The total variation measure can be decomposed as

$$\text{TV}_{x_0,x_1}(y) = \text{NDE}_{x_0,x_1}(y) - \text{NIE}_{x_1,x_0}(y) + (\text{Exp-SE}_{x_1}(y) - \text{Exp-SE}_{x_0}(y)). \tag{4.13}$$

Furthermore, the measures NDE, NIE, and Exp-SE are admissible with respect to Str-DE, Str-IE, and Str-SE, respectively. We write

$$\text{Str-DE-fair} \implies \text{NDE-fair} \tag{4.14}$$

[4] The first term $y_{x_1,W_{x_1}}$ is equivalently written as y_{x_1}, which follows from the consistency axiom (Pearl, 2000, Sec. 7.2).

[5] Following prior discussion and reversing the usual simplification back, based on the application of the consistency axiom, these contrasts can more explicitly be written as:

$$\text{NDE}_{x_0,x_1}(y) = \mathcal{C}(\{x_0, W_{x_0}\}, \{x_1, W_{x_0}\}, \emptyset, \emptyset) \tag{4.9}$$

$$\text{NIE}_{x_1,x_0}(y) = \mathcal{C}(\{x_1, W_{x_1}\}, \{x_1, W_{x_0}\}, \emptyset, \emptyset). \tag{4.10}$$

It's evident when considering the NDE that the variations through the mediator W, W_{x_0}, coincide in both sides of the contrast and end up canceling out, which means that all remaining variations are due to the direct change of X from x_0 to x_1 in the first component of the pair. On the other hand, the direct variations in the NIE are both equal to $X = x_1$, which cancel out, and Y changes are in response to the change in W, which varies differently depending on whether $X = x_1$ and $X = x_0$, or W_{x_1} versus W_{x_0}.

$$\text{Str-IE-fair} \implies \text{NIE-fair} \tag{4.15}$$
$$\text{Str-SE-fair} \implies \text{Exp-SE-fair.} \tag{4.16}$$

Therefore, the measures

$$(\mu_{DE}, \mu_{IE}, \mu_{SE}) = (\text{NDE}_{x_0,x_1}(y), \text{NIE}_{x_1,x_0}(y), \text{Exp-SE}_x(y))$$

solve the FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$).

After showing a solution to FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$), we make two important remarks. Firstly, the measures discussed so far admit a structural basis expansion (Thm. 3.1) and can be expanded as follows:

$$\text{TV}_{x_0,x_1}(y) = \sum_u y(u)\big[P(u|x_1) - P(u \mid x_0)\big] \tag{4.17}$$
$$\text{TE}_{x_0,x_1}(y) = \sum_u \big[y_{x_1}(u) - y_{x_0}(u)\big]P(u) \tag{4.18}$$
$$\text{Exp-SE}_x(y) = \sum_u y_x(u)\big[P(u \mid x) - P(u)\big] \tag{4.19}$$
$$\text{NDE}_{x_0,x_1}(y) = \sum_u \big[y_{x_1,W_{x_0}}(u) - y_{x_0}(u)\big]P(u) \tag{4.20}$$
$$\text{NIE}_{x_1,x_0}(y) = \sum_u \big[y_{x_1,W_{x_0}}(u) - y_{x_1}(u)\big]P(u). \tag{4.21}$$

The factorization in the display above connects the measures to the sampling-evaluation process discussed in Sec. 3.2, explaining the observed contrasts in terms of unit-level quantities. We revisit this point shortly. Secondly, one of the significant and practical implications of Thm. 4.2 appears through the Eq. 4.14's contrapositive (and Eqs. 4.15, 4.16), i.e.:

$$(\text{NDE}_{x_0,x_1}(y) \neq 0) \implies \neg\text{Str-DE-fair.} \tag{4.22}$$

Based on this, we have now a principled way of testing the following hypothesis:

$$H_0 : \text{NDE}_{x_0,x_1}(y) = 0. \tag{4.23}$$

If the H_0 hypothesis is rejected, the fairness analyst can conclude that the dataset provides evidence of direct discrimination under the assumptions encoded in the causal diagram. In contrast, any statistics or hypothesis test based on the TV are insufficient to test for the existence of a direct effect.

We display in Fig. 4.2 the measures TE, NDE, NIE, and Exp-SE along the population and mechanism axes of the explainability plane (Fig. 3.3). One may be tempted to surmise that the FPCFA is fully solved based on the results discussed so far. This is unfortunately not always the case, as illustrated next.

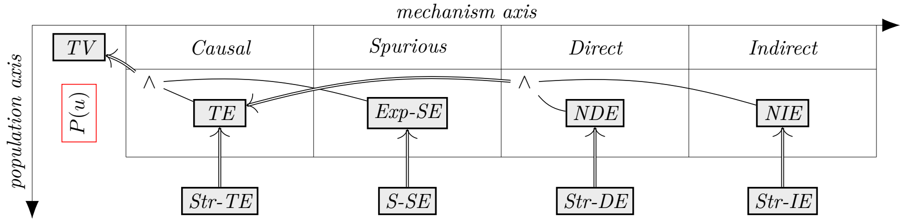


Figure 4.2: Placing the total, experimental spurious, natural direct, and natural indirect effects along the population and mechanism axes that were first introduced in Fig. 3.3.

Example 4.1 (Limitation of the NDE). A startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$ indicates whether the candidate is hired) is based on gender ($X \in \{0,1\}$ represents females and males, respectively), age ($Z \in \{0,1\}$, indicating younger and older applicants, respectively), and education level ($W \in \{0,1\}$ indicating whether the applicant has a PhD). The true SCM $\mathcal{M}$, unknown to the fairness analyst, is given by:

$$U \leftarrow N(0,1) \tag{4.24}$$

$$X \leftarrow \text{Bernoulli}(\mathit{expit}(U)) \tag{4.25}$$

$$Z \leftarrow \text{Bernoulli}(\mathit{expit}(U)) \tag{4.26}$$

$$W \leftarrow \text{Bernoulli}(0.3) \tag{4.27}$$

$$Y \leftarrow \text{Bernoulli}(\frac{1}{5}(X + Z - 2XZ) + \frac{1}{6}W), \tag{4.28}$$

where $\text{expit}(x) = \frac{e^x}{1+e^x}$. In this case, the NDE can be computed as:

$$\begin{aligned}
\text{NDE}_{x_0,x_1}(y) &= P(y_{x_1,W_{x_0}}) - P(y_{x_0}) && (4.29)\\
&= P(\text{Bernoulli}(\frac{1}{5}(1-Z) + \frac{1}{6}W) = 1) && (4.30)\\
&\quad - P(\text{Bernoulli}(\frac{1}{5}(Z) + \frac{1}{6}W) = 1)\\
&= \sum_{z\in\{0,1\}} \sum_{w\in\{0,1\}} P(z,w)[\frac{1}{5}(1-z) + \frac{1}{6}w - \frac{1}{5}z - \frac{1}{6}w] && (4.31)\\
&= \sum_{z\in\{0,1\}} \sum_{w\in\{0,1\}} P(z)P(w)[\frac{1}{5}(1-2z)] \quad \text{as } P(z,w) = P(z)P(w)\\
& && (4.32)\\
&= \sum_{z\in\{0,1\}} P(z)[\frac{1}{5}(1-2z)] = \frac{1}{2}\times\frac{1}{5} + \frac{1}{2}\times\frac{-1}{5} = 0. && (4.33)
\end{aligned}$$

In other words, the $\text{NDE}_{x_0,x_1}(y)$ is equal to zero. Still, perhaps surprisingly, the structural direct effect is present in this case, that is Str-DE-fair does not hold, since the outcome Y is a function of gender X, as evident from the structural Eq. 4.28. □

This example illustrates that even though the NDE is admissible with respect to structural direct effect, it may still be equal to 0 while structural direct effect exists. One can see through Eq. 4.33 that the NDE is an aggregate measure over two distinct sub-populations. Specifically, when considering junior applicants, females are 20% less likely to be hired (units with $(Z = 0, X = 0)$), whereas for senior applicants, males are 20% less likely to be hired (units with $(Z = 1, X = 1)$). Mixing these two groups together results in the cancellation of the two effects and the NDE equating to 0, in turn making it impossible for the analyst to detect discrimination using only the NDE. [6]

Another interesting way of understanding this phenomenon is through the structural basis expansion of the NDE. In Eq. 4.20, the posterior weighting term is $P(u)$, which means that both younger and older applicants are included in the contrast. The fact that this contrast mixes somewhat heterogeneous units of the population, with respect to the decision-making procedure f_y that determines Y, motivates another important notion in fairness analysis:

Definition 4.3 (Power). Let Ω be a space of SCMs. Let Q be a structural criterion and μ_1, μ_2 fairness measures defined on Ω. Suppose that μ_1, μ_2 are (Q, Ω)-admissible. We say that μ_2 is more powerful than μ_1 if

$$\forall \mathcal{M} \in \Omega : \mu_2(\mathcal{M}) = 0 \implies \mu_1(\mathcal{M}) = 0. \tag{4.34}$$

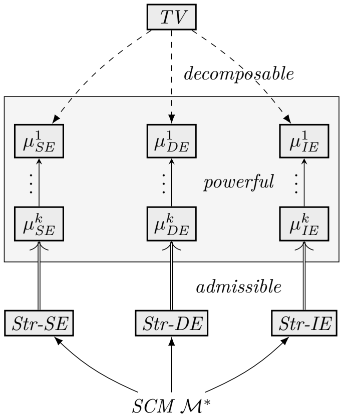


Figure 4.3: FPCFA with power.

The notion of power can be useful in the following context. Suppose there is an SCM $\mathcal{M}$ in the space Ω for which discrimination is present, $Q(\mathcal{M}) = 1$, while the measure μ_1 is admissible but unable to capture it, i.e., $\mu_1(\mathcal{M}) = 0$. Still, another measure may exist such that $\mu_2(\mathcal{M}) \neq 0$. If this is the case, we would say that discrimination qualitatively described by criterion Q can be detected using measure μ_2, but not using μ_1. We would then say that μ_2 is *more powerful* than μ_1. Putting it differently, what Ex. 4.1 showed was that the measure

$$\text{NDE}_{x_0,x_1}(y) = \mathcal{C}(x_0, \{x_1, W_{x_0}\}, \emptyset, \emptyset) \tag{4.35}$$

[6] This observation is structural, and despite of the number of samples available.

was not powerful enough. The reason is that for the NDE, the conditioning events are $E_0 = E_1 = \emptyset$, which is not refined enough to capture the discrimination in the aforementioned example. We next re-write the definition of FPCFA to account for the measures' power:

Definition 4.4 (FPCFA – continued with Power)**.** The Fundamental Problem of Causal Fairness Analysis is to find a collection of measures $\mu_1, \dots, \mu_k$ such that the following properties are satisfied:

(1) μ is decomposable w.r.t. $\mu_1, \dots, \mu_k$;

(2) $\mu_1, \dots, \mu_k$ are admissible w.r.t. the structural fairness criteria $Q_1, \dots, Q_k$.

(3) $\mu_1, \dots, \mu_k$ are as powerful as possible.

We provide in Fig. 4.3 an updated, visual representation of the FPCFA that accounts for the power relation across measures. In some sense, picking $(\text{NDE}_{x_0,x_1}(y), \text{NIE}_{x_1,x_0}(y)$, Exp-SE$_x(y))$ as the measures $(\mu^1_{DE}, \mu^1_{IE}, \mu^1_{SE})$ helped to solve the original problem, but the gap between TV and the structural measures is so substantive that certain critical instances were left undetected. In the updated definition, the requirement is to find measures that are as powerful as possible, or in other words, the closest possible to the corresponding structural ones, Str-{DE,IE,SE}. In the sequel, we discuss how to construct increasingly more powerful measures by using more specific events E.

4.1.1 X-specific Contrasts - $P(u \mid x)$

We will quantify the level of discrimination for a specific subgroup of the population for which $X(u) = x$ (for example, females) by considering contrasts with the conditioning event $E = \{X = x\}$. In fact, we are moving inwards in the population axis in Fig. 3.3, following the discussion in Sec. 3.2, and the sub-population we are focusing on is more specific. More formally, this can be seen through the structural basis expansion (Eq. 3.24) and the fact that the posterior after using the new E becomes $P(u \mid X = x)$, which generates a family of x-specific measures:

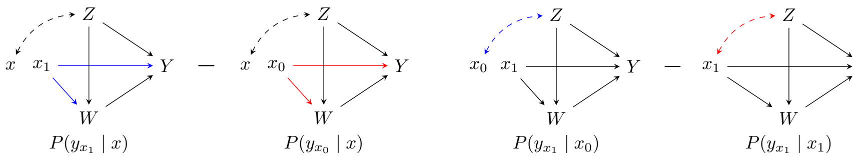
Definition 4.5 (x-specific TE, DE, IE, and SE)**.** The x-{total, direct, indirect, spurious} effects are defined as follows:

$$x\text{-TE}_{x_0,x_1}(y \mid x) = P(y_{x_1} \mid x) - P(y_{x_0} \mid x) \tag{4.36}$$

$$x\text{-DE}_{x_0,x_1}(y \mid x) = P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_0} \mid x) \tag{4.37}$$

$$x\text{-IE}_{x_1,x_0}(y \mid x) = P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_1} \mid x) \tag{4.38}$$

$$x\text{-SE}_{x_1,x_0}(y) = P(y_{x_1} \mid x_0) - P(y_{x_1} \mid x_1). \tag{4.39}$$



(a) $\text{ETT}_{x_0,x_1}(y \mid x)$. **(b)** $\text{Ctf-SE}_{x_1,x_0}(y)$.

Figure 4.4: Graphical representations of some x-specific causal fairness measures. The blue and red color highlight where the contrast between the quantities lies.

The x-TE is a well-known quantity and usually called the effect of treatment on the treated (ETT, for short), and appeared in (Heckman *et al.*, 1998), while the x-specific DE, IE, and SE are more recent quantities, introduced in (Zhang and Bareinboim, 2018b). [7] Some observations ensue from these definitions. Firstly, these measures can be written as their structural basis and unit-level factorization (Eqs. 3.24 and 3.25), that is

$$x\text{-TE}_{x_0,x_1}(y \mid x) = \sum_u [y_{x_1}(u) - y_{x_0}(u)] P(u \mid x) \tag{4.40}$$

$$x\text{-DE}_{x_0,x_1}(y \mid x) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)] P(u \mid x) \tag{4.41}$$

$$x\text{-IE}_{x_1,x_0}(y \mid x) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_1}(u)] P(u \mid x) \tag{4.42}$$

$$x\text{-SE}_{x_1,x_0}(y) = \sum_u y_{x_1}(u) [P(u \mid x_0) - P(u \mid x_1)]. \tag{4.43}$$

To simplify the notation and the comparison with the measures discussed earlier, we re-write them as factual and counterfactual contrasts, namely:

$$x\text{-TE}_{x_0,x_1}(y \mid x) = \mathcal{C}(x_0, x_1, x, x) \tag{4.44}$$

$$x\text{-DE}_{x_0,x_1}(y \mid x) = \mathcal{C}(x_0, \{x_1, W_{x_0}\}, x, x) \tag{4.45}$$

$$x\text{-IE}_{x_1,x_0}(y \mid x) = \mathcal{C}(x_1, \{x_1, W_{x_0}\}, x, x) \tag{4.46}$$

$$x\text{-SE}_{x_1,x_0}(y) = \mathcal{C}(x_1, x_1, x_1, x_0). \tag{4.47}$$

Secondly, we will consider each of the measures individually. Starting with the x-TE, we note that it is simply a conditional version of the total effect (TE) for the subset of units $\mathcal{U}$ for which $X(u) = x$. This can be easily seen by comparing the contrast representation of the TE (Eq. 4.3) versus the x-TE

[7] Zhang and Bareinboim, 2018b originally named these quantities the *counterfactual* DE, IE, and SE, but we highlight here that they are the x-specific counterparts of their marginal effects.

(Eq. 4.44), namely:

$$\begin{aligned} x\text{-TE}_{x_0,x_1}(y \mid x) = &\quad \mathcal{C}(x_0, x_1, \underline{x, x}) \\ \text{TE}_{x_0,x_1}(y) = &\quad \mathcal{C}(x_0, x_1, \underline{\emptyset, \emptyset}), \end{aligned}$$

which make it obvious that the former has $E_0 = E_1 = \emptyset$, whereas the latter has $E_0 = E_1 = x$. Both measures, however, use the same counterfactual clauses $C_0 = x_0$ and $C_1 = x_1$. In terms of the sampling-evaluation process discussed earlier, even though these measures evaluate each unit in the same way (due to the same counterfactual clauses), the TE draws units at random from the population, while the x-TE filters them out based on X's particular instantiation. The graphical visualization of the ETT is shown in Fig. 4.4a and can be compared with that of TE in Fig. 4.1a, for grounding the intuition. In words, note that the downstream effect of X on Y is the same, but now Z is no longer disconnected from X, but varies in accordance to the event $X = x$. As we will show later on, in the startup hiring example (Ex. 4.1), the gender will lead to an additional source of information about age, which can be used in the measure.

Thirdly, the counterfactual measures of direct and indirect effects, x-DE and x-IE, are conditional versions of the NDE and NIE, respectively. These observations are also reflected in Eqs. 4.41-4.42, in which the only difference compared to the general population measures is in the posterior weighting term $P(u \mid x)$, while for the NDE and NIE the weighting term is simply $P(u)$ (Eqs. 4.20-4.21). One difference relative to the natural DE and IE is that here a reference value, $X = x$, needs to be picked such that the baseline population can be selected. For instance, in the context of comparing the direct effect on Y from transitioning X from x_0 to x_1, one could more naturally set the baseline population to $X = x_0$.

Fourthly, we consider the x-SE and its graphical representation, as shown in Fig. 4.4b. This quantity also generalizes that of Exp-SE$_x(y)$ shown in Fig. 4.1b. The difference between these two quantities is in the weighting term, where $P(u) - P(u \mid x)$ in Exp-SE$_x(y)$ is replaced by $P(u \mid x_0) - P(u \mid x_1)$ in x-SE$_{x_1,x_0}(y)$. Despite its innocent appearance, this a substantive difference since the Exp-SE entails a comparison between the observational and interventional distributions, while x-SE is a counterfactual measure. [8]

Following the above, we can finally state the main result of this section, namely, that the quantities x-{DE, IE, SE} solve the FPCFA.

[8] In terms of the Pearl Causal Hierarchy (PCH), the Exp-SE entails assumptions only relative to associational and experimental quantities (PCH's layers 1 and 2), while the x-SE requires substantively stronger assumptions regarding counterfactuals (layer 3). For a more detailed discussion, refer to (Bareinboim *et al.*, 2022).

Theorem 4.3 (x-specific FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) Solution)**.** The total variation measure can be decomposed as

$$\text{TV}_{x_0,x_1}(y) = x\text{-DE}_{x_0,x_1}(y \mid x_0) - x\text{-IE}_{x_1,x_0}(y \mid x_0) - x\text{-SE}_{x_1,x_0}(y). \tag{4.48}$$

Further, the measures x-{DE, IE, SE} are admissible w.r.t. Str-{DE,IE,SE}, respectively. Moreover, the counterfactual family is more powerful than NDE, NIE, and Exp-SE, respectively. More formally, the admissibility relations can be written as:

$$\text{Str-DE-fair} \implies x\text{-DE-fair} \tag{4.49}$$

$$\text{Str-IE-fair} \implies x\text{-IE-fair} \tag{4.50}$$

$$\text{Str-SE-fair} \implies x\text{-SE-fair}, \tag{4.51}$$

and the power relations as:

$$x\text{-DE-fair} \circ\!\!\longrightarrow \text{NDE-fair}, \tag{4.52}$$

$$x\text{-IE-fair} \circ\!\!\longrightarrow \text{NIE-fair}, \tag{4.53}$$

$$x\text{-SE-fair} \circ\!\!\longrightarrow \text{Exp-SE-fair}. \tag{4.54}$$

Therefore, the measures

$$(\mu_{DE}, \mu_{IE}, \mu_{SE}) = (x\text{-DE}_{x_0,x_1}(y), x\text{-IE}_{x_1,x_0}(y), x\text{-SE}_{x_1,x_0}(y))$$

solve the FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$).

Similarly to the discussion in the general-population measures (i.e., $P(u)$), the significance, and practical implications of Thm. 4.3 appear through the Eq. 4.49's contrapositive (and Eqs. 4.50, 4.51), i.e.:

$$(x\text{-DE}_{x_0,x_1}(y) \neq 0) \implies \neg\text{Str-DE-fair}. \tag{4.55}$$

Based on this, we have now a principled way of testing the following hypothesis:

$$H_0 : x\text{-DE}_{x_0,x_1}(y) = 0. \tag{4.56}$$

If the H_0 hypothesis is rejected, the fairness analyst can conclude that the dataset provides evidence of direct discrimination (under the assumptions in the causal diagram). Naturally, similar tests can be performed regarding the indirect and spurious structural measures.

Example 4.2 (Revisiting Startup Hiring & NDE Lack of Power)**.** Consider the SCM $\mathcal{M}$ given in Eq. 4.24-4.28. For $X = x_0$ we compute the x-DE as:

$$\begin{aligned} x\text{-DE}_{x_0,x_1}(y \mid x_0) &= P(y_{x_1,W_{x_0}} \mid x_0) - P(y_{x_0} \mid x_0) && (4.57)\\ &= P(\text{Bernoulli}(\frac{1}{5}(1-Z) + \frac{1}{6}W) = 1 \mid x_0) && (4.58)\end{aligned}$$

$$- P(\text{Bernoulli}(\frac{1}{5}(Z) + \frac{1}{6}W) = 1 \mid x_0) \tag{4.59}$$

$$= \sum_{z \in \{0,1\}} \sum_{w \in \{0,1\}} P(w)P(z \mid x_0)[\frac{1}{5}(1-2z) + \frac{1}{6}w - \frac{1}{6}w] \tag{4.60}$$

$$= \sum_{z \in \{0,1\}} \frac{1}{5}(1-2z)P(z \mid x_0) = 0.036. \tag{4.61}$$

In words, when considering female applicants ($X = x_0$), they are 3.6% less likely to be hired than they would be, had they been male. In other words, direct discrimination is present in the company's hiring process. □

4.1.2 Z-specific Contrasts - $P(u \mid z)$

One might also be interested in capturing discrimination for a specific subset of $\mathcal{U}$ for which $Z(u) = z$, similarly as for the x-specific measures. Here, we will consider two possibilities in terms of sub-population selection, first when event $Z(u) = z$ and then when $Z(u) = z, X(u) = x$. Before introducing the corresponding z- and (x,z)-specific quantities, we clarify one major difference compared to the general and x-specific case, namely in the spurious effects. As noted in Sec. 3, spurious effects are captured by factual contrasts of the form

$$P(y_x \mid E_1) - P(y_x \mid E_0) = \sum_u y_x(u)[P(u \mid E_1) - P(u \mid E_0)], \tag{4.62}$$

which rely on comparing different units corresponding to events E_0, E_1. These spurious effects represent variations that causally precede X and Y. Interestingly enough, under the assumptions of the SFM (Sec. 2.3.1), conditioning on $Z(u) = z$ closes all backdoor paths between X and Y. In other words, fixing Z also fixes the possible spurious variations, and therefore on a z- or (x,z)-specific level spurious effects are always equal to zero[9]. Therefore, we can consider the following measures:

Definition 4.6 (z- and (x,z)-specific TE, DE, and IE). The z-specific and (x,z)-specific total, direct and indirect effects are defined as

$$z\text{-TE}_{x_0,x_1}(y \mid z) = P(y_{x_1} \mid z) - P(y_{x_0} \mid z) \tag{4.65}$$

[9] Experienced readers might notice that in the presence of unobserved confounders (UCs) we could have more explicitly defined the corresponding z-, (x,z)-specific notions

$$z\text{-SE}_x(y) = P(y \mid x, z) - P(y_x \mid z), \tag{4.63}$$

$$(x,z)\text{-SE}_{x_0,x_1}(y) = P(y_x \mid x_1, z) - P(y_x \mid x_0, z). \tag{4.64}$$

Naturally, this would account for the spurious variations brought about by the UCs. For a more comprehensive treatment of these issues, we refer readers to Sec. 6.

$$z\text{-DE}_{x_0,x_1}(y \mid z) = P(y_{x_1,W_{x_0}} \mid z) - P(y_{x_0} \mid z) \tag{4.66}$$

$$z\text{-IE}_{x_1,x_0}(y \mid z) = P(y_{x_1,W_{x_0}} \mid z) - P(y_{x_1} \mid z) \tag{4.67}$$

$$(x,z)\text{-TE}_{x_0,x_1}(y \mid z) = P(y_{x_1} \mid x, z) - P(y_{x_0} \mid x, z) \tag{4.68}$$

$$(x,z)\text{-DE}_{x_0,x_1}(y \mid z) = P(y_{x_1,W_{x_0}} \mid x, z) - P(y_{x_0} \mid x, z) \tag{4.69}$$

$$(x,z)\text{-IE}_{x_1,x_0}(y \mid z) = P(y_{x_1,W_{x_0}} \mid x, z) - P(y_{x_1} \mid x, z). \tag{4.70}$$

As before, the measures can be factorized using the corresponding unit-level outcomes:

$$z\text{-TE}_{x_0,x_1}(y \mid z) = \sum_u [y_{x_1}(u) - y_{x_0}(u)]P(u \mid z) \tag{4.71}$$

$$z\text{-DE}_{x_0,x_1}(y \mid z) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)]P(u \mid z) \tag{4.72}$$

$$z\text{-IE}_{x_1,x_0}(y \mid z) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_1}(u)]P(u \mid z) \tag{4.73}$$

$$(x,z)\text{-TE}_{x_0,x_1}(y \mid z) = \sum_u [y_{x_1}(u) - y_{x_0}(u)]P(u \mid x, z) \tag{4.74}$$

$$(x,z)\text{-DE}_{x_0,x_1}(y \mid z) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)]P(u \mid x, z) \tag{4.75}$$

$$(x,z)\text{-IE}_{x_1,x_0}(y \mid z) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_1}(u)]P(u \mid x, z). \tag{4.76}$$

These quantities can also be represented more explicitly as contrasts:

$$z\text{-TE}_{x_0,x_1}(y \mid z) = \mathcal{C}(x_0, x_1, z, z) \tag{4.77}$$

$$z\text{-DE}_{x_0,x_1}(y \mid z) = \mathcal{C}(x_0, \{x_1, W_{x_0}\}, z, z) \tag{4.78}$$

$$z\text{-IE}_{x_1,x_0}(y \mid z) = \mathcal{C}(x_1, \{x_1, W_{x_0}\}, z, z) \tag{4.79}$$

$$(x,z)\text{-TE}_{x_0,x_1}(y \mid z) = \mathcal{C}(x_0, x_1, \{x, z\}, \{x, z\}) \tag{4.80}$$

$$(x,z)\text{-DE}_{x_0,x_1}(y \mid x, z) = \mathcal{C}(x_0, \{x_1, W_{x_0}\}, \{x, z\}, \{x, z\}) \tag{4.81}$$

$$(x,z)\text{-IE}_{x_1,x_0}(y \mid x, z) = \mathcal{C}(x_1, \{x_1, W_{x_0}\}, \{x, z\}, \{x, z\}). \tag{4.82}$$

The z-TE, z-DE, and z-IE (and similarly the (x, z)- counterparts) are simply conditional versions of TE, NDE, and NIE, respectively, restricted to the subpopulation of $\mathcal{U}$ such that $Z(u) = z$ (or $Z(u) = z, X(u) = x$), which is reflected in the posterior weighting term which becomes $P(u \mid z)$ (or $P(u \mid x, z)$).

Several important remarks are due. Using the sampling of units analogy from before, we notice that z-specific effects filter on units which have $Z(u) = z$, which means they provide us with a more refined lens for detecting discrimination than the general population measures. Similarly, the (x, z)-specific measures can be seen as additionally filtering the units on $Z(u) = z$, after they were filtered based on $X(u) = x$, which is precisely what x-specific measures have done. Therefore, (x, z)-specific measures can be seen as more refined than x- and z- specific ones. The only uncertainty left in terms of power is about comparing x-specific and z-specific measures.

Interestingly, under the SFM, the (x, z)-specific measures are equal to the z-specific measures. This result cannot be deduced from the structural basis expansions above (Eq. 4.72-4.76), but requires the assumptions encoded in the SFM (namely the absence of backdoor paths from X to Y conditional on Z). This equivalence of z- and (x, z)-specific measures under the SFM shows that z-specific measures are in fact more powerful than the x-specific ones, although this need not be the case in general. Following this discussion, we are ready to present the main result regarding the measures introduced above (while, as discussed, for spurious effects we rely on general and x-specific notions):

Theorem 4.4 (z-specific FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) Solution)**.** The total variation measure can be decomposed as

$$\begin{aligned}
\text{TV}_{x_0,x_1}(y) &= \sum_z z\text{-DE}_{x_0,x_1}(y \mid z)P(z) - \sum_z z\text{-IE}_{x_1,x_0}(y \mid z)P(z) \\
&\quad - (\text{Exp-SE}_{x_0}(y) - \text{Exp-SE}_{x_1}(y)) && (4.83) \\
&= \sum_z (x, z)\text{-DE}_{x_0,x_1}(y \mid x, z)P(z \mid x) && (4.84) \\
&\quad - \sum_z (x, z)\text{-IE}_{x_1,x_0}(y \mid x, z)P(z \mid x) \quad - x\text{-SE}_{x_1,x_0}(y).
\end{aligned}$$

Further, the measures z-DE and (x, z)-DE are admissible w.r.t. Str-DE, whereas z-IE and (x, z)-IE are admissible w.r.t. Str-IE. Moreover, the following power relations hold:

$$(x, z)\text{-DE-fair} \;\circ\!\!\longrightarrow\; z\text{-DE-fair} \;\circ\!\!\longrightarrow\; \text{NDE-fair}, \tag{4.85}$$

$$(x, z)\text{-IE-fair} \;\circ\!\!\longrightarrow\; z\text{-IE-fair} \;\circ\!\!\longrightarrow\; \text{NIE-fair}, \tag{4.86}$$

and also

$$(x, z)\text{-DE-fair} \;\circ\!\!\longrightarrow\; x\text{-DE-fair}, \tag{4.87}$$

$$(x, z)\text{-IE-fair} \;\circ\!\!\longrightarrow\; x\text{-IE-fair}. \tag{4.88}$$

Additionally, under the SFM, we can say that:

$$z\text{-DE-fair} \;\circ\!\!\longrightarrow\; x\text{-DE-fair}, \tag{4.89}$$

$$z\text{-IE-fair} \;\circ\!\!\longrightarrow\; x\text{-IE-fair}. \tag{4.90}$$

Therefore, under the SFM, the measures

$$(\mu_{DE}, \mu_{IE}, \mu_{SE}) = (z\text{-DE}_{x_0,x_1}(y), z\text{-IE}_{x_1,x_0}(y), x\text{-SE}_{x_1,x_0}(y))$$

give a more powerful solution to FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) than the x-specific ones.

With z-specific measures in hand, we revisit Ex. 4.1, which showed that the NDE can equal 0 even though direct discrimination exists:

Example 4.3 (Revisiting Startup Hiring & NDE Lack of Power)**.** Consider the SCM $\mathcal{M}$ given in Eq. 4.24-4.28. For $Z = 0$ we compute the z-specific direct effects as:

$$z\text{-DE}(y \mid Z = 0) = P(y_{x_1, W_{x_0}} \mid Z = 0) - P(y_{x_0} \mid Z = 0) \tag{4.91}$$

$$= P(\text{Bernoulli}(\frac{1}{5}(1 - Z) + \frac{1}{6}W) = 1 \mid Z = 0) \tag{4.92}$$

$$- P(\text{Bernoulli}(\frac{1}{5}(Z) + \frac{1}{6}W) = 1 \mid Z = 0)$$

$$= \sum_{w \in \{0,1\}} P(w)[\frac{1}{5} + \frac{1}{6}w - \frac{1}{6}w] = \frac{1}{5}. \tag{4.93}$$

In words, when considering younger applicants ($Z = 0$), females are 20% less likely to be hired than their male counterparts. □

Interestingly, note that the z-specific DE is able to detect discrimination in the above example, and finds an even larger disparity transmitted through the direct mechanism compared to the x-specific DE measure in Ex. 4.2.

4.1.3 More informative contrasts ($V' \subseteq V$-specific).

In case even more detailed measures of fairness are needed, we can consider specific subsets of the observed variables, $V' \subseteq V$. For example, we might be interested in quantifying discrimination for specific units u that correspond to $Z(u) = z, W(u) = w$ (for example quantifying discrimination for a specific age group with a specific level of education). Other choices of V' than $\{Z, W\}$ are possible, but due to a large number of possibilities, we do not cover all of them here. Instead, we define generic v'-specific measures for an arbitrary choice of v':

Definition 4.7 ($V' \subseteq V$-specific TE, DE and IE)**.** Let $V' \subseteq V$ be a subset of the observables V. For any fixed value of $V' = v'$, we define the v'-specific total, direct, and indirect effects as:

$$v'\text{-TE}_{x_0,x_1}(y \mid v') = P(y_{x_1} \mid v') - P(y_{x_0} \mid v') \tag{4.94}$$

$$v'\text{-DE}_{x_0,x_1}(y \mid v') = P(y_{x_1, W_{x_0}} \mid v') - P(y_{x_0} \mid v') \tag{4.95}$$

$$v'\text{-IE}_{x_1,x_0}(y \mid v') = P(y_{x_1, W_{x_0}} \mid v') - P(y_{x_1} \mid v'). \tag{4.96}$$

Once more, these measures admit a structural basis expansion and can be written as contrasts:

$$v'\text{-DE}_{x_0,x_1}(y \mid v') = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)]P(u \mid v') \tag{4.97}$$

$$= \mathcal{C}(x_0, \{x_1, W_{x_0}\}, v', v') \tag{4.98}$$

$$v'\text{-IE}_{x_1,x_0}(y \mid v') = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_1}(u)]P(u \mid v') \tag{4.99}$$

$$= \mathcal{C}(x_1, \{x_1, W_{x_0}\}, v', v'). \tag{4.100}$$

Similarly as in the z-specific case, the notion of a spurious effect is lacking whenever $Z \subseteq V'$, so once again we rely on previously developed notions of spurious effects. Importantly, the v'-specific measures give an even stronger solution to FPCFA than the z- or (x, z)-specific measures:

Theorem 4.5 (v'-specific FPCFA(Str-{DE,IE,SE}, TV$_{x_0,x_1}(y)$) Solution)**.** Suppose $V' \subseteq V$ is a subset of the observables that contains both X and Z. The total variation measure can be decomposed as

$$\begin{aligned}\text{TV}_{x_0,x_1}(y) = &\sum_{v'} v'\text{-DE}_{x_0,x_1}(y \mid v')P(v' \mid x) \\ &- \sum_{v'} v'\text{-IE}_{x_1,x_0}(y \mid v')P(v' \mid x) - x\text{-SE}_{x_1,x_0}(y).\end{aligned} \tag{4.101}$$

Further, the measures v'-{DE, IE} are admissible w.r.t. Str-DE, Str-IE, respectively. Moreover, the v'-specific family is more powerful than the (x, z)-specific, namely:

$$v'\text{-DE-fair} \;\circ\!\!\longrightarrow\; (x, z)\text{-DE-fair}, \tag{4.102}$$

$$v'\text{-IE-fair} \;\circ\!\!\longrightarrow\; (x, z)\text{-IE-fair}. \tag{4.103}$$

Therefore, the measures

$$(\mu_{DE}, \mu_{IE}, \mu_{SE}) = (v'\text{-DE}_{x_0,x_1}(y), v'\text{-IE}_{x_1,x_0}(y), x\text{-SE}_{x_1,x_0}(y))$$

give a more powerful solution to FPCFA(Str-{DE,IE,SE}, TV$_{x_0,x_1}(y)$) than the z- or (x, z)-specific ones.

The next example illustrates why having more flexible, v'-specific measures can be informative, and therefore useful in some practical settings.

Example 4.4 (Startup Hiring – Version II)**.** A startup company is hiring employees. Let $X \in \{x_0, x_1\}$ denote female and male applicants respectively. The employment decision $Y \in \{0, 1\}$ is based on gender and education level W. The SCM $\mathcal{M}$ is given by:

$$X \leftarrow \text{Bernoulli}(0.5) \tag{4.104}$$

$$W \leftarrow \mathcal{N}(14, 4) \tag{4.105}$$

$$Y \leftarrow \text{Bernoulli}\big(0.1 + \frac{W}{50} + 0.1 \cdot X \cdot \mathbb{1}(W < 20)\big). \tag{4.106}$$

Since there are no confounders ($Z = \emptyset$), general, x-specific and z-specific effects are all equal:

$$\text{NDE}_{x_0,x_1}(y) = x\text{-DE}_{x_0,x_1}(y \mid x) = z\text{-DE}_{x_0,x_1}(y \mid z) = 9.2\%. \tag{4.107}$$

Therefore, there is clearly direct discrimination against female employees by the company.

The company argues in the legal proceedings that in the high-tech industry, they are mostly concerned with highly educated individuals. In words, they should be asked whether they discriminate highly educated female applicants, which is represented through the quantity $w\text{-DE}_{x_0,x_1}(y \mid w > 20)$. The quantity in fact equals

$$w\text{-DE}_{x_0,x_1}(y \mid w > 20) = 0\%, \tag{4.108}$$

In words, the company's claim was accurate since highly educated individuals were not discriminated against. □

What the example shows is that v'-specific measures can sometimes capture aspects of discrimination that otherwise cannot be quantified using general, x-specific, or z-specific measures.

Probabilities of causation. Remarkably, the v'-specific measures carry a fundamental connection to what is known in the literature as *probabilities of causation* (Pearl, 2000, Ch. 9). For example, by picking event $v' = \{x_0, y_0\}$, the measure v'-TE becomes

$$(x, y)\text{-TE}_{x_0,x_1}(y \mid x_0, y_0) = P(y_{x_1} \mid x_0, y_0) - P(y_{x_0} \mid x_0, y_0), \tag{4.109}$$

where y is a shortcut to $Y = 1$. First, note that $P(y_{x_0} \mid x_0, y_0) = P(y \mid x_0, y_0)$, since by the consistency axiom $Y = Y_{x_0}$ whenever $X = x_0$. Obviously, $P(y \mid x_0, y_0) = 0$ since $y_0 \neq 1$. Putting these together, the r.h.s. of Eq. 4.109 can be re-written as

$$(x, y)\text{-TE}_{x_0,x_1}(y \mid x_0, y_0) = P(y_{x_1} \mid x_0, y_0), \tag{4.110}$$

which is known as the *probability of sufficiency* (Pearl, 2000, Def. 9.2.2). The measure computes the probability that a change in attribute from $X = x_0$ to $X = x_1$ produces a change in outcome from $Y = y_0$ to $Y = y_1$, or, in words, how likely X's value is to be "sufficient" for producing y_1. Along similar lines, v'-TE for the event $v' = \{x_1, y_1\}$ can be written as

$$\begin{aligned}
(x, y)\text{-TE}_{x_0,x_1}(y \mid x_1, y_1) &= P(y_{x_1} \mid x_1, y_1) - P(y_{x_0} \mid x_1, y_1) && (4.111)\\
&= 1 - P(y_{x_0} \mid x_1, y_1) && (4.112)\\
&= P(y_{x_0} = 0 \mid x_1, y_1), && (4.113)
\end{aligned}$$

which is known as the *probability of necessity* (Pearl, 2000, Def. 9.2.1). The second line of the derivation follows from the consistency axiom, and the fact that $Y = 1$ in the factual world. The measure computes the probability that a change in attribute from $X = x_1$ to $X = x_0$ produces a change in outcome from $Y = y_1$ to $Y = y_0$, or how often X's value is "necessary" for producing y_1. These two types of variations usually appear together and may be modeled through what is known as the probability of necessity and sufficiency (PNS). We refer readers to (Pearl, 2000, Ch. 9) for a further discussion.

4.1.4 Unit-level Contrasts - δ_u

Finally, the most powerful measures to consider are unit-level measures, as defined next:

Definition 4.8 (Unit-level TE, DE, and IE)**.** Given a unit $U = u$, the unit-level total, direct, and indirect effects are given by

$$u\text{-TE}_{x_0,x_1}(y(u)) = y_{x_1}(u) - y_{x_0}(u) = \mathcal{C}(x_0, x_1, u, u) \tag{4.114}$$

$$u\text{-DE}_{x_0,x_1}(y(u)) = y_{x_1,W_{x_0}}(u) - y_{x_0}(u) = \mathcal{C}(x_0, \{x_1, W_{x_0}\}, u, u) \tag{4.115}$$

$$u\text{-IE}_{x_1,x_0}(y(u)) = y_{x_1,W_{x_0}}(u) - y_{x_1}(u) = \mathcal{C}(x_1, \{x_1, W_{x_0}\}, u, u). \tag{4.116}$$

For unit-level measures the posterior distribution that is used as a weighting term is δ_u, where δ is the Dirac delta function. The unit-level measures can be seen as the canonical basis from which all other measures are expanded. They also give the strongest theoretical solution to the FPCFA, once again, with the help of x-specific spurious effect developed earlier:

Theorem 4.6 (Unit-level FPCFA(Str-{DE,IE,SE}, TV$_{x_0,x_1}(y)$) Solution)**.** The total variation measure can be decomposed as

$$\begin{aligned} \text{TV}_{x_0,x_1}(y) = &\sum_u u\text{-DE}_{x_0,x_1}(y(u))P(u \mid x) \\ &- \sum_u u\text{-IE}_{x_1,x_0}(y(u))P(u \mid x) - x\text{-SE}_{x_1,x_0}(y). \end{aligned} \tag{4.117}$$

Further, the measures u-{DE, IE} are admissible w.r.t. Str-DE, Str-IE, respectively. Moreover, the u-specific family is more powerful than the v'-specific, namely:

$$u\text{-DE-fair} \implies v'\text{-DE-fair}, \tag{4.118}$$

$$u\text{-IE-fair} \implies v'\text{-IE-fair}. \tag{4.119}$$

Therefore, the measures

$$(\mu_{DE}, \mu_{IE}, \mu_{SE}) = (u\text{-DE}_{x_0,x_1}(y), u\text{-IE}_{x_1,x_0}(y), x\text{-SE}_{x_1,x_0}(y))$$

give the most powerful solution to FPCFA(Str-{DE,IE,SE}, TV$_{x_0,x_1}(y)$).

The unit-level measures represent the most refined level at which discrimination can be described. In fact, introducing these measures also brings us to the final level of the population axis of the explainability plane (Fig. 3.3). Recall, the population axis ranges from the general population measures (with a posterior $P(u)$), all the way to the deterministic measures which consider a single unit (with a posterior δ_u), eliciting a range of measures which may be useful for fairness analysis. We next move onto giving a systematic overview of the TV-family of measures that was introduced in this section.

4.2 Summary of the TV-family & the Fairness Map

To facilitate comparison and understanding after introducing the measures of the TV-family, we show how they can be more explicitly written as contrasts:

Lemma 4.7 (TV-family as Contrasts)**.** The TV-family of causal fairness measures is a collection of contrasts $\mathcal{C}(C_0, C_1, E_0, E_1)$ (Def. 3.7) that follow the specific instantiations of counterfactual and factual clauses, C_0, C_1, E_0, E_1, as described in Tab. 4.1.

	Measure	C_0	C_1	E_0	E_1
general	TV_{x_0,x_1}	$\emptyset$	$\emptyset$	x_0	x_1
	Exp-SE_x	x	x	$\emptyset$	x
	TE_{x_0,x_1}	x_0	x_1	$\emptyset$	$\emptyset$
	NDE_{x_0,x_1}	x_0	x_1, W_{x_0}	$\emptyset$	$\emptyset$
	NIE_{x_0,x_1}	x_0	x_0, W_{x_1}	$\emptyset$	$\emptyset$
$X = x$	$x\text{-TE}_{x_0,x_1}$	x_0	x_1	x	x
	$x\text{-SE}_{x_0,x_1}$	x_0	x_0	x_0	x_1
	$x\text{-TE}_{x_0,x_1}$	x_0	x_1	x	x
	$x\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	x	x
	$x\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	x	x
$Z = z$	$z\text{-TE}_{x_0,x_1}$	x_0	x_1	z	z
	$z\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	z	z
	$z\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	z	z
$V' \subseteq V$	$v'\text{-TE}_{x_0,x_1}$	x_0	x_1	v'	v'
	$v'\text{-TE}_{x_0,x_1}$	x_0	x_1	v'	v'
	$v'\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	v'	v'
	$v'\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	v'	v'
unit	$u\text{-TE}_{x_0,x_1}$	x_0	x_1	u	u
	$u\text{-TE}_{x_0,x_1}$	x_0	x_1	u	u
	$u\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	u	u
	$u\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	u	u

Table 4.1: Measures of fairness in the TV-family.

A few things are worth noting relative to this taxonomy. First, the measures are grouped into five categories, based on the granularity of the events E_0, E_1. For each of the contrasts, we define a criterion based on the resulting measure. Namely, we say Y is fair with respect to X in the x-TE measure if $x\text{-TE}_{x_0,x_1}(y \mid x) = 0 \; \forall x$. We write $x\text{-TE-fair}_X(Y)$ for this condition, or x-TE-fair, for short.

Further note that Tab. 4.1 has a distinct structure. In particular, the contrasts corresponding to TE, DE, and IE measures have repeating (equal) counterfactual clauses C_0 and C_1, whereas the conditioning event E changes. Contrasts corresponding to the SE measures, as was noted in Thm. 3.1 and in previous sections, are only possible at the population and x-specific level. Mathematically, the measures in the table, but for the spurious effects, can be written more succinctly as

$$\left\{\begin{aligned} &E\text{-TE}_{x_0,x_1}(y \mid E) &&= \mathcal{C}(x_0, x_1, E, E) \\ &E\text{-DE}_{x_0,x_1}(y \mid E) &&= \mathcal{C}(x_0, \{x_1, W_{x_0}\}, E, E) \\ &E\text{-IE}_{x_0,x_1}(y \mid E) &&= \mathcal{C}(x_0, \{x_0, W_{x_1}\}, E, E) \end{aligned}\right\} \text{for } E \in \{\emptyset, x, z, v', u\}. \tag{4.120}$$

Apart from the overarching structure underlying the measures, as described in Tab. 4.1, there is more structure across them as delineated in the next result, which comes under the rubric of the fairness map.

Theorem 4.8 (Fairness Map)**.** The total variation (TV) family of causal measures of fairness admits a number of relations of decomposability, admissibility, and power, which are represented in what we call the Fairness Map, as shown in Fig. 4.5.

In words, the measures of the TV-family satisfy an entire hierarchy of relations in terms of the properties discussed so far, namely, admissibility, decomposability, and power. This hierarchy is one of the main results of this manuscript. There are several observations worth making at this point. First, each arrow in Fig. 4.5 corresponds to an implication, and the full and more syntactic version of the map is provided in the Appendix A.1, including the proofs. There are different ways of reading the map, and perhaps the most natural one is to navigate along the two axes, mechanisms and population, which match the dimensions of the explainability plane discussed earlier (Fig. 3.3/Sec. 3.2).

Navigating the Map. Note that the mechanism axis is partitioned into two categories: *Composite* and *Atomic* measures, as indicated by the vertical line in the map. Atomic measures (direct, indirect, spurious) capture the most refined notions of fairness when working with the Standard Fairness Model (SFM). Composite measures, on the other hand, include measures of total (or causal) effect, and the total variation (TV) measure. Measures of total effect are composite since they capture both direct and indirect variations, whereas the TV measure is composite as it includes direct, indirect, and spurious variations.

In a complementary manner, the population axis can also be divided into two categories: *Structural* versus *Empirical* notions, as indicated by the horizontal line on the map. First, there are the elementary structural fairness criteria (as defined in Def. 3.2), representing idealized, qualitative notions of discrimination that can be directly evaluated using an SCM. Additionally, the unit-level measures, which quantify discrimination for each unit of the population, are the measures of fairness closest to the structural notions.

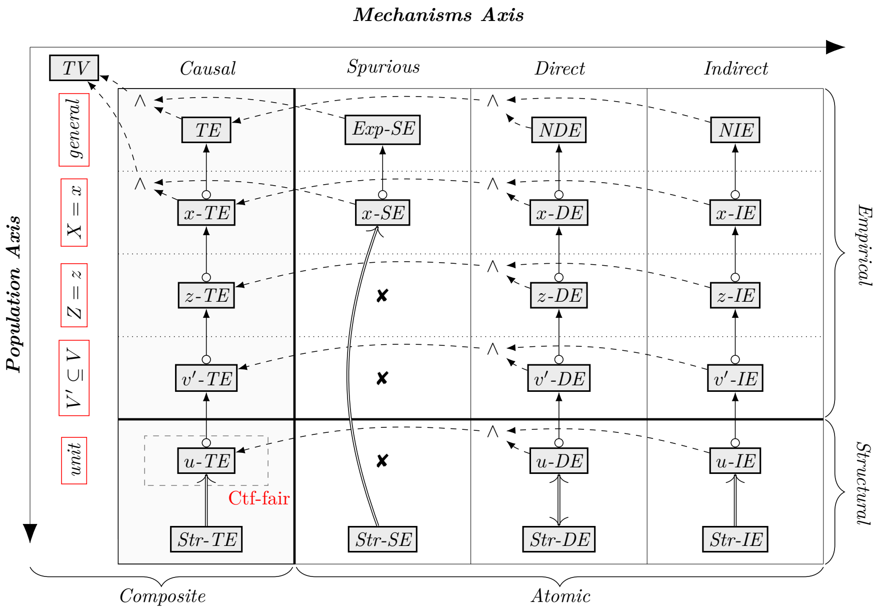


Figure 4.5: Fairness Map for the TV-family of measures. The horizontal axis represent the mechanisms (causal, spurious, direct, and indirect), and the vertical axis the events that capture increasingly more granular sub-populations, from general ($P(u)$) to unit level, and structural. The arrow $\Longrightarrow$ indicates relations of admissibility, $\circ\!\!\longrightarrow$ of power, and $\dashrightarrow$ of decomposability.

While these can be computed directly from the SCM, they are almost never obtainable in practice. Secondly, the empirical measures are positioned above the structural notions. These may be estimated from the available dataset combined with assumptions about the underlying generative processes.

Given this initial structure of the Map, we note this is a preliminary characterization, and then navigate through the axes in a more detailed manner, along each of them separately.

Population Axis (vertical) – Admissibility & Power Relations. When reading the map vertically, from bottom to top, one can find all power and admissibility relations from Thm. 4.2 to Thm. 4.6. For example, the last column of the map (“indirect”) shows that

$$\text{Str-IE} \Longrightarrow u\text{-IE} \circ\!\!\longrightarrow v'\text{-IE} \circ\!\!\longrightarrow z\text{-IE} \circ\!\!\longrightarrow x\text{-IE} \circ\!\!\longrightarrow \text{NIE}. \quad (4.121)$$

In words, this says that:

(i) unit IE is *admissible* w.r.t. structural IE;

(ii) unit IE is *more powerful than* v'-IE, which is *more powerful than* z-IE, which is *more powerful than* x-IE, which is *more powerful* than NIE;

(iii) by transitivity of the admissibility and power relations, it follows that every measure in the column is *admissible* w.r.t. structural IE.

The other columns of the map can be interpreted in a similar fashion.

Mechanisms Axis (horizontal) – Decomposability Relations. When reading the map horizontally, from the right to the left, the decomposability relations are encoded. For example, consider the first row of the map ("general"), it shows that

$$\text{TE} \dashrightarrow \text{NDE} \wedge \text{NIE} \tag{4.122}$$
$$\text{TV} \dashrightarrow \text{TE} \wedge \text{Exp-SE}, \tag{4.123}$$

In words, this says that:

(i) the total variation (TV) can be decomposed into the total (TE) and experimental spurious effects (Exp-SE),

(ii) the total effect (TE) can further be decomposed into natural direct effect (NDE) and natural indirect effect (NIE),

(iii) More explicitly, these relations can be combined and written as:

$$\text{TV} \dashrightarrow \text{NDE} \wedge \text{NIE} \wedge \text{Exp-SE}. \tag{4.124}$$

More strongly, this can be stated for every level of the population axis (i.e., the TE is decomposed into DE and IE at every level), as shown next:

Corollary 4.9 (Extended Mediation Formula)**.** The total effect admits a decomposition into its direct and indirect parts, at every level of granularity of event E in the Fairness Map in Fig. 4.5. Formally, we can say that

$$\text{TE}_{x_0,x_1}(y) = \text{NDE}_{x_0,x_1}(y) - \text{NIE}_{x_1,x_0}(y) \tag{4.125}$$
$$x\text{-TE}_{x_0,x_1}(y \mid x) = x\text{-DE}_{x_0,x_1}(y \mid x) - x\text{-IE}_{x_1,x_0}(y \mid x) \tag{4.126}$$
$$z\text{-TE}_{x_0,x_1}(y \mid z) = z\text{-DE}_{x_0,x_1}(y \mid z) - z\text{-IE}_{x_1,x_0}(y \mid z) \tag{4.127}$$
$$v'\text{-TE}_{x_0,x_1}(y \mid v') = v'\text{-DE}_{x_0,x_1}(y \mid v') - v'\text{-IE}_{x_1,x_0}(y \mid v') \tag{4.128}$$
$$u\text{-TE}_{x_0,x_1}(y(u)) = u\text{-DE}_{x_0,x_1}(y(u)) - u\text{-IE}_{x_1,x_0}(y(u)). \tag{4.129}$$

Furthermore, the TV measure admits different expansions into DE, IE, and SE measures (as shown in Thm. 4.2-4.6). The importance of these decompositions was already stated earlier, as they played a crucial role in solving the decomposability part of the FPCFA.

In summary, the Fairness Map represents a general, theoretical solution to the FPCFA, and shows how the gap between the observed (TV in the top left of the map) and the structural (bottom of the map) can be bridged from first principles. The map therefore, in principle, closes the problem pervasive throughout the literature, as formalized earlier in this manuscript.

4.3 The Identification Problem & the FPCFA in Practice

The Fairness Map introduced in Thm. 4.8 contains various admissible measures w.r.t. to different structural mechanisms. All these measures are well-defined and computable from the underlying data-generating model, the true SCM $\mathcal{M}$. However, $\mathcal{M}$ is not available in practice, which was the very motivation for engaging in the discussions so far, and finding proxies for the structural measures. One key consideration that follows is which of these measures can be computed in practice, given (1) a set of assumptions $\mathcal{A}$ about the underlying $\mathcal{M}$ and (2) data from past decisions generated by $\mathcal{M}$. This question indeed can be seen as a problem of *identifiability* (Pearl, 2000, Sec. 3.2.4). We formalize this notion considering the context of this discussion.

Definition 4.9 (Identifiability). Let $\mathcal{M} = \langle V, U, \mathcal{F}, P(u)\rangle$ be the true, generative SCM, $\mathcal{A}$ a set of assumptions, and $P(v)$ the observational distribution generated by $\mathcal{M}$. Let $\Omega_{\mathcal{A}}$ the space of all SCMs compatible with $\mathcal{A}$. Let ϕ be a query that can be computed from $\mathcal{M}$. The quantity ϕ is said to be identifiable from $\Omega_{\mathcal{A}}$ and the observational distribution $P(V)$ if

$$\forall \mathcal{M}_1, \mathcal{M}_2 \in \Omega_{\mathcal{A}} : \mathcal{A}^{\mathcal{M}_1} = \mathcal{A}^{\mathcal{M}_2} \text{ and} \tag{4.130}$$

$$P^{\mathcal{M}_1}(V) = P^{\mathcal{M}_2}(V) \implies \phi(\mathcal{M}_1) = \phi(\mathcal{M}_2). \tag{4.131}$$

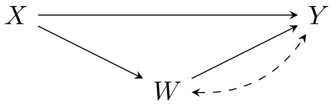
In words, if any two SCMs agree with the set of assumptions ($\mathcal{A}$) and also generate the same observational distribution ($P(v)$), then they should agree with the answer to the query ϕ.

A query ϕ is identifiable if it can be uniquely computed from the combination of qualitative assumptions and empirical data. In fact, the lack of identifiability means that one cannot compute the value of ϕ from the observational data and the set of assumptions, i.e., the gap between the true generative process, $\mathcal{M}$, and the feature that we are trying to obtain from it, ϕ, is too large, and cannot be bridged through the pair $\langle \mathcal{A}, P(v)\rangle$. In practice, one common way of articulating assumptions about $\mathcal{M}$ is through the use of causal diagrams. Whenever the causal diagram is known, we can then write the following:

$$\Omega^{\mathcal{G}} = \{\mathcal{M} : \mathcal{M} \text{ compatible with } \mathcal{G}\}, \tag{4.132}$$

where compatibility is related to sharing the same causal diagram, which encodes qualitative assumptions, following the construction in Def. 2.6[10].

Example 4.5 ((Non-)Identifiability of Measures)**.** Let $\Omega^{\mathcal{G}}$ be the space of SCMs that are compatible with the causal diagram $\mathcal{G}$



.

When considering the quantities $\text{TE}_{x_0,x_1}(y)$ and $\text{NIE}_{x_0,x_1}(y)$ in this context, we can say that:

(i) quantity $\text{TE}_{x_0,x_1}(y)$ is identifiable over $\Omega^{\mathcal{G}}$,

(ii) quantity $\text{NIE}_{x_0,x_1}(y)$ is not identifiable over $\Omega^{\mathcal{G}}$.

In fact, for any SCM in $\Omega^{\mathcal{G}}$, we have that $\text{TE}_{x_0,x_1}(y)$ is equal to

$$P(y \mid x_1) - P(y \mid x_0). \tag{4.133}$$

To show that $\text{NIE}_{x_0,x_1}(y)$ is not identifiable, consider the following two SCMs:

$$\mathcal{M}_1 := \begin{cases} X & \leftarrow U_X & (4.134) \\ W & \leftarrow \mathbb{1}(U_D < 0.2 + 0.4X + 0.4U_{WY}) & (4.135) \\ Y & \leftarrow \mathbb{1}(U_Y < 0.1X + \underline{0.7}W + \underline{0.1}U_{WY}), & (4.136) \end{cases}$$

$$\mathcal{M}_2 := \begin{cases} X & \leftarrow U_X & (4.137) \\ W & \leftarrow \mathbb{1}(U_D < 0.2 + 0.4X + 0.4U_{WY}) & (4.138) \\ Y & \leftarrow \mathbb{1}(U_Y < 0.2X + \underline{0.1}W + \underline{0.7}U_{WY}), & (4.139) \end{cases}$$

where U_X, U_D, U_{WY} and U_Y are independent, exogenous variables, with U_X, U_{WY} binary with $P(U_X = 1) = P(U_{WY} = 1) = \frac{1}{2}$, and U_D, U_Y distributed uniformly Unif$[0, 1]$. Both $\mathcal{M}_1, \mathcal{M}_2$ are compatible with $\mathcal{G}$ and hence are in $\Omega^{\mathcal{G}}$. The reader can verify that the two SCMs generate the same observational distribution. However, computing that

$$\text{NIE}^{\mathcal{M}_1}_{x_0,x_1}(y) = 28\% \neq \text{NIE}^{\mathcal{M}_2}_{x_0,x_1}(y) = 4\% \tag{4.140}$$

shows lack of identifiability in the given context. □

[10] For a more formal account of this notion, see discussion on CBNs in Bareinboim *et al.*, 2022, Sec. 1.3)

Following the discussion in Sec. 2.3, we noted that one SCM $\mathcal{M}$ induces a particular causal diagram $\mathcal{G}$. Still, specifying the precise $\mathcal{G}$ may be non-trivial in practice, and we hence introduced the standard fairness model (SFM). In this case, we will be particularly interested in the set of SCMs defined by the SFM projection of the causal diagram, which is called Ω^{SFM}. Reasoning within the Ω^{SFM} space has two interesting consequences. First, identification is in principle more challenging since this context is generally larger, containing more SCMs than the true $\Omega^{\mathcal{G}}$. Given that more SCMs implies the possibility of finding an alternative SCM that agrees with the assumptions and $P(v)$, and disagrees in the query, identifiability will in general be less frequent. Still, second, since the SFM projection encodes fewer assumptions than the specific causal diagram $\mathcal{G}$, from the fairness analyst's perspective, it will be in general easier to elicit such knowledge to construct a diagram. This situation is more visibly seen through Fig. 4.6.

We now extend the FPCFA to account for the identifiability issues discussed above:

Definition 4.10 (FPCFA continued with Identifiability)**.** [Ω, Q as before] Let $\mathcal{M}$ be the true, unobserved generative SCM, $\mathcal{A}$ a set of assumptions, and $P(v)$ the observational distribution generated by $\mathcal{M}$. Let $\Omega^{\mathcal{A}}$ the space of all SCMs compatible with $\mathcal{A}$. The Fundamental Problem of Causal Fairness Analysis is to find a collection of measures $\mu_1, \ldots, \mu_k$ such that the following properties are satisfied:

(1) μ is decomposable w.r.t. $\mu_1, \ldots, \mu_k$;

(2) $\mu_1, \ldots, \mu_k$ are admissible w.r.t. the structural fairness criteria $Q_1, \ldots, Q_k$.

(3) $\mu_1, \ldots, \mu_k$ are as powerful as possible.

(4) $\mu_1, \ldots, \mu_k$ are identifiable from the observational distribution $P(v)$ and class $\Omega^{\mathcal{A}}$.

The first question we ask is about solving Step (4) of FPCFA when having the full causal graph $\mathcal{G}$. To this end, we state the following theorem:

Theorem 4.10 (Identifiability over $\Omega^{\mathcal{G}}$)**.** Let $\mathcal{G}$ be a causal diagram compatible with the SFM and let $\Omega^{\mathcal{G}}$ be the context defined based on $\mathcal{G}$. Then,

(i) TE, NDE, NIE, and Exp-SE are identifiable,

(ii) x-TE, x-DE, x-IE, and x-SE are identifiable,

(iii) z-TE, z-DE, and z-IE are identifiable,

(iv) if $\{W, Y\} \cap V' \neq \emptyset$, then v'-TE , v'-DE , and v'-IE are not identifiable except in degenerate cases, and excluding the measures (x, w)-DE and (x, z, w)-DE which are identifiable,

(v) u-TE, u-DE, and u-IE are not identifiable except in degenerate cases.

By degenerate cases we refer to instances in which a measure is equal to 0 and identifiable from the absence of pathways (edges) in the graph.

For example, v'-DE or u-DE could be identifiable (and equal to 0) if the causal diagram $\mathcal{G}$ does not contain the arrow $X \to Y$ (this is a case we call degenerate in the above theorem). In summary, we can claim that general, x-specific, and z-specific measures are identifiable over $\Omega^{\mathcal{G}}$ whenever $\mathcal{G}$ is compatible with the SFM. However, v' or unit level measures are in general not identifiable, without additional assumptions.

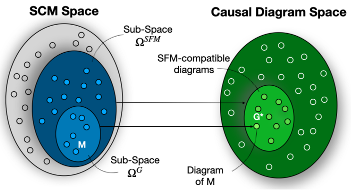


Figure 4.6: Spaces of SCMs (left) and Causal Diagrams (right). (Left) Each point corresponds to a fully instantiated SCM. The SCMs compatible with the diagram $\mathcal{G}$ are shown in light blue, and the ones with the SFM in dark blue. (Right) Each point corresponds to a causal diagram. The lightest green dot corresponds to the true diagram $\mathcal{G}$, while the ones in the light green area correspond to different diagrams compatible with the SFM assumption.

The important next question we ask is whether there is a gap in solving the FPCFA under the context Ω^{SFM} compared to $\Omega^{\mathcal{G}}$. In the first instance, as shown in the following theorem, the answer is negative, showing formally show why our definition of the SFM is indeed sensible in the context of FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$):

Theorem 4.11 (Identifiability over Ω^{SFM} & Soundness of SFM)**.** Under the Standard Fairness Model (SFM) the orientation of edges within possibly multidimensional variable sets Z and W does not change any of general, x-specific or z-specific measures. That is, if two diagrams $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same projection to the Standard Fairness Model, i.e.,

$$\Pi_{\text{SFM}}(\mathcal{G}_1) = \Pi_{\text{SFM}}(\mathcal{G}_2) \tag{4.141}$$

then any measure $\mu(P(v), \mathcal{G})$ will satisfy

$$\mu(P(v), \mathcal{G}_1) = \mu(P(v), \mathcal{G}_2) = \mu(P(v), \mathcal{G}_{\text{SFM}}). \tag{4.142}$$

That is, if measures $\mu_1, \ldots, \mu_k$ in Step (4) of FPCFA in Def. 4.10 are identifiable over the class of SCMs $\Omega^{\mathcal{G}}$ corresponding to a causal diagram $\mathcal{G}$, then they are also identifiable over the class of SCMs Ω^{SFM} corresponding to the diagram's SFM projection $\mathcal{G}_{\text{SFM}}$. The notation $\mu(P(v), \mathcal{G})$ indicates the measures are computed based on the observational distribution $P(v)$ and the causal diagram $\mathcal{G}$ (as opposed to being computed based on the SCM $\mathcal{M}$ as before).

The proofs of Thm. 4.10 and 4.11 are given in Appendix A.2, together with a discussion on relaxing the assumptions of the SFM, and a discussion on the estimation of measures. The theorem shows that the SFM projection of a diagram $\mathcal{G}_{\text{SFM}}$ is equally useful as the fully specified diagram $\mathcal{G}$ for computing any of the general, x-specific or z-specific measures in Lem. 4.7. That is, specifying more precisely the causal structure contained in multivariate nodes Z and W would not change the values of the different measures. The SFM projection $\mathcal{G}_{\text{SFM}}$ can be understood as a coarsening of the equivalence class of SCMs compatible with the graph $\mathcal{G}$. Perhaps surprisingly, this coarsening does not hurt the identifiability of some of the most interesting measures. Moreover, for computing the v'-specific and unit-level measures, additional assumptions would be necessary, even if the full diagram $\mathcal{G}$ was available (see Appendix A.2 for more details). The key observation is that v'-specific measures require the identification of the joint counterfactual distribution $P(v'_{x_0}, v'_{x_1})$, and these two potential outcomes are never observed simultaneously. Therefore, unless we are interested in v'-specific or unit-level measures, we can simply focus on constructing the $\mathcal{G}_{\text{SFM}}$ and not worry about full details of the diagram $\mathcal{G}$. The formulation of FPCFA with identifiability uncovers an interesting interplay of power and identifiability, in which increasingly strong assumptions are needed to identify more powerful measures.

Sensitivity Analyses. Identification results derived in the preceding section are based on the assumptions encoded in the SFM. However, if the lack-of-confounding assumptions of the SFM (encoded in the absence of bidirected edges) are violated, estimating effects based on the derived identification expressions may lead to incorrect results. In such settings, a possible approach is to perform a type of sensitivity analysis, in which we attempt to understand how much the effect estimates would change if unobserved confounding was actually present. For instance, we may be interested in how the estimate of, say, the direct effect would change for varying strengths of unobserved confounding between the attribute X and the outcome Y (this would correspond to a bidirected $X \dashleftarrow\dashrightarrow Y$ edge). This type of approach would allow one to quantify how robust the effect estimates are with respect to violations of the SFM assumptions. There is important previous literature on this topic, but the most common focus is on conditional total effects in a setting with no mediators (i.e., the z-TE quantity in Eq. 4.65 for an SFM with $W = \emptyset$) (Ding and VanderWeele, 2016). Other interesting works focus mainly on the linear setting (Cinelli and Hazlett, 2020; Cinelli *et al.*, 2019). Therefore, adapting the existing methods to the setting of estimating x-specific or population-level direct, indirect, and spurious in a non-parametric fashion represents an important technical challenge that we leave for future work.

4.4 Other relations with the literature

Equipped with the Fairness Map, which was the culmination of understanding the relationship between a multitude of measures, we can now analyze the connection of Causal Fairness Analysis with some influential previous works that articulated other measures in the literature. In particular, we will discuss the criteria of counterfactual fairness (Sec. 4.4.1), individual fairness (Sec. 4.4.2), and predictive parity (Sec. 4.4.3).

4.4.1 Criterion 1. Counterfactual Fairness

One criterion that has received considerable attention in the literature is called *counterfactual fairness* (Kusner *et al.*, 2017). Noteworthy in terms of terminology, the name counterfactual fairness is a misnomer, and may be misleading, as there are various measures that are counterfactual in nature and could be employed to reason about fairness, following the previous discussion and the Fairness Map (Fig. 4.5). In this section, we elaborate on some important limitations of the criterion.

To begin with, the definition of the proposed criterion is somewhat ambiguous in regard to whether it represents a unit-level quantity or a probabilistic-type of counterfactual[11]. To understand the issue, we list in the sequel three possible definitions compatible with the original paper, and then discuss their interpretations:

(i) Counterfactual Fairness – Unit-level ($\mathrm{Ctf}_{\mathrm{fair}}^{(\mathrm{u})}$):

$$y_{x'}(u) - y_x(u) = 0, \ \ \forall x, x', u \in \mathcal{U}. \tag{4.143}$$

(ii) Counterfactual Fairness – Unit-level/probabilistic version ($\mathrm{Ctf}_{\mathrm{fair}}^{(up)}$):

$$P(y_x(u) \mid X = x, W = w) = P(y_{x'}(u) \mid X = x, W = w), \ \ \forall x, x', w. \tag{4.144}$$

(iii) Counterfactual Fairness – Population-level ($\mathrm{Ctf}_{\mathrm{fair}}^{(\mathrm{p})}$):

$$P(y_x \mid X = x, W = w) = P(y_{x'} \mid X = x, W = w), \ \ \forall x, x', w. \tag{4.145}$$

In fact, the paper uses the unit-level probabilistic version ($\mathrm{Ctf}_{\mathrm{fair}}^{(up)}$) as its core definition (Kusner *et al.*, 2017, Def. 5), which is a direct translation to our notation so as to make the context and comparisons more transparent. [12]

[11] For various reasons, probabilistic measures tend to be discussed in the literature.

[12] In particular, the original paper uses A for the protected attribute, where we use X, and it uses X for the remaining attributes where we use W.

The authors "emphasize that counterfactual fairness is an individual-level definition, which is substantially different from comparing different individuals that happen to share the same "treatment" $X = x$ and coincide on the values of $W = w$" (Kusner *et al.*, 2017, Sec. 3). Interestingly, this seems a deliberate choice and suggest a unit-level definition of fairness. Importantly, the probabilistic unit-level ($\text{Ctf}_{\text{fair}}^{(\text{up})}$) and the unit-level definition ($\text{Ctf}_{\text{fair}}^{(\text{u})}$) are equivalent, as shown next:

Proposition 4.2 ($\text{Ctf}_{\text{fair}}^{(\text{up})} \iff \text{Ctf}_{\text{fair}}^{(\text{u})}$)**.** The unit-level counterfactual fairness (Eq. 4.143) and the unit-level/probabilistic counterfactual fairness (Eq. 4.144) criteria are equivalent.

This proposition suggests that the notation used in the original definition of the counterfactual fairness criterion, $\text{Ctf}_{\text{fair}}^{(up)}$, entails some confusion. In words, once the unit $U = u$ is specified, as originally stated in the criterion, $Y_x(u)$ is fully determined. It is therefore redundant, and there is no need for considering or conditioning on event $X = x, W = w$, as this is implied by the choice of the unit u.

However, the authors also state that "the distribution over possible predictions for an individual should remain unchanged in a world where an individual's protected attributes had been different" (Kusner *et al.*, 2017, Sec. 1). As explained above, if the unit $U = u$ is known, there are no probabilities involved, and the statements are deterministic. Therefore, under the alternative description the authors provide, a different formulation of the criterion is needed. In fact, if the goal is to have a probabilistic counterpart of Eq. 4.143, as the above statement might lead one to think, then the unit $U = u$ should be removed altogether, which leads more explicitly to $\text{Ctf}_{\text{fair}}^{(\text{p})}$ definition, as displayed in Eq. 4.145. Interestingly, using structural basis expansion from Thm. 3.1, we can show the relation of the unit- and the probabilistic-level definitions:

Proposition 4.3 ($\text{Ctf}_{\text{fair}}^{(\text{p})}$ is a probabilistic average of $\text{Ctf}_{\text{fair}}^{(\text{u})}$)**.** Consider the following measure:

$$(x, w)\text{-TE}_{x,x'}(y \mid x, w) = P(y_{x'} \mid X = x, W = w) - P(y_x \mid X = x, W = w). \tag{4.146}$$

Then, the $\text{Ctf}_{\text{fair}}^{(\text{p})}$ criterion is equivalent to $(x, w)\text{-TE}_{x,x'}(y \mid x, w) = 0, \quad \forall x, x', w$. Furthermore, the measure underlying the $\text{Ctf}_{\text{fair}}^{(\text{p})}$ criterion can be written as

$$(x, w)\text{-TE}_{x,x'}(y \mid x, w) = \sum_u [y_{x'}(u) - y_x(u)] P(u \mid x, w). \tag{4.147}$$

In words, Prop. 4.3 shows that probabilistic counterfactual fairness criterion takes an average of the unit level differences $y_{x'}(u) - y_x(u)$, weighted by the posterior $P(u \mid x, w)$, and requires the average to be equal to 0. Note the difference between this definition and the unit-level definition, which requires every unit-level difference $y_{x'}(u) - y_x(u)$ to be 0.

After explaining the difference between the two possible and qualitatively different interpretations of counterfactual fairness, and clearing up the notational confusion with respect to fixing a unit $U = u$, we now discuss somewhat more serious issues limitations of the criterion, including from a conceptual, technical, and practical viewpoints. In fact, the issues listed below apply to both the $\text{Ctf}^{(\text{u})}_{\text{fair}}$ and $\text{Ctf}^{(\text{p})}_{\text{fair}}$ interpretations of counterfactual fairness, with the three major points being:

1. inadmissibility of $\text{Ctf}^{(\text{u})}_{\text{fair}}$ and $\text{Ctf}^{(\text{p})}_{\text{fair}}$ with respect to Str-{DE,IE,SE},

2. lack of accounting for spurious effects, and

3. hardness/impossibility of identifiability.

Limitation 1. Inadmissibility w.r.t. Str-{DE,IE,SE}

As formally shown in the following result, the counterfactual fairness measure is inadmissible w.r.t. any of the structural criteria:

Proposition 4.4 (Unit-TE, (x, w)-TE not admissible)**.** The unit-level total effect (unit-$\text{TE}_{x_0,x_1}(y)$) and the (x, w)-specific total effect ((x, w)-$\text{TE}_{x_0,x_1}(y \mid x, w)$) are both not admissible w.r.t. the structural direct, indirect, and spurious criteria. Formally, we write

$$\text{Str-DE-fair} \not\Rightarrow \text{unit-TE-fair}, \; \text{Str-DE-fair} \not\Rightarrow (x, w)\text{-TE-fair} \tag{4.148}$$

$$\text{Str-IE-fair} \not\Rightarrow \text{unit-TE-fair}, \; \text{Str-IE-fair} \not\Rightarrow (x, w)\text{-TE-fair} \tag{4.149}$$

$$\text{Str-SE-fair} \not\Rightarrow \text{unit-TE-fair}, \; \text{Str-SE-fair} \not\Rightarrow (x, w)\text{-TE-fair}. \tag{4.150}$$

The importance of this result stems from the fact that even if one is able to ascertain that

$$y_{x_1}(u) - y_{x_0}(u) = 0 \;\; \forall u, \text{ or}$$
$$P(y_{x_1} \mid X = x, W = w) - P(y_{x_0} \mid X = x, W = w) = 0 \;\; \forall x, w,$$

it could still be that case that neither the direct nor the indirect (nor the spurious) effects are equal to 0. The broader discussion around the Fairness Map, and the idea of decomposability of measures into admissible ones was introduced precisely to avoid such situations. The next example highlights this issue more vividly.

Example 4.6 (Startup Hiring Continued - Salaries). The startup company from Ex. 4.1 has closed the hiring season. In the hiring process, the company achieved demographic parity, which means in this context that 50% of new hires were female. Now, the company needs to decide on each employee's salary. In an attempt of the company to be fair, each employee is evaluated on how well they perform their tasks. The salary Y is then determined based on this information, but, due to a subconscious bias of the executive determining the salaries, gender also affects how salaries are determined. The SCM $\mathcal{M}^*$ corresponding to this process is:

$$\mathcal{F}^*, P^*(U) : \begin{cases} X \leftarrow U_X & (4.151) \\ W \leftarrow -X + U_W & (4.152) \\ Y \leftarrow X + W + U_Y. & (4.153) \\ & \\ U_X \in \{0,1\}, P(U_X = 1) = 0.5, & (4.154) \\ U_W, U_Y \sim N(0,1). & (4.155) \end{cases}$$

For any unit $u = (u_x, u_w, u_y)$, we can compute that

$$y_{x_1}(u) - y_{x_0}(u) = \underbrace{(1 + (-1 + u_w) + u_y)}_{y_{x_1}(u)} - \underbrace{(0 + (-0 + u_w) + u_y)}_{y_{x_0}(u)} = 0, \tag{4.156}$$

showing that unit-level total effect is 0. Furthermore, for each choice of $X = x, W = w$, it is also true that

$$P(y_{x_1} \mid X = x, W = w) - P(y_{x_0} \mid X = x, W = w) = 0. \tag{4.157}$$

Therefore, both interpretations of the counterfactual fairness criterion are satisfied. However, direct discrimination against female employees still exists since the f_y mechanism in Eq. 4.153 assigns a higher salary to male employees. On the other hand, the mechanism f_w in Eq. 4.152 shows that female employees are better at performing their tasks, and should therefore be paid more. Nevertheless, the superior performance of female employees in performing their tasks is canceled out by the direct discrimination favoring males (as witnessed by Eq. 4.156). In effect, they are paid the same as they would be had they been male. □

The inability of total effect to detect direct and indirect effects stems from the fact that the total effect is decomposable (see Cor. 4.9). The example above illustrates the first critical shortcoming of the criterion proposed by Kusner *et al.*, 2017, as in any other composite measure, and any optimization procedure based on it, i.e., zeroing the Ctf_{fair} measure, may lead to unintended side effects and discrimination if implemented in the real world.[13]

[13] A formal result of this form is discussed in Thm. 5.1.

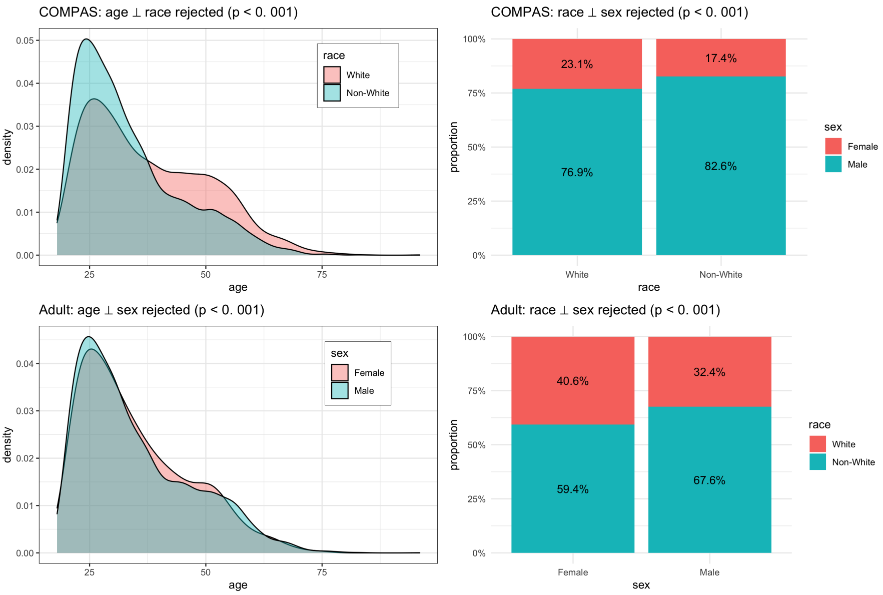
Limitation 2. Ancestral closure & Spurious effects

The purported criterion rules out, by construction, the possibility of existence of any spurious types of variations. In particular, the argument relies on the notion introduced in the paper called *ancestral closure* (AC, for short) w.r.t. the protected attribute set. The AC requires that all protected attributes and their parents, and all their ancestors, should be measured and included in the set of endogenous variables. This is obviously a very stringent requirement, which is hard to ascertain in practice. The paper then argues that "the fault should be at the postulated set of protected attributes rather than with the definition of counterfactual fairness, and that typically we should expect set X to be closed under ancestral relationships given by the causal graph. For instance, if Race is a protected attribute, and Mother's race is a parent of Race, then it should also be in X".

Conceptually speaking, we contrast this constraint over the space of models with the very existence of dashed-bidirected arrows in causal diagrams, as discussed earlier. These arrows in particular allow for the possibility that there are variations between X and Z that can be left unexplained in the model, or unmeasured confounders may exist. Practically speaking, assuming that no bidirected arrows exist is a strong assumption that does not hold in many settings. For instance, consider the widely recognized phenomenon in the fairness literature known as *redlining* (Zenou and Boccard, 2000; Hernandez, 2009). In some practical settings, the location where loan applicants live may correlate with their race. Applications might be rejected based on the zip code, disproportionately affecting certain minority groups in the real world.

It has been reported in the literature that correlation between gender and location, or religion and location, may possibly exist, and should therefore be acknowledged through modeling. For instance, the one-child policy affecting mainly urban areas in China had visible effects in terms of shifting the gender ratio towards males (Hesketh *et al.*, 2005; Ding and Hesketh, 2006). Beyond race or gender, religious segregation is also a recognized phenomenon in some urban areas (Brimicombe, 2007). Again, while we make no claim that location affects race (or religion), or vice-versa, the bidirected arrows give a degree of modeling flexibility that allows for the encoding of such co-variations. Still, this is without making any commitment to whatever historical processes and other complex dynamics took place and generated such imbalance in the first place. To corroborate this point, consider the following example:

Example 4.7 (Spurious associations in COMPAS & Adult datasets). A data scientist is trying to understand the correlation between the features in the COMPAS dataset. The protected attribute X is race, and the demographic variables Z_1, Z_2 are age and sex. The data scientist tests two hypotheses,



COMPAS: age ⊥ race rejected (p < 0. 001)

COMPAS: race ⊥ sex rejected (p < 0. 001)

Adult: age ⊥ sex rejected (p < 0. 001)

Adult: race ⊥ sex rejected (p < 0. 001)

Figure 4.7: Testing for independence of the protected attribute (X) and the confounders (Z) on the Adult and COMPAS datasets.

namely:

$$H_0^{(1)} : X \perp\!\!\!\perp Z_1, \tag{4.158}$$

$$H_0^{(2)} : X \perp\!\!\!\perp Z_2. \tag{4.159}$$

The association of X and Z_1, Z_2 are shown graphically in the bottom row of Fig. 4.7. Both of the hypotheses are rejected (p-values < 0.001). However, possible confounders of this relationship are not measured in the corresponding dataset.

Similarly, the same data scientist is now trying to understand the correlation of the features in the Adult dataset. The protected attribute X is gender, and the demographic variables Z_1, Z_2 are age and race. The data scientist tests the independence of sex and age ($X \perp\!\!\!\perp Z_1$), and sex and race ($X \perp\!\!\!\perp Z_2$), and both hypotheses are rejected (p-values < 0.001, see Fig. 4.7 top row). Again, possible confounders of this relationship are not measured in the corresponding dataset, meaning that the attribute X cannot be separated from the confounders Z_1, Z_2 using any of the observed variables. □

As the example illustrates, from both a conceptual and practical standpoint, disallowing the possibility of non-causal relationships and confounding induced by some historical or societal context, and the associated spurious effects, can be an major limitation to any type of fairness analysis.

Limitation 3. Lack of identifiability

An important practical property of any fairness measure is its identifiability under different sets of causal assumptions. We introduced the notion of identifiability in Sec. 4.3 to better understand when a fairness measure can be used in practice. We then discussed some necessary assumptions for measures in the Fairness Map to be identifiable. A significant implication of this prior discussion in the context of counterfactual fairness is highlighted by the following result:

Proposition 4.5 (Unit-TE, (x, w)-TE not identifiable)**.** Suppose that $\mathcal{M}$ is a Markovian model and that $\mathcal{G}$ is the associated causal diagram. Assume that the set of mediators between X and Y is non-empty, $W \neq \emptyset$. Then, the measures unit-$\text{TE}_{x_0,x_1}(y)$ and (x, w)-$\text{TE}_{x_0,x_1}(y \mid x, w)$ are not identifiable from observational data, even if the fully specified diagram $\mathcal{G}$ is known.

The proposition shows that the measures on which counterfactual fairness is based are never computable from observational data and the causal diagram, even for models in which Markovianity is assumed to hold. The issue with these quantities is that they require knowledge of the joint distribution of counterfactual outcomes Y_{x_1}, Y_{x_0}, which are never observed simultaneously [14].

The issue discussed above obviously curtails the generality of the proposed method, since the underlying measures are not identifiable immediately, as illustrated next.

Example 4.8 (Non-ID of $\text{Ctf}^{(\text{u})}_{\text{fair}}$, $\text{Ctf}^{(\text{p})}_{\text{fair}}$ - Startup Salaries Continued)**.** Consider the SCM $\mathcal{M}^*$ of the Startup Salaries example (Ex. 4.6) given in Eq. 4.151-4.153. In $\mathcal{M}^*$ we showed that

$$(x, w)\text{-TE}_{x_0,x_1}(y \mid x, w) = 0. \tag{4.160}$$

Consider now an alternative SCM $\mathcal{M}'$ given by:

$$\mathcal{F}', P'(U) : \begin{cases} X \leftarrow U_X & (4.161) \\ W \leftarrow -X + (-1)^X U_W & (4.162) \\ Y \leftarrow X + W + U_Y, & (4.163) \end{cases}$$

[14] Such quantities can be identified only under additional, stronger assumptions, such as monotonicity (Tian and Pearl, 2000; Plečko and Meinshausen, 2020).

with the same distribution $P(u)$ over the units as for $\mathcal{M}^*$. It's verifiable that that $\mathcal{M}'$ generates the same observational distribution as $\mathcal{M}^*$ and has the same causal diagram $\mathcal{G}$. However, notice that for $u = (1, u_w, u_y)$, we have

$$u\text{-TE}_{x_0,x_1}(y) = y_{x_1}(u) - y_{x_0}(u) = -2u_w \neq 0 \text{ whenever } u_w \neq 0. \tag{4.164}$$

Furthermore, we have that

$$(x, w)\text{-TE}_{x_0,x_1}(y \mid x, w > 0) \neq 0. \tag{4.165}$$

Therefore, $\mathcal{M}^*$ and $\mathcal{M}'$ generate the same observational distribution and have the same causal diagram, but differ substantially with respect to counterfactual fairness. □

The example constructed above is not atypical, but stems from the general non-identifiability result in Prop. 4.5. These results raise the question as to whether counterfactual fairness criteria – either $\text{Ctf}^{(\text{u})}_{\text{fair}}$ or $\text{Ctf}^{(\text{p})}_{\text{fair}}$ – can be used for the purpose of bias detection in any practical setting. In fact, to circumvent the identifiability issue discussed above, the proposal of the paper is that "the model $\mathcal{M}^*$ must be provided" (Kusner *et al.*, 2017, Sec. 4.2). This means that the fully specified causal model $\mathcal{M}^*$ is needed to assess the existence of discrimination. The assumptions put forward in our manuscript are concerned with constructing the causal diagram $\mathcal{G}$, or the simplified version of the diagram in the form of an SFM. In stark contrast, the assumptions needed to provide the model $\mathcal{M}^*$ are orders of magnitude stronger than those needed for constructing the causal diagram or the SFM. This level of knowledge requires reading the intentions and minds of decision-makers, or having access to the internal systems and strategic secrets of companies, which are usually not accessible to outsiders. On the more mathematical side, as alluded to earlier, inducing such a structural model from observational data alone is almost never possible (Bareinboim *et al.*, 2022, Thm. 1).

4.4.2 Criterion 2. Individual Fairness

In this section, we discuss a prominent measure introduced by Dwork *et al.*, 2012 called *individual fairness* (IF, for short). One of the most natural intuitions behind fairness is that if we constrain the population in a way that the units are the same but for the protected attribute, this would allow us to make claims about the impact of variations of this attribute. In fact, since nothing else remains to explain the observed disparities, the differences in outcome would be attributable to the change in the protected attribute.

To ground this intuition, we introduced in Sec. 3 the explainability plane (Fig. 3.3) spanned by the population and mechanism axes. In terms of the population axis, we noted that as the event $E = e$ is enlarged, the corresponding measure of fairness becomes more and more *individualized.* Formally, the restriction on the observed information translates into a more precise subpopulation of the space of unobservable units $\mathcal{U}$. Our earlier analysis relied on three observations that will be key to comparing other causal measures with the IF measure and trying to understand its causal implications. First, the plane is contingent on the assumptions encoded in the SFM. As we will show formally, assumptions about the underlying causal structure are also relevant in the framework of IF. Secondly, the explainability plane considers the admissibility and power of different measures, and we use these notions to place and understand the IF condition in the context of the Fairness Map. Thirdly, as highlighted by our analysis of the FPCFA, optimizing based on a specific composite criterion may in fact fail to remove bias that could be in principle detected when a more fine-grained analysis of the causal mechanisms generating the disparity is undertaken. We discuss conditions under which the IF framework can be given a causal interpretation, and show the framework is optimizing based on a composite measure. We further provide practical examples in which this may lead to unintended and potentially harmful side effects. We start with the definition of individual fairness:

Definition 4.11 (Individual Fairness)**.** Let d be a fairness metric on $\mathcal{X} \times \mathcal{Z} \times \mathcal{W}$. An outcome Y is said to satisfy individual fairness if

$$|P(y \mid x, z, w) - P(y \mid x', z', w')| \leq d((x, z, w), (x', z', w')), \tag{4.166}$$

$\forall\, x, x', w, w', z, z'$.

The framework of IF assumes the existence of a fairness metric d that computes the distance between two individuals described by attributes (x, z, w) and (x', z', w'), while the outcome y is not taken into account. In words, IF requires that individuals who are similar with respect to metric d need to have a similar outcome. This requirement is represented by a Lipschitz property in Eq. 4.166. If the distance between two values of the covariates, $d((x, z, w), (x', z', w'))$, is smaller than ϵ, then the criterion in Eq. 4.166 implies that individuals who coincide with these covariate values must have a similar probability of a positive outcome, that is

$$|P(y \mid x, z, w) - P(y \mid x', z', w')| \leq \epsilon. \tag{4.167}$$

We now look at the implications of the IF criterion, and observe some possible shortcomings that can result from ignoring the causal structure.

Limitation 1. IF is oblivious to causal structure

The IF definition in Eq. 4.166 is agnostic to the underlying causal structure that generated the data. We start with two examples of a hiring process that are on the surface similar, but differ with respect to the underlying causal

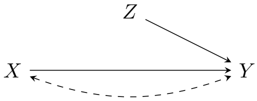
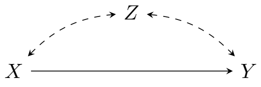
		Example 4.9A	Example 4.9B
SCM $\mathcal{M}$	$\mathcal{F}$	$X \leftarrow U_{XY}$ (4.168) $Z \leftarrow U_Z$ (4.169) $Y \leftarrow X - U_{XY} + Z + U_Y$ (4.170)	$X \leftarrow U_{XZ}$ (4.171) $Z \leftarrow U_{XZ} + U_{ZY}$ (4.172) $Y \leftarrow U_{ZY} + U_Y$ (4.173)
	$P(u)$	$U_{XY} \sim \text{Bernoulli}(0.5)$, $U_Z, U_Y \sim N(0,1)$	$U_{XZ} \sim \text{Bernoulli}(0.5)$, $U_{ZY}, U_Y \sim N(0,1)$
diagram	$\mathcal{G}$		

Table 4.2: An example of two situations in which the IF criterion has different meanings.

structure. As we will see, this will show that the implications of the IF criterion can be quite different based on the causal setting, highlighting the fact that causal structure cannot be dismissed when using this criterion.

Example 4.9 (Startup Hiring III). Suppose that two startup companies, A and B, are hiring employees. Let sex X represent the protected attribute (x_0 female, x_1 male), Z the candidate's performance on an aptitude test, and Y the overall score for job hiring Y. The set of mediators W is in this case empty. The hiring process is similar, yet there is a difference between the two companies. In both instances, we assume age is a latent, unobserved factor, which has shared information with gender. In company A, age affects the salary directly, whereas in company B, age affects the aptitude test result. Additionally, in company B the aptitude test result has shared information with the salary, represented by the unobserved variable which measures how much the candidate prepared for the interview day. The respective SCMs and causal diagrams are shown in Tab. 4.2. Suppose that the fairness metric d in both cases equals

$$d((x,z),(x',z')) = |z - z'|. \tag{4.174}$$

Then, the IF criterion can be written as

$$\big|\mathbb{E}[y \mid x, z] - \mathbb{E}[y \mid x', z']\big| \leq d((x,z),(x',z')) = |z - z'|. \ \forall x, x', z, z'. \tag{4.175}$$

Notice that in company A, we can compute that

$$\begin{aligned}
\mathbb{E}^{\mathcal{M}_A}[y \mid x, z] &= \mathbb{E}^{\mathcal{M}_A}[X - U_{XY} + Z + U_Y \mid x, z] && (4.176)\\
&= \underbrace{\mathbb{E}^{\mathcal{M}_A}[X - U_{XY} \mid x, z]}_{=0 \text{ as } X = U_{XY}} + \mathbb{E}^{\mathcal{M}_A}[Z \mid x, z] + \underbrace{\mathbb{E}^{\mathcal{M}_A}[U_Y \mid x, z]}_{\substack{=0 \text{ as } U_Y \sim N(0,1),\\ U_Y \perp\!\!\!\perp Z, X}} && (4.177)\\
&= z. && (4.178)
\end{aligned}$$

Therefore, we can conclude that

$$\left|\mathbb{E}^{\mathcal{M}_A}[y \mid x_1, z] - \mathbb{E}^{\mathcal{M}_A}[y \mid x_0, z']\right| = |z - z'|. \tag{4.179}$$

In company B, however, we can compute:

$$\begin{aligned}
\mathbb{E}^{\mathcal{M}_B}[y \mid x, z] &= \mathbb{E}^{\mathcal{M}_B}[U_{ZY} + U_Y \mid x, z] && (4.180)\\
&= \mathbb{E}^{\mathcal{M}_B}[Z - U_{XZ} \mid x, z] + \underbrace{\mathbb{E}^{\mathcal{M}_B}[U_Y \mid x, z]}_{\substack{=0 \text{ as } U_Y \sim N(0,1),\\ U_Y \perp\!\!\!\perp Z, X}} && (4.181)\\
&= \mathbb{E}^{\mathcal{M}_B}[Z - X \mid x, z] = z - x. && (4.182)
\end{aligned}$$

Therefore, the IF criterion is not satisfied, which can be shown by computing:

$$\left|\mathbb{E}^{\mathcal{M}_B}[y \mid x_1, z] - \mathbb{E}^{\mathcal{M}_B}[y \mid x_0, z']\right| = |z - 1 - z'|. \tag{4.183}$$

When assessing direct discrimination on a structural level, in company A, the mechanism f_y in Eq. 4.170 shows the presence of direct discrimination. In company B, however, the mechanism f_y in Eq. 4.173 shows no direct discrimination. We could pick a more empirical measure of DE, such as the NDE (Def. 4.2). Evaluating the NDE using the generated data:

$$\begin{aligned}
\text{NDE}^{\mathcal{M}_A}_{x_0,x_1}(y) &= 1, && (4.184)\\
\text{NDE}^{\mathcal{M}_B}_{x_0,x_1}(y) &= 0, && (4.185)
\end{aligned}$$

which is consistent with the observed discrimination at the structural level. □

Somewhat paradoxically, the example illustrates that in company A direct discrimination exists, yet the IF criterion is satisfied, whereas in company B the criterion is not fulfilled, but there is no direct discrimination. This example, even though perhaps surprising at first, is reflective of the fact that IF does not take the causal structure into account. Our conclusion is that without the causal diagram, the consequences of using IF may be unclear. Therefore, from this point forward, we assume the SFM structure, and look at the IF framework in this fixed context.

Limitation 2. IF captures the direct effect only under the SFM

We next show that under the assumptions of the standard fairness model, the IF condition given in Eq. 4.166 has causal implications. In other words, we investigate where the IF condition can be placed in the Fairness Map in Fig. 4.5. An initial difficulty arises from the fact that the IF criterion is not written in the form of a contrastive measure (which were studied in Sec. 3). Therefore, instead of using the exact IF criterion, we look at a criterion that is implied by the IF criterion, but is itself a contrastive measure. This criterion is based on the measure known as the observational direct effect:

Definition 4.12 (Observational direct effect)**.** The observational direct effect (Obs-DE, for short) is defined as

$$\text{Obs-DE}_{x_0,x_1}(y \mid z, w) = P(y \mid x_1, z, w) - P(y \mid x_0, z, w). \tag{4.186}$$

Based on this measure, we define the Obs-DE-fair criterion as:

$$\text{Obs-DE-fair} \iff \text{Obs-DE}_{x_0,x_1}(y \mid z, w) = 0 \;\; \forall z, w. \tag{4.187}$$

The Obs-DE-fair criterion is implied by IF whenever the fairness metric d satisfies

$$d((x_1, z, w), (x_0, z, w)) = 0 \;\; \forall z, w, \tag{4.188}$$

that is, when the metric d does not depend on the protected attribute X. The Obs-DE condition can then be obtained from Eq. 4.166 by setting $(x, z, w) = (x_1, z, w)$ and $(x', z', w') = (x_0, z, w)$. The Obs-DE criterion, which is implied by the IF condition under certain assumptions, is admissible with respect to structural direct criterion:

Proposition 4.6 (Admissibility of Obs-DE w.r.t. Str-DE and IF)**.** Suppose that the metric d does not depend on the X variable, that is

$$d((x, z, w), (x', z', w')) = d((z, w), (z', w')). \tag{4.189}$$

Then, the IF criterion in Eq. 4.166 implies the Obs-DE-fair criterion in Eq. 4.187. Furthermore, under the assumptions of the standard fairness model the Obs-DE measure is admissible with respect to Str-DE, that is

$$\text{Str-DE-fair} \implies \text{Obs-DE-fair}. \tag{4.190}$$

A further positive result shows that the Obs-DE criterion is in fact powerful in the context of detecting direct discrimination (again under suitable assumptions):

Proposition 4.7 (Power of IF w.r.t. Str-DE)**.** Suppose that the Obs-DE-fair criterion in Eq. 4.187 holds. Under the assumptions of the standard fairness model, the Obs-DE measure is more powerful than z-DE, x-DE and NDE:

$$\text{Obs-DE-fair} \;\circ\!\!\longrightarrow\; z\text{-DE-fair} \;\circ\!\!\longrightarrow\; x\text{-DE-fair} \;\circ\!\!\longrightarrow\; \text{NDE-fair}. \tag{4.191}$$

Under the SFM[15] $P(y \mid x_1, z, w) - P(y \mid x_0, z, w)$ equals what is known as the *controlled direct effect*

$$\text{CDE}_{x_0,x_1} := P(y_{x_1,z,w}) - P(y_{x_0,z,w}). \tag{4.192}$$

Therefore, under certain assumptions, the constraint implied by IF in fact precludes the existence of a direct effect and has a valid causal interpretation. Importantly, the assumptions that are needed are of a causal nature, and ignoring the causal diagram of the data generating model can lead to undesired consequences when using the IF condition (see Ex. 4.9).

To continue the discussion, we consider two distinct cases when choosing the fairness metric d, on which much of the IF framework relies:

(i) metric d is sparse, meaning that it does not depend on all variables in the sets Z, W,

(ii) metric d is complete, meaning that it depends on all variables in the sets Z, W.

We now consider these two cases separately, and point out their possible drawbacks. We emphasize that our goal is not to pick a metric but to shed light on the fundamental interplay between the arguments/properties of the fairness metric d and the underlying causal mechanisms.

Limitation 3. Sparse metrics d lead to lack of admissibility

From individual to global. Suppose that the IF condition in Eq. 4.166 holds. Under suitable causal assumptions, the condition precludes the existence of direct discrimination, as was shown above. However, even if the IF condition holds, the disparity between the groups corresponding to $X = x_0$ and $X = x_1$ (measured by the TV) could still be large, if the conditional distributions

$$Z, W \mid X = x_0 \text{ and } Z, W \mid X = x_1$$

differ. This observation leads to the second step of the framework of Dwork *et al.*, 2012. The authors provide the following significant result:

[15] The exact assumption needed here can be written as $Y_{x,z,w} \perp\!\!\!\perp X, Z, W$. This assumption is encoded in the SFM.

Proposition 4.8 (Optimal Transport bound on TV (Dwork *et al.*, 2012))**.** Let d be a fairness metric, and suppose that the individual fairness condition in Eq. 4.166 holds. Let the optimal transport cost between $Z, W \mid X = x_1$ and $Z, W \mid X = x_0$ be denoted by

$$\text{OTC}^{d}_{x_0,x_1}((Z, W)). \tag{4.193}$$

Then, the TV measure between the groups is bounded by the optimal transport cost up to a constant C_d dependent on the metric d only, namely

$$|\text{TV}_{x_0,x_1}(y)| \leq C_d \cdot \text{OTC}^{d}_{x_0,x_1}((Z, W)). \tag{4.194}$$

In words, if the optimal transport (OT) distance between distributions

$$Z, W \mid X = x_1 \text{ and } Z, W \mid X = x_0,$$

with the metric d measuring the transport cost, is small, the TV measure is consequently small as well. Here, however, there is an important nuance, stemming from the decomposability of the TV measure, as shown in the following proposition:

Proposition 4.9 (Inadmissibility of OTC)**.** The optimal transport cost measure $\text{OTC}^{d}_{x_0,x_1}((Z, W))$ is not admissible with respect to structural indirect and structural spurious criteria. Formally, we write that:

$$\text{Str-IE-fair} \;\not\Rightarrow\; \big(\text{OTC}^{d}_{x_0,x_1}((Z, W)) = 0\big), \tag{4.195}$$

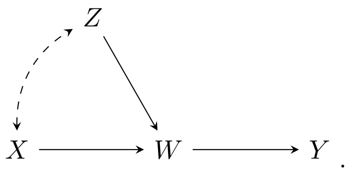
$$\text{Str-SE-fair} \;\not\Rightarrow\; \big(\text{OTC}^{d}_{x_0,x_1}((Z, W)) = 0\big). \tag{4.196}$$

To see the relevance of the proposition above, we proceed by means of an example, in which the above optimal transport distance is small and the TV is minimized, but in which indirect and spurious discrimination still exist.

Example 4.10 (Startup Hiring IV)**.** Suppose that a startup company is hiring accountants. Let sex X be the protected attribute (x_0 female, x_1 male), Z be the age of the candidate and W their performance on an accountancy test, upon which the job decision Y is based. The following SCM $\mathcal{M}^*$ describes the situation:

$$\mathcal{F}^*, P^*(U) : \begin{cases} X \leftarrow U_{XZ} & (4.197) \\ Z \leftarrow -U_{XZ} + U_Z & (4.198) \\ W \leftarrow X + Z + U_W & (4.199) \\ Y \leftarrow \mathbb{1}(U_Y < \text{expit}(W)), & (4.200) \\ & \\ U_{XZ} \in \{0, 1\}, P(U_{XZ} = 1) = 0.5, & (4.201) \\ U_Z, U_W, U_Y \sim \text{Unif}[0, 1], & (4.202) \end{cases}$$

where $\text{expit}(x) = \frac{e^x}{1+e^x}$. The f_w mechanism in Eq. 4.199 shows that older candidates perform better at the test, and that women perform better than men, given equal age. However, due to latent confounding, arising from a specific historical context, women tend to leave the profession at an earlier age (mechanisms f_x, f_z in Eq. 4.197 and 4.198 show that lower age is correlated with being female, through the U_{XZ} variable). The causal graph representing this situation is given by



Importantly, the marginal distributions $W \mid X = x_0$ and $W \mid X = x_1$ are equal in $\mathcal{M}^*$. An outside authority, which certifies whether discrimination is present in the decision-making process, decides that the metric d is given by:

$$d((x, z, w), (x', z', w')) = |w - w'|. \tag{4.203}$$

In this case, we have that

$$\begin{aligned} |P(y \mid x, z, w) - P(y \mid x', z', w')| =& |\text{expit}(w) - \text{expit}(w')| \qquad (4.204) \\ \leq& \frac{1}{4}|w - w'|, \qquad (4.205) \end{aligned}$$

where the last inequality follows from an application of the mean value theorem. Furthermore, the optimal transport cost is 0, because the marginal distributions of W are matching between the groups. There is no direct discrimination, since Y is not a function of X (Eq. 4.200). Therefore, the IF criterion is satisfied and the TV measure equals 0. However, when applying the x-specific decomposition of TV from Thm. 4.3, we have that

$$\begin{aligned} \text{TV}_{x_0,x_1}(y) &= x\text{-DE}_{x_0,x_1}(y \mid x_0) - x\text{-IE}_{x_1,x_0}(y \mid x_0) - x\text{-SE}_{x_1,x_0}(y) \qquad (4.206) \\ &= \underbrace{(0\%)}_{\text{direct}} - \underbrace{(14\%)}_{\text{indirect}} - \underbrace{(-14\%)}_{\text{spurious}}, \qquad (4.207) \end{aligned}$$

which indicates that even though the TV equals 0, the spurious and indirect effects are non-zero. □

Notice the following about the example. Women, who are naturally better at their jobs, are interviewed at a younger age. If the source of the confounding comes from the fact that women (willingly) advance to a different profession in later stages of their career, then the cancellation of spurious and indirect effects in Eq. 4.207 might be acceptable. If, however, the spurious effect stems from a confounding mechanism in which women abandon their careers for certain adverse reasons, then the situation could reasonably be deemed unfair. Without causal considerations, these two cases are indistinguishable. This example is inspired by an example of the original IF paper, which says that "the imposition of a metric already occurs in many classification processes, including credit scores for loan applications" (Dwork *et al.*, 2012, Sec. 6.1.1). Notice that such a metric is based on a single mediator W, similar to the metric in Ex. 4.10.

A possible objection to Ex. 4.10 is that the metric d does not include all confounders and mediators Z, W, which introduces a different issues, as discussed next.

Issue 4. Complete metrics d do not allow for business necessity

We now suppose that the fairness metric d includes all variables in Z, W. If this is the case, then the optimal transport condition implies the independence of X and the Z, W variables, as shown in the following proposition:

Proposition 4.10 (OTC $\implies X \perp\!\!\!\perp Z, W$). Suppose that the metric d is of the following form

$$d((x,z,w),(x',z',w')) = \|z - z'\| + \|w - w'\|, \tag{4.208}$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^d$. Then, we have that the optimal transport condition implies the independence of X and $\{Z, W\}$, namely:

$$\text{OTC}^d_{x_0,x_1}((Z,W)) = 0 \implies X \perp\!\!\!\perp Z, W. \tag{4.209}$$

Furthermore, if the metric d does not consider X then the IF condition implies the independence of X and Y conditional of Z, W.

Proposition 4.11 (IF $\implies X \perp\!\!\!\perp Y \mid Z, W$). Suppose that d is a fairness metric and suppose that the IF condition in Eq. 4.166 holds. Then, for a binary outcome Y, $X \perp\!\!\!\perp Y \mid Z, W$.

Finally, putting the above two propositions together implies that the variable X is independent from all other observables in V, as shown next:

Proposition 4.12 (OTC $\land$ IF $\implies X \perp\!\!\!\perp V \setminus \{X\}$). Suppose that the metric d is of the form $d((x,z,w),(x',z',w')) = \|z - z'\| + \|w - w'\|$, where $\|\cdot\|$ is any norm on $\mathbb{R}^d$. Suppose also that $\text{OTC}^d_{x_0,x_1}((Z,W)) = 0$ and the IF condition in Eq. 4.166 holds. Then we have that

$$X \perp\!\!\!\perp Z, W, Y. \tag{4.210}$$

The proposition shows that if (i) the metric d includes all variables in Z, W; (ii) the IF condition holds; (iii) the optimal transport distance is small, then the protected attribute X is independent from all other endogenous variables in the system. As we will discuss later in Sec. 5, this can be a very strong requirement in practice, which requires completely removing the influence of X, and is not compatible with considerations about business necessity under the disparate impact doctrine.

4.4.3 Criterion 3. Predictive Parity

In this section, we discuss another important criterion appearing in the literature. The notion of predictive parity, introduced by Chouldechova, 2017 [16], is defined as follows:

Definition 4.13 (Predictive Parity (Chouldechova, 2017)). Let X be the protected attribute, and Y the outcome. Let $\widehat{Y}$ be the predictor of Y. We say that $\widehat{Y}$ satisfies predictive parity (PP) with respect to X, Y if

$$P(y \mid x_1, \widehat{y}) = P(y \mid x_0, \widehat{y}) \quad \forall \widehat{y}. \tag{4.211}$$

Alternatively, the PP criterion can also be written as a conditional independence statement

$$Y \perp\!\!\!\perp X \mid \widehat{Y}. \tag{4.212}$$

Finally, define the predictive parity measure to be

$$\text{PPM}_{x_0,x_1}(y \mid \widehat{y}) = P(y \mid x_1, \widehat{y}) - P(y \mid x_0, \widehat{y}). \tag{4.213}$$

In words, the PP criterion ensures that for a group with the value of the predictor $\widehat{Y} = \widehat{y}$, both males and females have the same average outcome Y. Alternatively, the criterion can also be understood as saying that the attribute X provides no additional information about the outcome Y, given that we know the prediction $\widehat{Y}$. Our goal in this section will be to relate this criterion to our previous discussion on the FPCFA (see 3.6) and the decompositions of the TV measure, and show how predictive parity offers important insight when assessing business necessity arguments.

The first well-known result that is important when trying to understand the PP criterion is the following:

Proposition 4.13 (PP and Efficient Learning). Let $\mathcal{M}$ be an SCM compatible with the Standard Fairness Model (SFM). Suppose that the predictor $\widehat{Y}$ is

[16] The original paper (Chouldechova, 2017) distinguishes between predictive parity for a binary predictor $\widehat{Y}$, and calibration for a continuous score We are agnostic with respect to this distinction, and thus use predictive parity for both.

based on the features X, Z, W. Suppose also that $\widehat{Y}$ is an efficient learner, meaning that:

$$\widehat{Y}(x, z, w) = P(y \mid x, z, w) \; \forall x, z, w. \tag{4.214}$$

Then, it follows that $\widehat{Y}$ satisfies predictive parity w.r.t. X and Y.

Proof. Notice that we can write, for any $X = x$:

$$P(y \mid x, \widehat{y}) = \sum_{\substack{z,w: \\ \widehat{Y}(x,z,w)=\widehat{y}}} P(y \mid x, z, w, \widehat{y}) P(z, w \mid x, \widehat{y}) \tag{4.215}$$

$$= \sum_{\substack{z,w: \\ \widehat{Y}(x,z,w)=\widehat{y}}} P(y \mid x, z, w) P(z, w \mid x, \widehat{y}) \tag{4.216}$$

$$= \widehat{y} \sum_{\substack{z,w: \\ \widehat{Y}(x,z,w)=\widehat{y}}} P(z, w \mid x, \widehat{y}) = \widehat{y}. \tag{4.217}$$

Eq. 4.215 follows from the law of total probability, Eq. 4.216 follows from the fact that $Y \perp\!\!\!\perp \widehat{Y} \mid X, Z, W$, and Eq. 4.217 follows from the fact that $P(y \mid x, z, w) = \widehat{y}$ for all z, w with $\widehat{Y}(x, z, w) = \widehat{y}$ (due to efficiency). Therefore, it follows that $P(y \mid x_1, \widehat{y}) = P(y \mid x_0, \widehat{y})$, meaning that $\widehat{Y}$ satisfies PP. ■

Prop. 4.13 shows that PP is expected to hold for an efficient learner $\widehat{Y}$. In some sense, this means that $\widehat{Y}$ should "exhaust" and capture all the variations coming into the outcome Y in the factual, real world. In particular, $\widehat{Y}$ should also capture all the variations of X coming into Y.

Note the stark contrast between the PP notion and the TV measure discussed in the previous sections. For the TV measure to be equal to 0, the predictor $\widehat{Y}$ should not be associated at all with the attribute X, whereas $\widehat{Y}$ should contain all variations of X for the PP condition to hold. Perhaps unsurprisingly based on this discussion, the following theorem holds:

Theorem 4.12 (PP and DP impossibility (Barocas *et al.*, 2017)). The fairness criteria following predictive parity and demographic parity,

$$Y \perp\!\!\!\perp X \mid \widehat{Y}, \tag{4.218}$$

$$\widehat{Y} \perp\!\!\!\perp X, \tag{4.219}$$

are mutually exclusive expect for degenerate cases when $Y \perp\!\!\!\perp X$.

The above theorem might lead the reader to believe that PP and DP are criteria that come from two different realities, and bear no relation to each

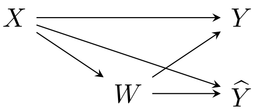


Figure 4.8: Standard Fairness Model with no confounders from Thm. 4.13, extended with the predictor $\widehat{Y}$.

other. After all, the theorem states that it is not possible for the predictor $\widehat{Y}$ to include *all variations* of X in Y and, simultaneously, include *no variations* of X in Y. This realization is the starting point for our discussion in the rest of this section.

First, note that a large amount of effort was needed in previous sections to formalize what the causal meaning of the conditional independence statement $\widehat{Y} \perp\!\!\!\perp X$ is. Along similar lines, we show next how to decompose the predictive parity measure in terms of the underlying mechanisms that transmit change:

Theorem 4.13 (Causal Decomposition of Predictive Parity)**.** Let $\mathcal{M}$ be an SCM compatible with the causal graph in Fig. 4.8. Then, it follows that the PPM can be decomposed into its causal and spurious, reverse-causal variations as follows:

$$P(y \mid x_1, \widehat{y}) - P(y \mid x_0, \widehat{y}) = P(y_{x_1} \mid x_1, \widehat{y}) - P(y_{x_0} \mid x_1, \widehat{y}) \tag{4.220}$$

$$+ P(y_{x_0} \mid \widehat{y}_{x_1}) - P(y_{x_0} \mid \widehat{y}_{x_0}). \tag{4.221}$$

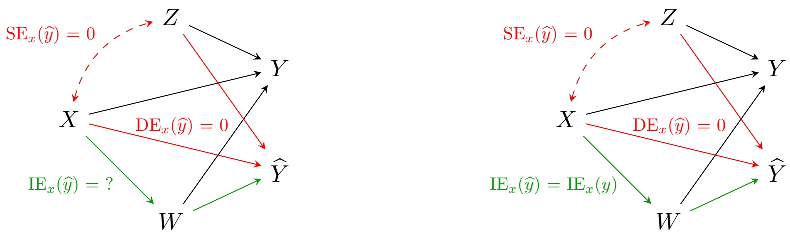
Corollary 4.14 (Linear Causal Decomposition of Predictive Parity)**.** Under the additional assumption that (i) the SCM $\mathcal{M}$ is linear and Y is continuous; (ii) the learner $\widehat{Y}$ is efficient, we have that:

$$\mathbb{E}(y_{x_1} \mid x_1, \widehat{y}) - \mathbb{E}(y_{x_0} \mid x_1, \widehat{y}) = \alpha_{XW}\alpha_{WY} + \alpha_{XY} \tag{4.222}$$

$$\mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_0}) = -(\alpha_{XW}\alpha_{WY} + \alpha_{XY}), \tag{4.223}$$

where $\alpha_{V_iV_j}$ is the linear coefficient between variables V_i, V_j.

The proof of the above theorem and corollary are given in Appendix A.7. In words, these results show that the predictive parity measure can be decomposed as its causal $P(y_{x_1} \mid x_1, \widehat{y}) - P(y_{x_0} \mid x_1, \widehat{y})$ and reverse-causal spurious $P(y_{x_0} \mid \widehat{y}_{x_1}) - P(y_{x_0} \mid \widehat{y}_{x_0})$ counterparts. The first term of the decomposition measures the causal variations induced from a transition $x_0 \rightarrow x_1$, for a fixed set of units. Interestingly, in the linear case, this effect does not depend on the constructed predictor $\widehat{Y}$, but only on the underlying system, i.e., it is not under the control of the predictor designer. To achieve the condition PPM $= 0$, the second term needs to be exactly the reverse of the causal effect, captured by the spurious


$\mathrm{SE}_x(\widehat{y}) = 0$
Z
Y
X
$\mathrm{DE}_x(\widehat{y}) = 0$
$\widehat{Y}$
$\mathrm{IE}_x(\widehat{y}) = ?$
W
$\mathrm{SE}_x(\widehat{y}) = 0$
Z
Y
X
$\mathrm{DE}_x(\widehat{y}) = 0$
$\widehat{Y}$
$\mathrm{IE}_x(\widehat{y}) = \mathrm{IE}_x(y)$
W

(a) SFM annotated with business necessity considerations (discriminatory pathways in red, allowed pathways in green).

(b) Implications of causal predictive parity under business necessity considerations, enforcing a constraint for the indirect effect.

Figure 4.9: Business Necessity and Causal Predictive Parity (CPP).

variations induced by changing $\widehat{y}_{x_0} \to \widehat{y}_{x_1}$ in the selection of units. The second term, which is in the control of the predictor $\widehat{Y}$ designer, needs to cancel out the causal effect measured by the first term, as determined by the underlying system. This key observation allows for a causal interpretation of the predictive parity criterion, and hence we introduce the following definition:

Definition 4.14 (Causal Predictive Parity). Let $\widehat{Y}$ be a predictor of the outcome Y, and let X be the protected attribute. Then, $\widehat{Y}$ is said to satisfy causal predictive parity (CPP, for short) with respect to a counterfactual contrast (C_0, C_1, E, E) if

$$\mathbb{E}[\widehat{y}_{C_1} \mid E = e] - \mathbb{E}[\widehat{y}_{C_0} \mid E = e] = \mathbb{E}[y_{C_1} \mid E = e] - \mathbb{E}[y_{C_0} \mid E = e] \; \forall e. \tag{4.224}$$

Furthermore, $\widehat{Y}$ is said to satisfy CPP with respect to a factual contrast (C, C, E_0, E_1) if

$$\mathbb{E}[\widehat{y}_C \mid E_1] - \mathbb{E}[\widehat{y}_C \mid E_0] = \mathbb{E}[y_C \mid E_1] - \mathbb{E}[y_C \mid E_0]. \tag{4.225}$$

The intuition behind causal predictive parity is simple – if a contrast $\mathcal{C}$ describes some amount of variation (factual or counterfactual) in the true outcome Y, then it should describe the same amount of variation in the predicted outcome $\widehat{Y}$.

Complementary to the notion of causal predictive parity, and based on the discussion in Sec. 4, we can also define the notion of causal statistical parity:

Definition 4.15 (Causal Statistical Parity). Let $\widehat{Y}$ be a predictor, and let X be the protected attribute. Then, $\widehat{Y}$ is said to satisfy causal statistical parity (CSP) with respect to a counterfactual contrast (C_0, C_1, E, E) if

$$\mathbb{E}[\widehat{y}_{C_1} \mid E = e] - \mathbb{E}[\widehat{y}_{C_0} \mid E = e] = 0 \; \forall e. \tag{4.226}$$

Furthermore, $\widehat{Y}$ is said to satisfy CSP with respect to a factual contrast (C, C, E_0, E_1) if

$$\mathbb{E}[\widehat{y}_C \mid E_1] - \mathbb{E}[\widehat{y}_C \mid E_0] = 0. \tag{4.227}$$

Once again, the intuition behind the above notion is that if a contrast $\mathcal{C}$ captures variations of X in the predictor $\widehat{Y}$ along a discriminatory pathway, then the causal effect transmitted along this pathway should equal 0.

We can now tie the notions of causal statistical parity and causal predictive parity through the concept of *business necessity*. Our proposal is illustrated graphically in Fig. 4.9 through an example. Firstly, in Fig. 4.9a, we highlight the causal pathways in the business necessity set in green, and those that are not in red. If a contrast $\mathcal{C}$ is associated with variations that are not in the business necessity set, then the value of this contrast should be $\mathcal{C}(X, \widehat{Y}) = 0$[17], in this case ensuring that direct and spurious effects should equal 0. This step can be understood as following from the principles of causal statistical parity.

On the other hand, if the variations associated with the contrast *are* in the business necessity set, then the principles of causal statistical parity are not sufficient for determining the value of the contrast in a principled fashion (indicated with $\text{IE}_x(\widehat{y}) = ?$ in Fig. 4.9a). In other words, causal statistical parity only helps us to determine that the contrasts not in the business necessity set should equal 0, but says nothing about the contrasts that are in the business necessity set. However, for the latter case, we can leverage the idea behind causal predictive parity. The value of a contrast that is not considered discriminatory should be equal for the predictor and the true outcome, i.e.,

$$\mathcal{C}(X, \widehat{Y}) = \mathcal{C}(X, Y), \tag{4.228}$$

which in our example ensures that $\text{IE}_x(\widehat{y}) = \text{IE}_x(y)$ as shown in Fig. 4.9b. The idea behind Eq. 4.228 is very intuitive, yet of crucial importance when considering arguments of business necessity. We also note that similar considerations are voiced within the legal literature on discrimination (Grimmelmann and Westreich, 2016), although not in a formal language provided here. Finally, we illustrate the above described notions through an example:

Example 4.11 (Statistical and Predictive Parity in Employment). A company recently decided to switch to a new AI human resources (HR) system for determining the salaries of their employees. Let Y denote the current salary, X gender (x_0 female, x_1 male), and W an indicator of certain job-related qualifications. The salary recommended by the AI system is denoted by $\widehat{Y}$. After this change, a group of male employees is suing the company, claiming

[17] Here, the notation $\mathcal{C}(A, B)$ indicates the value of an effect of A on B described by the contrast $\mathcal{C}$.

that the new salary system is unfair, since their salaries have been reduced, namely,

$$\mathbb{E}(\widehat{y} \mid x_1) < \mathbb{E}(y \mid x_1). \tag{4.229}$$

In a legal proceeding, the court finds that any kind of direct effect of gender on salary decisions is not permissible. In particular, they require that the company needs to prove that

$$\text{NDE}_{x_0,x_1}(\widehat{y}) = \text{NDE}_{x_1,x_0}(\widehat{y}) = 0. \tag{4.230}$$

However, due to a business necessity argument, the court allows the company to determine the salaries based on specific job-related qualifications. This, in particular, implies the opposite of Eq. 4.230, namely, that it is legally permissible that

$$\text{NIE}_{x_0,x_1}(\widehat{y}) \neq 0, \text{NIE}_{x_1,x_0}(\widehat{y}) \neq 0. \tag{4.231}$$

Nonetheless, the court rules that the values of NIE cannot be arbitrary, but need to be in line with the allocation of salaries prior to the deployment of the new AI system. Based on the notion of causal predictive parity, this means that:

$$\text{NIE}_{x_0,x_1}(\widehat{y}) = \text{NIE}_{x_0,x_1}(y), \tag{4.232}$$

$$\text{NIE}_{x_1,x_0}(\widehat{y}) = \text{NIE}_{x_1,x_0}(y). \tag{4.233}$$

□

The example demonstrates how considerations of statistical and predictive parity interact through business necessity requirements. Firstly, the direct effect, which is considered discriminatory by the court, needs to equal 0. The indirect effect of X on $\widehat{Y}$ was deemed permissible, and must therefore be equal to the indirect effect of X on the true outcome Y. For a further real-world example of applying the above concepts in practice, we refer the reader to Ex. 5.3 in Sec. 5.1 in which the principles of causal statistical and predictive parity are illustrated through an analysis of the COMPAS dataset.

The discussion in this section implies that the notion of statistical parity (i.e., $\text{TV}_{x_0,x_1} = 0$) will be satisfied when *none of the variations* are in the business necessity set. On the other hand, if *all of the variations* are in the business necessity, then the notion of predictive parity will be satisfied. The choice of the BN set can be thought of as interpolating between demographic parity and predictive parity, in the way described by Def. 4.14. Instead of being viewed as mutually exclusive notions, tied through the impossibility result from Thm. 4.12, these two criteria can be viewed as the extremes of a spectrum spanned by different business necessity requirements.

5

Fairness tasks

The main goal of this section is to develop tools to support causal fairness analysis in practice, building on the foundations introduced in previous sections. We classify fairness problems into three tasks, in increasing order of difficulty:

Task 1. **Bias detection and quantification:** the first and most basic task of fair ML. We may consider operating with a dataset $\mathcal{D}$ of past decisions, or in infinite samples with an observed distribution $P(V)$ over variables V. The task is to define a mapping

$$\mu : \mathcal{P} \to \mathbb{R},$$

where $\mathcal{P}$ is the set of possible distributions $P(V)$, and μ is viewed as a *fairness measure* and it is often constructed so that $\mu(P(V)) = 0$ suggests the absence of some form of discrimination.

Task 2. **Fair prediction:** The task of fair prediction, usually, relies on a certain measure of fairness. The task is to learn a distribution $P^*(V)$ while maximizing utility $U(P(V))$ and satisfying

$$|\mu(P^*(V))| \leq \epsilon,$$

where μ is a measure of fairness as discussed in Task 1. Fair classification and fair regression problems fall into this category[1].

[1] Different categories of fair prediction methods exist, namely pre-processing, in-processing, and post-processing, which are discussed in Sec. 5.2.

Task 3. **Fair decision-making:** In fair decision-making, the well-being of certain groups over time is considered. Notions of affirmative action also fall into this category. We might be interested in designing a policy π, which at every time step affects the observed distribution $P_t(V)$ (which now changes over time) so that we have

$$P_{t+1}(V) = \pi(P_t(V)),$$

and we are, perhaps, interested in controlling how $\mu(P_t(V))$ changes with t.

Note that these three tasks form a certain hierarchy, and are introduced in order of difficulty. Fair prediction often relies on a specific fairness measure; fair decision-making often relies on a fairness measure, a notion of utility, and possibly fair predictions. The three tasks are discussed in Sec. 5.1, 5.2, and 5.3, respectively.

5.1 Task 1: Bias Detection & Quantification

We distinguish two different, but closely related subtasks in the context of Task 1, that are referred to as bias detection and bias quantification. In bias detection, we are interested in providing a binary decision rule ψ that determines whether discrimination is present in the data-generating process or not. In bias quantification, we are interested in how strong the discrimination is, and therefore provide a real-valued number, instead of a Boolean yes/no decision. In what follows, we provide the mathematical formulation of the two subtasks, together with an approach for how to solve them.

Definition 5.1 (Bias Detection under SFM). Let Ω be a space of SCMs. Let Q be a structural fairness criterion, $Q : \Omega \to \{0, 1\}$ (Def. 3.1), determining whether a causal mechanism within the SCM $\mathcal{M} \in \Omega$ is active ($Q(\mathcal{M}) = 0$ if mechanism not active, $Q(\mathcal{M}) = 1$ if active). The task of bias detection is to test the hypothesis

$$H_0 : Q(\mathcal{M}) = 0, \tag{5.1}$$

that is, constructing a mapping $\psi(\mathcal{G}_{\text{SFM}}, \mathcal{D})$ into $\{0, 1\}$, which provides a decision rule for testing H_0, based on the standard fairness model $\mathcal{G}_{\text{SFM}}$ and the data on past decisions $\mathcal{D}$.

In words, we are interested whether direct, indirect, or spurious discrimination exists, corresponding to $Q \in$ Str-{DE,IE,SE}, as introduced in Def. 3.2. The null hypothesis H_0 assumes that there is no discrimination, and the decision rule ψ determines whether H_0 should be rejected based on the SFM and the available data. Notice, crucially, that ψ is a function of $\mathcal{G}_{\text{SFM}}$ and $\mathcal{D}$, due to the fact that the true SCM $\mathcal{M}$ is never available to the data scientist. Therefore, she/he cannot reason about $Q(\mathcal{M})$ directly, but instead needs to find an *admissible* measure μ that satisfies

$$Q(\mathcal{M}) = 0 \implies \mu(\mathcal{M}) = 0, \tag{5.2}$$

where $\mu(\mathcal{M})$ can be computed in practice. Recall the result from Prop. 3.1 showing that the TV measure is not admissible with respect to Str-{DE,IE,SE} and, therefore, should not be used for bias detection when one is interested in direct, indirect, and spurious effects. Moreover, we note that solving the bias detection task depends on solving the FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) from Def. 3.6, which we now restate in the form more suitable for Task 1:

Definition 5.2 (FPCFA continued for Task 1). [Ω, $(Q_i)_{i=1}^{k}$, μ as before] Let $\mathcal{M} = \langle V, U, P(u), \mathcal{F} \rangle$ be the true, unobserved generative SCM, $\mathcal{A}$ a set of assumptions, and $P(v)$ the observational distribution generated by $\mathcal{M}$. Let $\Omega^{\mathcal{A}}$ be the space of all SCMs compatible with $\mathcal{A}$. The Fundamental Problem of Causal Fairness Analysis is to find a collection of measures $\mu_1, \ldots, \mu_k$ such that the following properties are satisfied:

(1) μ is decomposable w.r.t. $\mu_1, \ldots, \mu_k$;

(2) $\mu_1, \ldots, \mu_k$ are admissible w.r.t. the structural fairness criteria $Q_1, \ldots, Q_k$.

(3) $\mu_1, \ldots, \mu_k$ are as powerful as possible.

(4) $\mu_1, \ldots, \mu_k$ are identifiable from the observational distribution $P(v)$ and class $\Omega^{\mathcal{A}}$.

The final step of FPCFA for Task 1 is

(5) estimate $\mu_1, \ldots, \mu_k$ and their $(1 - \alpha)$ confidence intervals from the observational data and the SFM projection of the causal diagram.

Upon solving FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) for Task 1, we obtain measures μ_i based on which the decision rule ψ can be constructed. In particular, the decision rule ψ will be constructed by computing the $(1 - \alpha)$ confidence interval for μ_i, for instance using bootstrap. If the interval excludes 0, the H_0 hypothesis is rejected.

The derived measures μ_i obtained from solving the FPCFA for Task 1 can also be used for the related task of bias quantification:

Definition 5.3 (Bias Quantification under SFM). Let Ω be a space of SCMs and let $(Q_i)_{i=1:3}$ = Str-{DE,IE,SE}. The task of bias quantification is to find a mapping $\phi(\mathcal{G}_{\text{SFM}}, \mathcal{D}) \mapsto \mathbb{R}^3$ where the i-th component ϕ_i is admissible with respect to Q_i.

In words, the amount of discrimination is summarized using a 3-dimensional statistic. Each component of the statistic corresponds to one of the direct, indirect, or spurious effects. The measures μ_i obtained from solving FPCFA(Str-$\{$DE,IE,SE$\}$, $\text{TV}_{x_0,x_1}(y)$) can be used to solve the task of bias quantification, by setting

$$\phi(\mathcal{M}) = \big(\mu_{\text{DE}}(\mathcal{M}), \mu_{\text{IE}}(\mathcal{M}), \mu_{\text{SE}}(\mathcal{M})\big). \tag{5.3}$$

We can now discuss a specific proposal for the measures μ_i.

Measures μ_i for Task 1. Following the x-specific solution of FPCFA from Thm. 4.3, we use the following measures:

$$\mu_{\text{DE}} \text{ is given by } x\text{-DE}_{x_0,x_1}(y \mid x_0) = P(y_{x_1,W_{x_0}} \mid x_0) - P(y_{x_0} \mid x_0) \tag{5.4}$$

$$\mu_{\text{IE}} \text{ is given by } x\text{-IE}_{x_1,x_0}(y \mid x_0) = P(y_{x_1,W_{x_0}} \mid x_0) - P(y_{x_1} \mid x_0) \tag{5.5}$$

$$\mu_{\text{SE}} \text{ is given by } x\text{-SE}_{x_1,x_0}(y) = P(y_{x_1} \mid x_0) - P(y_{x_1} \mid x_1). \tag{5.6}$$

Moreover, the solution (see Eq. 4.48) also showed that the TV can be decomposed as:

$$\text{TV}_{x_0,x_1}(y) = \underbrace{x\text{-DE}_{x_0,x_1}(y \mid x_0)}_{\mu_{\text{DE}}} - \underbrace{x\text{-IE}_{x_1,x_0}(y \mid x_0)}_{\mu_{\text{IE}}} - \underbrace{x\text{-SE}_{x_1,x_0}(y \mid x_0)}_{\mu_{\text{SE}}}. \tag{5.7}$$

In words, the TV equals the x-specific direct effect with a transition $x_0 \to x_1$, minus the x-specific indirect effect with the opposite transition $x_1 \to x_0$ and minus the x-specific spurious effect with the transition $x_1 \to x_0$. One critical point to note is that such a decomposition is not unique. In fact, the TV can also be decomposed as:

$$\text{TV}_{x_0,x_1}(y) = -x\text{-DE}_{x_1,x_0}(y \mid x_0) + x\text{-IE}_{x_1,x_0}(y \mid x_0) - x\text{-SE}_{x_1,x_0}(y). \tag{5.8}$$

Sometimes it may be desirable to achieve symmetry and avoid picking a specific order, so in such cases we propose using the average of the two available decompositions. In particular, we define the symmetric x-specific direct and indirect effects as follows:

Definition 5.4 (Symmetric x-specific direct and indirect effect)**.** The symmetric x-specific direct and indirect effects are defined as:

$$x\text{-DE}^{\text{sym}}_x(y \mid x) = \frac{1}{2}\big(x\text{-DE}_{x_0,x_1}(y \mid x) - x\text{-DE}_{x_1,x_0}(y \mid x)\big) \tag{5.9}$$

$$x\text{-IE}^{\text{sym}}_x(y \mid x) = \frac{1}{2}\big(x\text{-IE}_{x_0,x_1}(y \mid x) - x\text{-IE}_{x_1,x_0}(y \mid x)\big). \tag{5.10}$$

Therefore, we propose to use $x\text{-DE}^{\text{sym}}_{x}(y \mid x_0)$ and $x\text{-IE}^{\text{sym}}_{x}(y \mid x_0)$ instead of $x\text{-DE}_{x_1,x_0}(y \mid x_0)$ and $x\text{-IE}_{x_1,x_0}(y \mid x_0)$ for Task 1, which corresponds to using both of the TV decompositions and averaging them[2]. The benefit of these alternative measures is that no single transition has to be chosen for computing the direct/indirect effect. Instead, both $x_0 \to x_1$ and $x_1 \to x_0$ transitions are considered, by taking the average of the two. Such an approach offers measures of direct and indirect effect which are symmetric with respect to the change in the protected attribute, unlike their counterparts that consider a single transition. More generally, taking such averages over different transitions may be seen as related to a flow-based attribution approach (Singal *et al.*, 2021) based on Shapley values (Shapley *et al.*, 1953).

5.1.1 Legal Doctrines - A Formal Approach

Equipped with specific measures that can be used to perform bias detection and quantification, we develop a formal approach for assessing the legal doctrines of disparate impact and treatment. The operational approach is described in Alg. 5.1, and is one of the highlights of this manuscript. The algorithm takes the dataset $\mathcal{D}$, the SFM projection $\Pi_{\text{SFM}}(\mathcal{G})$ of the causal diagram, and the Business Necessity Set (BN-set) as an input. When using the SFM, the allowed BN-sets are $\emptyset, \{Z\}, \{W\}$, and $\{Z, W\}$[3].

We distinguish two slightly different cases of applying the Fairness Cookbook. The first, basic case is when the dataset $\mathcal{D}$ contains a single outcome variable, which is in Alg. 5.1 labeled as $\widehat{Y}$. In this case, we need to verify that the causal effect along any pathway that is not in the business necessity set equals 0. However, for the causal pathways that are in the business necessity set, we cannot prescribe a specific value for the effect transmitted along this pathway.

The second, more involved case is when the dataset $\mathcal{D}$ contains the true observed outcomes Y and the predicted outcomes $\widehat{Y}$. Again, we must verify that pathways not in the business necessity set transmit no variations from X to $\widehat{Y}$. However, if the true outcome Y is available, business necessity pathways require scrutiny, too, as discussed in Sec. 4.4.3. For the latter pathways, one additionally needs to ensure that a causal effect along them is not amplified for the predictor $\widehat{Y}$ compared to the true outcome Y. In other words, even if a causal pathway is considered as *non-discriminatory*, we must verify that

[2] VanderWeele, 2015 proposes a decomposition of the total effect (TE) that accounts for the direct and indirect effects, and also an explicit interaction term. Connecting our framework, a formal test for existence of an interaction of X and W in Y can be performed by testing the equality $x\text{-DE}^{\text{sym}}_{x}(y \mid x_0) = x\text{-DE}_{x_0,x_1}(y \mid x_0)$, or performing the analogous test for the indirect effect.

[3] Handling more involved BN-sets is discussed in detail in Sec. 6.

the discrimination along this pathway in not amplified to unacceptable levels – which is also reflected in Alg. 5.1.

5.1.2 Empirical Evaluation

We will start by applying the cookbook for the task of bias detection in the context of the US Census 2018 dataset. After that, we will apply the cookbook for the task of temporal bias quantification on a College Admissions dataset, and finish with an application on the COMPAS dataset.

Example 5.1 (US Government Census 2018). The United States Census of 2018 collected broad information about the US Government employees, including demographic information Z (Z_1 for age, Z_2 for race, Z_3 for nationality), gender X (x_0 female, x_1 male), marital and family status M, education information L, and work-related information R. In an initial analysis, a data scientist observed that male employees on average earn \$14000/year more than their female counterparts, that is

$$\mathbb{E}[y \mid x_1] - \mathbb{E}[y \mid x_0] \approx \$14000. \tag{5.16}$$

Following the Fairness Cookbook, the data scientist does the following:
SFM projection: the SFM projection of the causal diagram $\mathcal{G}$ of this dataset is given by

$$\Pi_{\text{SFM}}(\mathcal{G}) = \langle X = \{X\}, Z = \{Z_1, Z_2, Z_3\}, W = \{M, L, R\}, Y = \{Y\}\rangle. \tag{5.17}$$

In words, the set of confounders includes age, race, and nationality, while the set of mediators includes family status, education, and work-related information.
Disparate treatment: when considering disparate treatment, she computes $x\text{-DE}^{\text{sym}}_x(y \mid x_0)$ and its 95% confidence interval, i.e.,

$$x\text{-DE}^{\text{sym}}_x(y \mid x_0) = \$7210 \pm \$1049. \tag{5.18}$$

The hypothesis $H_0^{(x\text{-DE})}$ is thus rejected, providing evidence of disparate treatment against female employees.
Disparate impact: when considering disparate impact, the data scientist computes Ctf-SE, Ctf-IE and their respective 95% confidence intervals,

$$x\text{-IE}^{\text{sym}}_x(y \mid x_0) = \$5126 \pm \$778, \tag{5.19}$$

$$x\text{-SE}_{x_1,x_0}(y) = -\$1675 \pm \$955. \tag{5.20}$$

She then concludes that the differences in salary explained by the spurious correlation of gender with age, race, and nationality are not considered discriminatory. Therefore, she tests the hypothesis

$$H_0^{(x\text{-IE})} : x\text{-IE}^{\text{sym}}_x(y \mid x_0) = 0,$$

which is rejected, indicating evidence of disparate impact on government's female employees. Measures computed in the example are shown in Fig. 5.1. □

Algorithm 5.1 Fairness Cookbook for Task 1

```
• Inputs: Dataset $\mathcal{D}$, SFM $\Pi_{\text{SFM}}(\mathcal{G})$, Business Necessity Set BN-set.
Obtain the dataset $\mathcal{D}$.
Determine the Standard Fairness Model projection $\Pi_{\text{SFM}}(\mathcal{G})$.
Consider the existence of Disparate Treatment:
    • compute $x\text{-DE}^{\text{sym}}_{x}(\widehat{y} \mid x_0)$ and its 95% confidence interval (CI),
    • test the hypothesis
        $H_0^{(x\text{-DE})} : x\text{-DE}^{\text{sym}}_{x}(\widehat{y} \mid x_0) = 0. \quad (5.11)$
        – if $H_0^{(x\text{-DE})}$ rejected $\implies$ evidence of disparate treatment,
    • Additionally: if no evidence of disparate treatment in overall population, for $Z = z$ test the hypothesis $H_0^{(z\text{-DE})} : z\text{-DE}^{\text{sym}}_{x}(y \mid z) = 0$.
Consider the existence of Disparate Impact:
    • compute $x\text{-IE}^{\text{sym}}_{x}(\widehat{y} \mid x_0)$ and $x\text{-SE}_{x_1,x_0}(\widehat{y})$ and their 95% CIs,
    • if $Y \in \mathcal{D}$, compute $x\text{-IE}^{\text{sym}}_{x}(y \mid x_0)$ and $x\text{-SE}_{x_1,x_0}(y)$ with 95% CIs,
    • if $W \notin$ BN-set, test the hypothesis
        $H_0^{(x\text{-IE})} : x\text{-IE}^{\text{sym}}_{x}(\widehat{y} \mid x_0) = 0. \quad (5.12)$
        – if $H_0^{(x\text{-IE})}$ rejected $\implies$ evidence of disparate impact,
        – Additionally: if no evidence of disparate impact in overall population, for $Z = z$ test the hypothesis $H_0^{(z\text{-IE})} : z\text{-IE}^{\text{sym}}_{x}(y \mid z) = 0$,
    • if $W \in$ BN-set and outcome $Y$ is in the data $\mathcal{D}$, test the hypothesis
        $H_{0,\text{BN}}^{(x\text{-IE})} : x\text{-IE}^{\text{sym}}_{x}(y \mid x_0) = x\text{-IE}^{\text{sym}}_{x}(\widehat{y} \mid x_0). \quad (5.13)$
        – if $H_{0,BN}^{(x\text{-IE})}$ rejected, a possible violation of disparate impact,
    • if $Z \notin$ BN-set, test the hypothesis
        $H_0^{(x\text{-SE})} : x\text{-SE}_{x_1,x_0}(y) = 0. \quad (5.14)$
        – if $H_0^{(x\text{-SE})}$ rejected $\implies$ evidence of disparate impact,
    • if $Z \in$ BN-set and $Y \in \mathcal{D}$, test the hypothesis
        $H_{0,\text{BN}}^{(x\text{-SE})} : x\text{-SE}_{x_1,x_0}(y) = x\text{-SE}_{x_1,x_0}(\widehat{y}). \quad (5.15)$
        – if $H_{0,BN}^{(x\text{-SE})}$ rejected, a possible violation of disparate impact.
```

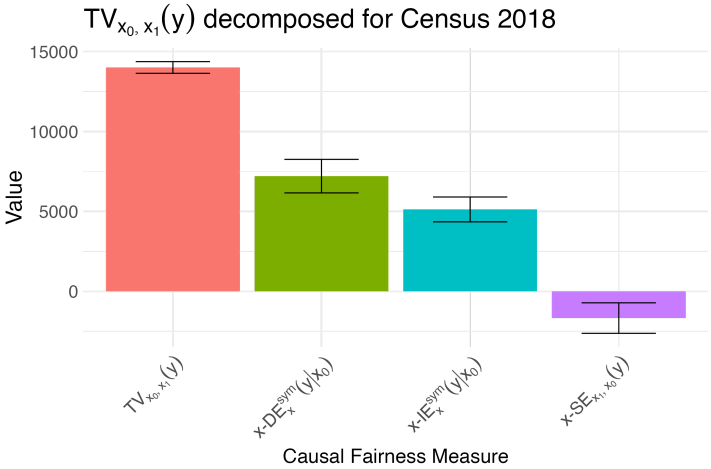


Figure 5.1: Measures obtained when applying the Fairness Cookbook for Task 1 on the Government Census 2018 dataset.

Example 5.2 (Bias Quantification in College Admissions). A university in the United States admits applicants every year through a competitive process. The university is facing strong regulatory pressure and aims to quantify discrimination in the admission process and track it over time, between 2011 and 2020.

The data generating process changes over time, described next. The set of observed variables is $V = \{X, Z, W, Y\}$, where X denotes gender (x_0 female, x_1 male), Z denotes age at time of application ($Z = 0$ under 20 years, $Z = 1$ over 20 years), and W denotes the department of application ($W = 0$ for arts&humanities, $W = 1$ for sciences). Finally, let Y denote the admission decision ($Y = 0$ rejection, $Y = 1$ acceptance). The application process changes over time and the mechanisms are given by

$$\mathcal{F}(t), P(U): \begin{cases} X \leftarrow \mathbb{1}(U_X < 0.5 + 0.1U_{XZ}) & (5.21)\\ Z \leftarrow \mathbb{1}(U_Z < 0.5 + \kappa(t)U_{XZ}) & (5.22)\\ W \leftarrow \mathbb{1}(U_W < 0.5 + \lambda(t)X) & (5.23)\\ Y \leftarrow \mathbb{1}(U_Y < 0.1 + \alpha(t)X + \beta(t)W + 0.1Z). & (5.24)\\ & \\ U_{XZ} \in \{0,1\}, P(U_{XZ} = 1) = 0.5, & (5.25)\\ U_X, U_Z, U_W, U_Y \sim \text{Unif}[0,1]. & (5.26) \end{cases}$$

Coefficient $\kappa(t)$ describes the spurious association of age and gender, while $\lambda(t)$ describes the difference in preference for science departments between genders. Coefficient $\alpha(t)$ describes direct discrimination, i.e., that a gender group is favored in an unjustified way. Coefficient $\beta(t)$ describes the increase in probability of admission when applying to a science department as opposed to an arts & humanities department. The coefficients change every year, and obey the following dynamics:

$$\kappa(t+1) = 0.9\kappa(t) \tag{5.27}$$
$$\lambda(t+1) = \lambda(t)(1 - \beta(t)) \tag{5.28}$$
$$\beta(t+1) = \beta(t)(1 - \lambda(t))f(t), f(t) \sim \text{Unif}[0.8, 1.2] \tag{5.29}$$
$$\alpha(t+1) = 0.8\alpha(t). \tag{5.30}$$

The equations can be interpreted as follows. The coefficient $\kappa(t)$ decreases over time, meaning that the overall age gap between the groups decreases. The coefficient $\lambda(t)$ decreases compared to the previous year, by an amount dependent on $\beta(t)$. In words, the rate of application to arts & humanities departments decreases if these departments have lower overall admission rates (i.e., students are less likely to apply to departments that are hard to get into). Further, $\alpha(t)$, which represents gender bias, decreases over time. Finally, $\beta(t)$ represent the increase in the probability of admission when applying to a science department. Its value depends on the value from the previous year, multiplied by $(1 - \lambda(t))$ and the random variable $f(t)$. Multiplication by the former factor describes the mechanism in which the benefit of applying to a science department decreases if a larger proportion of students apply for it. The latter factor describes a random variation over time which describes how well (in relative terms) the science departments are funded, and can be seen as depending on research and market dynamics in the sciences.

The head data scientist at the university decides to use the Fairness Cookbook for performing bias quantification of their admissions team. The SFM projection of the causal diagram $\mathcal{G}$ of the dataset is given by

$$\Pi_{\text{SFM}}(\mathcal{G}) = \langle X = \{X\}, Z = \{Z\}, W = \{W\}, Y = \{Y\}\rangle. \tag{5.31}$$

After that, the data scientist estimates the quantities

$$x\text{-DE}_x^{\text{sym}}(y \mid x_0), x\text{-DE}_x^{\text{sym}}(y \mid x_0), \text{ and } x\text{-SE}_{x_1,x_0}(y), \tag{5.32}$$

$\forall t \in \{2011, \dots, 2020\}$. The temporal dynamics of the estimated measures of discrimination (together with the ground truth values obtained from the SCM $\mathcal{M}_t$) are shown graphically in Fig. 5.2. As the figure illustrates, the policies put in place by the university have over time managed to mitigate the bias between groups that existed initially. □

Example 5.3 (COMPAS – continued)**.** Courts in Broward County, Florida use a machine learning algorithm, developed by Northpointe, to predict whether

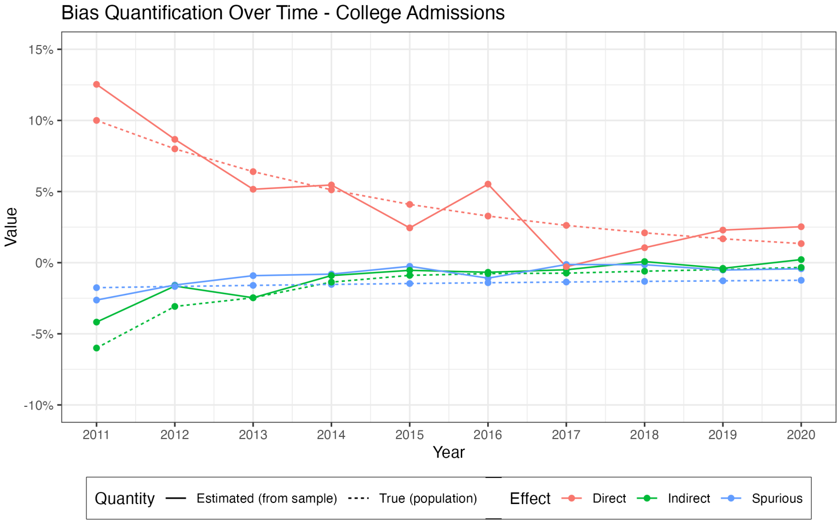


Figure 5.2: Tracking bias over time in the synthetic College Admissions dataset from Ex. 5.2, between years 2011 and 2020. Both the estimated values from simulated samples (solid line) and the true population values (dashed lines) are shown, for direct (red), indirect (green), and spurious (blue) effects.

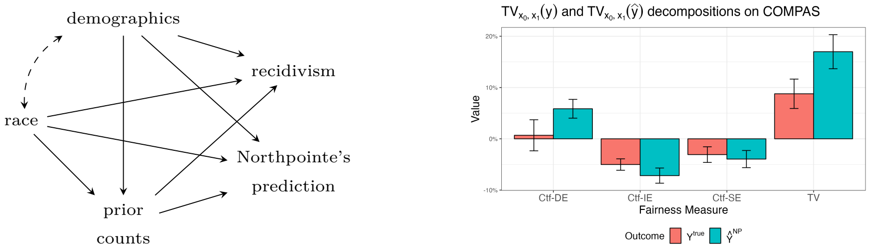
individuals released on parole are at high risk of re-offending within 2 years (Y for recidivism). The algorithm is based on the demographic information Z (Z_1 for gender, Z_2 for age), race X (x_0 denoting majority, x_1 minority), juvenile offense counts J, prior offense count P, and degree of charge D (mediators W). The predictions obtained from Northpointe's algorithm are labeled as $\widehat{Y}$.

The standard fairness model (SFM) for this example is shown in Fig. 5.3a. Suppose that in an initial hearing, the Broward County district court determines that the direct and indirect effects are not in the business necessity set, while the spurious effect is. In words, gender (Z_1) and age (Z_2) are allowed to be used to distinguish between the minority and majority groups when predicting recidivism, while other variables are not.

In light of this information, we proceed as follows. We first compute the TV decomposition for the true outcome Y. Then, we compute the Northpointe's predictions $\widehat{Y}^{NP}$. The comparison of the two decompositions is shown in Fig. 5.3b. For the direct effect, we have:

$$\text{Ctf-DE}_{x_0,x_1}(y \mid x_0) = -0.08\% \pm 2.59\%, \tag{5.33}$$
$$\text{Ctf-DE}_{x_0,x_1}(\widehat{y} \mid x_0) = 6\% \pm 2.96\%. \tag{5.34}$$

Since the direct effect is not in the business necessity set, Northpointe's predictions clearly violate the disparate treatment doctrine (turquoise bar for



(a) SFM for Ex. 5.3.

(b) TV decompositions for Ex. 5.3.

Figure 5.3: SFM and the TV decompositions for Ex. 5.3.

the Ctf-DE measure in Figure 5.3b), since they are significantly different from 0. Next, for the indirect effects, we obtain:

$$\text{Ctf-IE}_{x_1,x_0}(y \mid x_0) = -5.06\% \pm 1.24\%, \tag{5.35}$$

$$\text{Ctf-IE}_{x_1,x_0}(\widehat{y} \mid x_0) = -7.73\% \pm 1.53\%. \tag{5.36}$$

Once again, the indirect effect, which is not in the business necessity set, is different from 0 for the Northpointe's predictions (violating disparate impact, see turquoise bar for Ctf-IE in Figure 5.3b), but not statistically different from 0 for our predictions (blue bar). Interestingly, the indirect effect is different from 0 for the true outcome (red bar), indicating a bias in the current real world. The above provided reasoning was based on the principles of causal statistical parity.

Finally, for the spurious effects, which is in the business necessity set, we use the principles of causal predictive parity. In particular, we have computed that:

$$\text{Ctf-SE}_{x_1,x_0}(y) = -3.17\% \pm 1.53\%, \tag{5.37}$$

$$\text{Ctf-SE}_{x_1,x_0}(\widehat{y}) = -3.75\% \pm 1.58\%. \tag{5.38}$$

As each confidence interval contains the point estimate of the other quantity, no violations with respect to the spurious effect are found. □

5.2 Task 2. Fair Prediction

We are now ready to discuss Task 2, which builds on similar foundations as the previous task. The section is organized as follows.

(i) We first discuss previous literature on (fair) prediction; in particular, we discuss post-processing, in-processing, and pre-processing methods.

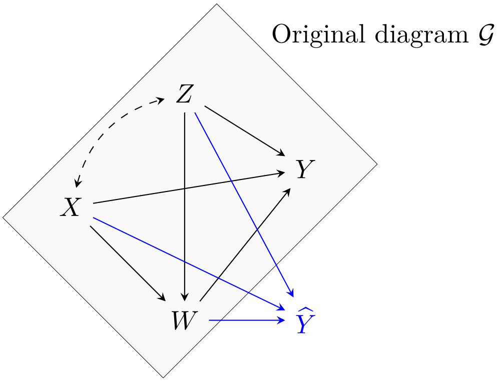


Figure 5.4: Standard Fairness Model (SFM) extended with a blue node $\widehat{Y}$, for the task of (fair) prediction.

(ii) We formalize the FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) for Task 2, which is the problem that needs to be solved such that causally meaningful fair predictions can be obtained.

(iii) We introduce the Fair Prediction Theorem (Thm. 5.1) that explains why standard methods for fair prediction, agnostic to the causal structure, fail in solving FPCFA.

(iv) We develop two alternative formulations of the fair prediction optimization problem capable of remedying the shortcomings of methods found in the literature.

5.2.1 Prediction

In the context of prediction, one is generally interested in constructing a predictor $\widehat{Y}$ of Y, which is a function of X, Z and W. More precisely, from a causal inference point of view, this process can be conceptualized as constructing an additional mechanism $\widehat{Y} \leftarrow f_{\widehat{Y}}(x, z, w)$ in the SCM, which is under our control, as shown in Fig. 5.4. A typical choice of $f_{\widehat{Y}}$ in the context of regression is the estimate of $\mathbb{E}[Y \mid X = x, Z = z, W = w]$, whereas for classification a rounded version of such an estimate is often considered. The key question from a fairness point of view is whether such an approach carries over the bias from the existing data into the prediction mechanism $f_{\widehat{Y}}$.

If the newly constructed prediction does exhibit undesired bias, one may be interested in designing *fair* predictions instead, by ensuring that the constructed $\widehat{Y}$ satisfies a fairness constraint. In the fairness literature, there are three broad categories for achieving this – post-processing, in-processing, and pre-processing. Although there are many possible target measures of fairness which the predictor $\widehat{Y}$ could satisfy, in this manuscript we focus on methods that aim to achieve the condition $\text{TV}_{x_0,x_1}(\widehat{y}) = 0$.

5.2.2 Post-processing

Post-processing methods are the simplest and most easily described. First, one constructs a predictor $f_{\widehat{Y}}$ without applying fairness constraints. The output of $f_{\widehat{Y}}(x, z, w)$ is then taken and transformed using a transformation T, such that the constructed predictor

$$\widehat{Y} \leftarrow T(f_{\widehat{Y}}(x, z, w)), \tag{5.39}$$

satisfies the condition $\text{TV}_{x_0,x_1}(\widehat{y}) = 0$. We illustrate the post-processing methods with an example. The *reject-option classification* of Kamiran *et al.*, 2012 starts by estimating the probabilities of belonging to the positive class, $P(y)$ (label the estimates with $f_{\widehat{Y}}(x, z, w)$). The classifier $\widehat{Y}$ is then constructed such that

$$\widehat{Y}(x, z, w) = \mathbb{1}(f_{\widehat{Y}}(x, z, w) > \theta_x),$$

where $\theta_{x_0}, \theta_{x_1}$ are group-specific thresholds chosen so that $\widehat{Y}$ satisfies the constraint $\text{TV}_{x_0,x_1}(\widehat{y}) = 0$, and that $\theta_{x_0}, \theta_{x_1}$ are as close as possible to 0.5 (to minimize the loss in accuracy). An important question is whether the $\widehat{Y}$ constructed in this way also behaves well from a causal perspective.

5.2.3 In-processing

In-processing methods take a different route, and instead of building on the unconstrained predictions, they attempt to incorporate a fairness constraint into the learning process. This in effect means that the mechanism $f_{\widehat{Y}}$ is no longer unconstrained, but is required to lie within a class of functions that satisfy the TV constraint. Broadly speaking, this is achieved by formulating an optimization problem of the form

$$\begin{aligned}
&\text{arg min}_{f_{\widehat{Y}}} && L\big(Y, f_{\widehat{Y}}(x, w, z)\big) && (5.40)\\
&\text{subject to} && \text{TV}_{x_0,x_1}(\widehat{y}) \leq \epsilon, && (5.41)\\
& && ||f_{\widehat{Y}}(x, w, z) - f_{\widehat{Y}}(x', w', z')|| \leq \tau((x, w, z), (x', w', z')). && (5.42)
\end{aligned}$$

where L is a suitable loss function[4] and τ is a metric on the covariates $V \setminus Y$. In the language of Dwork *et al.*, 2012, the TV minimization constraint in

[4] A common choice here is the loss $\mathbb{E}\big[Y - f_{\widehat{Y}}(x, w, z)\big]^2$.

Eq. 5.41 ensures *group fairness*, whereas the constraint in Eq. 5.42 ensures *covariate-specific fairness*[5], meaning that predictions for individuals with similar covariates x, z, w should be similar. Exactly formulating and efficiently solving problems such as in Eqs. 5.40-5.42 constitutes an entire field of research. Due to space limitations, we do not go into full detail on how this can be achieved, but rather list a few well-known examples. Zemel *et al.*, 2013 use a clustering-based approach, whereas Zhang *et al.*, 2018 use an adversarial network approach to solve such an optimization problem. Kamishima *et al.*, 2012 add a mutual information constraint to control for the TV constraint in parametric settings. Agarwal *et al.*, 2018 formulate a saddle-point problem with moment-based constraints to achieve the desired TV minimization. The mentioned methods differ in many practical details, but all attempt to satisfy the constraint $\text{TV}_{x_0,x_1}(\widehat{y}) = 0$ by constraining the learner $f_{\widehat{Y}}$. More fundamentally, the question arises again as to whether constructing the mechanism $f_{\widehat{Y}}$ so that TV equals 0 can provide guarantees about the causal behavior of the predictor and its fairness requirements.

5.2.4 Pre-processing

Pre-processing methods start from a distribution $P(x, w, z, y)$ and find its "fair version", labeled $\widetilde{P}(x, w, z, y)$ which is then used in the learning stage. Sometimes an exact mapping between $\tau : \mathcal{V} \to \mathcal{V}$ is constructed[6], where τ can be stochastic. In that case, the transformed distribution $\widetilde{P}$ is defined as:

$$\widetilde{P}(v) = \mathbb{E}_\tau\big[P \circ \tau(v)\big]. \tag{5.43}$$

The fair pre-processing methods formulate an optimization problem that attempts to find the optimal $\widetilde{P}(V)$, where optimality is defined as minimizing some notion of distance to the original distribution $P(V)$. There are two different approaches here, which have different causal implications, namely:

(a) the protected attribute X should be independent from the rest of observables $V \setminus X$ in the fair distribution $\widetilde{P}(V)$, written $X \perp\!\!\!\perp V \setminus X$,

(b) the protected attribute X should be independent from the the outcome Y in the fair distribution $\widetilde{P}(V)$, written $X \perp\!\!\!\perp Y$.

The first approach requires that the effect of the attribute X is entirely erased from the data. The second, less stringent option requires the independence $X \perp\!\!\!\perp Y$ in $\widetilde{P}(V)$, which is equivalent to having $\text{TV}_{x_0,x_1}(\widehat{y}) = 0$.

[5] This notion corresponds to *individual fairness* in the work of Dwork *et al.*, 2012. Causally speaking, this would be seen as a *covariate-specific* fairness constraint, as the term individual is overloaded.

[6] $\mathcal{V}$ here denotes the domain in which the observables V take values.

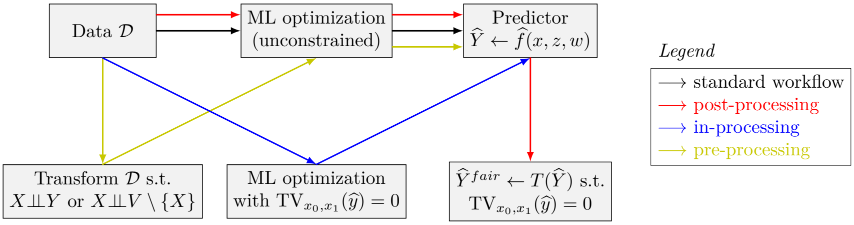

Data $\mathcal{D}$
ML optimization (unconstrained)
Predictor $\widehat{Y} \leftarrow \widehat{f}(x, z, w)$
Transform $\mathcal{D}$ s.t. $X \perp\!\!\!\perp Y$ or $X \perp\!\!\!\perp V \setminus \{X\}$
ML optimization with $\text{TV}_{x_0,x_1}(\widehat{y}) = 0$
$\widehat{Y}^{fair} \leftarrow T(\widehat{Y})$ s.t. $\text{TV}_{x_0,x_1}(\widehat{y}) = 0$
Legend
standard workflow
post-processing
in-processing
pre-processing

Figure 5.5: A schematic summary of the post-processing (red arrows), in-processing (blue), and pre-processing (yellow) fair prediction methods, compared to a typical ML workflow (black).

These two cases will be discussed separately in the remainder of the section. In Fig. 5.5 we provide a schematic representation of the three categories of fair prediction methods, and in particular how they relate to a typical (usually unfair) machine learning workflow. We next move onto formulating FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) for Task 2.

5.2.5 FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$) for Task 2.

Building on the previous definition of FPCFA(Str-{DE,IE,SE}, $\text{TV}_{x_0,x_1}(y)$), we can now state its version in the context of fair predictions:

Definition 5.5 (FPCFA continued for Task 2)**.** [Ω, $(Q_i)_{i=1}^k$, μ as before] Let $\mathcal{M} = \langle V, U, P(u), \mathcal{F} \rangle$ be the true, unobserved generative SCM, $\mathcal{A}$ a set of assumptions, and $P(v)$ the observational distribution generated by $\mathcal{M}$. Let $\Omega^{\mathcal{A}}$ be the space of all SCMs compatible with $\mathcal{A}$. The Fundamental Problem of Causal Fairness Analysis is to find a collection of measures $\mu_1, \ldots, \mu_k$ such that the following properties are satisfied:

(1) μ is decomposable w.r.t. $\mu_1, \ldots, \mu_k$;

(2) $\mu_1, \ldots, \mu_k$ are admissible w.r.t. the structural fairness criteria $Q_1, \ldots, Q_k$.

(3) $\mu_1, \ldots, \mu_k$ are as powerful as possible.

(4) $\mu_1, \ldots, \mu_k$ are identifiable from the observational distribution $P(v)$ and class $\Omega^{\mathcal{A}}$.

The final step of FPCFA for Task 2 is to construct an alternative SCM $\mathcal{M}'$ such that

(5) the measures $\mu_1, \ldots, \mu_k$ satisfy that

$$\mu_1(\mathcal{M}') = \cdots = \mu_k(\mathcal{M}') = 0. \tag{5.44}$$

Commonly, the alternative SCM $\mathcal{M}'$ will differ from the original SCM $\mathcal{M}$ in the $f_{\widehat{Y}}$ mechanism, which needs to be constructed in a fair way. More explicitly, we want to ensure that the constructed predictor $\widehat{Y}$ satisfies

$$x\text{-DE}_x^{\text{sym}}(\widehat{y} \mid x_0) = x\text{-IE}_x^{\text{sym}}(\widehat{y} \mid x_0) = x\text{-SE}_{x_1,x_0}(\widehat{y}) = 0, \tag{5.45}$$

instead of just requiring that $\text{TV}_{x_0,x_1}(\widehat{y}) = 0$. The question we address formally next is whether the conditions in Eq. 5.45 can be achieved by methods that focus on minimizing TV. For this purpose, we prove the Fair Prediction Theorem that is formulated for in-processing methods in the linear case:

Theorem 5.1 (Fair Prediction Theorem). Let $\text{SFM}(n_Z, n_W)$ be the standard fairness model with $|Z| = n_Z$ and $|W| = n_W$. Let E denote the set of edges of $\text{SFM}(n_Z, n_W)$. Further, let $\mathcal{S}^{linear}_{n_Z,n_W}$ be the space of linear structural causal models (with the exception of X variable which is Bernoulli) compatible with the $\text{SFM}(n_Z, n_W)$ and whose structural coefficients are drawn uniformly from $[-1,1]^{|E|}$. An SCM $\mathcal{M} \in \mathcal{S}^{linear}_{n_Z,n_W}$ is said to be ϵ-TV-compliant if

$$\widehat{f}_{\text{fair}} = \underset{f \text{linear}}{\arg\min} \quad \mathbb{E}[Y - f(X,Z,W)]^2 \tag{5.46}$$

$$\text{subject to} \quad TV_{x_0,x_1}(f) = 0 \tag{5.47}$$

also satisfies

$$|x\text{-DE}_{x_0,x_1}(\widehat{f}_{\text{fair}} \mid x_0)| \leq \epsilon, \tag{5.48}$$

$$|x\text{-IE}_{x_1,x_0}(\widehat{f}_{\text{fair}} \mid x_0)| \leq \epsilon, \tag{5.49}$$

$$|x\text{-SE}_{x_1,x_0}(\widehat{f}_{\text{fair}})| \leq \epsilon. \tag{5.50}$$

Under the Lebesgue measure over $[-1,1]^{|E|}$, the set of 0-TV-compliant SCMs in $\text{SFM}(n_Z, n_W)$ has measure 0. Furthermore, for any n_Z, n_W, there exists an $\epsilon = \epsilon(n_Z, n_W)$ such that

$$\mathbb{P}(\mathcal{M} \text{ is } \epsilon\text{-TV-compliant}) \leq \frac{1}{4}. \tag{5.51}$$

The proof is given in Appendix A.3. The theorem states that, for a random linear SCM, the optimal fair predictor with TV measure equal to 0 will almost never have the x-specific fairness measures equal to 0. The remarkable implication of this result is that minimizing the TV measure provides no guarantees whatsoever that the direct, indirect, and spurious effects are also minimized. In other words, the resulting classifier that is deemed fair may not be fair after all.

The Fair Prediction Theorem considers the linear case for in-processing methods, but we conjecture that it has implications for more complex settings too (see also empirical evidence on real data below). For example, note that in the optimization problem in Eqs. 5.46-5.47, we are searching over linear functions f of X, Z, and W. For pre-processing methods that achieve $X \perp\!\!\!\perp \widehat{Y}$, the space of allowed functions f would be even more flexible, but the underlying optimization problem would remain similar. These observations raise a serious concern about whether any of the fair prediction methods in the literature provide predictors that are well-behaved in a causal sense. We now exemplify this point empirically by applying several well-known fair prediction methods to the COMPAS dataset.

5.2.6 Empirical implications of the Fair Prediction Theorem

Consider the following example based on the COMPAS dataset.

Example 5.4 (COMPAS continued for Fair Prediction). A team of data scientists from ProPublica have shown that the COMPAS dataset from Broward County provides evidence of a strong racial bias against minorities (see Ex. 5.3). They are now interested in producing fair predictions $\widehat{Y}$ on the same dataset in order to replace the biased predictions. To this end they implement:

(i) **baseline:** a random forest classifier (Breiman, 2001; Wright *et al.*, 2020) trained without any fairness constraints,

(ii) **pre-processing:** *reweighing* approach of Kamiran and Calders, 2012, which introduces sample weights to guarantee TV minimization,

(iii) **in-processing:** *fair reductions* approach of Agarwal *et al.*, 2018 with a logistic regression base classifier,

(iv) **post-processing:** a random forest classifier trained without fairness constraints, with *reject-option* post-processing applied (Kamiran *et al.*, 2012).

The goal of fair prediction algorithms (ii), (iii), and (iv) is to ascertain that the TV measure equals 0. After constructing these predictors, the team compute the TV measure for each of them (represented in the red bar in each of the subplots in Fig. 5.6). Upon seeing the TV values, they conclude that the algorithms were successful at reducing the TV measure (at this stage, they are only aware of the TV values, indicated by a vertical dashed line in each subplot).

However, having heard of the Fair Prediction Theorem, the team also makes use of the Fairness Cookbook in Alg. 5.1 in order to see if discrimination is also removed from a causal viewpoint. Following the steps of the Fairness Cookbook, they compute measures of direct, indirect, and spurious discrimination. The

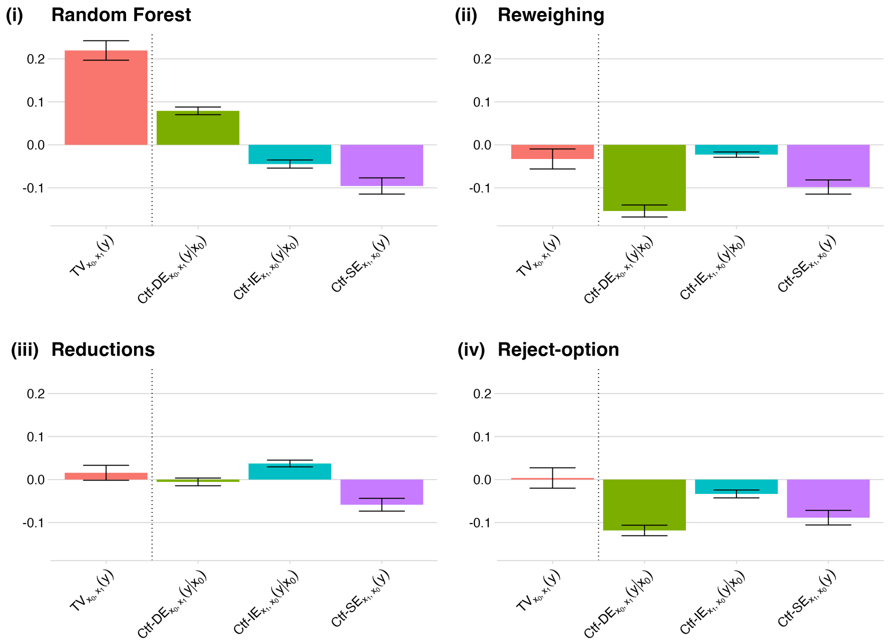


Figure 5.6: Causal Fairness Analysis applied to a standard prediction method (random forest, subfigure (i)) and three different fair prediction algorithms: reweighing (Kamiran and Calders, 2012) in subfigure (ii), reductions (Agarwal *et al.*, 2018) in subfigure (iii), and reject-option (Kamiran *et al.*, 2012) in subfigure (iv). All of the fair predictions methods reduce the TV measure, but fail to nullify the causal measures of fairness. Confidence intervals of the measures, obtained using bootstrap, are shown as vertical bars.

obtained decompositions of the TV measure are shown in Figures 5.6(ii), 5.6(iii), and 5.6(iv). The ProPublica team notes that even though all methods substantially reduce the $\text{TV}_{x_0,x_1}(\widehat{y})$, the measures of direct, indirect, and spurious effects are not necessarily reduced to 0, consistent with the Fair Prediction Theorem. They conclude that focusing only on the TV measure may in fact be misleading, since it does not guarantee that undesired forms of discrimination are removed from the predictions. □

One class of fair prediction methods that are not addressed in the discussion above are the pre-processing methods that achieve the independence of the protected attribute with all the observables, which is discussed next.

5.2.7 Pre-processing methods that achieve $X \perp\!\!\!\perp V \setminus \{X\}$

A prominent pre-processing method that achieves attribute independence ($X \perp\!\!\!\perp V \setminus \{X\}$) is the proposal of Dwork *et al.*, 2012, in which in the pre-processing step the distribution

$$V \setminus \{X\} \mid X = x_0 \text{ is transported onto } V \setminus \{X\} \mid X = x_1.$$

However, as witnessed by the following example, it may be difficult to provide guarantees that causal measure of fairness vanish for such an approach:

Example 5.5 (Failure of Optimal Transport Methods). A company is hiring prospective applicants for a new job position. Let X denote gender (x_0 for male, x_1 for female), W denotes a score on a test (taking to values, $\pm\epsilon$), Y the outcome of the application ($Y = 0$ for no job offer, $Y = 1$ for job offer). The following SCM $\mathcal{M}$ describes the data generating process:

$$\mathcal{F}, P(U) : \begin{cases} X \leftarrow U_X & (5.52) \\ W \leftarrow \epsilon(2U_W - 1) & (5.53) \\ Y \leftarrow \begin{cases} U_Y \vee \mathbb{1}(W > 0) \text{ if } X = x_0 \\ U_Y \vee \mathbb{1}(W < 0) \text{ if } X = x_1 \end{cases} & (5.54) \\ \\ U_X, U_Z, U_W, U_Y \sim \text{Bernoulli}(0.5). & (5.55) \end{cases}$$

After the first part of the selection process, the company goes through a certification process that ascertains that demographic parity is achieved, i.e.,

$$\text{TV}_{x_0,x_1}(y) = 0. \tag{5.56}$$

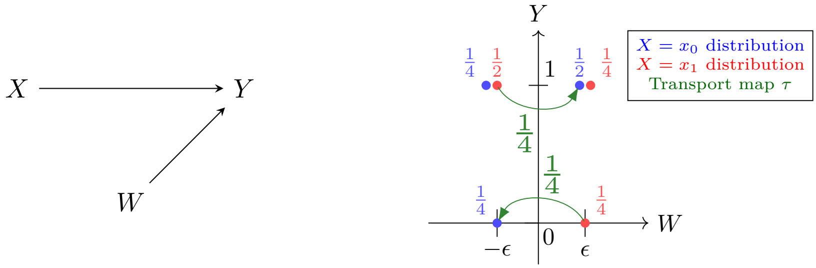
Still, the data science team is uncertain whether the process is causally fair with respect to the direct and indirect effects. For this reason, the company chooses to optimally transport the conditional distributions, namely

$$W, Y \mid x_1 \overset{\tau}{\mapsto} W, Y \mid x_0, \tag{5.57}$$

where τ denotes the optimal transport map between the two distributions. By doing so, the company aims to make sure that both the direct and the indirect effect are equal to 0.

The obtained optimal transport map τ can be described as follows:

$$\tau(w, y) = \begin{cases} (-\epsilon, 0) & \text{if } (w, y) = (\epsilon, 0) \\ (\epsilon, 1) & \text{if } (w, y) = (\epsilon, 1) \\ (\pm\epsilon, 1) \text{ w.p. } \frac{1}{2} & \text{if } (w, y) = (-\epsilon, 1) \end{cases} \tag{5.58}$$



(a) Causal graph corresponding to Example 5.5.

(b) Conditional distributions $W, Y \mid x_0$ (blue) $W, Y \mid x_1$ (red), and the optimal transport map τ (green) from Example 5.5.

Figure 5.7: Causal diagram and distribution from Ex. 5.5.

The conditional distributions $W, Y \mid x_0$ and $W, Y \mid x_1$ are shown in Fig. 5.7b, together with the optimal transport map.

Denote by $\widetilde{W}, \widetilde{Y}$ the transformed values of W, Y. After the transformation, the assignment of the value of $\widetilde{W}$ depends on $\widetilde{Y}$, which implies that the causal diagram in Fig. 5.7a is no longer valid for the transformed data (and, furthermore, the transformed data is no longer compatible with the SFM). Therefore, no causal guarantees can be provided after applying optimal transport. □

The reader may wonder about the underlying issue for why all of the discussed methods from previous literature fail from a causal perspective. We next move onto explaining the shortcomings of these methods, and give two possible formulations that can help when constructing causally meaningful predictors.

5.2.8 Towards the solution

Our next goal is to remedy the pitfalls of the fair prediction methods discussed so far. In particular, we outline a strategy for ensuring that direct, indirect, and spurious effects vanish (or a subset of them, in case of business necessity). There are two conditions that are needed to guarantee the causally fair behavior of our predictor:

(I) the causal structure of the SFM is preserved for the predictor $\widehat{Y}$,

(II) the identification expressions of x-DE, x-IE, and x-SE equal 0 in the new SCM $\mathcal{M}'$.

We first show formally that these two conditions translate into guarantees for the fairness of the constructed classifier $\widehat{Y}$:

Proposition 5.1 (Fair Predictor – Causal Conditions)**.** Let $\mathcal{M}$ be an SCM compatible with the SFM and let $\widehat{Y}$ be a predictor of the outcome Y satisfying:

(a) X, Z, W and $\widehat{Y}$ satisfy the assumptions of the SFM,

(b) the identification expressions for $x\text{-DE}^{\text{sym}}_x(y \mid x_0), x\text{-DE}^{\text{sym}}_x(y \mid x_0)$, and $x\text{-SE}_{x_1,x_0}(y)$ equal 0, by ensuring that

$$\sum_{z,w}[P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_1, z)P(z \mid x_0) = 0 \tag{5.59}$$

$$\sum_{z,w}[P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_0, z)P(z \mid x_0) = 0 \tag{5.60}$$

$$\sum_{z,w} P(y \mid x_0, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z \mid x) = 0 \tag{5.61}$$

$$\sum_{z,w} P(y \mid x_1, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z \mid x) = 0 \tag{5.62}$$

$$\sum_{z} P(y \mid x_1, z)[P(z \mid x_0) - P(z \mid x_1)] = 0. \tag{5.63}$$

Then, the predictor $\widehat{Y}$ satisfies:

$$x\text{-DE}^{\text{sym}}_x(\widehat{y} \mid x_0) = x\text{-IE}^{\text{sym}}_x(\widehat{y} \mid x_0) = x\text{-SE}_{x_1,x_0}(\widehat{y}) = 0. \tag{5.64}$$

Based on the proposition above, we provide two different strategies for constructing causally meaningful fair predictors, as discussed in the next sections.

5.2.9 Causally aware in-processing

The first strategy for constructing fair predictions that obey causal constraints is via in-processing. The natural idea is to replace the constraint $\text{TV}_{x_0,x_1}(\widehat{y}) = 0$ with a number of constraints that represent the identification expressions of the causal quantities that we wish to minimize, as described in Eqs. 5.59-5.63. After that, we can use the fact that the causal structure of the SFM is inherited for a predictor $\widehat{Y}$ constructed with in-processing. The formal statement showing the validity of the in-processing approach is given in the following result:

Theorem 5.2 (In-processing with Causal Constraints)**.** Let $\mathcal{M}$ be an SCM compatible with the SFM. Let $\widehat{Y}$ be constructed as the optimal solution to

$$\begin{aligned}
\widehat{Y} = \arg\min_{f} \quad & \mathbb{E}[Y - f(X, Z, W)]^2 && (5.65)\\
\text{subject to} \quad & x\text{-DE}^{\text{ID}}_{x_0,x_1}(\widehat{y} \mid x_0) = 0 && (5.66)\\
& x\text{-DE}^{\text{ID}}_{x_1,x_0}(\widehat{y} \mid x_0) = 0 && (5.67)\\
& x\text{-IE}^{\text{ID}}_{x_0,x_1}(\widehat{y} \mid x_0) = 0 && (5.68)\\
& x\text{-IE}^{\text{ID}}_{x_1,x_0}(\widehat{y} \mid x_0) = 0 && (5.69)\\
& x\text{-SE}^{\text{ID}}_{x_1,x_0}(\widehat{y}) = 0, && (5.70)
\end{aligned}$$

Algorithm 5.2 Causal Individual Fairness (Causal IF)

```
• Inputs: Dataset $\mathcal{D}$, SFM projection $\Pi_{\text{SFM}}(\mathcal{G})$, Business Necessity Set BN-set.
for $V' \in \{Z, W, Y\}$ do
    if $V' \notin$ BN-set then
        transport $V' \mid x_1, \tau^{\text{pa}(V')}(\text{pa}(V'))$ onto $V' \mid x_0, \text{pa}(V')$
        let $\tau^{V'}$ denote the transport map
    else if $V' \in$ BN-set then
        transport $V' \mid x, \tau^{\text{pa}(V')}(\text{pa}(V'))$ onto $V' \mid x, \text{pa}(V')$ for $x \in \{x_0, x_1\}$
        let $\tau^{V'}$ denote the transport map
    end if
end for
```

where x-DE$^{\text{ID}}$, x-IE$^{\text{ID}}$, and x-SE$^{\text{ID}}$ represent the identification expressions of the corresponding measures (as shown in Prop. 5.1). Then, the predictor $\widehat{Y}$ satisfies

$$x\text{-DE}^{\text{sym}}_x(\widehat{y} \mid x_0) = x\text{-IE}^{\text{sym}}_x(\widehat{y} \mid x_0) = x\text{-SE}_{x_1,x_0}(\widehat{y}) = 0. \tag{5.71}$$

The following remark shows that the result of the theorem holds even more broadly than just for the standard fairness model:

Remark 5.1 (Robustness of In-processing with Causal Constraints)**.** Thm. 5.2 is stated for an SCM that is compatible with the SFM. However, such an assumption can be relaxed. In particular, the result of the theorem remains true even if the bidirected edges $X \dashleftarrow\dashrightarrow Y$, $Z \dashleftarrow\dashrightarrow Y$, and $W \dashleftarrow\dashrightarrow Y$ are present in the original model.

5.2.10 Causally aware pre-processing

We now discuss a strategy based on pre-processing that is related to the optimal transport approach of Dwork *et al.*, 2012, which we call Causal Individual Fairness, described in Alg. 5.2. Let the business necessity set be denoted as BN-set, taking values

$$\text{BN-set} \in \big\{\emptyset, \{Z\}, \{W\}, \{Z, W\}\big\}. \tag{5.72}$$

The algorithm performs sequential optimal transport of the distributions of Z, W, and Y (in this topological ordering) conditional on the values of the parental set. In particular, Causal IF procedure starts by optimally transporting

$$Z \mid X = x_1 \text{ onto } Z \mid X = x_0,$$

unless Z is in the business necessity set. Let τ^Z denote the optimal transport map. Then, the distribution of W is transported in the next step, namely,

$$W \mid X = x_1, Z = \tau^Z(z) \text{ onto } W \mid X = x_0, Z = z \;\; \forall z.$$

In the final step, the distribution of Y is transported

$$Y \mid X = x_1, Z = \tau^Z(z), W = \tau^W(w) \text{ onto } Y \mid X = x_0, Z = z, W = w \;\; \forall z, w.$$

Theorem 5.3 (Soundness Causal Individual Fairness)**.** Let $\mathcal{M}$ be an SCM compatible with the SFM and τ^Y be the optimal transport map obtained when applying Causal IF (Alg. 5.2). Define an additional mechanism of the SCM $\mathcal{M}$ such that

$$\widetilde{Y} \leftarrow \tau^Y(Y; X, Z, W). \tag{5.73}$$

For the transformed outcome $\widetilde{Y}$, we can then claim:

$$\text{if } Z \notin \text{BN-set} \implies x\text{-SE}_{x_1,x_0}(\widetilde{y}) = 0. \tag{5.74}$$
$$\text{if } W \notin \text{BN-set} \implies x\text{-IE}_x^{\text{sym}}(\widetilde{y} \mid x_0) = 0. \tag{5.75}$$

Furthermore, the transformed outcome $\widetilde{Y}$ also satisfies

$$x\text{-DE}_x^{\text{sym}}(\widetilde{y} \mid x_0) = 0. \tag{5.76}$$

This theorem's proof can be found in Appendix A.4. After showing that the Causal IF procedure provides certain guarantees for the causal measures of fairness, we go back to Ex. 5.5 to understand why the method of joint optimal transport fails to produce a causally meaningful predictor:

Remark 5.2 (Why Joint Optimal Transport Fails)**.** Recall that in Ex. 5.5 we showed that the optimal transport map, in general, may break the causal structure in the data, and after applying it, no causal guarantees can be made. Causal IF (Alg. 5.2), on the other hand, applies optimal transport sequentially, by making sure that the causal structure encoded in the SFM is preserved – which in turn allows for causal fairness guarantees.

5.3 Task 3. Fair Decision-Making

After introducing and discussing Task 2, which focused on constructing fair predictions, we now move onto Task 3, which is concerned with fair decision-making. In particular, it is important to clarify what the gap between Tasks 2 and 3 is. In Task 2, we focused on producing predictions $\widehat{Y}$ for an outcome Y, such that they

(i) produce a small loss $L(Y, \widehat{Y})$ with respect to a loss function L,

(ii) satisfy a set of causal fairness constraints, which we denote with $\mathcal{C}(\widehat{Y})$.

In Task 3, on the other hand, we are interested not only in making predictions, but also in making decisions in practice. The gap between the two tasks will often be encapsulated in a user-specified utility function U, that may account for utility terms not directly specified in the dataset.[7]

Definition 5.6 (Utility Function). A utility function U is a function of the decision policy D, covariates X, Z, W, and Y to the real line $\mathbb{R}$. We write $U(D; X, Z, W, Y)$ for the function, and sometimes $U(D)$ when it is clear from the context which covariates are used.

For concreteness, we now give an example which attempts to highlight the distinction between the two tasks:

Example 5.6 (College Admissions I: Decision-Making). A university is deciding on admissions of prospective applicants. Let X denote gender (x_0 for female, x_1 for male), W_1 denotes the SAT score, W_2 denotes the student's score on the admission exam. Suppose the following SCM $\mathcal{M}^*$ describes this setting:

$$\mathcal{F}^*, P^*(U) : \begin{cases} X \leftarrow U_X & (5.77) \\ W_1 \leftarrow U_W + X & (5.78) \\ W_2 \leftarrow \begin{cases} W_1 + 2(1 - X) \text{ if } W_1 > 0.5, \\ W_1 \ \text{ otherwise.} \end{cases} & (5.79) \\ \widehat{Y} \leftarrow f_{\widehat{Y}}(X, W_1, W_2) & (5.80) \\ D \leftarrow f_D(X, W_1, W_2, \widehat{Y}), & (5.81) \\ \\ U_X \sim \text{Bernoulli}(0.8), U_W \sim \text{Unif}[0, 1], & (5.82) \end{cases}$$

where $\widehat{Y}$ is the predicted GPA of the student, and D the final admission decision. In words, $\mathcal{M}^*$ can be interpreted as follows. Due to a societal bias, females (x_0) are discouraged from STEM subjects, causing lower SAT scores (Eq. 5.78). However, female students who nonetheless perform well on the SAT ($W_1 > 0.5$) are additionally encouraged and perform even better on the admission test (Eq. 5.79). Male candidates perform the same on the admission test as they have on the SAT tests, regardless of SAT score.

Based on the above example, we can clarify the distinction between the tasks:

[7] One viewpoint on the distinction between the tasks is that Task 2 is a specific subcase of Task 3 in which the utility function U only depends on the prediction loss between Y and $\widehat{Y}$.

(T2) The aim of Task 2 is to produce predictions $\widehat{Y}$, that is, construct the $f_{\widehat{Y}}$ mechanism based on the available information X, W_1, W_2. Furthermore, the predictions should satisfy a fairness constraint, for example, contain no direct effect of gender X:

$$\mathbf{NDE}_{x_0,x_1}(\widehat{y}) = \mathbf{NDE}_{x_1,x_0}(\widehat{y}) = 0. \tag{5.83}$$

The prediction task is often tied to a specific mechanism, in this case f_Y (that exists in reality but is not exemplified in the SCM above), that encodes how individual's GPA depends on their test scores and gender.

(T3) The aim of Task 3 is to decide which applicants should get admitted, that is, to construct the decision mechanism f_D. The decision policy D may need to take into account different types of utility, such as the university long-term reputation (based on overall success rate of completing the studies), the total expected income from tuition fees, or representation of protected groups. Furthermore, the decision policy may also be required to satisfy a desired notion of fairness, for example,

$$\mathbf{NDE}_{x_0,x_1}(d) = \mathbf{NDE}_{x_1,x_0}(d) = 0. \tag{5.84}$$

□

After clarifying the gap between the tasks, we discuss some possible approaches for Task 3. One initial approach to solve Task 3 might be to simply use the predictions $\widehat{Y}$ constructed in Task 2, and construct fair decisions based on fair predictions. As it turns out, this approach is possible only under two special scenarios:

Proposition 5.2 (Chaining Fair Predictions into Fair Decisions)**.** Let μ be a fairness measure defined by a contrast $\mathcal{C}$ of the form (C_1, C_0, E_1, E_0). Suppose that a predictor $\widehat{Y}$ is fair w.r.t. μ, that is, $\mu(\widehat{y}) = 0$. Furthermore, suppose that a decision policy D is constructed as a transformation of $\widehat{Y}$, i.e.,

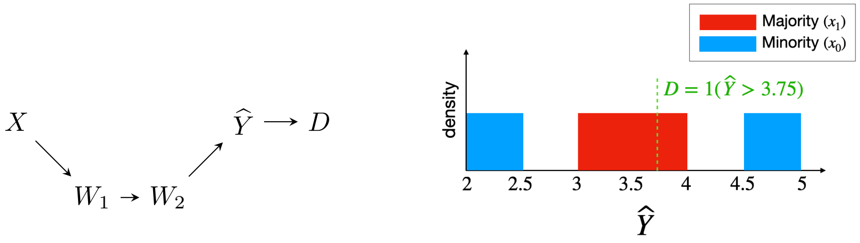
$$D := f_D(\widehat{Y}). \tag{5.85}$$

Then, D is fair with respect to μ if one of the following conditions hold:

(a) the function f_D is linear, or

(b) the measure μ is a unit level measure.

The proof of the proposition is given in Appendix A.5. As the next example illustrates, other transformations of fair predictions do not necessarily preserve the fairness condition satisfied by the baseline predictor:



(a) Causal diagram of the SCM in Ex. 5.7.

(b) Density of $\widehat{Y}$ and the decision policy D.

Figure 5.8: Causal diagram and the density of the GPA predictor $\widehat{Y}$ for Ex. 5.7.

Example 5.7 (College Admissions Continued: Thresholding)**.** Consider the SCM from Ex. 5.6, with the mechanisms $f_{\widehat{Y}}, f_D$ instantiated by the university as:

$$\widehat{Y} \leftarrow W_2 + 2 \tag{5.86}$$

$$D \leftarrow \mathbb{1}(\widehat{Y} > 3.75), \tag{5.87}$$

The mechanisms were obtained after the selection committee realized that the average admission test score is the same for male and female candidates, and therefore decided to create the fair predictor $\widehat{Y}$ as simply a function of W_2 (Eq. 5.86). The admission decision D is then simply based on thresholding the values of the predictor $\widehat{Y}$, and was chosen so that a fixed number of candidates are admitted. The natural indirect effect for the predictor $\widehat{Y}$ can be computed as

$$\text{NIE}_{x_1,x_0}(\widehat{y}) = 0. \tag{5.88}$$

However, we want to verify whether the decisions D obtained from the thresholding operation in Eq. 5.87 are also fair with respect to indirect effect. In particular, we want to compute the NIE

$$\text{NIE}_{x_1,x_0}(d) = P(d_{x_1,W_{x_0}}) - P(d_{x_1}). \tag{5.89}$$

We now compute the first term:

$$\begin{aligned}
P(d_{x_1,W_{x_0}}) &= P(\widehat{y}_{x_1,W_{x_0}} > 3.75) = P((W_2)_{x_0} > 1.75) && (5.90)\\
&= P((W_2)_{x_0} > 1.75 \mid (W_1)_{x_0} \leq 0.5)P((W_1)_{x_0} \leq 0.5) && (5.91)\\
&+ P((W_2)_{x_0} > 1.75 \mid (W_1)_{x_0} > 0.5)P((W_1)_{x_0} > 0.5) && (5.92)\\
&= P((W_1)_{x_0} > 1.75)\frac{1}{2} + P((W_1)_{x_0} + 2 > 1.75)\frac{1}{2} && (5.93)\\
&= P(U_W > 1.75) \cdot \frac{1}{2} + P(U_W > -0.25) \cdot \frac{1}{2} = \frac{1}{2}. && (5.94)
\end{aligned}$$

Similarly, we also compute the second term:

$$P(d_{x_1}) = P(\widehat{y}_{x_1} > 3.75) = P((W_2)_{x_1} > 1.75) \tag{5.95}$$

$$= P((W_1)_{x_1} > 1.75) = P(U_W > 0.75) = \frac{1}{4}. \tag{5.96}$$

Therefore, we have obtained that

$$\text{NIE}_{x_1,x_0}(d) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \tag{5.97}$$

The density of the scores for both groups is visualized in Fig. 5.8b, which shows that the mean predicted GPA is equal between the groups, while choosing a threshold of $\widehat{Y} > 3.75$ results in a non-zero indirect effect. □

After all, a fairness criterion satisfied by a predictor $\widehat{Y}$ is not necessarily satisfied by a policy D constructed following this predictor.

5.3.1 Decision-Making as In-processing

To remedy the issue highlighted above, we propose optimizing the utility function directly subject to causal constraints of fairness. This formulation of the problem is captured in the following theorem, closely related to Thm. 5.2:

Theorem 5.4 (In-processing with Causal Constraints)**.** Let $\mathcal{M}$ be an SCM compatible with the SFM. Let $U(D; X, Z, W, Y)$ be the utility function for a possible policy D. Let the decision policy D^* be constructed as the optimal solution to

$$D^* = \arg\max_{D} \quad \mathbb{E}[U(D; X, Z, W, Y)] \tag{5.98}$$

$$\text{subject to} \quad x\text{-DE}^{\text{ID}}_{x_0,x_1}(d \mid x_0) = 0 \tag{5.99}$$

$$x\text{-DE}^{\text{ID}}_{x_1,x_0}(d \mid x_0) = 0 \tag{5.100}$$

$$x\text{-IE}^{\text{ID}}_{x_0,x_1}(d \mid x_0) = 0 \tag{5.101}$$

$$x\text{-IE}^{\text{ID}}_{x_1,x_0}(d \mid x_0) = 0 \tag{5.102}$$

$$x\text{-SE}^{\text{ID}}_{x_1,x_0}(d) = 0 \tag{5.103}$$

where x-DE$^{\text{ID}}$, x-IE$^{\text{ID}}$, and x-SE$^{\text{ID}}$ represent the identification expressions of the corresponding measures (as shown in Prop. 5.1). Then the decision policy D^* satisfies

$$x\text{-DE}^{\text{sym}}_{x}(d^* \mid x_0) = x\text{-IE}^{\text{sym}}_{x}(d^* \mid x_0) = x\text{-SE}_{x_1,x_0}(d^*) = 0, \tag{5.104}$$

where the x-DE$^{\text{sym}}_x$, x-IE$^{\text{sym}}_x$ are the symmetric DE and IE effects from Def. 5.4.

This formulation guarantees to remove direct, indirect, and spurious effects from the decision policy D. In general, however, under considerations of business necessity, one might wish to relax this formulation, and only guarantee that a subset of these effects is equal to 0. In the next section, we focus on a specific utility function, that has received attention in the literature, in the context of decision-making.

5.3.2 Explicit Trade-off of Utility and Fairness

A common challenge described in the fairness literature is about achieving a high utility of a decision policy, while increasing the number of individuals from the discriminated group who are favored by the policy. This balance between the utility of a decision with respect to the true outcome, and the utility of favoring a discriminated group is sometimes called the fairness-accuracy trade-off (Friedler *et al.*, 2019)[8]. Motivated by this trade-off, a commonly considered class of utility functions that explicitly trade between the two types of utility take the form:

$$U(D; X, Z, W, Y) = R(D, Y) + \lambda \mathbb{1}(X = x_0)D. \tag{5.105}$$

The two terms in Eq. 5.105 represent (1) the utility of the decision with respect to true outcome Y, measured by the reward term $R(D, Y)$ and (2) the utility of favoring an individual belonging to the protected group x_0, measured by the term $\lambda \mathbb{1}(X = x_0)D$. The tuning parameter λ describes the trade-off between the two objectives. For $\lambda = 0$, the utility function is only focused on the reward with respect to outcome, whereas as $\lambda \to \infty$, positive decisions for individuals belonging to the protected group are increasingly important.

Example 5.8 (College Admissions: Explicit Trade-off)**.** Let X denote the protected attribute, Y the high school GPA, and D the university admission decision. Consider the utility function

$$U(D; X, Y) = -\underbrace{(Y-4)^2}_{\text{reward}} + \underbrace{\lambda \mathbb{1}(X = x_0 \wedge D = 1)}_{\text{minority group preference}} \tag{5.106}$$

that balances the reward associated with each student's success (measured by a squared loss from a perfect GPA) with the utility of admitting students from the minority group (giving a fixed utility of λ for each admitted student from the x_0 group). The first component of the utility ensures that students finish university with a good success rate (important for the university reputation), while the second term ensures diversity and minority representation. □

[8] In practice, the trade-off between utility and fairness may be small or negligible in certain settings, for instance in the context of equality of odds (Dutta *et al.*, 2020; Rodolfa *et al.*, 2021).

Even though utility functions discussed above can trade between reward and fairness, the question still remains on how the optimal decision policy D, with respect to a utility function U, behaves from a causal perspective. Here we recall the Fair Prediction Theorem (Thm. 5.1). The theorem can be leveraged in the context of decision-making, too. In particular, we give the following corollary of the theorem, and then explain its implications:

Corollary 5.5 (Fair Decision-Making). [SFM(n_Z, n_W), $\mathcal{S}^{linear}_{n_Z,n_W}$ as in Thm. 5.1] Let $\mathcal{M}$ be sampled from the space of linear SCMs with coefficients drawn from the symmetric hypercube (i.e., according to the model $\mathcal{S}^{linear}_{n_Z,n_W}$). Let the utility function be given as

$$U(D; X, Z, W, Y) = -[Y - D(X, Z, W)]^2 + \lambda \mathbb{1}(X = x_0)D. \tag{5.107}$$

Let $D^{\max}$ denote the maximum utility policy, which solves the problem

$$D^{\max} = \underset{D \text{ linear}}{\arg\min} \quad \mathbb{E}[U(D; X, Z, W, Y)]. \tag{5.108}$$

Further, let d^{CF} denote the causally fair maximum utility policy, which solves the problem

$$\begin{aligned} D^{\text{CF}} = \underset{D \text{ linear}}{\arg\min} \quad & \mathbb{E}[U(D; X, Z, W, Y)] && (5.109)\\ \text{subject to} \quad & x\text{-DE}_{x_0,x_1}(d \mid x_0) = 0 && (5.110)\\ & x\text{-IE}_{x_0,x_1}(d \mid x_0) = 0, && (5.111)\\ & x\text{-SE}_{x_0,x_1}(d) = 0. && (5.112) \end{aligned}$$

Then, we can claim that

$$\exists \epsilon(n_Z, n_W) > 0 \text{ s.t. } \mathbb{P}(U(D^{\max}) - U(D^{\text{CF}}) > \epsilon(n_Z, n_W)) \geq \frac{3}{4}. \tag{5.113}$$

In words, the corollary implies that for a randomly sampled SCM $\mathcal{M}$, the maximum utility policy $D^{\max}$ will have, with high probability, a higher utility (by some ϵ-margin) than the causally fair policy D^{CF}, for any choice of parameter λ. Interestingly, an analogous result was obtained by (Nilforoshan *et al.*, 2022) for a different class of models, namely models in which variables have discrete distributions over the unit simplex, rather than considering linear SCMs analyzed in this manuscript. Since the utility function already captures the benefit allocated to the minority group, and causal constraints seemingly impede this utility, the authors interpret this result to mean that causal fairness constraints are not useful in practice. This viewpoint, however, does not capture the entire complexity of the problem. As illustrated by the following example, our interpretation of the result is that admitting minority applicants is not the same as removing discrimination as it happened in the real world:

Example 5.9 (College Admissions III: Who is Who?). A university is deciding on admissions of prospective applicants. The information available to the selection committee is the following. Let X denote race (x_0 for minority groups, x_1 for majority group), W denotes the SAT score, Z denotes the socio-economic status of the family of the student ($Z = 0$ for low-income, $Z = 1$ for high-income). Let D be the decision whether to admit an applicant. Suppose that the following SCM $\mathcal{M}^*$ describes the situation:

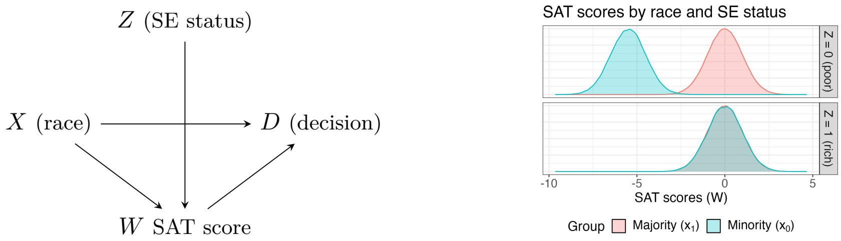
$$\mathcal{F}^*, P^*(U) : \begin{cases} X \leftarrow U_X & (5.114) \\ Z \leftarrow U_Z & (5.115) \\ W \leftarrow U_W - 5(1-Z)(1-X) & (5.116) \\ D \leftarrow f_D(X, W), & (5.117) \\ & \\ U_X \sim \text{Bernoulli}(0.8), & \\ U_Z \sim \text{Bernoulli}(0.3), & (5.118) \\ U_W \sim N(0,1). & \end{cases}$$

The SCM can be described as follows. The population consists of 80% of applicants from the majority group (x_1), and 30% of the applicants come from high socio-economic background ($Z = 1$, and for simplicity, we assume that socio-economic status and race are independent). Due to an existing societal bias regarding school funding, minority group applicants (x_0) from low income families ($Z = 0$) have lower SAT scores on average, compared to their majority group counterparts. For the high-income families, there is no difference in the SAT scores between the majority and minority groups, since high income allows access to better schooling. The causal diagram associated with the decision-making process is shown in Fig. 5.9a.

The difference between the groups in terms of the SAT scores (stratified by socio-economic status) is shown in Fig. 5.9b, indicating no racial discrimination within the high-income group, but evidence of discrimination within the low-income group. In this university, the probability of finishing the degree is proportional to the SAT score, and the utility function can be written as

$$U(D; X, Z, W) = W + \lambda \mathbb{1}(X = x_0 \wedge D = 1). \tag{5.119}$$

The university decides to choose the parameter λ such that exactly 10% of the applicants from each of the groups are admitted. For the majority group, this simply results in admitting all applicants who score better than 90% of peers in their group. However, for the minority group, there is a difference between the maximum utility policy and the causally fair policy. In particular, the maximum utility policy is based on the current, factual SAT scores W, while the causally fair policy considers the counterfactual SAT scores that would be



(a) Causal diagram of Ex. 5.9.

(b) SAT scores from Ex. 5.9.

Figure 5.9: SFM and SAT scores visualization in Ex. 5.9.

obtained in the absence of discrimination, W_{x_1}. Notice that the latter scores in particular equal

$$W_{x_1} = W + 5 \cdot \mathbb{1}(Z = 0)\mathbb{1}(X = 0), \tag{5.120}$$

meaning that the counterfactual values of the SAT scores would be higher for those in the minority group who are from low-income families. Therefore, the two resulting policies can be described as follows:

(1) $D^{\mathbf{max}}$: The policy admits all applicants from the minority group that come from a high socio-economic status, and no applicants from a low status,

(2) $D^{\mathbf{CF}}$: The policy admits the top 10% of applicants from a high socio-economic status, and also the top 10% from the low socio-economic status. Such a policy takes into account that the very best students from low-income families were actually discriminated, since in the absence of high family income, their race was a cause of poorer access to good schooling, also reflected in the fact that their counterfactual outcomes (in absence of racial discrimination) would show better performance on SATs, as indicated by Eq. 5.120.

Therefore, the $D^{\max}$ policy admits individuals from high-income backgrounds, regardless of the fact that they were not discriminated (due to privileged socio-economic status) and that they perform poorly compared to their peers of the same socio-economic background. The D^{CF} policy, however, does not admit applicants who perform poorly despite the fact they were not discriminated. Instead, it admits applicants who perform exceptionally well compared to their group, but were discriminated in their access to good schooling. □

The above example highlights the tension between admitting minority applicants and admitting applicants who were actually discriminated, which are not always the same objectives. This fundamental tension is captured in Cor. 5.5. The above example is a simplification of the issues that appear in practice, and is constructed to highlight the fundamental tension between minority representation and actual discrimination removal. Such issues need to be considered on a case-by-case basis and take into account further societal considerations.

5.3.3 Outcome Control & Principal Fairness

In this section, we introduce the setting of *outcome control*, characterized by a decision D which precedes the outcome of interest Y. The decision D (sometimes also referred to as *treatment*) is assumed to be binary, with d_1 indicating the more favorable decision, and d_0 the less favorable one. Similarly, the outcome Y is also assumed to be binary[9], with y_1 being more desirable than y_0. The setting of outcome control appears across a broad range of applications, from clinical decision-making (Hamburg and Collins, 2010) and public health (Insel, 2009), to criminal justice (Larson *et al.*, 2016) and various welfare interventions (Coston *et al.*, 2020). Formally, we are interested in the following decision-making task:

Definition 5.7 (Decision-Making Optimization)**.** Let $\mathcal{M}$ be an SCM compatible with the SFM. Given a fixed budget b, the optimal decision problem is defined as finding the (possibly stochastic) solution to the following optimization problem:

$$\begin{aligned} D^* = \operatorname*{arg\,max}_{D(x,z,w)} \quad & \mathbb{E}[Y_D] && (5.121) \\ \text{subject to} \quad & P(d) \le b. && (5.122) \end{aligned}$$

Outcome control is also characterized by the fact that the institution (i.e., the decision-maker) and the individual receiving treatment have the same utility function. In this section, we will focus on a cancer surgery example, in which both the clinician and the patient are interested in maximizing the patient's survival by means of performing a surgery. In contrast to this, a setting that does not fall under outcome control is that of issuing bank loans – the institution (in this case the bank) can safely ignore the utility of individuals who are not given a loan, while the clinician does not ignore the utility of patients who are not given surgery. Therefore, not all decision-making settings fall under outcome control, with loan approvals being an example.

In the sequel, we also discuss a fairness criterion called *principal fairness* (Imai and Jiang, 2020), that is intuitively appealing in the contexts of decision-making when the decision may influence the outcome of interest. However, as

[9] This assumption is made for simplicity of exposition, and most of the discussion in the section does not rely on it.

we will see shortly, there are several shortcomings of this definition, which can be circumvented by an alternative approach, which we call *benefit fairness*.

Oracle's Perspective

We start by introducing the cancer surgery example through the perspective of an all-knowing oracle:

Example 5.10 (Cancer Surgery – Oracle's Perspective). A clinical team has access to information about the sex of cancer patients ($X = x_0$ male, $X = x_1$ female) and their degree of illness severity determined from tissue biopsy ($W \in [0, 1]$). They wish to optimize the 2-year survival of each patient (Y), and the decision $D = 1$ indicates whether to perform surgery. The following SCM $\mathcal{M}^*$ describes the data generating mechanisms (unknown to the team):

$$\mathcal{F}^*, P^*(U): \begin{cases} X \leftarrow U_X & (5.123) \\ W \leftarrow \begin{cases} \sqrt{U_W} \text{ if } X = x_0, \\ 1 - \sqrt{1 - U_W} \text{ if } X = x_1 \end{cases} & (5.124) \\ D \leftarrow f_D(X, W) & (5.125) \\ Y \leftarrow \mathbb{1}(U_Y + \frac{1}{3}WD - \frac{1}{5}W > 0.5). & (5.126) \\ U_X \in \{0, 1\}, P(U_X = 1) = 0.5, & (5.127) \\ U_W, U_Y \sim \text{Unif}[0, 1], & (5.128) \end{cases}$$

where the f_D mechanism is constructed by the team.

The clinical team has access to an oracle that is capable of predicting the future perfectly. In particular, the oracle tells the team how each individual would respond to surgery. That is, for each unit $U = u$ (of the 500 units), the oracle returns the values of

$$Y_{d_0}(u), Y_{d_1}(u). \tag{5.129}$$

Having access to this information, the clinicians quickly realize how to use their resources. In particular, they notice that for units for whom $(Y_{d_0}(u), Y_{d_1}(u))$ equals $(0, 0)$ or $(1, 1)$, there is no effect of surgery, since they will (or will not) survive regardless of the decision. They also notice that surgery is harmful for individuals for whom $(Y_{d_0}(u), Y_{d_1}(u)) = (1, 0)$. These individuals would not survive if given surgery, but would survive otherwise. Therefore, they ultimately decide to treat 100 individuals who satisfy

$$(Y_{d_0}(u), Y_{d_1}(u)) = (0, 1), \tag{5.130}$$

since these individuals are precisely those whose death can be prevented by surgery. They learn there are 100 males and 100 females in the $(0, 1)$-group, and thus, to be fair with respect to sex, they decide to treat 50 males and 50 females. □

The space of units corresponding to the above example is represented in Fig. 5.10. The groups described by different values of $Y_{d_0}(u), Y_{d_1}(u)$ in the example are known as *canonical types* (Balke and Pearl, 1994) or *principal strata* (Frangakis and Rubin, 2002). Two groups cannot be influenced by the treatment decision (called “Safe” and “Doomed”, see Fig. 5.10). The third group represents those who are harmed by treatment (called “Harmed”). Finally, the last group represents exactly those for whom the surgery is life-saving, which is the main goal of the clinicians (this group is called “Helped”). This example illustrates how, in presence of perfect knowledge, the team can allocate resources efficiently. In particular, the consideration of fairness comes into play when deciding which of the individuals corresponding to the $(0, 1)$ principal stratum will be treated. Since the number of males and females in this group is equal, the team decides that half of those treated should be female. The approach described above can be seen as appealing in many applications unrelated to to the medical setting, and motivated the definition of principal fairness (Imai and Jiang, 2020).

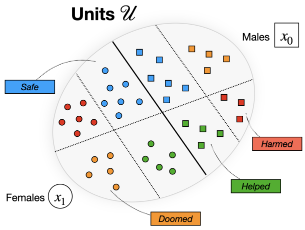


Figure 5.10: The space of units corresponding to Ex. 5.10.

Principal Fairness Definition

The discussion from above motivates a fairness criterion that was first introduced by Imai and Jiang, 2020, through the following definition:

Definition 5.8 (Principal Fairness (Imai and Jiang, 2020))**.** Let D be a decision that possibly affects the outcome Y. The pair (Y, D) is said to satisfy principal fairness if

$$P(d \mid y_{d_0}, y_{d_1}, x_1) = P(d \mid y_{d_0}, y_{d_1}, x_0), \tag{5.131}$$

for each principal stratum (y_{d_0}, y_{d_1}). That is, the decision rate should be independent of the attribute X in every principal stratum (y_{d_0}, y_{d_1}). Furthermore, define the principal fairness measure (PFM) as:

$$\text{PFM}_{x_0,x_1}(d \mid y_{d_0}, y_{d_1}) = P(d \mid y_{d_0}, y_{d_1}, x_1) - P(d \mid y_{d_0}, y_{d_1}, x_0). \tag{5.132}$$

The above notion aims to capture the intuition described in Ex. 5.10. However, unlike in the example, the definition needs to be evaluated under imperfect knowledge, when only the collected data is available[10]. An immediate cause for concern, in this context, is the joint appearance of the potential outcomes Y_{d_0}, Y_{d_1} in the definition of principal fairness. As is well known in the literature, the joint distribution of the potential outcomes Y_{d_0}, Y_{d_1} is in general impossible to obtain, which leads to the lack of identifiability of the principal fairness criterion:

Proposition 5.3 (Principal Fairness is Not Identifiable)**.** The Principal Fairness (PF) criterion from Eq. 5.131 is not identifiable from observational or experimental data.

The implication of the proposition is that principal fairness, in general, cannot be evaluated, even if an unlimited amount of data was available[11].

Monotonicity Assumption

To remedy the problem of non-identifiability of principal fairness, Imai and Jiang, 2020 propose the monotonicity assumption:

Definition 5.9 (Monotonicity)**.** We say that an outcome Y satisfies monotonicity with respect to a decision D if

$$Y_{d_1}(u) \geq Y_{d_0}(u). \tag{5.136}$$

In words, monotonicity says that for every unit, the outcome with the positive decision ($D = 1$) would not be worse than with the negative decision ($D = 0$).

[10] As discussed throughout this manuscript, and as implied by the definition of the SCM (Def. 2.1), we almost never have access to the unobserved sources of variation (u) that determine the identity of each unit.

[11] One way to see why PF is not identifiable is the following construction. Consider an SCM consisting of two binary variables $D, Y \in \{0, 1\}$ and the simple graph $D \to Y$. Suppose that we observe $P(d) = p_d$, and $P(y \mid d_1) = m_1$, $P(y \mid d_0) = m_0$ for some constants p_d, m_1, m_0 (additionally assume $m_0 \leq m_1$ w.l.o.g.). It is easy to show that these three values determine all of the observational and interventional distributions of the SCM. However, notice that for any $\lambda \in [0, 1 - m_1]$ the SCM given by

$$\begin{aligned}
D &\leftarrow U_D \quad &(5.133)\\
Y &\leftarrow \mathbb{1}(U_Y \in [0, m_0 - \lambda]) + D\mathbb{1}(U_Y \in [m_0 - \lambda, m_1]) + \quad &(5.134)\\
&\quad (1 - D)\mathbb{1}(U_Y \in [m_1, m_1 + \lambda]),\\
U_Y &\sim \text{Unif}[0, 1], U_D \sim \text{ Bernoulli}(p_d), \quad &(5.135)
\end{aligned}$$

satisfies $P(d) = p_d, P(y \mid d_1) = m_1$, and $P(y \mid d_0) = m_0$, but the joint distribution $P(y_{d_0} = 0, y_{d_1} = 1) = m_1 - m_0 + \lambda$ depends on the λ parameter and is therefore non-identifiable.

We now demonstrate how monotonicity aids the identifiability of principal fairness.

Proposition 5.4. Under the monotonicity assumption (Eq. 5.136), the principal fairness criterion is identifiable under the Standard Fairness Model (SFM).

Proof. The main challenge in PF is to obtain the joint distribution $P(y_{d_0}, y_{d_1})$, which is non-identifiable in general. Under monotonicity, however, we have that

$$Y_{d_0}(u) = 0 \wedge Y_{d_1}(u) = 0 \iff Y_{d_1}(u) = 0, \tag{5.137}$$
$$Y_{d_0}(u) = 1 \wedge Y_{d_1}(u) = 1 \iff Y_{d_0}(u) = 1. \tag{5.138}$$

Therefore, it follows from monotonicity that

$$P(y_{d_0} = 1, y_{d_1} = 0) = 0, \tag{5.139}$$
$$P(y_{d_0} = 0, y_{d_1} = 0) = P(y_{d_1} = 0), \tag{5.140}$$
$$P(y_{d_0} = 1, y_{d_1} = 1) = P(y_{d_0} = 1), \tag{5.141}$$
$$P(y_{d_0} = 0, y_{d_1} = 1) = 1 - P(y_{d_1} = 0) - P(y_{d_0} = 1), \tag{5.142}$$

thereby identifying the joint distribution whenever the interventional distributions $P(y_{d_0}), P(y_{d_1})$ are identifiable. ■

In the cancer surgery example, the monotonicity assumption would require that patients have strictly better survival outcomes when surgery is performed, compared to when it is not. Given the known risks of surgical procedures, the assumption may be rightfully challenged in such a setting. In the sequel, we argue that the assumption of monotonicity is not really necessary, and often does not help the decision-maker, even if it holds true. To fix this issue, we discuss a new definition of fairness in the setting of outcome control that suffers from neither of the above two problems but still captures the essential intuition behind PF.

Decision-Maker's Perspective

Very often we might be interested in constructing a decision-making system in which the PF criterion is satisfied (i.e., performing Task 3), rather than simply using the criterion to verify whether the decision policy is fair (i.e., performing Task 1). We now discuss a basic example of creating a decision-making policy.

Example 5.11 (Cancer Surgery - continued)**.** The team of clinicians constructs the causal diagram associated with the decision-making process, shown in

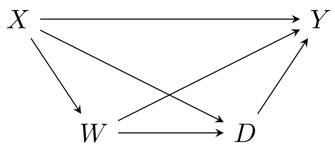


Figure 5.11: Causal diagram for Ex. 5.11 (breast cancer) and Ex. 5.13 (startup professional development courses).

Fig. 5.11. Using data from their electronic health records (EHR), they estimate the proportion of patients who benefit from the treatment based on x, w:

$$P(y_{d_1} = 1, y_{d_0} = 0 \mid w, x_1) = \frac{w}{3}. \tag{5.143}$$

$$P(y_{d_1} = 1, y_{d_0} = 0 \mid w, x_0) = \frac{w}{3}. \tag{5.144}$$

In words, at each level of illness severity $W = w$, the proportion of patients who benefit from the surgery is the same, regardless of sex. In light of this information, the clinicians decide to construct the decision policy f_D such that $f_D(W) = \mathbb{1}(W > \frac{1}{2})$. In words, if a patient's illness severity W is above $\frac{1}{2}$, the patient will receive treatment.

After implementing the policy and waiting for the 2-year follow-up period, clinicians estimate the probabilities of treatment within the stratum of those helped by surgery, and compute that

$$P(d \mid y_{d_0} = 0, y_{d_1} = 1, x_0) = \frac{7}{8}, \tag{5.145}$$

$$P(d \mid y_{d_0} = 0, y_{d_1} = 1, x_1) = \frac{1}{2}, \tag{5.146}$$

indicating that the allocation of the decision is not independent of sex. That is, within the group of those who are helped by surgery, males are more likely to be selected for treatment than females. □

Somewhat counter-intuitively, the decision-making policy introduced by the clinicians does not satisfy principal fairness, even though at each level of illness severity, the proportion of patients who benefit from the treatment is equal between both sexes. What is the issue at hand here?

The takeaway message from this example is best illustrated in Figure 5.12. In Figure 5.12a, we see the perspective under perfect knowledge. In particular, on the horizontal axis the noise variable u_y, which summarizes the patients' unobserved resilience and frailty, is available. Together with the value of illness severity (on the vertical axis) and the knowledge of the structural causal model, we can perfectly pick apart the different groups (i.e., principal strata) according to their potential outcomes (groups are indicated by color). In this

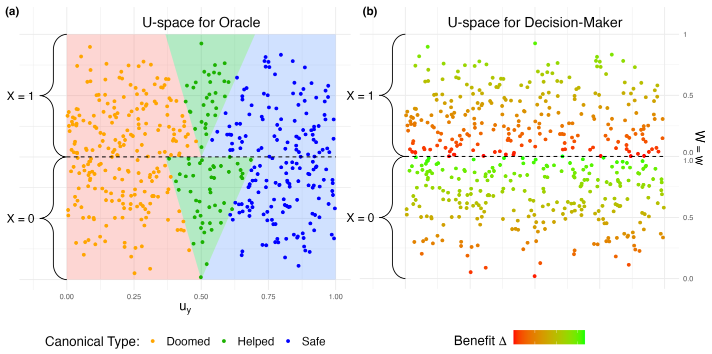


Figure 5.12: Difference in perspective between perfect knowledge of an oracle (left) and imperfect knowledge of a decision-maker in a practical application (right).

case, it is clear that our policy should allocate the treatment to the patients within the green area, since those are the ones who benefit from it.

In Figure 5.12b, however, we see the perspective of the decision-makers under imperfect knowledge. Firstly, the decision-makers have no knowledge about the values on the horizontal axis, since this represents variables that are outside their model. After computing the proportion of patients who benefit from treatment (corresponding to values in Eqs. 5.143-5.144), the decision-makers visualize the male and female groups (lighter color indicates larger increase in survival associated with surgery). It is visible from the figure that the estimated benefit from surgery is higher for the $X = x_0$ group than for $X = x_1$. Therefore, to the best of their knowledge, the decision-makers decide to treat more patients from the $X = x_0$ group.

The example illustrates why principal fairness might not be exactly the desired criterion, from the point of view of the decision-maker. In particular, the clinicians cannot determine (even with the monotonicity assumption holding true) exactly which patients belong to the group

$$Y_{d_0}(u) = 0, Y_{d_1}(u) = 1,$$

that is, who benefits from the treatment. Instead, the best thing they can do is to look at illness severity (W) as a proxy for treatment benefit. However, in the example, the degree of illness severity is sex-specific, since men are more severely ill on average. As a result, the optimal policy (from the viewpoint of the decision-maker), is excluded by principal fairness, i.e., it does not satisfy PF. The main issue lies in the fact that the intuition behind PF comes from the oracle case, in which we know deterministically who would benefit from the treatment, and then we wish the probability of treatment within this group not to depend on sex. In practice, however, our understanding of treatment benefit will almost always be probabilistic. As a consequence, the optimal policy (for example, in terms of overall patient survival) may be excluded by principal fairness.

Benefit Fairness

To remedy the issue described above, we propose an alternative definition, which takes the perspective of the decision-maker:

Definition 5.10 (Benefit Fairness). Let $\mathcal{M}$ be a structural causal model compatible with the SFM. Then, define the benefit relative to covariates (x, z, w) as:

$$\Delta(x, z, w) = P(y_{d_1} \mid x, z, w) - P(y_{d_0} \mid x, z, w). \tag{5.147}$$

We say that the pair (Y, D) satisfies the benefit fairness criterion (BFC, for short) if

$$P(d \mid \Delta = \delta, x_0) = P(d \mid \Delta = \delta, x_1) \;\; \forall \delta. \tag{5.148}$$

Notice that the BFC takes the perspective of the decision-maker who only has access to the unit's attributes (x, z, w), as opposed to the exogenous $U = u$[12]. In particular, the quantity $\Delta(x, z, w)$, which measures the benefit from the decision D for each set of covariates (x, z, w), which is *estimable* from the data, without invoking the monotonicity assumption. The BFC then requires that at each level of the degree of benefit, $\Delta(x, z, w) = \delta$, the rate of the decision does not depend on the protected attribute.

Note on Affirmative Versions of Benefit Fairness. The notion of fairness encoded in Def. 5.10 requires that the treatment allocation is independent of the protected attribute at any given, fixed value of the treatment's benefit Δ. In other words, the definition requires the decision-maker to be *agnostic* of the protected attribute and focus solely on the benefit Δ, and can thus be seen as a somewhat minimal fairness requirement. We now wish to highlight that alternative, stronger notions of fairness may be enforced in a similar way, best illustrated through an example:

Example 5.12. Let X denote sex of the patient (x_0 male, x_1 female) and Δ denote the benefit from medical treatment. Consider the SCM $\mathcal{M}$ given by:

$$X \leftarrow \text{Bernoulli}(0.1) \tag{5.149}$$

$$\Delta \leftarrow \text{Bernoulli}(0.5), \tag{5.150}$$

[12] Formally, having access to the exogenous instantiation of $U = u$ implies knowing which principal stratus from Fig. 5.10 the unit belongs to, since $U = u$ determines all of the variations of the model.

and suppose the budget is $\frac{1}{20}$ (a twentieth of all the individuals can be treated). In words, a tenth of the patients are female, and each patient either certainly benefits from treatment ($\Delta = 1$) or has no benefit from treatment ($\Delta = 0$). Each of these two cases happens with probability $\frac{1}{2}$ irrespective of the patient's sex. Suppose now that a team of clinicians has 2000 patients, with

- 1800 males ($X = x_0$), 900 with certain treatment benefit ($\Delta = 1$), and 900 with no treatment benefit ($\Delta = 0$),
- 200 females ($X = x_1$), 100 with certain treatment benefit ($\Delta = 1$), and 100 with no treatment benefit ($\Delta = 0$).

Suppose further that clinicians can (based on other information) infer exactly which patients benefit from treatment, and that they need to allocate the available 100 treatment slots to the cohort of 2000 patients. Following Def. 5.10, the clinicians would pick 90 male patients with certain benefit and 10 female patients with certain benefit, as this would ensure that:

$$P(d \mid \Delta = 1, x_0) = P(d \mid \Delta = 1, x_1) = \frac{1}{10}. \tag{5.151}$$

However, note that the policy of selecting 50 male patients with certain benefit and 50 female patients with certain benefit would also be optimal, but would not satisfy Def. 5.10. Instead, this policy would result in the equal allocation of resources between groups. □

The above example is intended to demonstrate that an "affirmative" version of benefit fairness could be articulated. The policy that selects an equal number of males and females for treatment, in the setting where there is a larger number of males requiring treatment, is not agnostic of sex, but instead aims to minimize the disparity in the amount of allocated resources between groups. In the sequel, we focus on Def. 5.10 but the above discussion is intended to inform the reader that stronger variants of this notion could also be considered – in particular, notions that still focus on the benefit, but aim to choose a policy, among all (nearly) optimal policies, that minimizes the disparity in resource allocation.

Canonical Types & Bounds

In decision-making, the goal is to treat as many units who are helped by the treatment, and as few who are harmed by it. As we demonstrate next, the benefit Δ is a function of the proportions of different canonical types:

Proposition 5.5 (Canonical Types Decomposition)**.** Let $\mathcal{M}$ be an SCM compatible with the SFM. Let D be a binary decision that possibly affects the outcome

Y. Denote by $(s, d, c, u)(x, z, w)$ the proportion of each of the canonical types safe, harmed, helped, and doomed, respectively, for a fixed set of covariates (x, z, w). It then holds that

$$P(y_{d_1} \mid x, z, w) = c(x, z, w) + s(x, z, w), \tag{5.152}$$

$$P(y_{d_0} \mid x, z, w) = d(x, z, w) + s(x, z, w). \tag{5.153}$$

Therefore, we have that

$$\begin{aligned} \Delta(x, z, w) &:= P(y_{d_1} \mid x, z, w) - P(y_{d_0} \mid x, z, w) & (5.154)\\ &= c(x, z, w) - d(x, z, w). & (5.155) \end{aligned}$$

Proof. Notice that we can write:

$$\begin{aligned} P(y_{d_1} \mid x, z, w) &= P(y_{d_1} = 1, y_{d_0} = 1 \mid x, z, w) & (5.156)\\ &\quad + P(y_{d_1} = 1, y_{d_0} = 0 \mid x, z, w) & \\ &= s(x, z, w) + c(x, z, w). & (5.157) \end{aligned}$$

where the first step follows from the law of total probability, and the second by definition. Similarly, we have that

$$\begin{aligned} P(y_{d_0} \mid x, z, w) &= P(y_{d_0} = 1, y_{d_1} = 1 \mid x, z, w) & (5.158)\\ &\quad + P(y_{d_0} = 1, y_{d_1} = 0 \mid x, z, w) & \\ &= s(x, z, w) + d(x, z, w), & (5.159) \end{aligned}$$

thereby completing the proof. ■

The proposition shows us that the degree of benefit $\Delta(x, z, w)$ captures exactly the difference between the proportion of those helped by the treatment, versus those who are harmed by it. From the point of view of the decision-maker, this is very valuable information since higher $\Delta(x, z, w)$ values indicate a higher utility of treating the group corresponding to covariates (x, z, w). As the next proposition shows, the values of $P(y_{d_1} \mid x, z, w), P(y_{d_0} \mid x, z, w)$ can also be used to bound the proportion of different canonical types:

Proposition 5.6 (Canonical Types Bounds and Tightness)**.** Let $(s, d, c, u)(x, z, w)$ be defined as in Prop. 5.5. Let $m_1(x, z, w) = P(y_{d_1} \mid x, z, w)$ and $m_0(x, z, w) = P(y_{d_0} \mid x, z, w)$ and suppose that $m_1(x, z, w) \geq m_0(x, z, w)$. We then have that (dropping (x, z, w) from the notation):

$$d \in [0, \min(m_0, 1 - m_1)], \tag{5.160}$$

$$c \in [m_1 - m_0, m_1]. \tag{5.161}$$

In particular, the above bounds are tight, meaning that there exists an SCM $\mathcal{M}$, compatible with the observed data, that attains each of the values within the interval. Under monotonicity, the bounds collapse to single points, with $d = 0$ and $c = m_1 - m_0$.

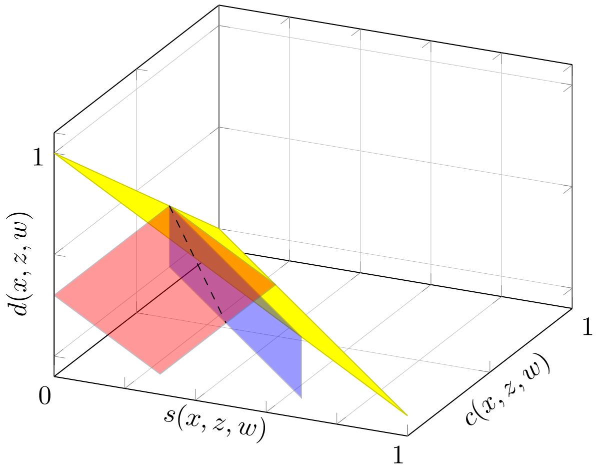


Figure 5.13: Canonical types solution space. The unit simplex is shown in yellow, the $s + c = m_1$ plane in blue, and the $s + d = m_0$ plane in red. The solution space for the possible values of $(s(x, z, w), c(x, z, w), d(x, z, w))$ lies at the intersection of the red and blue planes, indicated by the dashed black line.

Proof. There are three linear relations that the values s, d, c, u obey:

$$s + c = m_1, \tag{5.162}$$

$$s + d = m_0, \tag{5.163}$$

$$s + u + d + c = 1. \tag{5.164}$$

On top of this, we know that s, d, c, u are all non-negative. Based on the linear relations in Eqs. 5.162-5.164, we know that the following parametrization of the vector (s, d, c, u) holds

$$(s, d, c, u) = (m_0 - d, d, d + m_1 - m_0, 1 - m_1 - d), \tag{5.165}$$

which represents a line in the 3-dimensional space (s, d, c). In particular, we know that the values of (s, d, c) have to lie below the unit simplex in Fig. 5.13 (in yellow). In particular, the red and the blue planes represent the linear constraints from Eq. 5.162-5.163. The line parametrized in Eq. 5.165 lies at the intersection of the red and blue planes. Notice that $d \in [0, \min(m_0, 1 - m_1)]$ since each of the elements in Eq. 5.165 is positive. This bound on d also implies that $c \in [m_1 - m_0, m_1]$. Finally, we need to construct an f_Y mechanism which

Algorithm 5.3 Decision-Making with Benefit Fairness

```
• Inputs: Distribution $P(V)$, Budget $b$
1: Compute $\Delta(x, z, w) = \mathbb{E}[Y_{d_1} - Y_{d_0} \mid x, z, w]$ for all $(x, z, w)$.
2: If $P(\Delta > 0) \leq b$, set $D^* = \mathbb{1}(\Delta(x, z, w) > 0)$ and RETURN($D^*$).
3: Find $\delta_b > 0$ such that
       $$P(\Delta > \delta_b) \leq b, P(\Delta \geq \delta_b) > b. \tag{5.169}$$
4: Otherwise, define
       $$\mathcal{I} := \{(x, z, w) : \Delta(x, z, w) > \delta_b\}, \tag{5.170}$$
       $$\mathcal{B} := \{(x, z, w) : \Delta(x, z, w) = \delta_b\} \tag{5.171}$$
5: Construct the policy $D^*$ such that:
       $$D^* := \begin{cases} 1 \text{ for } (x, z, w) \in \mathcal{I}, \\ 1 \text{ with prob. } \frac{b - P(\mathcal{I})}{P(\mathcal{B})} \text{ for } (x_0, z, w) \in \mathcal{B}, \\ 1 \text{ with prob. } \frac{b - P(\mathcal{I})}{P(\mathcal{B})} \text{ for } (x_1, z, w) \in \mathcal{B}. \end{cases} \tag{5.172}$$
```

achieves any value within the bounds. To this end, define

$$f_Y(x, z, w, d, u_y) = \mathbb{1}(u_y \in [0, s]) + d \cdot \mathbb{1}(u_y \in [s, s + c]) + \tag{5.166}$$

$$(1 - d) \cdot \mathbb{1}(u_y \in [s + c, s + c + d]), \tag{5.167}$$

$$u_y \sim \text{Unif}[0, 1]. \tag{5.168}$$

which is both feasible and satisfies the proportion of canonical types to be (s, d, c, u). ■

Using the knowledge about canonical types and their bounds[13], we can formulate a solution for the decision-making task from Def. 5.7, given in Alg. 5.3. In particular, Alg. 5.3 takes as input the observational distribution $P(V)$, but its adaptation to inference from finite samples follows easily. In Step 2, we check whether we are operating under resource scarcity, and if not, the optimal policy simply treats everyone who stands to benefit from the treatment. Otherwise, we find the $\delta_b > 0$ which uses the entire budget (Step 3) and separate the interior $\mathcal{I}$ of those with the highest benefit (all of whom are treated), and the boundary $\mathcal{B}$ (those who are to be randomized) in Step 4. The budget remaining

[13] The canonical bounds in Eqs. 5.160-5.161 can be used the derive policies that perform best for the worst case proportions of $c(x, z, w)$ and $d(x, z, w)$ (Ben-Michael *et al.*, 2022).

to be spent on the boundary is $b - P(\mathcal{I})$, and thus individuals on the boundary are treated with probability $\frac{b-P(\mathcal{I})}{P(\mathcal{B})}$. Importantly, male and female groups are treated separately in this random selection process (Eq. 5.172), which in the finite sample case ensures that equal proportions of males and females on the boundary are selected. The BFC can be seen as a minimal fairness requirement in decision-making, and is often aligned with maximizing utility:

Proposition 5.7 (Alg. 5.3 Optimality)**.** Among all feasible policies D for the optimization problem in Eqs. 5.121-5.122, the result of Alg. 5.3 is optimal and satisfies benefit fairness.

A key extension we discuss next relates to the cases in which the benefit Δ itself may be deemed as discriminatory towards a protected group.

Fairness of the Benefit

Benefit fairness guarantees that at each fixed level of the benefit $\Delta = \delta$, the protected attribute plays no role in the treatment assignment. However, benefit fairness does not necessarily guarantee that the allocation rates of the treatment D are equal between groups, i.e., $P(d \mid x_1) = P(d \mid x_0)$, as shown by our example:

Example 5.12 (Cancer Surgery - continued)**.** After applying benefit fairness and implementing the optimal policy $D^* = \mathbb{1}\big(W > \frac{1}{2}\big)$, the clinicians compute that $P(d \mid x_1) - P(d \mid x_0) = -50\%$, that is, females are 50% less likely to be treated than males. □

In words, benefit fairness resulted in a disparity in resource allocation. Whenever this is the case, it implies that the benefit Δ differs between groups. In Alg. 5.4 we describe a formal procedure that helps the decision-maker to obtain a causal understanding of why that is, i.e., which underlying causal mechanisms (direct, indirect, spurious) lead to the difference in the benefit. We ground the idea behind Alg. 5.4 in our example:

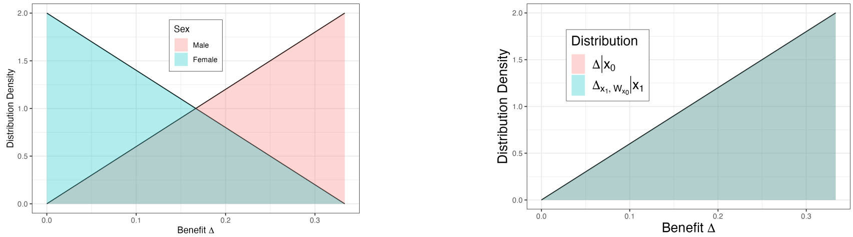
Example 5.13 (Decomposing the disparity)**.** Following Alg. 5.4, the clinicians first decompose the observed disparities into their direct, indirect, and spurious components:

$$P(d \mid x_1) - P(d \mid x_0) = \underbrace{0\%}_{\text{DE}} + \underbrace{-50\%}_{\text{IE}} + \underbrace{0\%}_{\text{SE}}, \tag{5.175}$$

$$\mathbb{E}(\Delta \mid x_1) - \mathbb{E}(\Delta \mid x_0) = \underbrace{0}_{\text{DE}} + \underbrace{-\frac{1}{9}}_{\text{IE}} + \underbrace{0}_{\text{SE}}, \tag{5.176}$$

Algorithm 5.4 Benefit Fairness Cookbook

```
• Inputs: Distribution $P(V)$, Benefit $\Delta(x, z, w)$, Decision policy $D$
1: Compute the causal decomposition (Thm. 4.3) of the resource allocation
   disparity into its direct, indirect, and spurious contributions:
       $P(d \mid x_1) - P(d \mid x_0) = \text{DE} + \text{IE} + \text{SE}.$    (5.173)
2: Compare the distributions $P(\Delta \mid x_1)$ and $P(\Delta \mid x_0)$.
3: Compute the causal decomposition (Thm. 4.3) of the benefit disparity
       $\mathbb{E}(\Delta \mid x_1) - \mathbb{E}(\Delta \mid x_0) = \text{DE} + \text{IE} + \text{SE}.$    (5.174)
4: Compute the counterfactual distribution $P(\Delta_C \mid x)$ for specific interventions
   $C$ that remove the direct, indirect, or total effect of X on the benefit
   $\Delta$.
```



(a) Density of benefit Δ by group.

(b) Density of counterfactual benefit Δ_C.

Figure 5.14: Elements of analytical tools from Alg. 5.4 in Ex. 5.13.

showing that the difference between groups is entirely explained by the levels of illness severity, that is, male patients are on average more severely ill than female patients (see Fig. 5.14a). Direct and spurious effects, in this example, do not explain the difference in benefit between the groups.

Based on these findings, the clinicians realize that the main driver of the disparity in the benefit Δ is the indirect effect. Thus, they decide to compute the distribution of the benefit $P(\Delta_{x_1, W_{x_0}} \mid x_1)$, which corresponds to the distribution of the benefit had X been equal to x_0 along the indirect effect. The comparison of this distribution, with the distribution $P(\Delta \mid x_0)$ is shown in Fig. 5.14b, indicating that the two distributions are in fact equal. □

In the example, the difference between groups is driven by the indirect effect, although generally, the situation may be more complex, with a combination of effects driving the disparity. Still, the tools of Alg. 5.4 equip the reader for analyzing such more complex cases. The key takeaway here is that the first step in analyzing a disparity in treatment allocation is to obtain a *causal understanding* of why the benefit differs between groups. Based on this understanding, the decision-maker may decide that the benefit Δ is unfair, which is what we discuss next.

Controlling the Gap

A causal approach. The first approach for controlling the gap in resource allocation takes a counterfactual perspective. We first define what it means for the benefit Δ to be causally fair:

Definition 5.11 (Causal Benefit Fairness). Suppose $\mathcal{C} = (C_0, C_1)$ describes a pathway from X to Y which is deemed unfair. The pair (Y, D) satisfies counterfactual benefit fairness (CBF) if

$$\mathbb{E}(y_{C_1,d_1} - y_{C_1,d_0} \mid x, z, w) = \mathbb{E}(y_{C_0,d_1} - y_{C_0,d_0} \mid x, z, w) \;\forall x, z, w \tag{5.177}$$

$$P(d \mid \Delta, x_0) = P(d \mid \Delta, x_1). \tag{5.178}$$

To account for discrimination along a specific causal pathway (after using Alg. 5.4), the decision-maker needs to compute an adjusted version of the benefit Δ, such that the protected attribute has no effect along the intended causal pathway $\mathcal{C}$. For instance, $\mathcal{C} = (\{x_0\}, \{x_1\})$ describes the total causal effect, whereas $\mathcal{C} = (\{x_0\}, \{x_1, W_{x_0}\})$ describes the direct effect. In words, CBF requires that treatment benefit Δ should not depend on the effect of X on Y along the causal pathway $\mathcal{C}$. Additionally, the decision policy D should satisfy BFC, meaning that at each degree of benefit $\Delta = \delta$, the protected attribute plays no role in deciding whether the individual is treated or not. This can be achieved using Alg. 5.5 with method = CF. In Step 1, the factual benefit values Δ, together with the adjusted, counterfactual benefit values Δ_C (that satisfy Def. 5.11) are computed. Then, δ_{CF} is chosen to match the budget b, and all patients with a counterfactual benefit above δ_{CF} are treated[14], as demonstrated in the following example:

Example 5.13 (Cancer Surgery - Counterfactual Approach). The clinicians realize that the difference in illness severity comes from the fact that female patients are subject to regular screening tests, and are therefore diagnosed earlier. The clinicians want to compute the adjusted benefit, by computing the counterfactual values of the benefit $\Delta_{x_1,W_{x_0}}(u)$ for all u such that $X(u) = x_1$. For the computation, they assume that the relative order of the illness severity for females in the counterfactual world would have stayed the same (which holds true in the underlying SCM). That is, they assume that for any u

[14] In this section, for clarity of exposition we assume that the distribution of the benefit admits a density, although the methods are easily adapted to the case when this does not hold.

Algorithm 5.5 Causal Discrimination Removal for Outcome Control

```
• Inputs: Distribution $P(V)$, Budget $b$, Intervention $C$, Max. Disparity $M$, Method $\in \{CF, UT\}$
Compute $\Delta(x, z, w), \Delta_C(x, z, w)$ for all $(x, z, w)$.
If $P(\Delta > 0) \leq b$, set $D = \mathbb{1}(\Delta(x, z, w) > 0)$ and RETURN($D$).
Find $\delta_{CF} > 0$ such that
    $P(\Delta_C \geq \delta_{CF}) = b. \tag{5.179}$
If Method is CF, set $D^{CF} = \mathbb{1}(\Delta_C(x, z, w) \geq \delta_{CF})$ and RETURN($D^{CF}$).
If $M$ not pre-specified, compute the disparity
    $M := P(\Delta_C \geq \delta_{CF} \mid x_1) - P(\Delta_C \geq \delta_{CF} \mid x_0). \tag{5.180}$
Find $\delta_{UT}$ such that $P(\Delta \geq \delta_{UT}) = b$. If
    $|P(\Delta \geq \delta_{UT} \mid x_1) - P(\Delta \geq \delta_{UT} \mid x_0)| \leq M,$
    set $D^{UT} = \mathbb{1}(\Delta(x, z, w) \geq \delta_{UT})$ and RETURN($D^{UT}$).
Otherwise, suppose w.l.o.g. that $P(\Delta \geq \delta_b \mid x_1) - P(\Delta \geq \delta_b \mid x_0) = M + \epsilon$ for $\epsilon > 0$. Define $l := \frac{P(x_1)}{P(x_0)}$, and let $\delta_{lb}^{(x_0)}$ be such that
    $P(\Delta \geq \delta_{lb}^{(x_0)} \mid x_0) = P(\Delta \geq \delta_b \mid x_0) + \epsilon \frac{l}{1 + l}.$
    Set $\delta^{(x_0)} = \max(\delta_{lb}^{(x_0)}, 0)$, and $\delta^{(x_1)}$ s.t. $P(\Delta \geq \delta^{(x_1)} \mid x_1) = \frac{b}{P(x_1)} - \frac{1}{l} P(\Delta \geq \delta^{(x_0)} \mid x_0)$.
Construct and RETURN the policy $D^{UT}$:
    $D^{UT} := \begin{cases} 1 \text{ for } (x_1, z, w) \text{ s.t. } \Delta(x_1, z, w) \geq \delta^{(x_1)}, \\ 1 \text{ for } (x_0, z, w) \text{ s.t. } \Delta(x_0, z, w) \geq \delta^{(x_0)}, \\ 0 \text{ otherwise.} \end{cases} \tag{5.181}$
```

with $X(u) = x_1$, a factual value $W(u)$ with a relative quantile $Q(u)$ in the $W \mid x_1$ distribution would map to a counterfactual value $W_{x_0}(u)$ that has the same quantile $Q(u)$ in the $W \mid x_0$ distribution. Implicitly, they construct the mapping

$$W_{x_0}(u) = \sqrt{1 - (1 - W(u))^2}, \tag{5.182}$$

$$\Delta_{x_1, W_{x_0}}(u) = \frac{1}{3} W_{x_0}(u), \tag{5.183}$$

for each unit u with $X(u) = x_1$. After applying Alg. 5.3 with the counterfactual benefit values Δ_C, the resulting policy $D^{CF} = \mathbb{1}(\Delta_C > \frac{1}{4})$ has a resource allocation disparity of 0. □

The above example illustrates the core of the causal counterfactual approach to discrimination removal. The BFC was not appropriate in itself, since the clinicians are aware that the benefit of the treatment depends on sex in a way they deemed unfair. Therefore, to solve the problem, they first remove the undesired effect from the benefit Δ, by computing the counterfactual benefit Δ_C. After this, they apply Alg. 5.5 with the counterfactual method (CF) to construct a fair decision policy.

Remark 5.3 (Direct Effect on Benefit is Computable)**.** Under the assumptions of the SFM, the potential outcome $\Delta_{x_1, W_{x_0}}(u)$ is identifiable for any unit u with $X(u) = x_0$ for which the attributes $Z(u) = z, W(u) = w$ are observed. However, the same is not true for the indirect effect. While the counterfactual distribution of the benefit when the indirect effect is manipulated, written $P(\Delta_{x_0, W_{x_1}} \mid x_0)$, is identifiable under the SFM, the covariate-level values $\Delta_{x_0, W_{x_1}}(x_0, z, w)$ are not identifiable without further assumptions.

A utilitarian/factual approach. An alternative, utilitarian (or factual) approach to reduce the disparity in resource allocation uses the factual benefit $\Delta(u)$, instead of the counterfactual benefit $\Delta_C(u)$. This approach is also described in Alg. 5.5, with the utilitarian (UT) method. Firstly, in Step 5, the counterfactual values Δ_C are used to compute the disparity that would arise from the optimal policy in the hypothetical, counterfactual world:

$$M := |P(\Delta_C \geq \delta_{CF} \mid x_1) - P(\Delta_C \geq \delta_{CF} \mid x_0)|. \tag{5.184}$$

The idea then is to introduce different thresholds $\delta^{(x_0)}, \delta^{(x_1)}$ for x_0 and x_1 groups, such that they introduce a disparity of at most M. Thus, the utilitarian approach uses the counterfactual values Δ_C to determine the maximum allowed disparity M, but then focuses on the factual benefit values Δ to select individuals for treatment. In Step 6 we first check whether the overall optimal policy introduces a disparity bounded by M (if yes, we are done). If the disparity is larger than M by an ϵ, in Step 7 we determine how much slack the disadvantaged group requires, by finding thresholds $\delta^{(x_0)}, \delta^{(x_1)}$ that either treat everyone in the disadvantaged group, or achieve a disparity bounded by M. The maximum allowed disparity M can also be pre-determined, as is done in the following example:

Example 5.14 (Cancer Surgery - Utilitarian Approach)**.** Due to regulatory purposes, clinicians decide that $M = 20\%$ is the maximum allowed disparity that can be introduced by the new policy D. Using Alg. 5.5, they construct D^{UT} and find that for $\delta^{(x_0)} = 0.21, \delta^{(x_1)} = 0.12$,

$$P(\Delta > \delta^{(x_0)} \mid x_0) \approx 60\%, P(\Delta > \delta^{(x_1)} \mid x_1) \approx 40\%, \tag{5.185}$$

which yields $P(d^{UT}) \approx 50\%$, and $P(d^{UT} \mid x_1) - P(d^{UT} \mid x_0) \approx 20\%$, which is in line with the hospital resources and the maximum disparity allowed by the regulators. □

Finally, we describe the theoretical guarantees for the methods in Alg. 5.5 (proof given in Appendix A.8):

Theorem 5.6 (Alg. 5.5 Guarantees)**.** The policy D^{CF} is optimal among all policies with a budget $\leq b$ that in the counterfactual world described by the intervention C. The policy D^{UT} is optimal among all policies with a budget $\leq b$ that either introduce a bounded disparity in resource allocation $|P(d \mid x_1) - P(d \mid x_0)| \leq M$ or treat everyone with a positive benefit in the disadvantaged group.

Equivalence of the utilitarian and causal approach. We remark that policies D^{CF} and D^{UT} do not necessarily treat the same individuals in general. A natural question to ask is whether there are conditions under which the two methods in Alg. 5.5 yield the same decision policy in terms of the individuals that are selected for treatment. To examine this issue, we first define the notion of counterfactual crossing:

Definition 5.12 (Counterfactual Crossing)**.** Two units of the population u_1, u_2 are said to satisfy counterfactual crossing with respect to an intervention C if

(i) u_1, u_2 belong to the same protected group, $X(u_1) = X(u_2)$.

(ii) unit u_1 has a higher factual benefit than u_2, $\Delta(u_1) > \Delta(u_2)$,

(iii) unit u_1 has a lower counterfactual benefit than u_2 under the intervention C, $\Delta_C(u_1) < \Delta_C(u_2)$.

In words, two units satisfy counterfactual crossing if u_1 has a higher benefit than u_2 in the factual world, while in the counterfactual world the benefit is larger for the unit u_2. Based on this notion, we can give a condition under which the causal and utilitarian approaches are equivalent:

Proposition 5.8 (Causal and Utilitarian Equivalence)**.** Suppose that no two units of the population satisfy counterfactual crossing with respect to an intervention C, and suppose that the distribution of the benefit Δ admits a density. Then, the policies D^{CF} and D^{UT} from Alg. 5.5 select the same set of units for treatment.

Proof. The policy D^{UT} treats individuals who have the highest benefit Δ in each group. The D^{CF} policy treats individuals with the highest counterfactual benefit Δ_C. Importantly, the policies treat the same number of individuals in the x_0 and x_1 groups. Note that, in the absence of counterfactual crossing, the relative ordering of the values of Δ, Δ_C does not change, since

$$\Delta(u_1) > \Delta(u_2) \iff \Delta_C(u_1) > \Delta_C(u_2). \tag{5.186}$$

Thus, since both policies pick the same number of individuals, and the relative order of Δ, Δ_C is the same, D^{UT} and D^{CF} will treat the same individuals. ∎

Interestingly, when the intervention C is changing the protected attribute X along the direct causal pathway, the implications of the above proposition are *testable* in practice, under the assumptions of the SFM (see Remark 5.3 for more details).

To close the section, we summarize the discussion by giving a recap of the main steps for fair decision-making in the outcome-control setting:

(1) Apply the principles of benefit fairness, using Alg. 5.3, to design the policy D^*,

(2) Analyze the disparity introduced by the policy D^*, using Alg. 5.4, to better understand which causal pathways drive the difference in resource allocation,

(3) If a causal pathway that drives the difference in resource allocation between the groups is deemed unfair (or if the disparity is deemed too large):

  (a) use Alg. 5.5 with method CF to find the policy D^{CF} that satisfies Causal Benefit Fairness from Def. 5.11, or

  (b) use Alg. 5.5 with method UT to find the policy D^{UT} that uses the factual benefit and allows distinct thresholds $\delta^{(x_0)}, \delta^{(x_1)}$ for the two groups, but sets the disparity in resource allocation according to the counterfactual world.

6

Disparate Impact and Business Necessity

In this section, we generalize the analysis introduced earlier, including the Fairness Cookbook (Alg. 5.1), to consider more refined settings described by an arbitrary causal diagram. The main motivation for doing so comes from the observation that when analyzing disparate impact, quantities such as $\text{Ctf-DE}_{x_0,x_1}(y \mid x_0)$, $\text{Ctf-IE}_{x_0,x_1}(y \mid x_0)$, and $\text{Ctf-SE}_{x_0,x_1}(y)$ are insufficient to account for certain *business necessity* requirements. For concreteness, consider the following example.

Example 6.1 (COMPAS continued with Business Necessity)**.** The courts at Broward County, Florida, were using machine learning to predict whether individuals released on parole are at high risk of re-offending within 2 years. The algorithm is based on the demographic information Z (Z_1 for gender, Z_2 for age), race X (x_0 denoting White, x_1 Non-White), juvenile offense counts J, prior offense count P, and degree of charge D.

A causal analysis using the Fairness Cookbook revealed that:

$$\text{Ctf-IE}_{x_1,x_0}(y \mid x_1) = -5.7\% \pm 0.5\%, \tag{6.1}$$

$$\text{Ctf-SE}_{x_1,x_0}(y) = -4.0\% \pm 0.9\%, \tag{6.2}$$

potentially indicating presence of disparate impact. Based on this information, a legal team of ProPublica filed a lawsuit to the district court, claiming discrimination w.r.t. the Non-White subpopulation based on the doctrine of disparate impact. After the court hearing, the judge rules that using the attributes age (Z_2), prior count (P), and charge degree (D) is not discriminatory, but using the attributes juvenile count (J) and gender (Z_1) is discriminatory. The

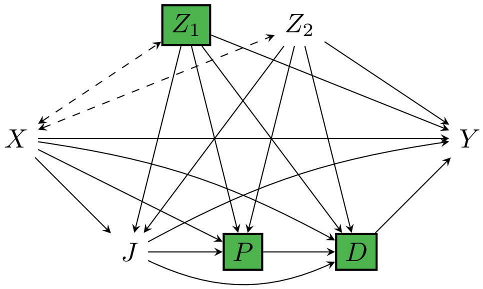


Figure 6.1: Causal diagram of the COMPAS dataset. Business necessity variables are highlighted in green.

causal diagram with a visualization of which variables are included in the business-necessity set is given in Fig. 6.1. Data scientists at ProPublica need to consider how to proceed in light of this new requirement for discounting the allowable attributes in the quantitative analysis. □

The difficulty in this example is that the quantity $\text{Ctf-SE}_{x_1,x_0}(y)$ measures the spurious discrimination between the attribute X and outcome Y as generated by *both confounders* Z_1 and Z_2. Since using the confounder Z_2 is not considered discriminatory, but using the confounder Z_1 is, the quantity $\text{Ctf-SE}_{x_1,x_0}(y)$ needs to be refined such that the spurious variations based on the different confounders are disentangled. In particular, one might be interested in finding a decomposition of the spurious effect such that

$$\text{Ctf-SE}_{x_1,x_0}(y) = \underbrace{\text{Ctf-SE}^{Z_1}_{x_1,x_0}(y)}_{\text{gender variations}} + \underbrace{\text{Ctf-SE}^{Z_2}_{x_1,x_0}(y)}_{\text{age variations}}, \tag{6.3}$$

which would allow the data analyst to further distinguish the variations explained by each of the confounders. A similar challenge is present for the $\text{Ctf-IE}_{x_1,x_0}(y \mid x_1)$ measure, since it has contributions explained by juvenile offense counts J, prior counts P, and the charge degree D. Therefore, we might be interested in decomposing the indirect effect into

$$\begin{aligned}\text{Ctf-IE}_{x_1,x_0}(y \mid x_1) = &\underbrace{\text{Ctf-IE}^{J}_{x_1,x_0}(y \mid x_1)}_{\text{juvenile count variations}} + \underbrace{\text{Ctf-IE}^{P}_{x_1,x_0}(y \mid x_1)}_{\text{prior count variations}} \\ &+ \underbrace{\text{Ctf-IE}^{D}_{x_1,x_0}(y \mid x_1)}_{\text{charge degree variations}} .\end{aligned} \tag{6.4}$$

Again, such a decomposition would allow the data analyst to better understand the contribution of each of the mediators to the totality of the indirect effect. In situations when some mediating variables are in the business necessity set, while others are not, such a decomposition would allow for the assessment of disparate impact claims.

In the following sections we discuss how such more refined decompositions of the spurious and indirect effects can be obtained, which allow us to reason about disparate impact under business necessity on a more granular level. Before doing so, we discuss some foundational techniques of causal inference, upon which our decomposition machinery will be built.

6.1 Causal Inference Procedures

In this section, we cover important inferential techniques that are used to compute causal (and statistical) queries. In particular, we will focus on various associational, interventional, and counterfactual queries, which correspond to the three layers of the Pearl's Causal Hierarchy (PCH) (Bareinboim *et al.*, 2022).

6.1.1 Abduction and Prediction

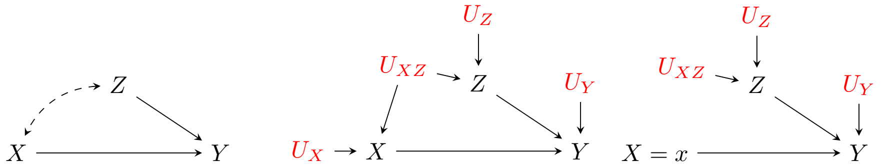
The first and basic method of inference is the *abduction-prediction* method. Given an SCM $\langle \mathcal{F}, P(u) \rangle$, a variable of interest Y and some observed evidence $E = e$, the method can be summarized as follows:

Algorithm 6.1 (Abduction and Prediction)**.** Given an SCM $\langle \mathcal{F}, P(u) \rangle$, the conditional probability $P(Y = y \mid E = e)$ of an event $Y = y$ upon observing the evidence $E = e$, can be evaluated using the following two steps:

(i) **Abduction** – update $P(u)$ by the evidence e to obtain $P(u \mid e)$,

(ii) **Prediction** – use the model $\langle \mathcal{F}, P(u \mid e) \rangle$ to compute the probability of $Y = y$.

The abduction-prediction procedure may be seen as one of the basic building blocks of Bayesian inference. In the first step, the probabilities of the exogenous variables U are updated according to the observed evidence $E = e$. After this, the updated model $\langle \mathcal{F}, P(u \mid e) \rangle$ is used to compute the conditional probability $P(y \mid e)$. Importantly, such a procedure can help one handle queries in the first, associational layer of the PCH. Queries handling questions about what would happen if X was set to x by intervention, or what value would Y had taken had we imagined (contrary to the fact) that $X = x$, are not within the scope of the procedure. Therefore, the abduction-prediction method can be used for computing statistical queries, but cannot be used for computing causal queries. We now demonstrate the method on an example:



(a) Causal diagram of the SCM in Ex. 6.2.

(b) Extended representation of Ex. 6.2, latent variables in red.

(c) Diagram of Ex. 6.2 under $do(X = x)$ intervention.

Figure 6.2: Graphical representations of the SCM in Ex. 6.2.

Example 6.2 (Abduction-Prediction)**.** Consider the following SCM

$$\mathcal{F} := \begin{cases} X \leftarrow f_X(U_X, U_{XZ}) & (6.5) \\ Z \leftarrow f_Z(U_Z, U_{XZ}) & (6.6) \\ Y \leftarrow f_Y(X, Z, U_Y), & (6.7) \end{cases}$$

with $P(U_X, U_{XZ}, U_Z, U_Y)$ the distribution over the exogenous variables. The causal diagram of the model is shown in Fig. 6.2a, and a more detailed representation with an explicit representation of the exogenous variables is shown in Fig. 6.2b.

We are interested in the query $P(y \mid x)$ in the given model. Based on the abduction-prediction procedure, we can simply compute that:

$$P(y \mid x) = \sum_u \mathbb{1}(Y(u) = y)P(u \mid x) \tag{6.8}$$

$$= \sum_u \mathbb{1}(Y(u) = y)P(u_z, u_y)P(u_x, u_{xz} \mid x), \tag{6.9}$$

where the first step follows from the definition of the observational distribution, and the second step follows from noting the independence $U_Z, U_Y \perp\!\!\!\perp U_X, U_{XZ}, X$. In the abduction step, we can compute the probabilities $P(u_x, u_{xz} \mid x)$. In the prediction step, we compute the probability $P(y \mid x)$ based on Eq. 6.9. □

6.1.2 Abduction, Action, and Prediction

As mentioned, the abduction-prediction procedure was limited to the first layer of the PCH. Often, one may be interested in the higher layers of the hierarchy, and thus the original procedure needs to be adapted. To do so, we make use of the definition of a submodel (Def. 2.2). Recall that the SCM $\mathcal{M}_x$ is obtained from $\mathcal{M}$ by replacing all equations in $\mathcal{F}$ related to variables X by equations that set X to a specific value x. This corresponds to "intervening" (layer 2 of the hierarchy) or "imagining" (layer 3) that $X = x$. Equipped with this definition, we can introduce the three step procedure that allows one to handle queries in all layers of the PCH:

Algorithm 6.2 (Abduction, Action, and Prediction (Pearl, 2000)). Given an SCM $\langle \mathcal{F}, P(u) \rangle$, the conditional probability $P(Y_C \mid E = e)$ of a counterfactual sentence "if it were C then Y", upon observing the evidence $E = e$, can be evaluated using the following three steps:

(i) **Abduction** – update $P(u)$ by the evidence e to obtain $P(u \mid e)$,

(ii) **Action** – modify $\mathcal{F}$ by the action $do(C)$, where C is an antecedent of Y, to obtain $\mathcal{F}_C$,

(iii) **Prediction** – use the model $\langle \mathcal{F}_C, P(u \mid e) \rangle$ to compute the probability of Y_C.

We now contrast Alg. 6.2 with Alg. 6.1. The newly introduced procedure adds an important *action* step between the abduction and prediction, which allows for a great deal of additional flexibility. In the first step, we update the $P(u)$ according to the available evidence. After this, however, we modify the model $\mathcal{M}$ to a submodel $\mathcal{M}_C$. This action step is precisely what allows for the additional flexibility in our inference – corresponding to interventions or imaginative, counterfactual thinking. The final step of prediction is again shared with the basic abduction-prediction procedure. We note that whenever the abduction step is empty, but the action step is not, the procedure handles *interventional* queries in the second layer of the PCH. We refer to this special case as *action-prediction*. Similarly, when the action step is empty, we recover the basic procedure of abduction-prediction, showing that Alg. 6.1 is contained within Alg. 6.2. The following example illustrates the usage of Alg. 6.2:

Example 6.3 (Abduction, Action, Prediction). Consider the model in Eq. 6.5-6.7. We are interested in computing the query $P(y_x)$ (see Fig. 6.2c):

$$\begin{aligned} P(y_x) &= \sum_u \mathbb{1}(Y_x(u) = y)P(u) && (6.10) \\ &= \sum_u \mathbb{1}(Y(x, u_{xz}, u_z, u_y) = y)P(u). && (6.11) \end{aligned}$$

where the first step follows from the definition of an interventional distribution, and the second step follows from noting that Y_x does not depend on u_x. In this case, the abduction step is void, since we are not considering any specific evidence $E = e$. The value of $Y(x, u_{xz}, u_z, u_y)$ can be computed from the submodel $\mathcal{M}_x$. Finally, using Eq. 6.11 we can perform the prediction step. We remark that

$$\mathbb{1}(Y(x, u_{xz}, u_z, u_y) = y) = \sum_{u_x} \mathbb{1}(Y(u_x, u_{xz}, u_z, u_y) = y) P(u_x \mid x, u_{xz}, u_z, u_y), \tag{6.12}$$

by the law of total probability and noting that X is a deterministic function of u_x, u_{xz}. Thus, $P(y_x)$ also admits an alternative representation

$$P(y_x) = \sum_{u} \mathbb{1}(Y(u_x, u_{xz}, u_z, u_y) = y) P(u_x \mid x, u_{xz}, u_z, u_y) P(u_{xz}, u_z, u_y) \tag{6.13}$$

$$= \sum_{u} \mathbb{1}(Y(u) = y) P(u_x \mid x, u_{xz}) P(u_{xz}, u_z, u_y), \tag{6.14}$$

where Eq. 6.14 follows from using the independencies among U and X in the graph in Fig. 6.2b. We revisit the representation in Eq. 6.14 in Ex. 6.4. □

6.1.3 Foundations of Decomposing Spurious Variations

After getting familiar with the abduction-action-prediction procedure, our next task is to introduce a new procedure that allows us to decompose spurious effects. First, we define the concept of a *partially abducted submodel*:

Definition 6.1 (Partially Abducted Submodel). Let $U_1, U_2 \subseteq U$ be a partition of the exogenous variables. Let the partially abducted (PA, for short) submodel with respect to the exogenous variables U_1 and evidence $E = e$ be defined as:

$$\mathcal{M}^{U_1, E=e} := \langle \mathcal{F}, P(u_1) P(u_2 \mid u_1, E = e) \rangle. \tag{6.15}$$

In words, in the PA submodel, the typically obtained posterior distribution $P(u \mid e)$ is replaced by the distribution $P(u_2 \mid u_1, e)$. Effectively, the exogenous variables U_1 are *not updated according to evidence*. The main motivation for introducing the PA model is that spurious variations arise whenever we are comparing units of the population that are different, a realization dating back to Pearson in the 19th century (Pearson, 1899). To give a formal discussion on what became known as *Pearson's shock*, consider two sets of differing evidence $E = e$ and $E = e'$. After performing the abduction step, the variations between posterior distributions $P(u \mid e)$ and $P(u \mid e')$ will be explained by *all the exogenous variables that precede the evidence* E. In a PA submodel, however, the posterior distribution $P(u_1)P(u_2 \mid u_1, e)$ will differ from $P(u_1)P(u_2 \mid u_1, e')$ only in variables that are in U_2, while the variables in U_1 will induce no spurious variations. Thus, the PA submodel allows us to choose which subset of the exogenous confounders introduce spurious variations. When the set $U_1 = \emptyset$, then the spurious variations are generated from all the U. On the other, when $U_1 = U$, the PA submodel will introduce no spurious variations. Different choices of U_1, ranging from the $\emptyset$ to U, as we will see shortly, provide a way for decomposing spurious variations in general.

We next demonstrate how the definition of a PA submodel can be used to obtain partially abducted conditional probabilities:

Proposition 6.1 (PA Conditional Probabilities)**.** Let $P(Y = y \mid E = e^{U_1})$ denote the conditional probability of the event $Y = y$ conditional on evidence $E = e$, while the exogenous variables U_1 are not updated according to the evidence. Then, we have that:

$$P(Y = y \mid E = e^{U_1}) = \sum_{u_1} P(U_1 = u_1) P(Y = y \mid E = e, U_1 = u_1). \quad (6.16)$$

6.1.4 Partial Abduction and Prediction

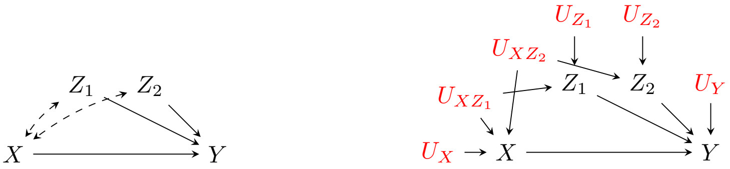
Based on the notion of a PA submodel, we can introduce the partial-abduction and prediction procedure:

Algorithm 6.3 (Partial Abduction and Prediction)**.** Given an SCM $\langle \mathcal{F}, P(u) \rangle$, the conditional probability $P(Y = y \mid E = e^{U_1})$ of an event $Y = y$ upon observing the evidence e, in a world where variables U_1 are unresponsive to evidence, can be evaluated using the following two steps:

(i) **Partial Abduction** – update $P(u)$ by the evidence e to obtain the posterior $P(u_1)P(u_2 \mid u_1, e)$, where (u_1, u_2) is a partition of the exogenous variables u,

(ii) **Prediction** – use the model $\langle \mathcal{F}, P(u_1)P(u_2 \mid u_1, e) \rangle$ to compute the probability of $Y = y$.

In the first step of the algorithm, we only perform *partial abduction*. The exogenous variables U_2 are updated according to the available evidence $E = e$, while the variables U_1 retain their original distribution $P(u_1)$ and remain unresponsive to evidence. As mentioned earlier, this allows us to consider queries in which only a subset of the exogenous variables respond to the available evidence. We next explain what kind of queries this entails, beginning with an example:



(a) Causal diagram corresponding to the SCM in Ex. 6.5.

(b) Extended graphical representation of the SCM in Ex. 6.5, latent variables in red.

Figure 6.3: Graphical representations of the SCM in Ex. 6.3.

Example 6.4 (Partial Abduction and Prediction)**.** Consider the model in Eq. 6.5-6.7. We are interested in computing the query:

$$P(y \mid x^{U_{xz},U_z}) = \sum_u \mathbb{1}(Y(u) = y)P(u_{xz}, u_z)P(u_x, u_y \mid u_{xz}, u_x, x) \tag{6.17}$$

$$= \sum_u \mathbb{1}(Y(u) = y)P(u_{xz}, u_z)P(u_y)P(u_x \mid u_{xz}, u_x, x) \tag{6.18}$$

$$= \sum_u \mathbb{1}(Y(u) = y)P(u_{xz}, u_z, u_y)P(u_x \mid u_{xz}, u_x, x), \tag{6.19}$$

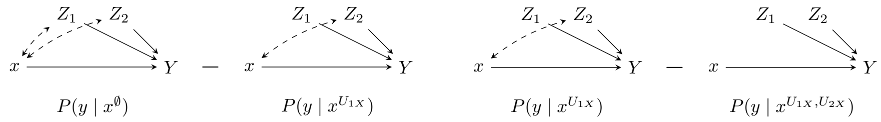
where the first step follows from Prop. 6.1, and the remaining steps from conditional independencies between the U variables and X. Crucially, the query yields the same expression as in Eq. 6.14 that we obtained for $P(y_x)$ in Ex. 6.3. Therefore, the conditional probability $P(y \mid x^{U_{xz},U_z})$ in a world where U_{XZ}, U_Z are unresponsive to evidence is equal to the interventional probability $P(y_x)$. □

As the example illustrates, we have managed to find another procedure that mimics the behavior of the interventional ($do(X = x)$) operator in the given example. Interestingly, however, in this procedure, we have not made use of the submodel $\mathcal{M}_x$ that was used in the abduction-action-prediction procedure. The idea from Ex. 6.4 can be extended in order to decompose spurious variations in causal models, as shown in the following example:

Example 6.5 (Spurious Decomposition)**.** Consider an SCM compatible with the graphical representation in Fig. 6.3b (with exogenous variables U shown explicitly in red), and the corresponding Semi-Markovian causal diagram in Fig. 6.3a. We note that, based on the partial abduction-prediction procedure, the following two equalities hold:

$$P(y \mid x) = P(y \mid x^{\emptyset}) \tag{6.20}$$

$$P(y_x) = P(y \mid x^{U_{xz_1},U_{xz_2}}), \tag{6.21}$$



(a) $\text{Exp-SE}_x^{\emptyset, U_1X}(y)$. **(b)** $\text{Exp-SE}_x^{U_1X, \{U_1X, U_2X\}}(y)$.

Figure 6.4: Graphical representation of the Exp-SE decomposition in Ex. 6.5.

which shows that

$$\text{Exp-SE}_x(y) = P(y \mid x^{\emptyset}) - P(y \mid x^{U_{xz_1}, U_{xz_2}}). \tag{6.22}$$

The experimental spurious effect can be written as a difference of conditional probabilities $y \mid x$ in a world where all variables U are responsive to evidence vs. a world in which U_{XZ_1}, U_{XZ_2} are unresponsive to evidence. Furthermore, we can also consider a refinement that decomposes the effect

$$\text{Exp-SE}_x(y) = \underbrace{P(y \mid x^{\emptyset}) - P(y \mid x^{U_{xz_1}})}_{\text{variations of } U_{xz_1}} + \underbrace{P(y \mid x^{U_{xz_1}}) - P(y \mid x^{U_{xz_1}, U_{xz_2}})}_{\text{variations of } U_{xz_2}}, \tag{6.23}$$

allowing for an additive, non-parametric decomposition of the experimental spurious effect. □

The first term in Eq. 6.23 is shown in Fig. 6.7a. On the left side, $P(y \mid x^{\emptyset})$, both confounders U_{XZ_1}, U_{XZ_1} respond to the evidence $X = x$, whereas on the right, P, only U_{XZ_2} responds to evidence. Therefore, the first term captures spurious variations explained by U_{XZ_1}. The second term is shown in Fig. 6.4b. On the left, we have $P(y \mid x^{U_{XZ_1}})$ where only U_{XZ_2} responds to evidence. This is contrasted with $P(y \mid x^{U_{XZ_1}, U_{XZ_2}})$, where neither of the confounders are responsive to evidence. Therefore, the second term thus captures spurious variations explained by U_{XZ_2}.

For an overview, in Tab. 6.1 we summarize the different inferential procedures discussed so far, indicating the structural causal models associated with them. Our next task is to develop general spurious decomposition results, and connect them to the framework of Causal Fairness Analysis.

6.2 Refining spurious discrimination

Following the framework from Sec. 3, we start by defining a structural criterion of spurious discrimination under business necessity.

Definition 6.2 (Structural Spurious Criterion under Business Necessity). Let $\mathcal{M}$ be a Semi-Markovian SCM. Write $\text{an}^{\text{ex}}(\cdot)$ for the set of exogenous ancestors,

Procedure	SCM	Queries
Abduction-Prediction	$\langle \mathcal{F}, P(u \mid E) \rangle$	Layer 1
Action-Prediction	$\langle \mathcal{F}_x, P(u) \rangle$	Layer 2
Abduction-Action-Prediction	$\langle \mathcal{F}_x, P(u \mid E) \rangle$	Layers 1, 2, 3
Partial Abduction-Prediction	$\langle \mathcal{F}, P(u_1)P(u_2 \mid E) \rangle$	Layers 1, 2, 3

Table 6.1: Summary of the different procedures and the corresponding probabilistic causal models.

and let $\mathcal{G}_{\underline{X}}$ denote the diagram $\mathcal{G}$ with the outgoing edges from X removed. Let $\text{an}^{\text{ex}}(X) \subseteq U$ be the subset of exogenous variables U which have a causal path to the protected attribute X. Let U_{BN}, U_{BN}^{C} be a partition of the set $\text{an}^{\text{ex}}(X)$, where U_{BN} is the subset of the exogenous variables which fall under business necessity. We then define the structural spurious criterion under business necessity as

$$\text{Str-SE}(U_{BN})_X(Y) = \mathbb{1}(\text{an}^{\text{ex}}_{\mathcal{G}_{\underline{X}}}(Y) \cap \text{an}^{\text{ex}}(X) \cap U_{BN}^{C} = \emptyset). \tag{6.24}$$

In words, the criterion $\text{Str-SE}(U_{BN})_X(Y)$ precludes the existence of backdoor paths between X and Y for all exogenous variables that do not fall under business necessity. Having this definition in mind, our task is to find fairness measures that are admissible with respect to $\text{Str-SE}(U_{BN})_X(Y)$. To achieve this goal, we study the problem of decomposing the spurious effects between the attribute X and outcome Y into variations explained by the confounding latent variables $U_1, \dots, U_k$.

6.2.1 Markovian case

We begin the treatment of general spurious decompositions by considering the Markovian (fully observed) causal models and start with several definitions.

Definition 6.3 (Set-specific Experimental Spurious Effect). Let Z be the set of confounders between variables X and Y, and let U_Z be the corresponding latent variables. Suppose $U_A, U_B \subseteq U_Z$ are two nested subsets of U_Z, that is $U_A \subseteq U_B$. We then define the experimental spurious effect with respect to sets U_A, U_B as

$$\text{Exp-SE}_x^{U_A, U_B}(y) = P(y \mid x^{U_A}) - P(y \mid x^{U_B}). \tag{6.25}$$

We provide some intuition for the notion $\text{Exp-SE}_x^{U_A, U_B}(y)$. Suppose that the set of latent confounders $U_Z = (U_1, \dots, U_k)$ is split into three parts $U_Z =$

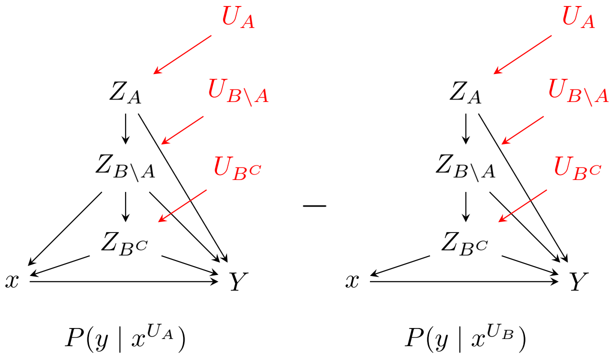


Figure 6.5: Quantity $\text{Exp-SE}_x^{A,B}(y)$ as a graphical contrast.

$(U_A, U_{B\setminus A}, U_{B^C})$. Consider the graphical representation shown in Fig. 6.5. The quantity $P(y \mid x^{U_A})$ computes the variations in Y explained by conditioning on $X = x$ when the confounders U_A are not responding to evidence. This is reflected in the absence of the arrow between U_A (and the corresponding observables Z_A) and $X = x$. In $P(y \mid x^{U_B})$, U_A is still disconnected from, or "unaware" of the fact that $X = x$. Additionally, $U_{B\setminus A}$ is now also disconnected from $X = x$. On both left and the right, the U_{B^C} is connected to $X = x$, which means that U_{B^C} is updated based on the evidence $X = x$. Therefore, the difference of the two quantities captures the spurious variations *explained by the variables* $U_{B\setminus A}$. We now show how this notion allows us to decompose the experimental spurious effect into variable-specific contributions:

Theorem 6.1 (Latent Spurious Decomposition for Markovian models). Let $\mathcal{M}$ be a Markovian model. Let $Z_1, \ldots, Z_k$ be the confounders between variables X and Y sorted in any valid topological order, and denote the corresponding exogenous variables as $U_1, \ldots, U_k$, respectively. Let $Z_{[i]} = \{Z_1, \ldots, Z_i\}$ and $U_{[i]} = \{U_1, \ldots, U_i\}$. The experimental spurious effect $\text{Exp-SE}_x(y)$ can be decomposed into latent variable-specific contributions as follows:

$$\text{Exp-SE}_x(y) = \sum_{i=0}^{k-1} \text{Exp-SE}_x^{U_{[i]}, U_{[i+1]}}(y) \tag{6.26}$$

$$= \sum_{i=0}^{k-1} P(y \mid x^{U_{[i]}}) - P(y \mid x^{U_{[i+1]}}). \tag{6.27}$$

We next provide an illustrative example of applying the theorem:

Example 6.6 (Latent Variable Attribution in a Markovian Model). Consider the

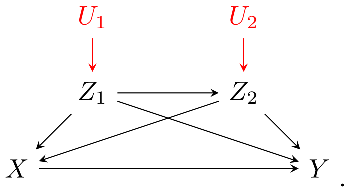


Figure 6.6: Causal diagram from Ex. 6.6 with explicitly drawn latents U_1, U_2.

following SCM $\mathcal{M}^*$:

$$\mathcal{M}^* : \begin{cases} Z_1 \leftarrow B(0.5) & (6.28) \\ Z_2 \leftarrow B(0.4 + 0.2Z_1) & (6.29) \\ X \leftarrow B(0.3 + 0.2Z_1 + 0.2Z_2) & (6.30) \\ Y \leftarrow X + Z_1 + Z_2, & (6.31) \end{cases}$$

and the causal diagram in Fig. 6.6. We wish to decompose the quantity $\text{Exp-SE}_x(y)$ into the variations attributed to the latent variables U_1, U_2. Following the decomposition from Thm. 6.1 we can write

$$\text{Exp-SE}_x(y \mid x_1) = \underbrace{\mathbb{E}(y \mid x_1) - \mathbb{E}(y \mid x_1^{U_1})}_{U_1 \text{ contribution}} + \underbrace{\mathbb{E}(y \mid x_1^{U_1}) - \mathbb{E}(y \mid x_1^{U_1,U_2})}_{U_2 \text{ contribution}}. \tag{6.32}$$

We now need to compute the terms appearing in Eq. 6.32. In particular, we know that

$$\begin{aligned} \mathbb{E}(y \mid x_1^{U_1,U_2}) &= \mathbb{E}(y \mid do(x_1)) & (6.33) \\ &= 1 + \mathbb{E}(Z_1 \mid do(x_1)) + \mathbb{E}(Z_2 \mid do(x_1)) & (6.34) \\ &= 1 + \mathbb{E}(Z_1) + \mathbb{E}(Z_2) = 1 + 0.5 + 0.5 = 2. & (6.35) \end{aligned}$$

Similarly, we can also compute

$$\mathbb{E}(y \mid x_1) = 1 + P(Z_1 = 1 \mid x_1) + P(Z_2 = 1 \mid x_1), \tag{6.36}$$

where $P(Z_1 = 1 \mid x_1)$ can be expanded as

$$P(Z_1 = 1 \mid x_1) = \frac{P(Z_1 = 1, X = 1)}{P(X = 1)} \tag{6.37}$$

$$= \frac{P(Z_1 = 1, X = 1, Z_2 = 1) + P(Z_1 = 1, X = 1, Z_2 = 0)}{P(X = 1)} \tag{6.38}$$

$$= \frac{0.5 \cdot 0.6 \cdot 0.7 + 0.5 \cdot 0.4 \cdot 0.5}{0.5} = 0.62. \tag{6.39}$$

The value of $P(Z_2 = 1 \mid x_1)$ is computed analogously and also equals 0.62, implying that $\mathbb{E}(y \mid x_1) = 1 + 0.62 + 0.62 = 2.24$. Finally, we want to compute $\mathbb{E}(y \mid x_1^{U_1})$, which equals

$$\mathbb{E}(y \mid x_1^{U_1}) = 1 + P(Z_1 = 1 \mid x_1^{U_1}) + P(Z_2 = 1 \mid x_1^{U_1}). \tag{6.40}$$

By definition, $P(Z_1 = 1 \mid x_1^{U_1}) = P(Z_1 = 1) = 0.5$. For $P(Z_2 = 1 \mid x_1^{U_1})$ we write

$$P(Z_2 = 1 \mid x_1^{U_1}) = \sum_{u_1} P(Z_2 = 1 \mid x_1, u_1)P(u_1) \tag{6.41}$$

$$= \sum_{z_1} \sum_{u_1 : Z(u_1) = z_1} P(Z_2 = 1 \mid x_1, u_1)P(u_1) \tag{6.42}$$

$$= \sum_{z_1} P(Z_2 = 1 \mid x_1, z_1)P(z_1) \tag{6.43}$$

$$= \frac{1}{2}\Big[\frac{P(Z_2 = 1, X = 1, Z_1 = 1)}{P(X = 1, Z_1 = 1)} + \frac{P(Z_2 = 1, X = 1, Z_1 = 0)}{P(X = 1, Z_1 = 0)}\Big] \tag{6.44}$$

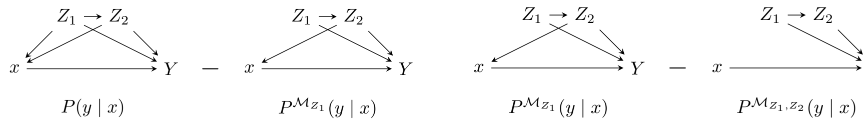
$$= \frac{1}{2}\Big[\frac{0.21}{0.31} + \frac{0.21}{0.31}\Big] \approx 0.68, \tag{6.45}$$

implying that $\mathbb{E}(y \mid x_1^{U_1}) = 2.18$. Putting everything together, we found that

$$\underbrace{\text{Exp-SE}_x(y \mid x_1)}_{=0.24} = \underbrace{\text{Exp-SE}_x^{\emptyset, U_1}(y \mid x_1)}_{=0.06 \text{ from } U_1} + \underbrace{\text{Exp-SE}_x^{U_1, \{U_1, U_2\}}(y \mid x_1)}_{=0.18 \text{ from } U_2}. \tag{6.46}$$

□

The terms appearing on the r.h.s. of Eq. 6.46 are shown as graphical contrasts in Fig. 6.7. On the left side of Fig. 6.7a, U_1, U_2 are responding to the conditioning $X = x$, compared against the right side where only U_2 is responding to the conditioning $X = x$. In the second term, in Fig. 6.7b, on the left only U_2 responds to $X = x$, compared against the right side in which neither U_1 nor U_2 respond to $X = x$ conditioning.


$Z_1 \to Z_2$
x
Y
$P(y \mid x)$
—
$Z_1 \to Z_2$
x
Y
$P^{\mathcal{M}_{Z_1}}(y \mid x)$
$Z_1 \to Z_2$
x
Y
$P^{\mathcal{M}_{Z_1}}(y \mid x)$
—
$Z_1 \to Z_2$
x
Y
$P^{\mathcal{M}_{Z_1,Z_2}}(y \mid x)$

(a) Exp-SE$_x^{\emptyset, U_1}(y)$. **(b)** Exp-SE$_x^{U_1, \{U_1, U_2\}}(y)$.

Figure 6.7: Graphical representation of Exp-SE effect decomposition in Ex. 6.6.

Identification of spurious effects in Markovian models. In Ex. 6.6, however, we used the exclusive knowledge of the SCM $\mathcal{M}^*$. In practice this knowledge is never available, so we need to compute such decompositions based on the observational data and the causal diagram (known as an *identifiability* problem (Pearl, 2000), see Sec. 4.3). In fact, when variables are added to the PA submodel in topological order, the attribution of variations to the latents U_i is identifiable, as we prove next:

Theorem 6.2 (Spurious Decomposition Identification in Topological Ordering). The quantity $P(y \mid x^{U_{[i]}})$ can be computed from observational data using the expression

$$P(y \mid x^{U_{[i]}}) = \sum_z P(y \mid z, x) P(z_{-[i]} \mid z_{[i]}, x) P(z_{[i]}), \tag{6.47}$$

rendering each term of decomposition in Eq. 6.27 identifiable from the observational distribution $P(v)$.

Proof. Notice that fixing a specific value for the variables $(U_1, \dots, U_k) = (u_1, \dots, u_k)$ also gives a unique value for the variables $(Z_1, \dots, Z_k) = (z_1, \dots, z_k)$. Therefore, we can write

$$\begin{aligned}
P(y \mid x^{U_{[i]}}) &= \sum_{u_{[i]}} P(u_{[i]}) P(y \mid x, u_{[i]}) && (6.48)\\
&= \sum_{u_{[i]}} P(u_{[i]}) P(y \mid x, u_{[i]}, z_{[i]}(u_{[i]})) && (6.49)\\
&= \sum_{z_{[i]}} \sum_{u_{[i]}} P(u_{[i]}) \mathbb{1}(Z_{[i]}(u_{[i]}) = z_{[i]}) P(y \mid x, z_{[i]}) && (6.50)\\
&= \sum_{z_{[i]}} P(z_{[i]}) P(y \mid x, z_{[i]}) && (6.51)\\
&= \sum_{z} P(y \mid x, z) P(z_{-[i]} \mid x, z_{[i]}) P(z_{[i]}). && (6.52)
\end{aligned}$$

■

The above proof is based on an important correspondence of the latent variables U_i, and the observed variables Z_i. In particular, there is a 1-to-1 correspondence between the two, since each Z_i is associated with a single U_i. Furthermore, a fixed value of $(u_1, \dots, u_i)$ corresponds to a fixed value of $(z_1, \dots, z_i)$, and this observation can be used to replace the appearances of u_i in Eq. 6.48 by z_i in Eq. 6.51. This observation also demonstrates that a spurious variation explained by $U_1, \dots, U_i$ can also, almost equivalently, be thought of as explained by $Z_1, \dots, Z_i$.

We next show how the ID expression in Eq. 6.47 can be used to compute the effects in Ex. 6.6:

Example 6.7 (Markov Spurious Decomposition with ID)**.** The terms $\text{Exp-SE}_x^{\emptyset, U_1}(y)$, $\text{Exp-SE}_x^{U_1, \{U_1, U_2\}}(y)$ in Ex. 6.6 can be computed from observational data as

$$\text{Exp-SE}_x^{\emptyset, U_1}(y) = \sum_{z_1, z_2} \mathbb{E}(y \mid z_1, z_2, x) P(z_2 \mid z_1, x)[P(z_1 \mid x) - P(z_1)], \tag{6.53}$$

$$\text{Exp-SE}_x^{U_1, \{U_1, U_2\}}(y) = \sum_{z_1, z_2} \mathbb{E}(y \mid z_1, z_2, x)[P(z_2 \mid z_1, x) - P(z_2 \mid z_1)] P(z_1). \tag{6.54}$$

□

For decompositions that follow a topological ordering, we stated a positive identification result in Thm. 6.2. However, when considering decompositions that do not follow a topological ordering, we lose the identifiability of the corresponding effects, as shown in the following example:

Example 6.8 (Non-identification of Latent Spurious Decomposition)**.** Consider two SCMs $\mathcal{M}_1, \mathcal{M}_2$. Both SCMs have the same set of assignment equations $\mathcal{F}$, given by

$$\mathcal{F} := \begin{cases} Z_1 \leftarrow U_1 & (6.55) \\ Z_2 \leftarrow \begin{cases} Z_1 & \text{if } U_2 = 1 \\ 1 - Z_1 & \text{if } U_2 = 2 \\ 1 & \text{if } U_2 = 3 \\ 0 & \text{if } U_2 = 4 \end{cases} & (6.56) \\ X \leftarrow (Z_1 \wedge U_{X1}) \vee (Z_2 \wedge U_{X2}) \vee U_X & (6.57) \\ Y \leftarrow X + Z_1 + Z_2, & (6.58) \end{cases}$$

and the causal diagram given in Fig. 6.6. The two SCMs differ in the distribution over the latent variables. In particular, for $\mathcal{M}_1$ we have

$$P^{\mathcal{M}_1}(U) : \begin{cases} U_1, U_{X1}, U_{X2}, U_X \sim \text{Bernoulli}(0.5) & (6.59) \\ U_2 \sim \text{Multinom}(4, 1, (0, \frac{1}{4}, \frac{1}{2}, \frac{1}{4})), & (6.60) \end{cases}$$

and for $\mathcal{M}_2$

$$P^{\mathcal{M}_2}(U) : \begin{cases} U_1, U_{X1}, U_{X2}, U_X \sim \text{Bernoulli}(0.5) & (6.61) \\ U_2 \sim \text{Multinom}(4, 1, (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0)). & (6.62) \end{cases}$$

That is, the only difference between $P^{\mathcal{M}_1}(U)$ and $P^{\mathcal{M}_2}(U)$ is in how U_2 attains its value. In fact, one can check that the observational distributions $P^{\mathcal{M}_1}(V)$ and $P^{\mathcal{M}_2}(V)$ are the same. However, when computing $\mathbb{E}^{\mathcal{M}}(y \mid x_0^{U_2})$ we have that

$$\mathbb{E}^{\mathcal{M}_1}(y \mid x_0^{U_2}) = 1 \tag{6.63}$$

$$\mathbb{E}^{\mathcal{M}_2}(y \mid x_0^{U_2}) = 0.93, \tag{6.64}$$

showing that the quantity $\mathbb{E}^{\mathcal{M}}(y \mid x_0^{U_2})$ is non-identifiable. □

The example illustrates that even in the Markovian case, when the variables are not considered in a topological order (in the example above, the variable U_2 was considered without the variable U_1 being added first), we might not be able to identify the decomposition of the spurious effects.

6.2.2 Semi-Markovian Models

As discussed above, in the Markovian case that we considered until now, there was a one-to-one correspondence between the observed confounders Z_i and their latent variables U_i. This, however, is no longer the case in Semi-Markovian models. In particular, it can happen that there exist exogenous variables U_j that induce common variations between X, Y, but affect more than one confounder Z_i. We are interested in $U_j \subseteq U$ that have causal (directed) paths to both X, Y, described by the following definition:

Definition 6.4 (Trek). Let $\mathcal{M}$ be an SCM corresponding to a Semi-Markovian model. Let $\mathcal{G}$ be the causal diagram of $\mathcal{M}$. A trek τ in $\mathcal{G}$ (from X to Y) is an ordered pair of causal paths (g_l, g_r) with a common exogenous source $U_i \in U$. That is, g_l is a causal path $U_i \to \cdots \to X$ and g_r is a causal path $U_i \to \cdots \to Y$. The common source U_i is called the top of the trek (ToT for short), denoted $top(g_l, g_r)$. A trek is called spurious if g_r is a causal path from U_i to Y that is not intercepted by X.

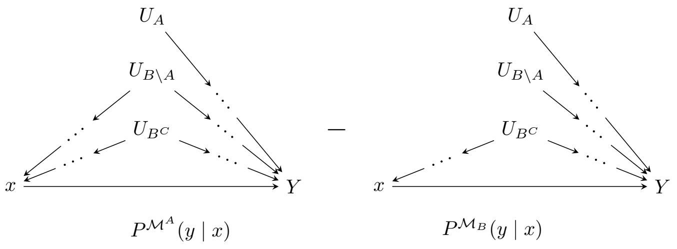


Figure 6.8: Quantity Exp-SE$_x^{U_A,U_B}(y)$ as a graphical contrast. Dots $\cdots$ indicate arbitrary observed confounders along the indicated pathway.

Example 6.9 (Spurious Treks). Consider the causal diagram in Fig. 6.6. In the diagram, latent variables U_1, U_2 both lie on top of a spurious trek because:

$$X \leftarrow Z_1 \leftarrow U_1 \rightarrow Z_1 \rightarrow Y \text{ is a spurious trek with top } U_1$$
$$X \leftarrow Z_2 \leftarrow U_2 \rightarrow Z_2 \rightarrow Y \text{ is a spurious trek with top } U_2.$$

Note also that there are other spurious treks with U_1 on top, such as the trek $X \leftarrow Z_1 \leftarrow U_1 \rightarrow Z_1 \rightarrow Z_2 \rightarrow Y$. □

When decomposing spurious effects, we are in fact interested in all the exogenous variables U_i that lie on top of a spurious trek between X and Y. It is precisely these exogenous variables that induce common variations between X and Y. Using any subset of the variables that are top of spurious treks, we define a set-specific notion of a spurious effect:

Definition 6.5 (Exogenous Set-specific Spurious Effect). Let $U_{sToT} \subseteq U$ be the subset of exogenous variables that lie on top of a spurious trek between X and Y. Suppose $U_A, U_B \subseteq U_{sToT}$ are two nested subsets of U_{sToT}, that is $U_A \subseteq U_B$. We then define the exogenous experimental spurious effect with respect to sets U_A, U_B as

$$\text{Exp-SE}_x^{U_A,U_B}(y) = P(y \mid x^{U_A}) - P(y \mid x^{U_B}). \tag{6.65}$$

The above definition is analogous to Def. 6.3, but we are now fixing different subsets of the tops of spurious treks. We present the quantity Exp-SE$_x^{U_A,U_B}(y)$ as a graphical contrast in Fig. 6.8. In particular, the set of tops of spurious treks U_{sToT} is partitioned into three parts $(U_A, U_{B\setminus A}, U_{B^C})$. The causal diagram in the figure is informal, and the dots $(\cdots)$ represent arbitrary possible observed confounders that lie along indicated pathways. On the l.h.s. of the figure, the set U_A does not respond to the conditioning $X = x$, whereas $U_{B\setminus A}, U_{B^C}$ do. This is contrasted with the r.h.s., in which neither U_A nor $U_{B\setminus A}$ respond to $X = x$, whereas U_{B^C} still does respond to the $X = x$ conditioning. The

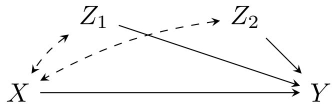


Figure 6.9: Causal diagram of the SCM in Ex. 6.10.

described contrast thus captures the spurious effect explained by the tops of spurious treks in $U_{B\setminus A}$.

Analogous to Thm. 6.1, we next state a variable-specific decomposition of the spurious effect, which is now with respect to exogenous variables that are top of spurious treks:

Theorem 6.3 (Semi-Markovian Spurious Decomposition)**.** Let $U_{sToT} \subseteq U$ be the subset $\{U_1, \dots, U_m\}$ of exogenous variables that lie on top of a spurious trek between X and Y. Let $U_{[i]}$ denote the variables $U_1, \dots, U_i$ ($U_{[0]}$ denotes the empty set $\emptyset$). The experimental spurious effect Exp-SE$_x(y)$ can be decomposed into variable-specific contributions as follows:

$$\text{Exp-SE}_x(y) = \sum_{i=0}^{m-1} \text{Exp-SE}_x^{U_{[i]}, U_{[i+1]}}(y) \tag{6.66}$$

$$= \sum_{i=0}^{k-1} P(y \mid x^{U_{[i]}}) - P(y \mid x^{U_{[i+1]}}). \tag{6.67}$$

We next given an example demonstrating the Semi-Markovian decomposition:

Example 6.10 (Semi-Markovian Spurious Decomposition)**.** Consider the following SCM $\mathcal{M}$:

$$\mathcal{F}, P(U) : \begin{cases} Z_1 \leftarrow U_1 \wedge U_{1X} & (6.68) \\ Z_2 \leftarrow U_2 \vee U_{2X} & (6.69) \\ X \leftarrow U_X \wedge (U_{1X} \vee U_{2X}) & (6.70) \\ Y \leftarrow X + Z_1 + Z_2 & (6.71) \\ \\ U_1, U_2, U_{1X}, U_{2X}, U_X \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(0.5). & (6.72) \end{cases}$$

The causal diagram $\mathcal{G}$ associated with $\mathcal{M}$ is given in Fig. 6.9. The exogenous variables that lie on top of a spurious trek are U_{1X}, U_{2X}. Therefore, following the decomposition from Thm. 6.3, we can attribute spurious variations to these two variables:

$$\begin{aligned}\text{Exp-SE}_x(y \mid x_1) = &\underbrace{\mathbb{E}(y \mid x_1) - \mathbb{E}(y \mid x_1^{U_{1X}})}_{U_{1X}\text{ contribution}} \qquad (6.73)\\ &+ \underbrace{\mathbb{E}(y \mid x_1^{U_{1X}}) - \mathbb{E}(y \mid x_1^{U_{1X}, U_{2X}})}_{U_{2X}\text{ contribution}}.\end{aligned}$$

We now compute the terms appearing in Eq. 6.73. In particular, we know that

$$\mathbb{E}(y \mid x_1^{U_{1X}, U_{2X}}) = \mathbb{E}(y \mid do(x_1)) = 1 + \mathbb{E}(Z_1 \mid do(x_1)) + \mathbb{E}(Z_1 \mid do(x_1)) \qquad (6.74)$$

$$= 1 + \mathbb{E}(Z_1) + \mathbb{E}(Z_2) = 1 + 0.25 + 0.75 = 2. \qquad (6.75)$$

Similarly, we can also compute

$$\mathbb{E}(y \mid x_1) = 1 + P(Z_1 = 1 \mid x_1) + P(Z_2 = 1 \mid x_1), \qquad (6.76)$$

Now, $P(Z_1 = 1 \mid x_1) = \frac{P(Z_1=1,x_1)}{P(x_1)}$, and we know that $X = 1$ if and only if $U_X = 1$ and $U_{1X} \vee U_{2X} = 1$, which happen independently with probabilities $\frac{1}{2}$ and $\frac{3}{4}$, respectively. Next, $Z_1 = 1, X = 1$ happens if and only if $U_X = 1, U_{1X} = 1$ and $U_1 = 1$, which happens with probability $\frac{1}{8}$. Therefore, we can compute

$$P(Z_1 = 1 \mid x_1) = \frac{\frac{1}{8}}{\frac{1}{2}\cdot\frac{3}{4}} = \frac{1}{3}. \qquad (6.77)$$

Furthermore, we similarly compute that $Z_2 = 1, X = 1$ happens if either $U_X = 1, U_{2X} = 1$ or $U_X = 1, U_{2X} = 0, U_2 = 1, U_{1X} = 1$ which happens disjointly with probabilities $\frac{1}{4}$, $\frac{1}{16}$, respectively. Therefore,

$$P(Z_2 = 1 \mid x_1) = \frac{\frac{1}{4} + \frac{1}{16}}{\frac{1}{2}\cdot\frac{3}{4}} = \frac{5}{6}. \qquad (6.78)$$

Putting everything together we obtain that

$$\mathbb{E}(y \mid x_1) = 1 + \frac{1}{3} + \frac{5}{6} = \frac{13}{6}. \qquad (6.79)$$

Finally, we want to compute $\mathbb{E}(y \mid x_1^{U_{1X}})$, which equals

$$\mathbb{E}(y \mid x_1^{U_{1X}}) = 1 + P(Z_1 = 1 \mid x_1^{U_{1X}}) + P(Z_2 = 1 \mid x_1^{U_{1X}}). \qquad (6.80)$$

Now, to evaluate these expressions, we distinguish two cases, namely (i) $U_{1X} = 1$ and (ii) $U_{1X} = 0$. In the first case, $P(Z_1 \mid x_1) = \frac{1}{2}$ and $P(Z_2 = 1 \mid x_1) = \frac{3}{4}$.

In the second case, $P(Z_1 \mid x_1) = 0$ and $P(Z_2 = 1 \mid x_1) = 1$. Therefore, we can compute

$$P(Z_1 = 1 \mid x_1^{U_{1X}}) = \frac{1}{2}P_{U_{1X}=1}(Z_1 \mid x_1) + \frac{1}{2}P_{U_{1X}=0}(Z_1 \mid x_1) = \frac{1}{4} \tag{6.81}$$

$$P(Z_2 = 1 \mid x_1^{U_{1X}}) = \frac{1}{2}P_{U_{1X}=1}(Z_2 \mid x_1) + \frac{1}{2}P_{U_{1X}=0}(Z_2 \mid x_1) = \frac{7}{8}, \tag{6.82}$$

which implies that $\mathbb{E}(y \mid x_1^{U_{1X}}) = \frac{17}{8}$. Finally, this implies that

$$\underbrace{\text{Exp-SE}_x(y \mid x_1)}_{=\frac{1}{6}} = \underbrace{\text{Exp-SE}_x^{\emptyset, U_{1X}}(y \mid x_1)}_{=\frac{1}{24}\text{ from } U_{1X}} + \underbrace{\text{Exp-SE}_x^{U_{1X}, \{U_{1X}, U_{2X}\}}(y \mid x_1)}_{=\frac{1}{8}\text{ from } U_{2X}}. \tag{6.83}$$

□

The terms appearing on the r.h.s. of Eq. 6.83 are shown as graphical contrasts in Fig. 6.4. On the left side of Fig. 6.4a, U_{1X}, U_{2X} are responding to the conditioning $X = x$, compared against the right side where only U_{2X} is responding to the conditioning $X = x$. In the second term, in Fig. 6.4b, on the left only U_{2X} responds to $X = x$, compared against the right side in which neither U_{1X} nor U_{2X} respond to $X = x$ conditioning.

After introducing the Semi-Markovian spurious decomposition, we can now show its importance in the context of Causal Fairness Analysis, namely its admissibility to the structural spurious criterion under business necessity:

Proposition 6.2 (Admissibility of Exogenous Spurious Effects)**.** Let $U_{BN} \subseteq U$ be a subset of the exogenous confounders of X, Y that fall under business necessity. Let U_{BN}^C denote the exogenous ancestors of X that do not fall under business necessity, that is $U_{BN}^C = \text{an}^{\text{ex}}(X) \setminus U_{BN}$. Then the measures $\text{Exp-SE}_x^{\emptyset, U_{BN}^C}(y), \text{Exp-SE}_x^{U_{BN}, U}(y)$ are admissible with respect to the structural criterion $\text{Str-SE}(U_{BN})_X(Y)$, that is

$$(\text{Str-SE}(U_{BN})_X(Y) = 0) \implies (\text{Exp-SE}_x^{\emptyset, U_{BN}^C}(y) = 0) \tag{6.84}$$

$$(\text{Str-SE}(U_{BN})_X(Y) = 0) \implies (\text{Exp-SE}_x^{U_{BN}, U}(y) = 0). \tag{6.85}$$

Proof. Assume that $\text{Str-SE}(U_{BN})_X(Y)$, meaning that the set of exogenous variables not in the business necessity set that lie on top of a spurious trek between X and Y is empty. We can expand the two quantities of interest using the definition as

$$\text{Exp-SE}_x^{\emptyset, U_{BN}^C}(y) = P(y \mid x^{\emptyset}) - P(y \mid x^{U_{BN}^C}) \tag{6.86}$$

$$\text{Exp-SE}_x^{U_{BN}, U}(y) = P(y \mid x^{U_{BN}}) - P(y \mid x^{U}). \tag{6.87}$$

Note that $P(y \mid x^{\emptyset}) = P(y \mid x)$, so to prove that $\text{Exp-SE}_x^{\emptyset, U^C_{BN}}(y) = 0$, we want to show that $P(y \mid x^{U^C_{BN}}) = P(y \mid x)$. Using Prop. 6.1, we compute:

$$P(y \mid x^{U^C_{BN}}) = \sum_{u^C_{BN}} P(u^C_{BN}) P(y \mid x, u^C_{BN}) \tag{6.88}$$

$$= \sum_{u^C_{BN}} P(u^{C,x}_{BN}) P(u^{C,y}_{BN}) P(y \mid x, u^{C,x}_{BN}, u^{C,y}_{BN}), \tag{6.89}$$

where $U^{C,x}_{BN}$ are the exogenous variables not in BN-set with a path to X, and $U^{C,y}_{BN}$ with a path to Y (these two sets are disjoint by assumption). Now, we know that the following independence relations hold (using the assumption):

$$U^{C,x}_{BN} \perp\!\!\!\perp Y \mid X, U^{C,y}_{BN}, \tag{6.90}$$

$$U^{C,y}_{BN} \perp\!\!\!\perp X. \tag{6.91}$$

Hence we can write

$$P(y \mid x^{U^C_{BN}}) = \sum_{u^{C,x}_{BN}} \sum_{u^{C,y}_{BN}} P(u^{C,x}_{BN}) P(u^{C,y}_{BN}) P(y \mid x, u^{C,x}_{BN}, u^{C,y}_{BN}) \tag{6.92}$$

$$= \sum_{u^{C,x}_{BN}} \sum_{u^{C,y}_{BN}} P(u^{C,x}_{BN}) P(u^{C,y}_{BN}) P(y \mid x, u^{C,y}_{BN}) \quad \text{(Eq. 6.90)} \tag{6.93}$$

$$= \sum_{u^{C,x}_{BN}} P(u^{C,x}_{BN}) \sum_{u^{C,y}_{BN}} P(u^{C,y}_{BN}) P(y \mid x, u^{C,y}_{BN}) \tag{6.94}$$

$$= \sum_{u^{C,y}_{BN}} P(u^{C,y}_{BN}) P(y \mid x, u^{C,y}_{BN}) \tag{6.95}$$

$$= \sum_{u^{C,y}_{BN}} P(u^{C,y}_{BN} \mid x) P(y \mid x, u^{C,y}_{BN}) \quad \text{(Eq. 6.91)} \tag{6.96}$$

$$= P(y \mid x). \tag{6.97}$$

To show that $\text{Exp-SE}_x^{U_{BN}, U}(y) = 0$, notice that $P(y \mid x^U) = P(y_x)$. Therefore, we need to show that $P(y \mid x^{U_{BN}}) = P(y_x)$. By definition, we have that

$$P(y \mid x^{U_{BN}}) = \sum_{u_{BN}} P(u_{BN}) P(y \mid x, u_{BN}). \tag{6.98}$$

Take any observed variable that lies on top of an open back-door path between X and Y, say Z_i. Then, every exogenous ancestor of Z_i must lie in the U_{BN} set (otherwise the assumption is violated). Furthermore, for any open spurious trek with top U_i, U_i must also be in U_{BN}. This implies that

$$Y_x \perp\!\!\!\perp X \mid U_{BN}. \tag{6.99}$$

Hence we can write

$$\begin{aligned}
P(y \mid x^{U_{BN}}) &= \sum_{u_{BN}} P(u_{BN}) P(y \mid x, u_{BN}) && (6.100)\\
&= \sum_{u_{BN}} P(u_{BN}) P(y_x \mid x, u_{BN}) \quad \text{(Consistency axiom)} && (6.101)\\
&= \sum_{u_{BN}} P(u_{BN}) P(y_x \mid u_{BN}) \quad \text{(Eq. 6.99)} && (6.102)\\
&= P(y_x). && (6.103)
\end{aligned}$$

■

Identification of spurious effects in Semi-Markovian Models. In practice, however, we need to compute the spurious decompositions from observational data and the causal diagram. The first difficulty comes from the fact that exogenous variables U are not drawn out explicitly in the causal diagram $\mathcal{G}$. Therefore, to determine the exogenous variables that can possibly lie on top of a spurious trek, we give the following definition:

Definition 6.6 (Top of trek from the causal diagram). Let $\mathcal{M}$ be a Semi-Markovian model and let $\mathcal{G}$ be the associated causal diagram. The set of variables U_{sToT} can be constructed from the causal diagram in the following way:

(I) initialize $U_{sToT} = \emptyset$,

(II) for each bidirected edge $V_i \dashleftarrow\dashrightarrow V_j$ consider the associated exogenous variable U_{ij}; if there exists a spurious trek from U_{ij} to X and Y, add U_{ij} to U_{sToT},

(III) for each observed confounder Z_i, consider the associated exogenous variable U_i; if there exists a spurious trek from U_i to X and Y, add U_i to U_{sToT}.

Example 6.11 (continued - U_{sToT} Construction). We continue with Ex. 6.10 and the causal graph in Fig. 6.9 and perform the steps as follows:

(i) initialize $U_{sToT} = \emptyset$,

(ii) consider bidirected edges $X \dashleftarrow\dashrightarrow Z_1$ and $X \dashleftarrow\dashrightarrow Z_2$:

- U_{1X} associated with $X \dashleftarrow\dashrightarrow Z_1$ lies on top of a spurious trek,
- U_{2X} associated with $X \dashleftarrow\dashrightarrow Z_2$ lies on top of a spurious trek,

(iii) consider the observed confounders Z_1, Z_2 and their associated latent variables U_1, U_2:

- U_1, U_2 do not lie on top of spurious treks between X and Y.

Therefore, we have constructed the set $U_{sToT} = \{U_{1X}, U_{2X}\}$. □

After defining the explicit construction of the set U_{sToT}, we define the notion of the anchor set:

Definition 6.7 (Anchor Set)**.** Let U_{sToT} be the subset of the exogenous variables that lie on top of a spurious trek between X and Y. Let $U_1, \ldots U_l \subseteq U_{sToT}$ be a subset of these variables. We define the anchor set $\mathrm{AS}(U_1, \ldots, U_l)$ of $(U_1, \ldots, U_l)$ as the subset of observables V that are different from X and are directly influenced by any of the U_is,

$$\mathrm{AS}(U_1, \ldots, U_l) = \bigcup_{i=1}^{l} \mathrm{ch}(U_i) \setminus X. \tag{6.104}$$

Example 6.12 (continued - Anchor Set)**.** For the set $U_{sToT} = \{U_{1X}, U_{2X}\}$ associated with the causal diagram in Fig. 6.9, the anchor sets can be computed as follows:

$$\mathrm{AS}(U_{1X}) = Z_1, \tag{6.105}$$

$$\mathrm{AS}(U_{2X}) = Z_2, \tag{6.106}$$

$$\mathrm{AS}(U_{1X}, U_{2X}) = \{Z_1, Z_2\}. \tag{6.107}$$

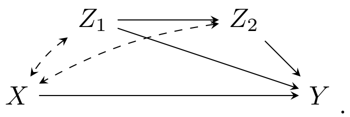
□

Another important definition is that of anchor set exogenous ancestral closure:

Definition 6.8 (Anchor Set Exogenous Ancestral Closure)**.** Let $U_s \subseteq U_{sToT}$ be a subset of the exogenous variables on top of a spurious trek between X and Y. Let $\mathrm{AS}(U_s)$ denote the anchor set of U_s, and let $\mathrm{an}^{\mathrm{ex}}_{sToT}(AS(U_s))$ denote all exogenous variables in U_{sToT} that have a path causal path to any variable in $\mathrm{AS}(U_s)$. U_s is said to satisfy anchor set exogenous ancestral closure (ASEAC) if

$$U_s = \mathrm{an}^{\mathrm{ex}}_{sToT}(AS(U_s)). \tag{6.108}$$

Example 6.13 (continued - Anchor Set Exogenous Ancestral Closure)**.** Consider the following causal diagram



With respect to the diagram, we have that

$$\begin{aligned} \text{an}^{\text{ex}}_{sToT}(\text{AS}(U_{1X})) &= U_{1X}, && (6.109)\\ \text{an}^{\text{ex}}_{sToT}(\text{AS}(U_{2X})) &= \{U_{1X}, U_{2X}\}, && (6.110)\\ \text{an}^{\text{ex}}_{sToT}(\text{AS}(\{U_{1X}, U_{2X}\})) &= \{U_{1X}, U_{2X}\}. && (6.111)\end{aligned}$$

Therefore, U_{1X} and $\{U_{1X}, U_{2X}\}$ satisfy anchor set exogenous ancestral closure, whereas U_{2X} does not, since U_{2X} has Z_2 in its anchor set, but Z_2 has U_{1X} as its ancestor. □

Based on the above, we provide a sufficient condition for identification in the Semi-Markovian case (the theorem's proof is given in Appendix A.9):

Theorem 6.4 (ID of Variable Spurious Effects in Semi-Markovian Models)**.** Let $U_s \subseteq U_{sToT}$. The quantity $P(y \mid x^{U_s})$ is identifiable from observational data $P(V)$ if the following hold:

(i) $Y \notin \text{AS}(U_s)$,

(ii) U_s satisfies anchor set exogenous ancestral closure, $U_s = \text{an}^{\text{ex}}_{sToT}(AS(U_s))$.

Example 6.14 (continued - Decomposition ID)**.** Consider the causal diagram in Fig. 6.9. We previously derived that the tops of spurious treks are given $U_{sToT} = \{U_{1X}, U_{2X}\}$ and computed the anchor sets as:

$$\text{AS}(U_{1X}) = Z_1, \text{AS}(U_{2X}) = Z_2, \text{AS}(U_{1X}, U_{2X}) = \{Z_1, Z_2\}. \qquad (6.112)$$

Furthermore, we can compute that $\text{an}^{\text{ex}}_{sToT}(\text{AS}(U_{1X})) = U_{1X}$. Similarly, we have $\text{an}^{\text{ex}}_{sToT}(\text{AS}(\{U_{1X}, U_{2X}\})) = \{U_{1X}, U_{2X}\}$. Thus, both U_{1X} and $\{U_{1X}, U_{2X}\}$ satisfy ASEAC from Def. 6.7. Therefore, $\mathbb{E}(y \mid x_1^{U_{1X}})$ and $\mathbb{E}(y \mid x_1^{U_{1X}, U_{2X}})$ are identifiable by the conditions in Thm. 6.4. In particular, in this case we can derive the expressions:

$$\text{Exp-SE}_x^{\emptyset, U_{1X}}(y) = \sum_{z_1, z_2} \mathbb{E}(y \mid z_1, z_2, x) P(z_2 \mid x)[P(z_1 \mid x) - P(z_1)], \qquad (6.113)$$

$$\text{Exp-SE}_x^{U_{1X}, \{U_{1X}, U_{2X}\}}(y) = \sum_{z_1, z_2} \mathbb{E}(y \mid z_1, z_2, x)[P(z_2 \mid x) - P(z_2)]P(z_1). \qquad (6.114)$$

□

6.2.3 x-specific Spurious Decompositions

So far, we focused on decomposing the Exp-SE$_x(y)$ quantity. However, recall that in Sec. 4 and the Fairness Map (Thm. 4.8) we also considered an x-specific analogue of the Exp-SE$_x(y)$. This measure was called Ctf-SE$_{x_0,x_1}(y)$ or x-SE$_{x_0,x_1}(y)$ (see Def. 4.5), and was defined as:

$$\text{Ctf-SE}_{x_0,x_1}(y) = P(y_{x_0} \mid x_1) - P(y \mid x_0). \tag{6.115}$$

To derive the decomposition of the Ctf-SE quantity, we need to define a slightly more flexible notion of an integrated submodel, namely:

Definition 6.9 (Doubly Partially Abducted Submodel)**.** The doubly partially abducted submodel with respect to prior evidence $E_0 = e_0$, latent variables U_1, and evidence $E_1 = e_1$ is defined as:

$$\mathcal{M}^{E_0=e_0,U_1,E_1=e_1} = \langle \mathcal{F}, P(u_1 \mid E_0 = e_0)P(u_2 \mid u_1, E_1 = e_1)\rangle. \tag{6.116}$$

The definition of the DPA submodel is an extension of the PA submodel definition. In a PA submodel, the variables U_1 do not respond to the evidence. In a DPA submodel, the variables U_1 have previously been updated according to other, prior evidence. For example, DPA submodel $\mathcal{M}^{x_0,U_1,x_1}$ can be described as follows. First, latent variables U_1 are updated according to $X = x_0$. After this, we update the remaining variables U_2 according to $X = x_1$, while U_1 does not respond to this update. The following proposition shows how a DPA submodel is used to compute conditional probabilities:

Proposition 6.3 (DPA Submodel Conditional Probabilities)**.** Let $P(Y = y \mid E = e^{U_1,E_0=e_0})$ denote the conditional probability of the event $Y = y$ conditional on evidence $E_1 = e_1$, while the exogenous variables U_1 are updated according to prior evidence $E_0 = e_0$. Then, $P(Y = y \mid E = e^{U_1,E_0=e_0})$ equals:

$$\sum_{u_1} P(U_1 = u_1 \mid E_0 = e_0)P(Y = y \mid E_1 = e_1, U_1 = u_1). \tag{6.117}$$

Based on the DPA submodel notion, we can also define a notion of an x-specific spurious effect:

Definition 6.10 (x-specific Spurious Effects)**.** Let $U_{sToT} \subseteq U$ be the subset of exogenous variables that lie on top of a spurious trek between X and Y. Suppose $U_A, U_B \subseteq U_{sToT}$ are two nested subsets of U_{sToT}, that is $U_A \subseteq U_B$. We define the counterfactual spurious effect with respect to sets U_A, U_B as:

$$\text{Ctf-SE}^{U_A,U_B}_{x_0,x_1}(y) = P(y \mid x_0^{U_A,x_1}) - P(y \mid x_0^{U_B,x_1}). \tag{6.118}$$

Based on the above notion, a variable-specific decomposition of the Ctf-SE$_{x_0,x_1}(y)$ quantity can be obtained, which is analogous to the decomposition of the quantity Exp-SE$_x(y)$ from Thm. 6.3. Therefore, we state the decomposition result as a corollary of Thm. 6.3:

Corollary 6.5 (Ctf-SE Decomposition). Let $U_{sToT} = \{U_1, \dots, U_m\} \subseteq U$ be the subset of exogenous variables that lie on top of a spurious trek between X and Y. Let $U_{[i]}$ denote the variables $U_1, \dots, U_i$ ($U_{[0]}$ denotes the empty set $\emptyset$). The x-specific spurious effect, Ctf-SE$_{x_0,x_1}(y)$ can be decomposed into variable-specific contributions as follows:

$$\begin{aligned}\text{Ctf-SE}_{x_0,x_1}(y) &= P(y_{x_0} \mid x_1) - P(y \mid x_0) && (6.119)\\ &= \sum_{i=0}^{m-1} \text{Ctf-SE}_{x_0,x_1}^{U_{[i]},U_{[i+1]}}(y) && (6.120)\\ &= \sum_{i=0}^{k-1} P(y \mid x_0^{U_{[i]},x_1}) - P(y \mid x_0^{U_{[i+1]},x_1}). && (6.121)\end{aligned}$$

The admissibility results for the decomposition presented above are similar as for the Exp-SE decomposition. Furthermore, a very similar identifiability result also holds. These results are, in the interest of space, not explicitly mentioned again (see Prop. 6.2 and Thm. 6.4 for reference).

6.3 Refining Indirect Effects

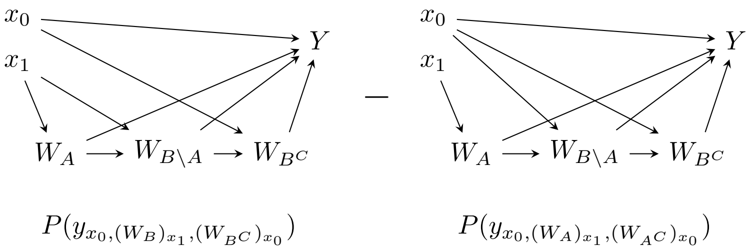
After introducing the conceptual underpinnings for decomposing spurious effects, we now handle indirect effects in a similar fashion. Here, however, we build on some previous literature. We start by defining a structural criterion of indirect fairness, under business necessity.

Definition 6.11 (Structural Indirect Criterion under Business Necessity). Let $\mathcal{M}$ be a Semi-Markovian SCM. Let W be the set of mediators between X and Y. Let W_{BN}, W_{BN}^C be a partition of the mediators, where W_{BN} is the subset of the variables which fall under business necessity. We then define the structural indirect criterion under business necessity as

$$\text{Str-IE}(W_{BN})_X(Y) = \mathbb{1}(\text{an}(Y) \cap \text{ch}(X) \cap W_{BN}^C = \emptyset). \qquad (6.122)$$

Similarly to Def. 3.2, the criterion Str-IE-BN$_X(Y)$ is an idealized notion that qualitatively describes whether there is discrimination based on the variables which do not fall under business necessity (set W_{BN}^C). Similarly as before, we want to find measures of fairness that are admissible with respect to the criterion Str-IE$(W_{BN})_X(Y)$. To do so, we first define the notion of an indirect effect for a subset of the mediators W:

Definition 6.12 (Set-specific Indirect Effects). Let W_A, W_B be nested subsets of the mediators W, so that $W_A \subseteq W_B$. Let W_{A^C} and W_{B^C} denote the



$P(y_{x_0,(W_B)_{x_1},(W_{B^C})_{x_0}})$ $P(y_{x_0,(W_A)_{x_1},(W_{A^C})_{x_0}})$

Figure 6.10: Quantity $E\text{-IE}^{W_A,W_B}_{x_0,x_1}(y)$ as a graphical contrast.

complements of W_A, W_B in W. We then define the E-specific indirect effect with respect to sets W_A, W_B as

$$E\text{-IE}^{W_A,W_B}_{x_0,x_1}(y) = P(y_{x_0,(W_B)_{x_1},(W_{B^C})_{x_0}} \mid E) - P(y_{x_0,(W_A)_{x_1},(W_{A^C})_{x_0}} \mid E). \tag{6.123}$$

The graphical contrast for the quantity $E\text{-IE}^{W_A,W_B}_{x_0,x_1}(y)$ is shown in Fig. 6.10. In particular, on the r.h.s. we have that W_A "listens to" $X = x_1$, whereas $W_{B\setminus A}, W_{B^C}$ listen to $X = x_0$. This is contrasted with the l.h.s. in which $W_A, W_{B\setminus A}$ listen to $X = x_1$, while W_{B^C} stills listens to $X = x_0$. Hence, the contrast captures the change in outcome mediated by variables $W_{B\setminus A}$ obtained by changing $x_0 \to x_1$. Such a notion allows us to find a variable-level decomposition of any E-specific indirect effect:

Theorem 6.6 (Variable Decomposition of Indirect Effects)**.** Let $W_1, \dots, W_k$ denote the set of mediators, sorted in a topological order. Define $W_{[i]}$ as the set $\{W_1, \dots, W_i\}$ and $W_{-[i]}$ as $\{W_{i+1}, \dots, W_k\}$. The E-specific indirect effect can then be decomposed as

$$\begin{aligned}
E\text{-IE}_{x_0,x_1}(y) &= P(y_{x_0,W_{x_1}} \mid E) - P(y_{x_0} \mid E) && (6.124)\\
&= \sum_{i=0}^{k-1} E\text{-IE}^{W_{[i]},W_{[i+1]}}_{x_0,x_1}(y) && (6.125)\\
&= \sum_{i=0}^{k-1} \big[P(y_{x_0,(W_{[i+1]})_{x_1},(W_{-[i+1]})_{x_0}} \mid E) && (6.126)\\
&\qquad - P(y_{x_0,(W_{[i]})_{x_1},(W_{-[i]})_{x_0}} \mid E)\big].
\end{aligned}$$

The theorem allows us to attribute the variations in the indirect effect to specific variables that explain them. For concreteness, we show the decomposition for the natural direct effect introduced in Def. 4.2:

Corollary 6.7 (Variable decomposition of NIE)**.** By choosing the event $E = \emptyset$ in Thm. 6.6 we obtain the decomposition for the natural indirect effect into

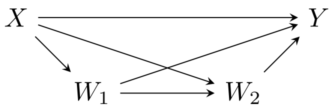


Figure 6.11: Causal diagram of Ex. 6.15.

variable specific contributions

$$\begin{aligned}
\text{NIE}_{x_0,x_1}(y) &= P(y_{x_0,W_{x_1}}) - P(y_{x_0}) && (6.127)\\
&= \sum_{i=0}^{k-1} \text{NIE}_{x_0,x_1}^{W_{[i]},W_{[i+1]}}(y) && (6.128)\\
&= \sum_{i=0}^{k-1} \big[P(y_{x_0,(W_{[i+1]})_{x_1},(W_{-[i+1]})_{x_0}}) && (6.129)\\
&\qquad - P(y_{x_0,(W_{[i]})_{x_1},(W_{-[i]})_{x_0}})\big].
\end{aligned}$$

We now work out the NIE decomposition on a specific example:

Example 6.15 (Indirect Effect decomposition)**.** Consider the following SCM:

$$\mathcal{M}^* : \begin{cases} X \leftarrow B(0.5) & (6.130)\\ W_1 \leftarrow B(0.4 + 0.2X) & (6.131)\\ W_2 \leftarrow W_1 + B(0.4 + 0.2X) & (6.132)\\ Y \leftarrow X + W_1 + W_2 & (6.133) \end{cases}$$

The causal diagram $\mathcal{G}$ associated with $\mathcal{M}^*$ is shown in Fig. 6.11. Following the decomposition from Thm. 6.6 we can write

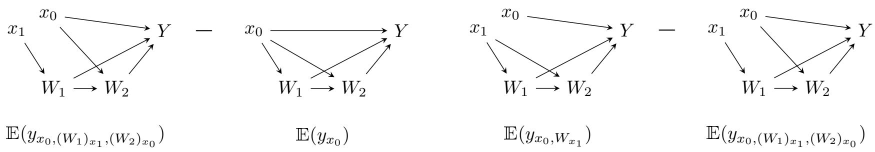
$$\begin{aligned}
\text{NIE}_{x_0,x_1}(y) = &\underbrace{\mathbb{E}(y_{x_0,W_{x_1}}) - \mathbb{E}(y_{x_0,(W_1)_{x_1},(W_2)_{x_0}})}_{W_2 \text{ contribution}} && (6.134)\\
&+ \underbrace{\mathbb{E}(y_{x_0,(W_1)_{x_1},(W_2)_{x_0}}) - \mathbb{E}(y_{x_0})}_{W_1 \text{ contribution}}.
\end{aligned}$$

We now need to compute the terms appearing in Eq. 6.134. In particular, we know that

$$\begin{aligned}
\mathbb{E}(y_{x_0}) = \mathbb{E}(y \mid x_0) &= \mathbb{E}(W_1 \mid x_0) + \mathbb{E}(W_2 \mid x_0) && (6.135)\\
&= 0.4 + 0.8 = 1.2 && (6.136)
\end{aligned}$$

Similarly, we can also compute

$$\begin{aligned}
\mathbb{E}(y_{x_0,W_{x_1}}) &= \mathbb{E}(W_1 \mid x_1) + \mathbb{E}(W_2 \mid x_1) && (6.137)\\
&= 0.6 + 1.2 = 1.8. && (6.138)
\end{aligned}$$


x_0
x_1
Y
W_1
W_2
$-$
$\mathbb{E}(y_{x_0,(W_1)_{x_1},(W_2)_{x_0}})$
$\mathbb{E}(y_{x_0})$
$\mathbb{E}(y_{x_0,W_{x_1}})$
$\mathbb{E}(y_{x_0,(W_1)_{x_1},(W_2)_{x_0}})$

(a) $\text{NIE}^{\emptyset,W_1}(y)$. **(b)** $\text{NIE}^{W_1,\{W_1,W_2\}}(y)$.

Figure 6.12: Graphical representation of how the NIEis decomposed in Ex. 6.15.

Finally, again using the f_y mechanism in the SCM $\mathcal{M}$ we compute that

$$\mathbb{E}(y_{x_0,(W_1)_{x_1},(W_2)_{x_0}}) = \mathbb{E}((W_1)_{x_1}) + \mathbb{E}((W_2)_{x_0,(W_1)_{x_1}}) \tag{6.139}$$

$$= \mathbb{E}(W_1 \mid x_1) + \mathbb{E}(\mathbb{E}(W_2 \mid x_0, (W_1)_{x_1})) \tag{6.140}$$

$$= 0.6 + 1 = 1.6. \tag{6.141}$$

Putting everything together, we found that

$$\underbrace{\text{NIE}_{x_0,x_1}(y)}_{=0.6} = \underbrace{\text{NIE}^{\emptyset,W_1}_{x_0,x_1}(y)}_{=0.2 \text{ from } W_1} + \underbrace{\text{NIE}^{W_1,\{W_1,W_2\}}_{x_0,x_1}(y)}_{=0.4 \text{ from } W_2}. \tag{6.142}$$

□

The two terms appearing the decomposition of the NIE in Ex. 6.15 are represented as graphical contrasts in Fig. 6.12. The first term, $\text{NIE}^{\emptyset,W_1}_{x_0,x_1}(y)$ compares the outcome Y when W_1 responds to $X = x_1$ and W_2, Y respond to $X = x_0$, against the outcome when all of W_1, W_2, and Y respond to $X = x_0$. This quantity captures the effect explained by the variable W_1, which behaves differently between the two terms. The second term $\text{NIE}^{W_1,\{W_1,W_2\}}_{x_0,x_1}(y)$ compares the outcome Y when W_1, W_2 respond to $X = x_1$ and Y responds to $X = x_0$, against the outcome Y when W_1 responds to x_1 and W_2, Y respond to $X = x_0$. This quantity captures the variations explained by W_2, which behaves differently between the potential response $y_{x_0,W_{x_1}}$ on the left, and the potential response $y_{x_0,(W_1)_{x_1},(W_2)_{x_0}}$ on the right. This decomposition shows us how we can distinguish between different variations within the indirect effect. We now show formally why the set-specific measures of indirect effect from Def. 6.12 are useful in practice, by proving their admissibility with respect to the structural criterion Str-IE-BN$_X(Y)$ from Def. 6.11:

Lemma 6.8 (Admissibility of Set-specific Indirect Effects)**.** Let $W_{BN} \subseteq W$ be a subset of the mediators that fall under business necessity. Then the measure $E\text{-IE}^{\emptyset,W^C_{BN}}_{x_0,x_1}(y)$ is admissible with respect to the structural criterion Str-IE$(W_{BN})_X(Y)$, that is

$$(\text{Str-IE-BN}_X(Y) = 0) \implies (E\text{-IE}^{\emptyset,W^C_{BN}}_{x_0,x_1}(y) = 0), \tag{6.143}$$

$$(\text{Str-IE-BN}_X(Y) = 0) \implies (E\text{-IE}^{W_{BN},W}_{x_0,x_1}(y) = 0). \tag{6.144}$$

Proof. Suppose that Str-IE-BN$_X(Y) = 0$. Note that the measure equals

$$E\text{-IE}^{W_{BN}^C, W}_{x_0,x_1}(y) = P(y_{x_0,(W_{BN}^C)_{x_1},(W_{BN})_{x_0}} \mid E) - P(y_{x_0} \mid E). \tag{6.145}$$

Since $\text{ch}(X) \cap W_{BN}^C = \emptyset$, we note that

$$y_{x_0,(W_{BN}^C)_{x_1},(W_{BN})_{x_0}} = y_{x_0,(W_{BN})_{x_0}} = y_{x_0,(W_{BN}^C)_{x_0},(W_{BN})_{x_0}} = y_{x_0}. \tag{6.146}$$

Using this expression in Eq. 6.145 shows our result. ■

After showing the admissibility of the measures, the final point to consider is when these target quantities are identifiable from the observational data and the causal diagram. Here, known complete algorithms can be used for identification, and we refer the reader to (Shpitser and Pearl, 2007) or more recently to (Correa *et al.*, 2021b).

Addressing the lack of symmetry. In Def. 5.4 of Sec. 5.1 we introduced the symmetric notions of the direct and indirect effects. The lack of symmetry arises from the fact that we can either consider a transition of the protected attribute $x_0 \to x_1$, or a reverse transition $x_1 \to x_0$. The same issue appears when considering variable-specific (or set-specific) decompositions. In particular, as discussed in Def. 5.4, we might be interested in considering the Ctf-IE$_{x_0,x_1}(y \mid x)$ or Ctf-IE$_{x_1,x_0}(y \mid x)$. However, each of these two measures can be decomposed in two different ways:

$$\text{Ctf-IE}_{x_0,x_1}(y \mid x) = \underbrace{\text{Ctf-IE}^{\emptyset, W_{BN}^C}_{x_0,x_1}(y \mid x)}_{\text{discriminatory}} + \underbrace{\text{Ctf-IE}^{W_{BN}^C, W}_{x_0,x_1}(y \mid x)}_{\text{BN variations}} \tag{6.147}$$

$$= \underbrace{\text{Ctf-IE}^{\emptyset, W_{BN}}_{x_0,x_1}(y \mid x)}_{\text{BN variations}} + \underbrace{\text{Ctf-IE}^{W_{BN}, W}_{x_0,x_1}(y \mid x)}_{\text{discriminatory}}, \tag{6.148}$$

and analogously for Ctf-IE$_{x_1,x_0}(y \mid x)$. For the spurious effect, we can also average over two distinct decompositions. To address this issue, we once again introduce measures that average out the effect of the measures that are admissible with respect to the Str-SE-BN and Str-IE-BN structural criteria:

Definition 6.13 (Symmetric Set-specific Measures under Business Necessity). Define the x-specific indirect and spurious measures under business necessity as:

$$x\text{-IE}^{\text{sym}}_{W_{BN}}(y \mid x) = \frac{1}{4}\Big(\text{Ctf-IE}^{\emptyset, W_{BN}^C}_{x_1,x_0}(y \mid x) + \text{Ctf-IE}^{W_{BN}, W}_{x_1,x_0}(y \mid x) - \tag{6.149}$$

$$\text{Ctf-IE}^{\emptyset, W_{BN}^C}_{x_0,x_1}(y \mid x) - \text{Ctf-IE}^{W_{BN}, W}_{x_0,x_1}(y \mid x)\Big) \tag{6.150}$$

$$x\text{-SE}^{\text{sym}, U_{BN}}_{x_1,x_0}(y) = \frac{1}{2}\Big(\text{Ctf-SE}^{\emptyset, U_{BN}^C}_{x_1,x_0}(y) + \text{Ctf-SE}^{U_{BN}, U}_{x_1,x_0}(y)\Big). \tag{6.151}$$

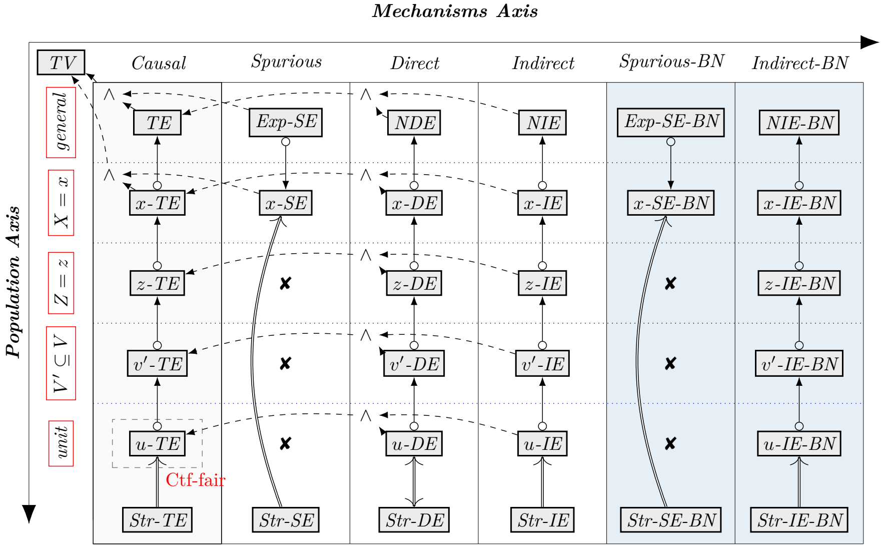


Figure 6.13: Fairness Map for the TV-family of measures. The horizontal axis represents the mechanisms (causal, spurious, direct, and indirect), and the vertical axis the events that capture increasingly more granular sub-populations, from general ($P(u)$) to unit-level. The bottom row contains structural measures. The arrow $\Longrightarrow$ indicates relations of admissibility, $\circ\!\!\longrightarrow$ of power, and $\dashrightarrow$ of decomposability.

The measures in the above definition capture the variations that are considered to be discriminatory, after accounting for corresponding business necessity sets W_{BN} and U_{BN}.

6.4 Extended Fairness Map

The set-specific and variable-specific decompositions of spurious and indirect effects give us an additional toolkit for solving fairness problems in practice, when considering business necessity. In particular, they extend the Fairness Map that was introduced in Fig. 4.5. The extension of the Map is shown in Fig. 6.13. In particular, we extend the map with two additional columns, corresponding to spurious-BN and indirect-BN variations, for arbitrary business necessity sets. These two columns expand the *mechanism axis*, as they further refine which mechanisms are included in the specific measures.

An important analogy can be drawn with the first instance of the Fairness Map. Our initial task was to decompose the variations within the TV, into those going through spurious, indirect, and direct pathways. For doing so, we only needed access to the SFM, whose “resolution” or granularity was sufficient for performing this task. However, once we wanted to better understand the variations within the spurious and indirect effects, to accommodate for business necessity, the granularity of the SFM was no longer sufficient, and we had to move to the fully specified causal diagram $\mathcal{G}$. Once $\mathcal{G}$ is considered, spurious and indirect effects become *decomposable*, in a similar fashion as how the total effect (TE) or the total variation (TV) were decomposable when using the SFM. Once again, this behavior reflects two important facts:

(a) increasingly strong causal assumptions allow for increasingly powerful decomposition of variations; in particular, the causal diagram $\mathcal{G}$ provides a stronger tool for decomposing variations than just the SFM[1], but is on the other hand much more difficult to construct,

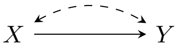
(b) the principles of decomposability, admissibility, and power apply more broadly than just for the SFM; in particular, after specifying the full diagram $\mathcal{G}$, the indirect and spurious effects become decomposable, and their decompositions are included in the Extended Fairness Map in Fig 6.13.

Before providing a practical algorithm for assessing the legal doctrines, we discuss another possibility for extending the mechanism axis.

6.4.1 Connection to Path-specific Notions

The measures described in this section so far can be called *set-specific* or *variable-specific*. However, a further level of granularity is possible, which allows for a path-specific analysis of causal effects (Pearl, 2001; Avin *et al.*, 2005). Path-specific definitions of fairness have also been considered in the fairness literature (Nabi and Shpitser, 2018; Chiappa, 2019; Wu *et al.*, 2019). We next draw some connections to the path-specific notions of fairness.

[1] We give here an interesting remark for the curious reader. There is an analogy to be drawn when we are transitioning from the SFM to the causal diagram by explicating additional causal assumptions. In fact, constructing the SFM can be seen as an expanded version of the basic bow graph



that contains no causal assumptions apart from the fact that X causally precedes Y (which is almost a given in any fairness application). In the bow graph, the direct effect equals the total causal effect, and we can already see that the TV measure can be expanded into causal vs. non-causal contributions. The transition from the bow graph to the SFM allows us to separate the causal variations into direct and indirect, similarly to how transitioning from the SFM to $\mathcal{G}$ allows us to decompose the indirect and spurious effects further.

For the purposes of the discussion, we will assume a Markovian model with k mediators $W_1, \dots, W_k$ between X and Y. Further, the causal diagram $\mathcal{G}$ is assumed to be fully connected. A causal path π from X to Y is any sequence of nodes and forward edges starting at X and ending at Y. Any path π can be encoded with a vector of length k, with the i-th entry indicating whether the respective mediator W_i lies on the path. For instance, $(0, \dots, 0)$ encodes the direct path $X \to Y$, while $(1, \dots, 1)$ encodes $X \to W_1 \to \cdots \to W_k \to Y$. Further, we note there are 2^k paths between X and Y, and a path π encoded by a vector $(s_1, \dots, s_k)$ is given the index $\sum_i s_i 2^{i-1}$. We exemplify the above notions on a two-mediator example that will be used in this section:

Example 6.16 (Paths in a Two Mediator Graph). Consider the causal diagram with two mediators W_1, W_2, shown in Fig. 6.11. There are four pathways in the graph, namely:

(1) direct path $X \to Y$, encoded by $(0, 0)$ and given index 0,

(2) path $X \to W_1 \to Y$, encoded by $(1, 0)$ and given index 1,

(3) path $X \to W_2 \to Y$, encoded by $(0, 1)$ and given index 2,

(4) path $X \to W_1 \to W_2 \to Y$, encoded by $(1, 1)$ and given index 3.

The paths are labeled $\pi_0, \pi_1, \pi_2, \pi_3$, respectively. □

For defining path-specific potential outcomes, we borrow the notation from (Zhang and Bareinboim, 2018c), and the defined notions are similar to previous works (Pearl, 2001; Avin *et al.*, 2005; Shpitser and Tchetgen, 2016). Let $W_i \in \text{pa}(Y)$, and let C be a vector of length 2^k with entries $\{x_0, x_1\}$ indicating values of X along each causal pathway π. The edge $W_i \to Y$ defines a funnel operator $\triangleleft_{W_i \to Y}(C)$ in the following way. Entries of C corresponding to paths of the form $X \to \cdots \to W_i \to Y$ are kept, and these values correspond to paths $X \to \cdots \to W_i$ after applying the funnel operator. All other entries of C are dropped by the operator. This allows us to define a notion of a path-specific potential response:

Definition 6.14 (Path-specific Potential Response). Let C be a vector of length 2^k with entries $\{x_0, x_1\}$, with each entry indicating the value of X along the j-th pathway. Let $C_{X \to Y}$ denote the value of X along the direct pathway $X \to Y$, and let $S = \text{pa}(Y) \setminus X$. The potential outcome $Y_C(u)$ is defined as

$$Y_C(u) = Y_{C_{X \to Y}, S_{\triangleleft_{S \to Y}(C)(u)}}(u), \tag{6.152}$$

where $S_{\triangleleft_{S \to Y}(C)}(u)$ is a set of path-specific potential responses $\{S_{i_{\triangleleft_{S_i \to Y}(C)}}(u) : S_i \in S\}$.

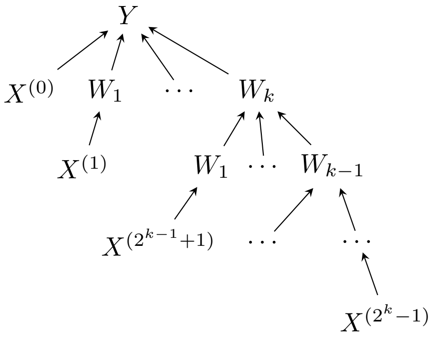


Figure 6.14: Visualization of how a path-specific potential response is obtained.

We provide an instructive visualization of how the potential response $Y_C(u)$ can be obtained (see Fig. 6.14). We start from the leaf node Y, and in the first row below it, we place all the parents of Y. Further, all parents of Y that are not root nodes of the causal diagram $\mathcal{G}$ (i.e., all parents different from X) are further expanded based on their parents, and the process is repeated recursively. This creates a tree $\mathcal{T}$, where each root node corresponds to X. Notice that any directed path from the root X to the leaf node Y in $\mathcal{T}$ corresponds to a causal path π_j in the causal diagram $\mathcal{G}$. The value assigned to X in the root node of $\mathcal{T}$, labeled $X^{(j)}$, determines how X behaves along the respective pathway π_j. Once all root nodes are specified (contained in the vector C of length 2^k), each node is evaluated based on its parents and the unit u, $V_i \leftarrow f_i(\text{pa}(V_i), u_i)$, until the path-specific potential outcome $Y_C(u)$ is obtained.

Example 6.16 (Two Mediators – Path-specific Potential Response). The path-specific potential response corresponding to the vector $C = (x_0, x_1, x_0, x_1)$ can be written as

$$Y_{x_0,(W_1)_{x_1},(W_2)_{x_0,(W_1)_{x_1}}}, \tag{6.153}$$

that is X equals x_0 along π_0, x_1 along π_1, x_0 along π_3, and x_1 along π_4. □

We next recall the definition of a counterfactual contrast (Def. 3.7), and show how such a contrast can be represented based on the underlying pathways:

Definition 6.15 (Counterfactual Contrast – Path Representation). Let $\mathcal{C} = (C_0, C_1, \emptyset, \emptyset)$ be a counterfactual contrast. Such a contrast $\mathcal{C}$ is represented by two vectors C_0, C_1 of length 2^k.

Furthermore, a structural measure of fairness (recall Def. 3.1) can also be represented based on paths:

Definition 6.16 (Structural Criterion – Path Representation). A path representation of a structural criterion Q is a vector q of length 2^k with $\{0, 1\}$ entries, with each entry indicating whether the respective path π_j is included in the structural criterion.

The definition can also be demonstrated to our running example:

Example 6.16 (Two Mediators – Path Structural Criteria). Using path-specific language, structural criteria Str-DE and Str-IE can be written as

$$\text{Str-DE} = (1, 0, 0, 0) \tag{6.154}$$

$$\text{Str-IE} = (0, 1, 1, 1). \tag{6.155}$$

Str-DE includes the path π_0, whereas Str-IE includes π_1, π_2, and π_3. □

From this notation, it becomes clear that many other structural criteria are possible, too. The notion of admissibility from the FPCFA (Def. 3.6) can also be written using the path-specific language:

Definition 6.17 (Admissibility – Path Representation). A measure μ based on a contrast $\mathcal{C} = (C_0, C_1)$ is admissible with respect to a structural criterion q if

$$(C_0)_j \neq (C_1)_j \implies q_j = 1. \tag{6.156}$$

The inequality $(C_0)_j \neq (C_1)_j$ means that there are variations transmitted along the pathway π_j within the contrast $\mathcal{C}$. The implication $(C_0)_j \neq (C_1)_j \implies q_j = 1$ requires that all the pathways captured by the contrast $\mathcal{C}$ are also contained in the structural criterion q.

Example 6.16 (Two Mediators – Path-specific Admissibility). The $\text{NIE}_{x_0,x_1}(y)$ measure corresponds to the path-specific contrast $\mathcal{C} = (C_0, C_1)$, where $C_0 = (x_0, x_0, x_0, x_0)$, and $C_1 = (x_0, x_1, x_1, x_1)$. The Str-TE measure corresponds to the vector $q = (1, 1, 1, 1)$. From Def. 6.17 it follows that NIE is admissible with respect to Str-TE. □

Notice, however, that the structural total effect captures more than the natural indirect effect. Therefore, another additional notion, which has not been strongly emphasized in the discussion so far, is needed for the discussion on path-specific effects:

Definition 6.18 (Tightness). A measure μ based on a contrast $\mathcal{C} = (C_0, C_1)$ is tight with respect to a structural criterion q if

$$q_j = 1 \implies (C_0)_j \neq (C_1)_j. \tag{6.157}$$

Example 6.16 (Two Mediators – Path-specific Tightness)**.** The $\text{NIE}_{x_0,x_1}(y)$ measure is not tight with respect to Str-TE. However, the $\text{NIE}_{x_0,x_1}(y)$ is both tight and admissible with respect to Str-IE, which is represented by $q = (0, 1, 1, 1)$. □

For tightness, the implication is reversed compared to the notion of admissibility. Intuitively, tightness requires that all variations within q are also captured within $\mathcal{C}$, while admissibility requires that variations captured within $\mathcal{C}$ are within q (in this sense, the two notions are complementary). We next define the notion of when a structural criterion *shatters* a collection of pathways:

Definition 6.19 (Shattering)**.** Let q be a structural criterion. Let $\Pi = \bigcup_{j \in J} \pi_j$ be a collection of pathways, with J the index set. Let q^1 be the set of indices for which $q_j = 1$, and q^0 defined analogously. We say q shatters Π (or J) if

$$J \cap q^1 \neq \emptyset, J \cap q^0 \neq \emptyset. \tag{6.158}$$

The notion of shattering is intuitive – a structural criterion q shatters a collection of pathways Π if there are pathways in Π that are contained in q, and also pathways that are not contained in q. In other words, Π is shattered by q if it is not fully contained within q, nor its complement $1 - q$. We next define an important notion of first-entry mapping:

Definition 6.20 (First-Entry Set)**.** Let W_i be a mediator. For a path π, the first entry $fe(\pi)$ is the first mediator along the path π (we assign the direct path to the node X, which we also label W_0 for convenience). The first-entry set J^i_{fe} of W_i is defined as

$$J^i_{fe} = \{j : fe(\pi_j) = W_i\}. \tag{6.159}$$

We now instantiate the notions of first entry and shattering on our running example:

Example 6.16 (Two Mediators – First-Entry and Shattering)**.** The mediator W_1 has the first entry set $\{\pi_1, \pi_3\}$ (or equivalently $J^1_{fe} = \{1, 3\}$) corresponding to paths $X \to W_1 \to Y$, and $X \to W_1 \to W_2 \to Y$. The mediator W_2 has the first entry set $\{\pi_2\}$ (or $J^2_{fe} = \{2\}$), corresponding to $X \to W_2 \to Y$.

The structural criterion $q_A = (0, 1, 0, 0)$ shatters J^1_{fe}, since it includes π_1 but not π_3. The structural criterion $q_B = (0, 1, 0, 1)$ does not shatter J^1_{fe}. □

Based on the first-entry mapping, and the notion of shattering, we can state a key identifiability result:

Theorem 6.9 (Tightness, Admissibility, Identifiability – Shattering)**.** For a structural criterion q, there exists a contrast $\mathcal{C}$ that is tight, admissible, and identifiable if and only if q does not shatter any of the J^i_{fe} for $i \in \{0, \ldots, k\}$.

We note that the result in Thm. 6.9 is related to the recanting witness criterion of (Avin *et al.*, 2005). We next apply the theorem to our example:

Example 6.16 (Two Mediators – Shattering and Identification)**.** Consider the structural criteria $q_A = (0,1,0,0)$, $q_B = (0,1,0,1)$ from before. There is no tight, admissible, and identifiable contrast w.r.t. q_A because q_A shatters J^1_{fe}. For q_B, there exists a contrast that is tight, admissible, and identifiable, since q_B does not shatter any J^i_{fe}. For example, the contrast

$$\mathcal{C}_B = P(y_{x_0,(W_1)_{x_1},(W_2)_{x_0,(W_1)_{x_1}}}) - P(y_{x_0,(W_1)_{x_0},(W_2)_{x_0,(W_1)_{x_0}}}) \tag{6.160}$$

is one such contrast. Its identification expression is given by

$$\sum_{w_1,w_2} P(y \mid x_0, w_1, w_2) P(w_2 \mid x_0, w_1) \big[P(w_1 \mid x_1) - P(w_1 \mid x_0) \big]. \tag{6.161}$$

Verifying tightness and admissibility is left as an exercise for the reader. □

The next question we ask is about maps from the set of paths π to the set of variables W_i. We define the notion of a valid path-to-variable mapping:

Definition 6.21 (Valid Path to Variable Mappings)**.** Let ν be a mapping from pathways to the set of mediators, $\pi_j \mapsto W_i$. A mapping ν is said to be valid if

(i) each π_j is mapped to some W_i that lies along π_j,

(ii) for each structural criterion $q^{(i)}$ corresponding to $\nu^{-1}(W_i)$ there exists a tight, admissible, and identifiable contrast $\mathcal{C}^{(i)}$.

We say that a mapping is valid under two conditions. Firstly, each pathway π needs to be mapped to a variable that lies along π; mappings that do not satisfy this property are not very meaningful. The second property ensures the ability to perform an analysis of variations. The value of $\nu^{-1}(W_i)$ contains all the paths assigned to W_i, and $q^{(i)}$ is the corresponding structural criterion. The existence of a tight, admissible, and identifiable contrast $\mathcal{C}^{(i)}$ with respect to $q^{(i)}$ ensures that there is a fairness measure computable in practice that captures variations along paths $\nu^{-1}(W_i)$ transmitted through variable W_i. As it turns out, there is a single valid path-to-variable mapping:

Corollary 6.10 (Unique Valid Path to Variable Mapping)**.** There exists a unique valid path to variable mapping ν, and it is given by the first entry map $fe(\pi)$.

Example 6.16 (Two Mediators – Path-to-variable Mappings)**.** The path-to-variable mapping ν_A, defined by

$$\big\{ \nu_A(\pi_0) = X, \nu_A(\pi_1) = W_1, \nu_A(\pi_2) = W_2, \nu_A(\pi_3) = W_2 \big\}, \tag{6.162}$$

is not a valid path-to-variable mapping. The structural measure $q_{A,W_2} = (0,1,1,0)$ corresponding to the set $\nu_A^{-1}(W_2) = \{\pi_2, \pi_3\}$ does not allow for a tight, admissible, and identifiable contrast. However, the map ν_B, defined by

$$\big\{\nu_B(\pi_0) = X, \nu_B(\pi_1) = W_1, \nu_B(\pi_2) = W_2, \nu_B(\pi_3) = W_1\big\}, \tag{6.163}$$

is valid path-to-variable mapping, and it is unique. □

Our results show formally that the first-entry mapping is the only valid assignment of causal pathways to variables that guarantees the existence of contrasts that are tight, admissible, and identifiable (TAI, for short). In other words, it is the only mapping that allows us to have a practical, variable-specific approach to fairness analysis, without requiring strong additional assumptions. Clearly, contrasts $\mathcal{C}^{(i)}$ that are TAI with respect to the $q^{(i)} = \nu^{-1}(W_i)$ need to be found in practice. Note that, in Thm. 6.6, we already developed one possible solution, which also satisfies the decomposability property (Def. 3.5).

Set-specific viewpoint. After relating the variable-specific decompositions to the path-specific ones, we want to do the same for set-specific decompositions. Again, we assume that the set of mediators $W_1, \dots, W_k$ is partitioned into BN mediators W_{BN} and non-BN mediators W_{BN}^C. One possible way of determining which causal paths π fall under business necessity with respect to the partition W_{BN}, W_{BN}^C is to use the first-entry map, that is,

$$q_{\text{fe-BN}} = \bigcup_{W_i \in W_{BN}} fe^{-1}(W_i), \tag{6.164}$$

where $q_{\text{fe-BN}}$ is the structural criterion corresponding to BN paths determined by the first-entry map. Interestingly, our prior discussion implies that there exist TAI contrasts $\mathcal{C}_{BN}, \mathcal{C}_{BN}^C$ with respect to $q_{\text{fe-BN}}$ and its complement $1 - q_{\text{fe-BN}}$, respectively. However, there exist two other ways in which the partition W_{BN}, W_{BN}^C can be mapped to structural criteria:

Definition 6.22 (Minimal and Maximal Variable-Specific Business Necessity)**.** Let W_{BN} be the business necessity set. The maximal business necessity (max-BN) path mapping states that π_j is in the business necessity set if any W_i along π_j is in W_{BN}. The minimal business necessity (min-BN) path mapping states that π_j is in the business necessity set if all W_i along π_j are in W_{BN}.

The definitions of max-BN and min-BN represent two extremes. The max-BN mapping says that if any variable along a path is BN, then the path is BN, thereby labeling as many paths as possible as BN. Contrary to this, min-BN labels as few paths as possible as BN, since a single non-BN variable along the path is sufficient to label the path non-BN. Let $q_{\text{max-BN}}, q_{\text{min-BN}}$ be the structural criteria obtained in this way.

Example 6.16 (Two Mediators – Minimal and Maximal BN). Suppose that the set $W_{BN} = \{W_2\}$. Then, the min-BN mapping states that only the path $\pi_2 = X \to W_2 \to Y$ is in the BN set. The max-BN mapping states that paths π_2, π_3 are both in the BN set. □

The natural question is when TAI contrasts with respect to the $q_{\text{max-BN}}$ and $q_{\text{min-BN}}$ criteria exist. To give a condition for this, we introduce the notions of forward and backward BN closure:

Definition 6.23 (Forward and Backward BN Closure). Let W_{BN} be the set of BN mediators. The set W_{BN} satisfies forward closure if

$$W_i \in W_{BN} \implies \big(\text{ch}(W_i) \cap W\big) \subset W_{BN}. \tag{6.165}$$

The set W_{BN} satisfies backward closure if

$$W_i \in W_{BN} \implies \big(\text{pa}(W_i) \cap W\big) \subset W_{BN}. \tag{6.166}$$

We ground these notions in our running example:

Example 6.16 (Two Mediators – Forward and Backward BN closure). The BN set $\{W_2\}$ does not satisfy backward closure (since $W_1 \in \text{pa}(W_2)$ is not in the BN set), but satisfies forward closure. The BN set $\{W_1\}$ satisfies backward, but not forward closure. The BN sets $\{W_1, W_2\}$ and $\emptyset$ satisfy both forward and backward closure. □

Based on the above definitions, we can state an important result for the $q_{\text{max-BN}}$ and $q_{\text{min-BN}}$ criteria:

Theorem 6.11 (Identification Under Closure). There exists a tight, admissible, and identifiable contrast $\mathcal{C}_{\text{max-BN}}$ for $q_{\text{max-BN}}$ if and only if W_{BN} satisfies backward closure. Further, there exists a tight, admissible, and identifiable contrast $\mathcal{C}_{\text{min-BN}}$ for $q_{\text{min-BN}}$ if and only if W_{BN} satisfies forward closure.

The above result shows that the max-BN mapping can be tested empirically if and only if W_{BN} satisfies backward closure. Similarly, min-BN mapping can be tested if and only if W_{BN} satisfies forward closure. Thus, we have established conditions under which alternative solutions to the first-entry mapping can be used to evaluate business necessity requirements.

We summarize the main insights from the discussion in this section. Our results imply that general path-specific attributions are not feasible in practice, without invoking further strong assumptions. Therefore, variable-specific or set-specific attributions can strike a good balance of modeling flexibility and practical usefulness. Firstly, we showed that a variable-specific attribution of variations is possible, and there is a unique mapping of pathways to variables that allows for this (the first-entry mapping). We further discussed several ways of conceptualizing set-specific attributions, based on the structural criteria $q_{\text{fe-BN}}$, $q_{\text{max-BN}}$, and $q_{\text{min-BN}}$. The $q_{\text{fe-BN}}$ criterion can be evaluated in practice for any Markovian model and selection of BN set, whereas $q_{\text{max-BN}}$ and $q_{\text{min-BN}}$ additionally require the notions of backward and forward BN closure, respectively. These results illuminate what are the most fine-grained analyses for assessing claims of business necessity that are feasible in practice. Furthermore, they also demonstrate that the approach developed in this section offers the most fine-grained analysis possible without invoking further assumptions, and that more detailed path-specific analyses may not be realistic in practice.

6.5 Extended Fairness Cookbook

Armed with the set-specific and variable-specific decompositions of spurious and indirect effects, we can now provide an extended version of the Fairness Cookbook in Alg. 5.1, which is able to accommodate for considerations of arbitrary business necessity sets within the disparate impact doctrine. The Extended Fairness Cookbook is presented in Alg. 6.4. The Extended Fairness Cookbook is intended to be used when one of the following holds:

(a) the hypothesis $H_0^{(\text{Ctf-SE})}$ in the Fairness Cookbook was rejected, but the U_{sToT} contains both variables that fall under Business Necessity ($U_{BN} \neq \emptyset$), and also variables that do not fall under Business Necessity ($U_{BN}^C \neq \emptyset$),

(b) the hypothesis $H_0^{(\text{Ctf-IE})}$ in the Fairness Cookbook was rejected, but the W-set contains both variables that fall under Business Necessity ($W_{BN} \neq \emptyset$), and also variables that do not fall under Business Necessity ($W_{BN}^C \neq \emptyset$).

We are now ready to revisit Ex. 6.1 which motivated the discussion in this section, and demonstrate the usage of the Extended Fairness Cookbook in practice:

Example 6.17 (COMPAS with Business Necessity Continued)**.** The data scientists at ProPublica applied the Fairness Cookbook to the COMPAS dataset and demonstrated that:

$$\text{Ctf-IE}_{x_1,x_0}(y \mid x_1) = -5.7\% \pm 0.5\%, \tag{6.171}$$

$$\text{Ctf-SE}_{x_1,x_0}(y) = -4.0\% \pm 0.9\%, \tag{6.172}$$

After the court hearing, the judge ruled that using the attributes age (Z_2), prior count (P), and charge degree (D) was not discriminatory, but using the attributes juvenile count (J) and gender (Z_1) was. In light of these business

Algorithm 6.4 Extended Fairness Cookbook

```
• Inputs: Dataset $\mathcal{D}$, SFM projection $\Pi_{\text{SFM}}(\mathcal{G})$, arbitrary BN-set.
1: Obtain the dataset $\mathcal{D}$ and construct the causal diagram $\mathcal{G}$.
2: Determine Business Necessity considerations.
     • determine which variables $U_{sToT}$ lie on top of spurious treks according to Def. 6.6,
     • determine which variables in $U_{sToT}$ and $W$ fall under business necessity; denote the sets as $U_{BN}$, $W_{BN}$, respectively, and denote by $U^C_{BN}$, $W^C_{BN}$ their respective complements in $U_{sToT}, W$.
3: Consider Disparate Impact under Business Necessity:
     • test the following two hypotheses:
           $H_0^{(\text{Ctf-SE}),\neg BN} : \text{Ctf-SE}^{\text{sym},U_{BN}}_{x_1,x_0}(\widehat{y}) = 0,$   (6.167)
           $H_0^{(\text{Ctf-IE}),\neg BN} : \text{Ctf-IE}^{\text{sym}}_{W_{BN}}(\widehat{y} \mid x_0) = 0.$   (6.168)
         – if either of the hypotheses is rejected $\implies$ there is evidence of disparate impact.
     • if $Y \in \mathcal{D}$, further test the following hypotheses:
           $H_0^{(\text{Ctf-SE}),BN} : \text{Ctf-SE}^{\text{sym},U^C_{BN}}_{x_1,x_0}(\widehat{y}) = \text{Ctf-SE}^{\text{sym},U^C_{BN}}_{x_1,x_0}(y),$   (6.169)
           $H_0^{(\text{Ctf-IE}),BN} : \text{Ctf-IE}^{\text{sym}}_{W^C_{BN}}(\widehat{y} \mid x_0) = \text{Ctf-IE}^{\text{sym}}_{W^C_{BN}}(y \mid x_0).$   (6.170)
         – if either of the hypotheses is rejected $\implies$ there is evidence of disparate impact.
```

necessity requirements, the ProPublica team realizes that their originally presented measures are not sufficient to prove discrimination under the disparate impact doctrine. Therefore, the team decides to use the Extended Fairness Cookbook and computes that:

$$\begin{aligned}\text{Ctf-IE}_{x_1,x_0}(y \mid x_1) &= \text{Ctf-IE}^{\emptyset,J}_{x_1,x_0}(y \mid x_1) + \text{Ctf-IE}^{J,\{J,P,D\}}_{x_1,x_0}(y \mid x_1) \qquad (6.173)\\ &= \underbrace{(-4.9\% \pm 0.4\%)}_{\text{juvenile count variations}} + \underbrace{(-0.9\% \pm 0.2\%)}_{\text{priors and charge variations}} . \qquad (6.174)\end{aligned}$$

Similarly, the team also finds that $U_{sToT} = \{U_{1X}, U_{2X}\}$ associated with bidirected edges $X \dashleftarrow\!\!\dashrightarrow Z_1, X \dashleftarrow\!\!\dashrightarrow Z_2$, respectively. Based on this they

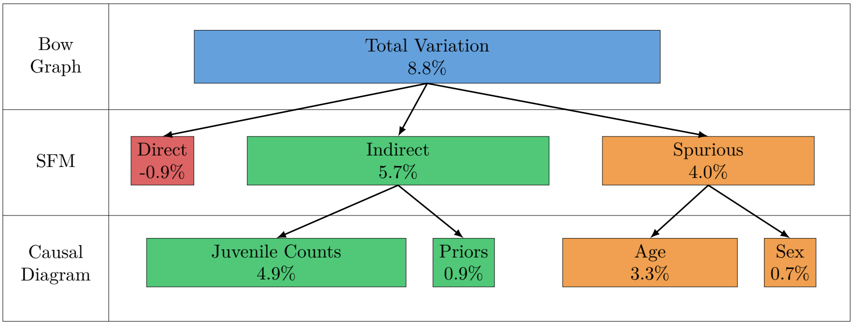


Figure 6.15: Steps of the analysis of the COMPAS dataset with different levels of causal assumptions. In the first step, we compute the TV measure and assume only the bow graph. In the second step, we use the SFM to obtain a causal decomposition of the TV measure. In the third step, which is performed when necessary, we obtain the further decompositions of the spurious and indirect effects, based on the full causal diagram $\mathcal{G}$.

compute that:

$$\text{Ctf-SE}_{x_1,x_0}(y) = \text{Ctf-SE}^{\emptyset,U_{1X}}_{x_1,x_0}(y) + \text{Ctf-IE}^{U_{1X},\{U_{1X},U_{2X}\}}_{x_1,x_0}(y) \tag{6.175}$$

$$= \underbrace{(-0.7\% \pm 0.7\%)}_{\text{gender variations}} + \underbrace{(-3.3\% \pm 0.8\%)}_{\text{age variations}}. \tag{6.176}$$

The steps of the entire analysis of the COMPAS dataset are also shown in Fig. 6.15. Based on this evidence, the court concludes that with respect to spurious effect there is no disallowed discrimination (the gender variations are not significantly different than 0). However, with respect to indirect variations, there is evidence of disparate impact, since the juvenile count variations are significantly different from 0, and the variable J is not included in the business necessity set. The ProPublica team presents this evidence, and the court finds that the disparate impact doctrine was violated, and that the future algorithm predictions need to be adapted to account for this violation. □

6.6 Extended Fair Prediction

In the previous section we discussed how to perform Task 1 (bias detection and quantification) under the extended model, that is the causal full causal diagram $\mathcal{G}$. We now discuss how to handle Task 2 (fair prediction) when the full causal diagram is specified. Recall that under the SFM, we described the procedure of Causal IF (see Alg. 5.2), which performed optimal transport between group-specific distributions of possibly multivariate sets Z, W, and the

Algorithm 6.5 Extended Fair Data Adaptation (EFDA)

```
• Inputs: Dataset $\mathcal{D}$, Causal Diagram $\mathcal{G}$, BN Sets $U_{BN}, W_{BN}$.

for $Z_i \in Z$ in topological order do                ▷ Adapting the confounders
    if $Z_i \in AS(U_{BN}) \wedge Z_i \in AS(U^C_{BN})$ then
        return FAIL: adaptation cannot be performed
    else if $Z_i \in AS(U_{BN}) \wedge Z_i \notin AS(U^C_{BN})$ then
        transport $Z_i \mid x_1, \tau^{\mathrm{pa}(Z_i)}(\mathrm{pa}(Z_i))$ onto $Z_i \mid x_0, \mathrm{pa}(Z_i)$
        let $\tau^{Z_i}$ denote the obtained transport map
    else if $Z_i \notin AS(U_{BN}) \wedge Z_i \in AS(U^C_{BN})$ then
        transport $Z_i \mid x, \tau^{\mathrm{pa}(Z_i)}(\mathrm{pa}(Z_i))$ onto $Z_i \mid x, \mathrm{pa}(Z_i)$ for $x \in \{x_0, x_1\}$
        let $\tau^{Z_i}$ denote the obtained transport map
    end if
end for

for $V_i \in \{W, Y\}$ in topological order do          ▷ Adapting the mediators
    let $\mathrm{set}(V_i) = W_{BN}$ if $V_i \in W_{BN}$ and $\mathrm{set}(V_i) = W^C_{BN}$ otherwise
    if $\exists W_j \in \mathrm{pa}(V_i)$ s.t. $W_j \in \mathrm{set}(V_i)^C \wedge W_j \dashleftarrow\!\!\dashrightarrow V_i$ then
        return FAIL (adaptation cannot be performed)
    end if
    if $V_i \notin W_{BN}$ then
        transport $V_i \mid x_1, \tau^{\mathrm{pa}(V_i)}(\mathrm{pa}(V_i))$ onto $V_i \mid x_0, \mathrm{pa}(V_i)$
        let $\tau^{V_i}$ denote the obtained transport map
    else if $V_i \in W_{BN}$ then
        transport $V_i \mid x, \tau^{\mathrm{pa}(V_i)}(\mathrm{pa}(V_i))$ onto $V_i \mid x, \mathrm{pa}(V_i)$ for $x \in \{x_0, x_1\}$
        let $\tau^{V_i}$ denote the transport map
    end if
end for
```

outcome Y. When using the causal diagram $\mathcal{G}$, we propose a similar approach, but on a higher level of granularity. In particular, we perform optimal transport sequentially on singleton variables, in a valid topological ordering. The formal description of the procedure is given in the following definition:

Definition 6.24 (Extended Fair Data Adaptation (EFDA)). Let $\mathcal{M}$ be an SCM and let $\mathcal{G}$ be the corresponding causal diagram. Assume further that $\mathcal{G}$ can be projected onto the standard fairness model (SFM) from Def. 2.7. Let $U_{BN} \subseteq U_{sToT}$ be the subset of the exogenous confounders which fall under business necessity, and let $AS(\cdot)$ denote the anchor set operator. Let $W_{BN} \subseteq W$ be the subset of the mediators that fall under business necessity. The extended fair data adaptation (EFDA) proceeds as described in Algorithm 6.5.

As can be noted, the procedure can sometimes fail, that is, the adaptation might not be possible. The failure conditions are related precisely to the identification conditions in Thm. 6.4, or the general identification conditions for indirect effects (Shpitser and Pearl, 2007; Correa *et al.*, 2021b). The procedure might fail if there is a confounder Z_i which is simultaneously in the anchor set of business necessity variables $AS(U_{BN})$ and the anchor set of the complement $AS(U^C_{BN})$. In this case, it is not possible to determine how the data should be adapted to accommodate for business necessity requirements. Similarly, if there exists a mediator W_i for which exists a parent W_j such that (i) there is a bidirected edge $W_j \dashleftarrow\!\!\dashrightarrow W_i$ and (ii) either $W_i \in W_{BN}, W_j \in W^C_{BN}$ or $W_i \in W^C_{BN}, W_j \in W_{BN}$ (i.e., W_j, W_i are in different partitions of W according to business necessity), then the procedure also fails, since we cannot determine how the data should be adapted.

The procedure described in Def. 6.24 is based on the fair data adaptation procedure from (Plečko and Meinshausen, 2020; Plečko *et al.*, 2021), with two key differences. Firstly, the EFDA procedure described here allows for the adaptation of confounding variables that causally precede the protected attribute X. Secondly, there is a slight difference in how the downstream effects of business necessity variables are handled. Plečko and Meinshausen, 2020 call these variables *resolving*, according to the definition of Kilbertus *et al.*, 2017, and consider any pathway from the protected attribute X to Y mediated by such variables to be in the business necessity set (corresponding to the $q_{\text{max-BN}}$ criterion from Sec. 6.4.1). The above procedure is an alternative, and considers pathways of the form $X \to W_i \to \cdots \to Y$, for a BN variable W_i, to be in the business necessity set (corresponding the the $q_{\text{fe-BN}}$ criterion). As discussed previously, the criterion $q_{\text{fe-BN}}$ gives a stronger definition of fairness than the $q_{\text{max-BN}}$ criterion.

Finally, to conclude, we show that after the adaptation procedure, the effects associated with variables outside the business necessity sets vanish:

Theorem 6.12 (Soundness of EFDA)**.** Let $\mathcal{M}$ be an SCM and let $\mathcal{G}$ be the corresponding causal diagram. Let $U_{BN} \subseteq U_{sToT}$ be the subset of the exogenous confounders which fall under business necessity, and let $W_{BN} \subseteq W$ be the subset of the mediators that fall under business necessity. Let $\widetilde{\mathcal{M}}$ denote the SCM after the EFDA procedure is performed. It then follows that

$$\text{Ctf-IE}^{\emptyset, W^C_{BN}}_{x_1,x_0}(y \mid x) = \text{Ctf-IE}^{W_{BN}, W}_{x_1,x_0}(y \mid x) = 0 \tag{6.177}$$

$$\text{Ctf-SE}^{\emptyset, U^C_{BN}}_{x_1,x_0}(y) = \text{Ctf-SE}^{U_{BN}, U}_{x_1,x_0}(y) = 0. \tag{6.178}$$

The theorem is given without proof, as it follows very closely (Plečko and Meinshausen, 2020, Thm. 1) and the proof of Thm. 5.3.

7

Conclusions

Modern automated decision-making systems based on AI are fueled by data, which encodes many complex historical processes and past, potentially discriminatory practices. Such data, imprinted with undesired biases, cannot by itself be used and expected to produce fair systems, regardless of the level of statistical sophistication of the methods used or the amount of data available. In light of this limitation in which more data or clever methods are not the solution, the AI designer is left to search for a new notion of what a fair reality should look like. By and large, the literature on fair machine learning attempts to address this question by formulating (and then optimizing) statistical notions about how fairness should be measured. Still, as many of the examples in this manuscript demonstrated, statistical notions fall short of providing a satisfactory answer for what a fair reality should entail. Using decision systems that arise when only considering statistical notions of fairness may be causally meaningless or even have unintended and possibly catastrophic consequences.

We combined in this manuscript two ingredients to address this challenge, (i) the language of causality and (ii) legal doctrines of discrimination as a sound basis for imagining what a fair reality should look like and how society's norms and expectations should be represented. This formalization of the fairness problem will allow communication between the key stakeholders involved in developing such systems in practice, including computer scientists, statisticians, and data scientists on the one hand and social, legal, and ethical experts on the other. A key observation is that mapping social, ethical, and legal norms onto statistical measures is a challenging task. A formulation we propose explicitly in the form of the Fundamental Problem of Causal Fairness Analysis is to map such social norms onto the underlying causal mechanisms and causal measures of fairness associated with these particular mechanisms. We believe such an approach can help data scientists to be more transparent when measuring discrimination and can also help social scientists to ground their principles and ideas in a formal mathematical language that is amenable to implementation.

The final important distinction introduced in this manuscript is between the different fairness tasks, namely (i) bias detection and quantification, (ii) fair prediction, and (iii) fair decision-making. The first task helps us to understand how much (and if any) bias exists in our data. The task of fair prediction allows us to correct for (parts or entirety) of this bias and envisage a more fair world in which such bias is removed. For the third task, we developed a set of preliminary results for a single time-step setting, but we leave for future work the extensions to a multi-step process. This will require interfacing the principles introduced in this manuscript with key ideas in economics and econometrics, which we view as an essential next step in designing fair systems.

7.1 Acknowledgments

We thank Dana Mackenzie and Kai-Zhan Lee for the feedback and help in improving this manuscript. This work was done in part while Drago Plecko was visiting the Causal AI lab at Columbia University. This research was supported in part by the Alfred P. Sloan Foundation, NSF, ONR, AFOSR, DoE, Amazon, and JP Morgan.

Appendices

A

Proofs of Main Theorems & Derivations

In this section, we provide the proofs of the main theorems presented in the manuscript. In particular, we give the proof for the Fairness Map theorem (Thm. 4.8), soundness of the SFM theorem (Thm. 4.11), the Fair Prediction theorem (Thm. 5.1), and the soundness of the Causal Individual Fairness procedure (Thm. 5.3).

A.1 Proof of Thm. 4.8

The proof of Thm. 4.8 is organized as follows. The full list of implications contained in the Fairness Map in Fig. 4.5 is given in in Tab. A.1. For each implication, we indicate the lemma in which the implication proof is given.

Lemma A.1 (Power relations of causal effects). The total, direct, and indirect effects admit the following relations of power (assuming that that $Z \subset V'$):

$$\begin{aligned}
\text{unit-TE}_{x_0,x_1}(y(u)) = 0 \;\forall u &\implies v'\text{-TE}_{x_0,x_1}(y \mid v') = 0 \;\forall v' && \text{(A.1)}\\
&\implies z\text{-TE}_{x_0,x_1}(y \mid z) = 0 \;\forall z && \text{(A.2)}\\
&\implies \text{ETT}_{x_0,x_1}(y \mid x) = 0 \;\forall x && \text{(A.3)}\\
&\implies \text{TE}_{x_0,x_1}(y) = 0, && \text{(A.4)}
\end{aligned}$$

	Implication	Proof
power	Unit-TE $\implies$ v'-TE $\implies$ z-TE $\implies$ ETT $\implies$ TE	Lem. A.1
	Unit-DE $\implies$ v'-DE $\implies$ z-DE $\implies$ Ctf-DE $\implies$ NDE	Lem. A.1
	Unit-IE $\implies$ v'-IE $\implies$ z-IE $\implies$ Ctf-IE $\implies$ NIE	Lem. A.1
	Exp-SE $\iff$ Ctf-SE	Lem. A.2
admissibility	S-SE $\implies$ Ctf-SE	Lem. A.5
	S-DE $\implies$ unit-DE	Lem. A.3
	S-IE $\implies$ unit-IE	Lem. A.4
decomposability	NDE $\wedge$ NIE $\implies$ TE	Lem. A.6
	Ctf-DE $\wedge$ Ctf-IE $\implies$ ETT	Lem. A.6
	z-DE $\wedge$ z-IE $\implies$ z-TE	Lem. A.6
	v'-DE $\wedge$ v'-IE $\implies$ v'-TE	Lem. A.6
	unit-DE $\wedge$ unit-IE $\implies$ unit-TE	Lem. A.6
	TE $\wedge$ Exp-SE $\implies$ TV	Lem. A.7
	ETT $\wedge$ Ctf-SE $\implies$ TV	Lem. A.7

Table A.1: List of implications in the Fairness Map in Fig. 4.5.

$$\begin{aligned}
\text{unit-DE}_{x_0,x_1}(y(u)) = 0 \;\forall u \implies & v'\text{-DE}_{x_0,x_1}(y \mid v') = 0 \;\forall v' && \text{(A.5)}\\
\implies & z\text{-DE}_{x_0,x_1}(y \mid z) = 0 \;\forall z && \text{(A.6)}\\
\implies & \text{Ctf-DE}_{x_0,x_1}(y \mid x) = 0 \;\forall x && \text{(A.7)}\\
\implies & \text{NDE}_{x_0,x_1}(y) = 0, && \text{(A.8)}\\
\text{unit-IE}_{x_0,x_1}(y(u)) = 0 \;\forall u \implies & v'\text{-IE}_{x_0,x_1}(y \mid v') = 0 \;\forall v' && \text{(A.9)}\\
\implies & z\text{-IE}_{x_0,x_1}(y \mid z) = 0 \;\forall z && \text{(A.10)}\\
\implies & \text{Ctf-IE}_{x_0,x_1}(y \mid x) = 0 \;\forall x && \text{(A.11)}\\
\implies & \text{NIE}_{x_0,x_1}(y) = 0. && \text{(A.12)}
\end{aligned}$$

Proof. We prove the statement for total effects (direct and indirect cases are analogous). We start by showing that ETT is more powerful than TE.

$$\begin{aligned}
\text{TE}_{x_0,x_1}(y) =& P(y_{x_1}) - P(y_{x_0})\\
=& \sum_x \big[P(y_{x_1} \mid x) - P(y_{x_0} \mid x)\big] P(x)\\
=& \sum_x \text{ETT}_{x_0,x_1}(y \mid x) P(x).
\end{aligned}$$

Therefore, if $\text{ETT}_{x_0,x_1}(y \mid x) = 0 \; \forall x$ then $\text{TE}_{x_0,x_1}(y) = 0$. Next, we can write

$$\begin{aligned}
\text{ETT}_{x_0,x_1}(y \mid x) &= P(y_{x_1} \mid x) - P(y_{x_0} \mid x) \\
&= \sum_z \big[P(y_{x_1} \mid x, z) - P(y_{x_0} \mid x, z)\big] P(z \mid x) \\
&= \sum_z \big[P(y_{x_1} \mid z) - P(y_{x_0} \mid z)\big] P(z \mid x) && Y_x \perp\!\!\!\perp X \mid Z \text{ in SFM} \\
&= \sum_z z\text{-TE}_{x_0,x_1}(y \mid z) P(z \mid x).
\end{aligned}$$

Therefore, if $z\text{-TE}_{x_0,x_1}(y \mid z) = 0 \; \forall z$ then $\text{ETT}_{x_0,x_1}(y \mid x) = 0 \; \forall x$. Next, for a set $V' \subseteq V$ such that $Z \subseteq V'$, we can write

$$\begin{aligned}
z\text{-TE}_{x_0,x_1}(y) &= P(y_{x_1} \mid z) - P(y_{x_0} \mid z) \\
&= \sum_{v' \setminus z} \big[P(y_{x_1} \mid z, v' \setminus z) - P(y_{x_0} \mid z, v' \setminus z)\big] P(v' \setminus z \mid z) \\
&= \sum_{v' \setminus z} v'\text{-TE}_{x_0,x_1}(y \mid v') P(v' \setminus z \mid z).
\end{aligned}$$

Therefore, if $v'\text{-TE}_{x_0,x_1}(y \mid v') = 0 \; \forall v'$ then $z\text{-TE}_{x_0,x_1}(y \mid z) = 0 \; \forall z$. Next, notice that

$$\begin{aligned}
v'\text{-TE}_{x_0,x_1}(y) &= P(y_{x_1} \mid v') - P(y_{x_0} \mid v') \\
&= \sum_u \big[y_{x_1}(u) - y_{x_0}(u)\big] P(u \mid v') \\
&= \sum_u \text{unit-TE}_{x_0,x_1}(y(u)) P(u \mid v').
\end{aligned}$$

Therefore, if $\text{unit-TE}_{x_0,x_1}(y(u)) = 0 \; \forall u$ then $v'\text{-TE}_{x_0,x_1}(y \mid v') = 0 \; \forall v'$. ■

Lemma A.2 (Power relations of spurious effects)**.** The criteria based on Ctf-SE and Exp-SE are equivalent in the case of binary X. Formally,

$$\text{Exp-SE}_x(y) = 0 \; \forall x \iff \text{Ctf-SE}_{x,x'}(y) = 0 \; \forall x \neq x'. \tag{A.13}$$

Proof.

$$\begin{aligned}
\text{Exp-SE}_x(y) &= P(y \mid x) - P(y_x) \\
&= P(y \mid x) - P(y_x \mid x)P(x) - P(y_x \mid x')P(x') \\
&= P(y \mid x)[1 - P(x)] - P(y_x \mid x')P(x') \\
&= P(y \mid x)P(x') - P(y_x \mid x')P(x') \\
&= -P(x')\text{Ctf-SE}_{x',x}(y).
\end{aligned}$$

Assuming $P(x') > 0$, the claim follows. ■

We remark that, in general (for multi-valued X), the criterion based on Ctf-SE is stronger than that based on Exp-SE.

Lemma A.3 (Admissibility w.r.t. structural direct)**.** The structural direct effect criterion ($X \notin \text{pa}(Y)$) implies the absence of unit-level direct effect. Formally:

$$\text{S-DE} \implies \text{unit-DE}_{x_0,x_1}(y(u)) = 0 \; \forall u. \tag{A.14}$$

Proof. Suppose that $X \notin \text{pa}(Y)$. Note that:

$$\begin{aligned}
\text{unit-DE}_{x_0,x_1}(y(u)) &= y_{x_1,W_{x_0}}(u) - y_{x_0}(u) \\
&= f_Y(x_1, W_{x_0}(u), Z(u), u_Y) - f_Y(x_0, W_{x_0}(u), Z(u), u_Y) \\
&= f_Y(W_{x_0}(u), Z(u), u_Y) \\
&\quad - f_Y(W_{x_0}(u), Z(u), u_Y) \qquad X \notin \text{pa}(Y) \\
&= 0.
\end{aligned}$$

■

Lemma A.4 (Admissibility w.r.t. structural indirect)**.** The structural indirect effect criterion ($\text{de}(X) \cap \text{pa}(Y) = \emptyset$) implies the absence of unit-level indirect effect. Formally:

$$\text{S-IE} \implies \text{unit-IE}_{x_1,x_0}(y(u)) = 0 \; \forall u. \tag{A.15}$$

Proof. Let $W_{de} \subseteq W$ be the subset of mediators W which are in $\text{de}(X)$, and let W_{de}^C be its complement in W. Then, by assumption, $W_{de} \cap \text{pa}(Y) = \emptyset$. We can write:

$$\begin{aligned}
\text{unit-IE}_{x_1,x_0}(y(u)) &= y_{x_1,W_{x_0}}(u) - y_{x_1}(u) \\
&= f_Y(x_1, (W_{de}^C)_{x_0}(u), Z(u), u_Y) \\
&\quad - f_Y(x_1, (W_{de}^C)_{x_1}(u), Z(u), u_Y) \\
&= f_Y(x_1, W_{de}^C(u), Z(u), u_Y) \\
&\quad - f_Y(x_1, W_{de}^C(u), Z(u), u_Y) \qquad W_{de}^C \notin \text{de}(X) \\
&= 0.
\end{aligned}$$

■

Lemma A.5 (Admissibility w.r.t. structural spurious)**.** The structural spurious effect criterion ($U_X \cap \text{an}(Y) = \emptyset$ and $\text{an}(X) \cap \text{an}_{\mathcal{G}_{\underline{X}}}(Y) = \emptyset$) implies counterfactual spurious effect is 0. Formally:

$$\text{S-SE} \implies \text{Ctf-SE}_{x_0,x_1}(y) = 0 \; \forall u. \tag{A.16}$$

Proof. Note that S-SE implies there is no open backdoor path between X and Y. As a consequence, we know that

$$Y_x \perp\!\!\!\perp X.$$

Furthermore, the absence of backdoor paths also implies we can use the 2nd rule of do-calculus (Action/Observation Exchange). Therefore, we can write:

$$\begin{aligned}
\text{Ctf-SE}_{x_0,x_1}(y) &= P(y_{x_0} \mid x_1) - P(y \mid x_0) \\
&= P(y_{x_0}) - P(y \mid x_0) && \text{since } Y_x \perp\!\!\!\perp X \\
&= P(y_{x_0}) - P(y_{x_0}) && \text{Action/Observation Exchange} \\
&= 0.
\end{aligned}$$

■

Lemma A.6 (Extended Mediation Formula). The total effect can be decomposed into its direct and indirect contributions on every level of the population axes in the explainability plane. Formally, we write:

$$\begin{aligned}
\text{TE}_{x_0,x_1}(y) &= \text{NDE}_{x_0,x_1}(y) - \text{NIE}_{x_1,x_0}(y) && \text{(A.17)} \\
\text{ETT}_{x_0,x_1}(y \mid x) &= \text{Ctf-DE}_{x_0,x_1}(y \mid x) - \text{Ctf-IE}_{x_1,x_0}(y \mid x) && \text{(A.18)} \\
z\text{-TE}_{x_0,x_1}(y \mid z) &= z\text{-DE}_{x_0,x_1}(y \mid z) - z\text{-IE}_{x_1,x_0}(y \mid z) && \text{(A.19)} \\
v'\text{-TE}_{x_0,x_1}(y \mid v') &= v'\text{-DE}_{x_0,x_1}(y \mid v') - v'\text{-IE}_{x_1,x_0}(y \mid v') && \text{(A.20)} \\
\text{unit-TE}_{x_0,x_1}(y(u)) &= \text{unit-DE}_{x_0,x_1}(y(u)) - \text{unit-IE}_{x_1,x_0}(y(u)). && \text{(A.21)}
\end{aligned}$$

Proof. The proof follows from the structural basis expansion from Eq. 3.24. In particular, note that

$$\begin{aligned}
E\text{-TE}_{x_1,x_0}(y \mid E) &= P(y_{x_1} \mid E) - P(y_{x_0} \mid E) && \text{(A.22)} \\
&= P(y_{x_1} \mid E) - P(y_{x_1,W_{x_0}} \mid E) && \text{(A.23)} \\
&\quad + P(y_{x_1,W_{x_0}} \mid E) - P(y_{x_0} \mid E) \\
&= -E\text{-IE}_{x_1,x_0}(y \mid E) + E\text{-DE}_{x_1,x_0}(y \mid E). && \text{(A.24)}
\end{aligned}$$

By using different events E the claim follows. ■

Lemma A.7 (TV Decompositions). The total variation (TV) measure admits the following two decompositions

$$\begin{aligned}
\text{TV}_{x_0,x_1}(y) &= \text{Exp-SE}_{x_1}(y) + \text{TE}_{x_0,x_1}(y) - \text{Exp-SE}_{x_0}(y) && \text{(A.25)} \\
&= \text{ETT}_{x_0,x_1}(y \mid x_0) - \text{Ctf-SE}_{x_1,x_0}. && \text{(A.26)}
\end{aligned}$$

Proof. We write

$$\begin{aligned}\mathrm{TV}_{x_0,x_1}(y) &= P(y \mid x_1) - P(y \mid x_0)\\ &= P(y \mid x_1) - P(y_{x_1}) + P(y_{x_1}) - P(y_{x_0}) + P(y_{x_0}) - P(y \mid x_0)\\ &= \text{Exp-SE}_{x_1}(y) + \mathrm{TE}_{x_0,x_1}(y) - \text{Exp-SE}_{x_0}(y).\end{aligned}$$

Alternatively, we can write

$$\begin{aligned}\mathrm{TV}_{x_0,x_1}(y) &= P(y \mid x_1) - P(y \mid x_0)\\ &= P(y \mid x_1) - P(y_{x_1} \mid x_0) + P(y_{x_1} \mid x_0) - P(y \mid x_0)\\ &= \mathrm{ETT}_{x_0,x_1}(y \mid x_0) - \text{Ctf-SE}_{x_1,x_0}(y),\end{aligned}$$

which completes the proof. ■

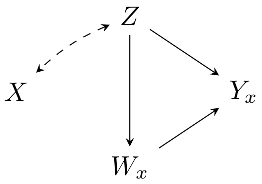
A.2 Soundness of the SFM: Proof of Thm. 4.10 and 4.11

Proof. The proof consists of two parts. In the first part, we show that the quantities where the event E is either of $\emptyset, \{x\}, \{z\}$ (corresponding to the first three rows of the fairness map) are identifiable under the assumptions of the Standard Fairness Model. We in particular show that $\mathrm{TE}_{x_0,x_1}(y)$, Exp-SE$_x(y)$, $\mathrm{TE}_{x_0,x_1}(y \mid z)$, $\mathrm{ETT}_{x_0,x_1}(y \mid x)$, and Ctf-DE$_{x_0,x_1}(y \mid x)$ are identifiable (it follows from very similar arguments that all other quantities are also identifiable). Additionally, we also show that (x,w)-DE$_{x_0,x_1}(y \mid x,w)$ and (x,z,w)-DE$_{x_0,x_1}(y \mid x,z,w)$ are identifiable (being the only identifiable v'-specific measures with $W \subseteq V'$). From this, it follows that for any graph $\mathcal{G}$ compatible with $\mathcal{G}_{\text{SFM}}$, the quantities of interest are (i) identifiable; (ii) their identification expression is the same. This in turn shows that using $\mathcal{G}_{\text{SFM}}$ instead of the full $\mathcal{G}$ does not hurt identifiability of these quantities. In the second part of the proof, we show that any contrast defined by an event E which contains either $W = w$ or $Y = y$ (excluding (x,w)-DE and (x,z,w)-DE) is not identifiable under some very mild conditions (namely the existence of a path $X \to W_{i_1} \to \ldots \to W_{i_k} \to Y$). This part of the proof, complementary to the first part, shows that for contrasts with event E containing post-treatment observations (i.e., descendants of the protected attribute which is manipulated), even having the full graph $\mathcal{G}$ would not make the expression identifiable. All of the proofs here need to be derived from first principles, since the graph $\mathcal{G}_{\text{SFM}}$ contains "groups" of variables Z and W, making the standard identification machinery (Pearl, 2000) not directly applicable.

Part I: Note that for identifying $\mathrm{TE}_{x_0,x_1}(y)$ we need to identify $P(y_x)$. We can write

$$\begin{aligned} P(y_x) &= P(y \mid do(x)) \\ &= \sum_z P(y \mid do(x), z) P(z \mid do(x)) && \text{Law of Total Probability} \\ &= \sum_z P(y \mid x, z) P(z) && (Y \perp\!\!\!\perp X \mid Z)_{\mathcal{G}_{\underline{X}}}, (X \perp\!\!\!\perp Z)_{\mathcal{G}_{\overline{X}}} \end{aligned}$$

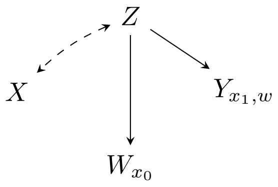
from which it follows that $\text{TE}_{x_0,x_1}(y) = \sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)] P(z)$. Note that the identifiability of $\text{TE}_{x_0,x_1}(y \mid z)$ also follows from the above derivation, namely $\text{TE}_{x_0,x_1}(y \mid z) = \sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)]$, and so does $\text{Exp-SE}_x(y) = \sum_z P(y \mid x, z)[P(z \mid x) - P(z)]$. We are now left with showing that $\text{ETT}_{x_0,x_1}(y \mid x)$ and $\text{Ctf-DE}_{x_0,x_1}(y \mid x)$ are also identifiable. These are Layer 3, counterfactual quantities and therefore rules of do-calculus will not suffice. To be able to use independence statements of counterfactual variables, we will make use of the *make-cg* algorithm of Shpitser and Pearl, 2007 for construction of counterfactual graphs, which extends the twin-network approach of Balke and Pearl, 1994. Therefore, when considering an expression of the form $Y_x = y, X = x'$, we obtain the following counterfactual graph



from which we can see that $Y_x \perp\!\!\!\perp X \mid Z$. Therefore,

$$\begin{aligned} \text{ETT}_{x_0,x_1}(y) &= P(y_{x_1} \mid x) - P(y_{x_0} \mid x) \\ &= \sum_z [P(y_{x_1} \mid x, z) - P(y_{x_0} \mid x, z)] P(z \mid x) && \text{Law of Tot. Prob.} \\ &= \sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)] P(z \mid x) && Y_x \perp\!\!\!\perp X \mid Z. \end{aligned}$$

Finally, for identifying $\text{Ctf-DE}_{x_0,x_1}(y \mid x)$ we use make-cg applied to $\mathcal{G}_{\text{SFM}}$ and $y_{x_1,w}, w_{x_0}, x, z$ to obtain



from which we can say that $Y_{x_1,w} \perp\!\!\!\perp (W_{x_0}, X) \mid Z$. Therefore, we know that Ctf-DE$_{x_0,x_1}(y \mid x)$ equals to

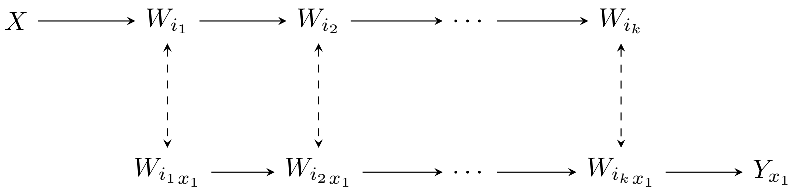
$$\begin{aligned}
&P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_0,W_{x_0}} \mid x) \\
&= \sum_{z} [P(y_{x_1,W_{x_0}} \mid x, z) - P(y_{x_0,W_{x_0}} \mid x, z)]P(z \mid x) && \text{Law of Tot. Prob.} \\
&= \sum_{z,w} [P(y_{x_1,w}, w_{x_0} \mid x, z) - P(y_{x_0,w}, w_{x_0} \mid x, z)]P(z \mid x) && \text{Ctf. unnesting} \\
&= \sum_{z,w} [P(y_{x_1,w} \mid x, z) - P(y_{x_0,w} \mid x, z)]P(w_{x_0} \mid z)P(z \mid x) && Y_{x_1,w} \perp\!\!\!\perp W_{x_0} \mid Z \\
&= \sum_{z,w} [P(y_{x_1,w} \mid x, z) - P(y_{x_0,w} \mid x, z)]P(w \mid x_0, z)P(z \mid x) && W_{x_0} \perp\!\!\!\perp X \mid Z \\
&= \sum_{z,w} [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_0, z)P(z \mid x) && Y_{x,w} \perp\!\!\!\perp X \mid Z.
\end{aligned}$$

From the above, one can also show that

$$\begin{aligned}
(x,w)\text{-DE}_{x_0,x_1}(y \mid x, w) &= \sum_{z} [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)] \\
&\qquad \cdot P(w \mid z, x)P(z \mid x), \\
(x,z,w)\text{-DE}_{x_0,x_1}(y \mid x, z, w) &= P(y \mid x_1, z, w) - P(y \mid x_0, z, w),
\end{aligned}$$

completing the first part of the proof.

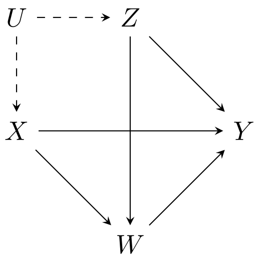
Part II: We next need to show that any contrast with either $W = w$ or $Y = y$ in the event E (excluding (x,w)-DE and (x,z,w)-DE) is not identifiable, even if using the full graph $\mathcal{G}$. We show this for the quantity $P(y_{x_1} \mid x_0, w)$, since other similar quantities work analogously. Assume for simplicity that (i) variable $Z = \emptyset$; (ii) there are no bidirected edges between the W variables. The latter assumption clearly makes the identifiability task easier, since adding bidirected edges can never help identification of quantities. To avoid degenerate cases (and trivial identifiability due to a lack of directed paths), assume that a path $X \to W_{i_1} \to ... \to W_{i_k} \to Y$ exists. Then, when applying make-cg to $\mathcal{G}$ and y_{x_1}, x_0, w the resulting counterfactual graph will contain



as a subgraph and therefore when applying the ID* algorithm of Shpitser and Pearl, 2007, we will encounter a C-component $\{W_i, W_{i_{x_1}}\}$ which will result in non-identifiability of the overall expression. Therefore, even having access to the full $\mathcal{G}$ will not help us identify contrasts that include observations of post-treatment variables, completing the proof. ■

A.3 Proof of Theorem 5.1

Proof. Considering the following SFM



for which we can write the linear structural causal model as follows:

$$U \leftarrow N(0,1) \tag{A.27}$$
$$X \leftarrow \text{Bernoulli}(\text{expit}(U)) \tag{A.28}$$
$$Z \leftarrow a_{UZ}U + a_{ZZ}Z\epsilon_Z \tag{A.29}$$
$$W \leftarrow a_{XW}X + a_{ZW}Z + a_{WW}W + \epsilon_W \tag{A.30}$$
$$Y \leftarrow a_{XY}X + a_{ZY}Z + a_{WY}W + \epsilon_Y \tag{A.31}$$

where matrices a_{ZZ}, a_{WW} are upper diagonal, making the above SCM non-recursive, in the sense that no variable is a functional argument of itself. For simplicity, we assume $\epsilon_Z \sim N(0, I_{n_Z})$, $\epsilon_W \sim N(0, I_{n_W})$ and $\epsilon_Y \sim N(0,1)$. The coefficients a of the above model are assumed to be drawn uniformly from $[-1,1]^{|E|}$, where $|E|$ is the number of edges, with each edge corresponding to a linear coefficient.

Based on the above SCM, the outcome Y can be written

$$Y = \sum_{V_i \in X,Z,W} a_{V_iY}V_i + \epsilon_Y,$$

and the linear predictor of Y, labeled f can be written as

$$f(X,Z,W) = \sum_{V_i \in X,Z,W} \tilde{a}_{V_iY}V_i.$$

The objective of the optimization (i.e., the MSE) can then be written as

$$\begin{aligned}\mathbb{E}[Y - f(X,Z,W)]^2 &= \mathbb{E}\Big[\sum_{V_i \in X,Z,W} (a_{V_iY} - \widetilde{a}_{V_iY})V_i + \epsilon_Y\Big]^2 \\ &= \mathbb{E}[\epsilon_Y^2] + \mathbb{E}\Big[\sum_{V_i,V_j \in X,Z,W} (a_{V_iY} - \widetilde{a}_{V_iY})(a_{V_jY} - \widetilde{a}_{V_jY})V_iV_j\Big] \\ &= 1 + (a_{VY} - \widetilde{a}_{VY})^T\mathbb{E}[VV^T](a_{VY} - \widetilde{a}_{VY}),\end{aligned}$$

when written as a quadratic form with the characteristic matrix $\mathbb{E}[VV^T]$. Here, (with slight abuse of notation) the set V includes X, Z, W, but not Y. Further, the constraint $\text{TV}_{x_0,x_1}(f) = 0$ is in fact a linear constraint on the coefficients $\widetilde{a}_{VY}$, since we have that

$$TV_{x_0,x_1}(f) = (\mathbb{E}[V \mid x_1] - \mathbb{E}[V \mid x_0])^T\widetilde{a}_{VY}.$$

We write

$$c = \mathbb{E}[V \mid x_1] - \mathbb{E}[V \mid x_0], \tag{A.32}$$

$$\Sigma = \mathbb{E}[VV^T] \tag{A.33}$$

and note that our optimization problem can be written as

$$\underset{\widetilde{a}_{VY}}{\arg\min} \qquad (a_{VY} - \widetilde{a}_{VY})^T\Sigma(a_{VY} - \widetilde{a}_{VY}) \tag{A.34}$$

$$\text{subject to} \qquad c^T\widetilde{a}_{VY} = 0. \tag{A.35}$$

The objective is a quadratic form centered at a_{VY}. Geometrically, the solution to the optimization problem is the meeting point of an ellipsoid centered at a_{VY} with the characteristic matrix Σ and the hyperplane through the origin with the normal vector c. After a change of basis (substituting $t = \Sigma^{\frac{1}{2}}(a_{VY} - \widetilde{a}_{VY})$), the solution can be derived explicitly as

$$\widehat{a}_{VY} = a_{VY} - \frac{c^T a_{VY}\Sigma^{-1}c}{c^T\Sigma^{-1}c}.$$

We next analyze the constraints

$$\text{Ctf-DE}_{x_0,x_1}(\widehat{f}_{\text{fair}} \mid x_0) = \text{Ctf-IE}_{x_1,x_0}(\widehat{f}_{\text{fair}} \mid x_0) = \text{Ctf-SE}_{x_1,x_0}(\widehat{f}_{\text{fair}}) = 0.$$

The first constraint $\text{Ctf-DE}_{x_0,x_1}(\widehat{f}_{\text{fair}} \mid x_0)$ can be simply written as $\widehat{a}_{XY}(x_1 - x_0) = 0$, and since $x_1 - x_0 = 1$, the constraint can be written as $c_1^T\widehat{a}_{VY} = 0$ where $c_1 = (1, 0, \ldots, 0)^T$. Similarly, but more involved, the Ctf-IE constraint can be written as $c_2^T\widehat{a}_{VY} = 0$ where entries of c_2 corresponding to W_i variables are

$$\mathbb{E}[(W_i)_{x_0} \mid x_0] - \mathbb{E}[(W_i)_{x_1} \mid x_0],$$

and 0 everywhere else. Finally, the Ctf-SE constraint can be written as $c_3^T \widehat{a}_{VY} = 0$ where entries of c_3 corresponding to W_i variables are

$$\mathbb{E}[(W_i)_{x_1} \mid x_0] - \mathbb{E}[(W_i)_{x_1} \mid x_1],$$

and the entries corresponding to Z_i variables

$$\mathbb{E}[Z_i \mid x_1] - \mathbb{E}[Z_i \mid x_0].$$

Notice also that $c_1 - c_2 - c_3 = c$ (following from the decomposition result in Eq. 4.48). We further note that by inverting Eq. A.29 and using linearity of expectations

$$\mathbb{E}[Z \mid x_0] - \mathbb{E}[Z \mid x_1] = -(I - a_{ZZ})^{-1} a_{UZ} \delta_u^{01}$$

where $\delta_u^{01} = \mathbb{E}[U \mid x_1] - \mathbb{E}[U \mid x_0]$ is a constant. Similarly,

$$\mathbb{E}[W_{x_1} \mid x_0] - \mathbb{E}[W_{x_1} \mid x_1] = -(I - a_{WW})^{-1} a_{ZW} (I - a_{ZZ})^{-1} a_{UZ} \delta_u^{01}.$$

Furthermore, for the indirect effect, we have that

$$\mathbb{E}[W_{x_0} \mid x_0] - \mathbb{E}[W_{x_1} \mid x_0] = -(I - a_{WW})^{-1} a_{XW}.$$

Therefore, we can now see how the three constraints can be expressed in terms of the structural coefficients a. What remains is understanding the entries of the Σ matrix. Note that $\mathbb{E}[V_i V_j]$ can be computed by considering all *treks* from V_i to V_j. A trek is a path that first goes backwards from V_i until a certain node, and then forwards to V_j. The slight complication comes from the treks with the turning point at U that pass through X, as the SCM is not linear along the bidirected $U \dashleftarrow\dashrightarrow X$ edge. Nonetheless, in this case the contribution to the covariance of V_i and V_j equals the product of the coefficients on the trek multiplied by $\mathbb{E}[XU]$. Therefore, we note that

$$\mathbb{E}[V_i V_j] = \sum_{\substack{\text{treks } T_s \\ \text{from } V_i \text{ to } V_j}} \lambda(T_s) \prod_{\substack{\text{edges } V_k \to V_l \\ \in T_s}} a_{V_k V_l}$$

where the weighing factor $\lambda(T_s)$ is either 1 or $\mathbb{E}[XU]$ depending on the trek T_s. To conclude the argument, notice the following. The entries of the Σ matrix are polynomial functions of the structural coefficients a. The same also therefore holds for Σ^{-1}. Furthermore, the coefficient c is also a polynomial function of coefficients in a. Therefore, the condition $c_1^T \widehat{a}_{VY} = 0$ can be written as

$$c_1^T (a_{VY} - \frac{c^T a_{VY} \Sigma^{-1} c}{c^T \Sigma^{-1} c}) = 0, \tag{A.36}$$

where the left hand side is a polynomial expression in the coefficients of a. Therefore, the above expression defines an algebraic hypersurface. Any such hypersurface has measure 0 in the space $[-1, 1]^{|E|}$, proving that the set of 0-TV-compliant SCMs is in fact of measure 0. Intuitively, the result is saying that the meeting point of an ellipsoid centered at a_{VY} with the characteristic matrix Σ and the hyperplane through the origin with the normal vector c with measure 0 also lies on a random hyperplane defined by the normal vector c_1 and passing through the origin.

To extend the result for an $\epsilon > 0$, we proceed as follows. Let $\mathcal{H}(\epsilon)$ be the set of ϵ-TV-compliant SCMs. Let $\mathcal{H}^{DE}(\epsilon)$ be the set of SCMs for which the direct effect is bounded by ϵ for the $\widehat{f}_{\text{fair}}$. Let $\mathcal{H}^{IE}(\epsilon)$, $\mathcal{H}^{SE}(\epsilon)$ be defined analogously for the indirect and spurious effects. We then analyze the degrees of the terms appearing in Eq. A.36, which defines the hypersurface $\mathcal{H}^{DE}(0)$. In particular, notice that

$$deg(c_1^T(a_{VY} - \frac{c^T a_{VY}\Sigma^{-1}c}{c^T\Sigma^{-1}c})) \leq deg(c_1) + deg(a_{VY}) + deg(\frac{c^T a_{VY}\Sigma^{-1}c}{c^T\Sigma^{-1}c}) \tag{A.37}$$

and also that

$$\begin{align} deg(\frac{c^T a_{VY}\Sigma^{-1}c}{c^T\Sigma^{-1}c}) &\leq deg(c^T a_{VY}\Sigma^{-1}c) + deg(c^T\Sigma^{-1}c) \tag{A.38}\\ &\leq 2deg(c) + deg(a_{VY}) + deg(\Sigma^{-1}) + 2deg(c) + deg(\Sigma^{-1}). \tag{A.39}\end{align}$$

Now, one can observe the following bounds, where $p = |V|$:

$$\begin{align} deg(c) &\leq p \text{ from Eq. A.32}, \tag{A.40}\\ deg(a_{VY}) &= 1 \text{ by definition}, \tag{A.41}\\ deg(\Sigma^{-1}) &\leq p^2 \cdot \max_{i,j} deg(\Sigma_{ij}) = p^4 \text{ from Eq. A.33}. \tag{A.42}\end{align}$$

from which it follows that the degree of the hypersurface of 0-TV-compliant SCMs, labeled $\mathcal{H}(0)$, is bounded by $2 + 4p + 2p^2$. Lojasiewicz's inequality (Ji *et al.*, 1992, Thm. 1) states that if K is a compact set, f a real analytic function on $\mathbb{R}^n$, and $Z = \{x \in \mathbb{R}^n : f(x) = 0\}$ is the locus of f, then there exist positive constants k_1, k_2 such that

$$\inf_{z\in Z} \|x - z\|_2 \leq k_1|f(x)|^{k_2} \; \forall x \in K. \tag{A.43}$$

Therefore, there exist constants k_1, k_2 such that:

$$\begin{align} \text{vol}(\mathcal{H}^{DE}(\epsilon)) &= \text{vol}\{a \in [-1, 1]^{|E|} \mid |c_1^T(a_{VY} - \frac{c^T a_{VY}\Sigma^{-1}c}{c^T\Sigma^{-1}c})| \leq \epsilon\} \tag{A.44}\\ &= \text{vol}\{a \in [-1, 1]^{|E|} \mid d(a, \mathcal{H}^{DE}(0)) \leq k_1\epsilon^{k_2}\}, \tag{A.45}\end{align}$$

where the second line follows from Lojasiewicz's inequality with the choice $f = \text{Ctf-DE}_{x_0,x_1}(\widehat{f}_{\text{fair}} \mid x_0)$, $Z = \mathcal{H}^{DE}(0)$, and setting $K = \mathcal{H}^{DE}(\epsilon)$. By an application of the Crofton's formula (Guth, 2009, p. 1975), for a real algebraic hypersurface $\mathcal{H}$ of a degree d, its volume in the unit n-ball can be bounded above by

$$\text{vol}(H) \leq C(n)d, \tag{A.46}$$

where the constant C only depends on the dimension n. By a rescaling argument, the volume in the n-ball of radius R can be bounded by $R^n C(n)d$. Therefore, the volume in Eq. A.45 can be bounded above by

$$\text{vol}(\mathcal{H}^{DE}(\epsilon)) \leq k_1 \epsilon^{k_2} |E|^{|E|/2} C(|E|) deg(\mathcal{H}^{DE}(0)), \tag{A.47}$$

by using the inequality Eq. A.46 with the choice $\mathcal{H} = \mathcal{H}^{DE}(0)$, scaling factor $R = \sqrt{|E|}$ (which ensures that the hypercube $[-1, 1]^n$ is contained in the $|E|$-ball of radius R), and noting that the maximal thickness of $\mathcal{H}^{DE}(\epsilon)$ compared to $\mathcal{H}^{DE}(0)$ is bounded above by $k_1\epsilon^{k_2}$ (see Eq. A.45). Finally, we can write that for a random M sampled from $\mathcal{S}^{linear}_{n_Z,n_W}$ we have that

$$\mathbb{P}(M \in \mathcal{H}^{DE}(\epsilon)) = \frac{\text{vol}(\mathcal{H}^{DE}(\epsilon))}{2^{|E|}}. \tag{A.48}$$

By noting that $|E| = p(p+1)$ and setting

$$\epsilon = \Big(\frac{2^{p(p+1)}}{8k_1 C(|E|)(p+1)^2 (p(p+1))^{\frac{p(p+1)}{2}}} \Big)^{1/k_2} \tag{A.49}$$

we obtain that $\mathbb{P}(M \in \mathcal{H}^{DE}(\epsilon)) \leq \frac{1}{4}$. Since we know that

$$\mathcal{H}(\epsilon) = \mathcal{H}^{DE}(\epsilon) \cap \mathcal{H}^{IE}(\epsilon) \cap \mathcal{H}^{SE}(\epsilon) \tag{A.50}$$

$$\implies \mathbb{P}(M \in \mathcal{H}(\epsilon)) \leq \mathbb{P}(M \in \mathcal{H}^{DE}(\epsilon)) \tag{A.51}$$

$$\implies \mathbb{P}(M \in \mathcal{H}(\epsilon)) \leq \frac{1}{4}, \tag{A.52}$$

for such an ϵ. Intuitively, any SCM in $\mathcal{H}(\epsilon)$ must also be in $\mathcal{H}^{DE}(\epsilon)$. Any SCM in $\mathcal{H}^{DE}(\epsilon)$ must be close to $\mathcal{H}^{DE}(0)$. The maximal deviation of an SCM in $\mathcal{H}^{DE}(\epsilon)$ from $\mathcal{H}^{DE}(0)$ can be bounded using Lojasiewicz's inequality, whereas the surface area of $\mathcal{H}^{DE}(0)$ can be bounded above by an application of Crofton's formula. Putting together, we get a bound on the measure of ϵ-TV-compliant SCMs. ■

The behavior of the ϵ term given in Eq. A.49 cannot be theoretically analyzed further, since the constants arising from the Lojasiewicz's inequality

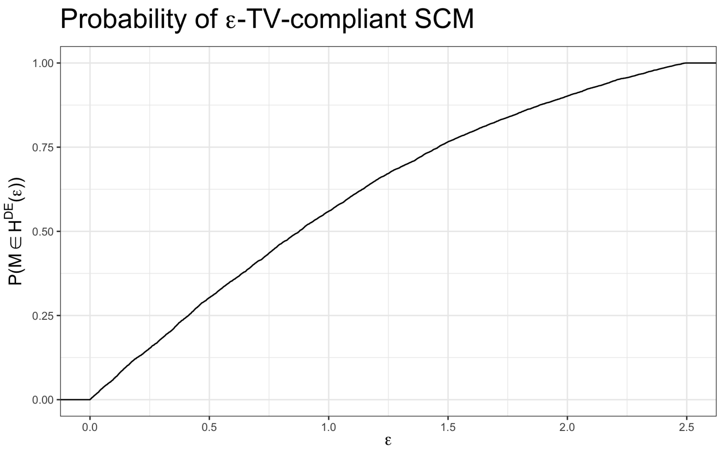


Figure A.1: Estimating empirically the probability that a random SCM in $\mathcal{S}^{linear}_{n_Z,n_W}$, for $n_Z = n_W = 5$, has a direct effect smaller than ϵ after ensuring that TV equals 0.

are dimension dependent. To this end, for $n_Z = n_W = 5$ we empirically estimate

$$\mathbb{P}(M \in \mathcal{H}^{DE}(\epsilon)) \tag{A.53}$$

for a range of ϵ values, and obtain the plot in Fig. A.1.

A.4 Proof of Thm. 5.3

Proof. We prove the result for the case BN-set= $\emptyset$ (the other cases of BN-sets follow analogously), in the population level case. Based on the standard fairness model, we are starting with an SCM $\mathcal{M}$ given by:

$$X \leftarrow f_X(u_x, u_z) \tag{A.54}$$
$$Z \leftarrow f_Z(u_x, u_z) \tag{A.55}$$
$$W \leftarrow f_W(X, Z, u_w) \tag{A.56}$$
$$Y \leftarrow f_Y(X, Z, W, u_y). \tag{A.57}$$

The noise variables u_x, u_z are not independent, but the variables u_w, u_y are mutually independent, and also independent from u_x, u_z.

We now explain how the sequential optimal transport steps extend the original SCM $\mathcal{M}$ (to which we do not have access). Firstly, the conditional distribution $Z \mid X = x_1$ is transported onto $Z \mid X = x_0$. Write τ^Z for the transport map. On the level of the SCM, this corresponds to extending the equations by an additional mechanism

$$\widetilde{Z} \leftarrow \begin{cases} f_Z(u_x, u_z) & \text{if } f_X(u_x, u_z) = x_0 \\ f_Z(\pi^Z(u_x, u_z)) & \text{if } f_X(u_x, u_z) = x_1 \end{cases}. \tag{A.58}$$

Here, there is an implicit (possibly stochastic) mapping π^Z that we cannot observe. For simplicity, we assume that the variable Z is continuous and that π^Z is deterministic. We can give an optimization problem to which π^Z is the solution, namely:

$$\begin{aligned} \pi^Z := \arg\min_{\pi} \quad & \int_{\mathcal{U}_X \times \mathcal{U}_Z} \|f_Z(\pi(u_z, u_x)) - f_Z(u_z, u_x)\|^2 du_{xz}^{X=x_1} \\ \text{s.t.} \quad & \underset{u_x, u_z \sim U_X, U_Z | X = x_1}{f_Z(\pi(u_z, u_x))} \overset{d}{=} \underset{u_x, u_z \sim U_X, U_Z | X = x_0}{f_Z(u_z, u_x)} \quad . \end{aligned} \tag{A.59}$$

The measure $du_{xz}^{X=x_1}$ in the objective is the probability measure associated with the distribution $P(u_x, u_z \mid X = x_1)$. The constraint ensures that after the transport, $\widetilde{Z} \mid X = x_1$ is equal in distribution to $\widetilde{Z} \mid X = x_0$. In the second step of the procedure, we are transporting the distribution of W. This results in adding the mechanism:

$$\widetilde{W} \leftarrow \begin{cases} f_W(x_0, \widetilde{Z}, u_w) & \text{if } X = x_0 \\ f_W(x_0, \widetilde{Z}, \pi^W(u_w)) & \text{if } X = x_1 \end{cases}. \tag{A.60}$$

Similarly as for π^Z, π^W is a mapping that solves following optimization problem:

$$\begin{aligned} \pi^W := \arg\min_{\pi} \quad & \int_{\mathcal{U}_W} \|f_W(x_0, \widetilde{z}, \pi(u_w)) - f_W(x_1, \widetilde{z}, u_w)\|^2 du_w \\ \text{s.t.} \quad & f_W(x_0, \widetilde{z}, \pi(u_w)) \overset{d}{=} f_W(x_0, \widetilde{z}, u_w). \end{aligned} \tag{A.61}$$

The above optimization problem is thought of being solved separately for each value of $\widetilde{Z} = \widetilde{z}$. Finally, in the last step, we are constructing the additional mechanism:

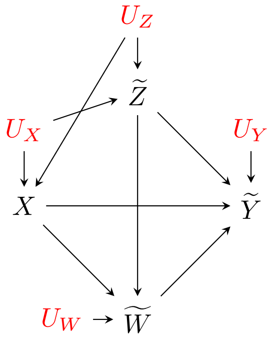
$$\widetilde{Y} \leftarrow \begin{cases} f_Y(x_0, \widetilde{Z}, \widetilde{W}, u_y) & \text{if } X = x_0 \\ f_Y(x_0, \widetilde{Z}, \widetilde{W}, \pi^Y(u_y)) & \text{if } X = x_1 \end{cases} \tag{A.62}$$

Again, the implicit mapping π^Y is constructed so that it is the solution to

$$\begin{aligned} \pi^Y := \arg\min_{\pi} \quad & \int_{\mathcal{U}_Y} \|f_Y(x_0, \widetilde{z}, \widetilde{w}, \pi(u_y)) - f_Y(x_1, \widetilde{z}, \widetilde{w}, u_y)\|^2 du_y \\ \text{s.t.} \quad & f_Y(x_0, \widetilde{z}, \widetilde{w}, \pi(u_y)) \overset{d}{=} f_Y(x_0, \widetilde{z}, \widetilde{w}, u_y). \end{aligned} \tag{A.63}$$

where the problem is solved separately for each fixed choice of parents $\widetilde{Z} = \widetilde{z}$, $\widetilde{W} = \widetilde{w}$.

After constructing the additional mechanisms $\widetilde{Z}, \widetilde{W}$, and $\widetilde{Y}$, we draw the explicit causal diagram corresponding to the new variables, which includes the unobservables U_X, U_Z, U_W, and U_Y (marked in red), given as follows:



.

Note that by marginalizing out the unobserved variables U_X, U_Z, U_W, U_Y, we obtain the new causal diagram, which is given by the standard fairness model over the variables $X, \widetilde{Z}, \widetilde{W}, \widetilde{Y}$. Therefore, it follows that the identification expressions for the spurious, indirect, and direct effects are known, and given by:

$$x\text{-DE}_{x_0,x_1}(\widetilde{y} \mid x_0) = \sum_{\widetilde{z},\widetilde{w}} [P(\widetilde{y} \mid x_1, \widetilde{z}, \widetilde{w}) - P(\widetilde{y} \mid x_0, \widetilde{z}, \widetilde{w})] P(\widetilde{w} \mid x_0, \widetilde{z}) P(\widetilde{z} \mid x) \tag{A.64}$$

$$x\text{-IE}_{x_0,x_1}(\widetilde{y} \mid x_0) = \sum_{\widetilde{z},\widetilde{w}} P(\widetilde{y} \mid x_0, \widetilde{z}, \widetilde{w}) [P(\widetilde{w} \mid x_1, \widetilde{z}) - P(\widetilde{w} \mid x_0, \widetilde{z})] P(\widetilde{z} \mid x) \tag{A.65}$$

$$x\text{-SE}_{x_1,x_0}(\widetilde{y}) = \sum_{\widetilde{z}} P(\widetilde{y} \mid x_1, \widetilde{z}) [P(\widetilde{z} \mid x_0) - P(\widetilde{z} \mid x_1)]. \tag{A.66}$$

To finish the proof, notice that by construction (the matching of distributions via optimal transport), we have that

$$P(\widetilde{y} \mid x_1, \widetilde{z}, \widetilde{w}) = P(\widetilde{y} \mid x_0, \widetilde{z}, \widetilde{w}) \tag{A.67}$$
$$P(\widetilde{w} \mid x_1, \widetilde{z}) = P(\widetilde{w} \mid x_0, \widetilde{z}) \tag{A.68}$$
$$P(\widetilde{z} \mid x_0) = P(\widetilde{z} \mid x_1), \tag{A.69}$$

implying that all three effects in Eq. A.64-A.66 are equal to 0 (the argument for showing that x-DE$_{x_1,x_0}(\widetilde{y} \mid x_0)$ and x-DE$_{x_1,x_0}(\widetilde{y} \mid x_0)$ are also equal to 0 is the same). ■

A.5 Proof of Prop. 5.2

Proof. Suppose that the contrast (C_0, C_1, E_0, E_1) is a counterfactual one, meaning that $C_1 \neq C_0$, $E_1 = E_0$ (the proof for factual contrasts with $C_1 = C_0, E_1 \neq E_0$ is the same). Using the structural basis expansion from Thm. 3.1, the fairness condition $\mu(\widehat{y}) = 0$ implies that

$$\sum_u [\widehat{y}_{C_1}(u) - \widehat{y}_{C_0}(u)] P(u \mid E) = 0. \tag{A.70}$$

For part (a), assume that the policy D is a linear function of $\widehat{Y}$, i.e., $f_D(\widehat{y}) = a\widehat{y} + b$. Then we simply have that:

$$\begin{aligned} \mu(d) &= \sum_u [d_{C_1}(u) - d_{C_0}(u)] P(u \mid E) && \text{(A.71)} \\ &= a \cdot \sum_u [\widehat{y}_{C_1}(u) - \widehat{y}_{C_0}(u)] P(u \mid E) && \text{(A.72)} \\ &= a\mu(\widehat{y}) = 0. && \text{(A.73)} \end{aligned}$$

For part (b), assume that the measure μ is a unit-level measure (the event $E = \{U = u\}$). Then, the fairness condition implies that $\widehat{y}_{C_1}(u) = \widehat{y}_{C_0}(u)\ \forall u$, from which it follows that

$$d_{C_1}(u) = f_D(\widehat{y}_{C_1}(u)) = f_D(\widehat{y}_{C_0}(u)) = d_{C_0}(u)\ \forall u. \tag{A.74}$$

■

A.6 Ex. 5.11 Computation

Here we provided the expanded computation from Ex. 5.11, showing why Eq. 5.143 hold. Notice that for $x \in \{x_0, x_1\}$ we can compute the probability of the joint distribution of the potential responses as follows:

$$\begin{aligned} P(y_{d_0} = 0, y_{d_1} = 1 \mid w, x) &= P(U_Y - \frac{w}{5} < 0.5, U_Y + \frac{w}{3} - \frac{w}{5} > 0.5) && \text{(A.75)} \\ &= P(U_Y < 0.5 + \frac{w}{5}, U_Y > 0.5 + \frac{w}{5} - \frac{w}{3}) && \text{(A.76)} \\ &= P(0.5 + \frac{w}{5} - \frac{w}{3} < U_Y < 0.5 + \frac{w}{5}) && \text{(A.77)} \\ &= \frac{w}{3} \quad (\text{using } U_Y \sim \text{Unif}[0,1]), && \text{(A.78)} \end{aligned}$$

from which Eq. 5.143 follows.

A.7 Proof of Thm. 4.13 and Cor. 4.14

Proof. For the theorem proof, consider that:

$$\begin{aligned}
\mathbb{E}(y \mid x_1, \widehat{y}) - \mathbb{E}(y \mid x_0, \widehat{y}) &= \mathbb{E}(y_{x_1} \mid x_1, \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_0, \widehat{y}_{x_0}) && \text{(A.79)}\\
&= \underbrace{\mathbb{E}(y_{x_1} \mid x_1, \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_1})}_{\text{Term (I)}} && \text{(A.80)}\\
&+ \underbrace{\mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_0})}_{\text{Term (II)}} && \text{(A.81)}\\
&+ \underbrace{\mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_0}) - \mathbb{E}(y_{x_0} \mid x_0, \widehat{y}_{x_0})}_{\text{Term (III)}}. && \text{(A.82)}
\end{aligned}$$

Since by assumption no backdoor paths between X and $Y, \widehat{Y}$ exist, Term (III) vanishes. By noting that $\mathbb{E}(y_x \mid x_1, \widehat{y}_{x_1}) = \mathbb{E}(y_x \mid x_1, \widehat{y}) \;\; \forall x$ by consistency (and applying it to Term (I)), and also that $Y_x \perp\!\!\!\perp X$ (and applying it to Term (II)) gives us the required result.

For Cor. 4.14, we further assume that the SCM is linear, and that the predictor $\widehat{Y}$ is efficient, i.e., $\widehat{Y}(x, w) = \mathbb{E}[Y \mid x, w]$. In the linear case, the efficiency simply translates to the fact that

$$\alpha_{W\widehat{Y}} = \alpha_{WY}, \tag{A.83}$$

$$\alpha_{X\widehat{Y}} = \alpha_{XY}. \tag{A.84}$$

Due to linearity, for every unit u, we have that

$$y_{x_1}(u) - y_{x_0}(u) = \alpha_{XW}\alpha_{WY} + \alpha_{XY}, \tag{A.85}$$

and since Term (I) can be written as $\sum_u [y_{x_1}(u) - y_{x_0}(u)] P(u \mid x_1, \widehat{y})$, Eq. 4.222 follows. We next look at Term (II), which can be expanded as

$$\sum_u \widehat{y}_{x_0}(u) [P(u \mid \widehat{y}_{x_1}) - P(u \mid \widehat{y}_{x_0})]. \tag{A.86}$$

We now look at units u which are compatible with $\widehat{Y}_{x_1}(u) = \widehat{y}$ and $\widehat{Y}_{x_0}(u) = \widehat{y}$. We can expand $\widehat{Y}_{x_1}(u)$ as

$$\widehat{Y}_{x_1}(u) = \alpha_{X\widehat{Y}} + \alpha_{XW}\alpha_{W\widehat{Y}} + \alpha_{W\widehat{Y}} u_W. \tag{A.87}$$

Thus, we have that

$$\widehat{Y}_{x_1}(u) = \widehat{y} \implies \alpha_{W\widehat{Y}} u_W = \widehat{y} - \alpha_{X\widehat{Y}} + \alpha_{XW}\alpha_{W\widehat{Y}}. \tag{A.88}$$

Similarly, we also obtain that

$$\widehat{Y}_{x_0}(u) = \widehat{y} \implies \alpha_{W\widehat{Y}} u_W = \widehat{y}. \tag{A.89}$$

Due to the efficiency of learning which implies that $\alpha_{W\widehat{Y}} = \alpha_{WY}$ and $\alpha_{X\widehat{Y}} = \alpha_{XY}$, Eq. A.88 and A.89 imply

$$y_{x_0}(u) = \widehat{y} - (\alpha_{XY} + \alpha_{XW}\alpha_{WY}) \;\; \forall u \text{ s.t. } \widehat{Y}_{x_1}(u) = \widehat{y}, \tag{A.90}$$

$$y_{x_0}(u) = \widehat{y} \;\; \forall u \text{ s.t. } \widehat{Y}_{x_0}(u) = \widehat{y}, \tag{A.91}$$

which in turn shows that

$$\mathbb{E}(y_{x_0} \mid \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid \widehat{y}_{x_0}) = -\alpha_{XY} - \alpha_{XW}\alpha_{WY}. \tag{A.92}$$

■

A.8 Proof of Thm. 5.6

Proof. The first part of the theorem states the optimality of the D^{CF} policy in the counterfactual world. Given that the policy uses the true benefit values from the counterfactual world, we apply the argument of Prop. 5.7 to prove its optimality.

We next prove the optimality of the D^{UT} policy from Alg. 5.5. In Step 2 we check whether all individuals with a positive benefit can be treated. If yes, then the policy D^{UT} is the overall optimal policy. If not, in Step 6 we check whether the overall optimal policy has a disparity bounded by M. If this is the case, D^{UT} is the overall optimal policy for a budget $\leq b$, and cannot be strictly improved. For the remainder of the proof, we may suppose that D^{UT} uses the entire budget b (since we are operating under scarcity), and that D^{UT} has introduces a disparity $\geq M$. We also assume that the benefit Δ admits a density, and that probability $P(\Delta \in [a, b] \mid x) > 0$ for any $[a, b] \subset [0, 1]$ and x.

Let $\delta^{(x_0)}, \delta^{(x_1)}$ be the two thresholds used by the D^{UT} policy. Suppose that $\widetilde{D}^{UT}$ is a policy that has a higher expected utility and introduces a disparity bounded by M, or treats everyone in the disadvantaged group. Then there exists an alternative policy $\overline{D}^{UT}$ with a higher or equal utility that takes the form

$$\overline{D}^{UT} = \begin{cases} 1 \text{ if } \Delta(x_1, z, w) > \delta^{(x_1)'}, \\ 1 \text{ if } \Delta(x_0, z, w) > \delta^{(x_0)'}, \\ 0 \text{ otherwise.} \end{cases} \tag{A.93}$$

with $\delta^{(x_0)'}, \delta^{(x_1)'}$ non-negative (otherwise, the policy can be trivially improved). In words, for any policy $\overline{D}^{UT}$ there is a threshold based policy that is no worse. The policy D^{UT} is also a threshold based policy. Now, if we had

$$\delta^{(x_1)'} < \delta^{(x_1)} \tag{A.94}$$

$$\delta^{(x_0)'} < \delta^{(x_0)} \tag{A.95}$$

it would mean policy $\overline{D}^{UT}$ is using a larger budget than D^{UT}. However, D^{UT} uses a budget of b, making $\overline{D}^{UT}$ infeasible. Therefore, we must have that

$$\delta^{(x_1)'} < \delta^{(x_1)}, \delta^{(x_0)'} > \delta^{(x_0)} \text{ or} \tag{A.96}$$

$$\delta^{(x_1)'} > \delta^{(x_1)}, \delta^{(x_0)'} < \delta^{(x_0)}. \tag{A.97}$$

We first handle the case in Eq. A.96. In this case, the policy $\overline{D}^{UT}$ introduces a larger disparity than D^{UT}. Since the disparity of D^{UT} is at least M, the disparity of $\overline{D}^{UT}$ is strictly greater than M. Further, note that $\delta^{(x_0)'} > \delta^{(x_0)} \geq 0$, showing that $\overline{D}^{UT}$ does not treat all individuals with a positive benefit in the disadvantaged group. Combined with a disparity of $> M$, this makes the policy $\overline{D}^{UT}$ infeasible.

For the second case in Eq. A.97, let $U(\delta_0, \delta_1)$ denote the utility of a threshold based policy:

$$U(\delta_0, \delta_1) = \mathbb{E}[\Delta \mathbb{1}(\Delta > \delta_0)\mathbb{1}(X = x_0)] + \mathbb{E}[\Delta \mathbb{1}(\Delta > \delta_1)\mathbb{1}(X = x_1)]. \tag{A.98}$$

Thus, we have that $U(\delta^{(x_0)}, \delta^{(x_1)}) - U(\delta^{(x_0)'}, \delta^{(x_1)'})$ equals

$$\begin{aligned}
&\mathbb{E}[\Delta \mathbb{1}(\Delta \in [\delta^{(x_1)}, \delta^{(x_1)'}])\mathbb{1}(X = x_1)] && \text{(A.99)}\\
&\quad - \mathbb{E}[\Delta \mathbb{1}(\Delta \in [\delta^{(x_0)'}, \delta^{(x_0)}])\mathbb{1}(X = x_0)] && \text{(A.100)}\\
\geq &\delta^{(x_1)}\mathbb{E}[\mathbb{1}(\Delta \in [\delta^{(x_1)}, \delta^{(x_1)'}])\mathbb{1}(X = x_1)] && \text{(A.101)}\\
&\quad - \delta^{(x_0)}\mathbb{E}[\mathbb{1}(\Delta \in [\delta^{(x_0)'}, \delta^{(x_0)}])\mathbb{1}(X = x_0)] && \text{(A.102)}\\
\geq &\delta^{(x_0)}\big(\mathbb{E}[\mathbb{1}(\Delta \in [\delta^{(x_1)}, \delta^{(x_1)'}])\mathbb{1}(X = x_1)] && \text{(A.103)}\\
&\qquad - \mathbb{E}[\mathbb{1}(\Delta \in [\delta^{(x_0)'}, \delta^{(x_0)}])\mathbb{1}(X = x_0)]\big) && \text{(A.104)}\\
= &\delta^{(x_0)}\big(P(\Delta \in [\delta^{(x_1)}, \delta^{(x_1)'}], x_1) && \text{(A.105)}\\
&\qquad - P(\Delta \in [\delta^{(x_0)'}, \delta^{(x_0)}], x_0)\big) && \text{(A.106)}\\
\geq &0, && \text{(A.107)}
\end{aligned}$$

where the last line follows from the fact that $\overline{D}^{UT}$ has a budget no higher than D^{UT}. Thus, this case also gives a contradiction.

Therefore, we conclude that policy D^{UT} is optimal among all policies with a budget $\leq b$ that either introduce a bounded disparity in resource allocation $|P(d \mid x_1) - P(d \mid x_0)| \leq M$ or treat everyone with a positive benefit in the disadvantaged group. ■

A.9 Proof of Thm. 6.4

Proof. Suppose that $U_s \subseteq U_{sToT}$, and suppose that (i) $Y \notin \mathrm{AS}(U_s)$; (ii) $U_s = \mathrm{an}^{ex}_{sToT}(AS(U_s))$. Let Z_s be the anchor set of U_s, with Y excluded. Note that we have by definition that

$$P(y \mid x^{U_s}) = \sum_{u_s} P(u_s) P(y \mid x, u_s). \tag{A.108}$$

Denote all the exogenous ancestors of the set Z_s by $\mathrm{an}^{ex}(Z_s)$. The set $\mathrm{an}^{ex}(Z_s)$ can be partitioned into three subsets:

$$U_s^x\text{, latents with a causal path to } X\text{, but which are not in } U_{sToT}, \tag{A.109}$$

$$U_s^y\text{, latents with a causal path to } Y\text{, but which are not in } U_{sToT}, \tag{A.110}$$

$$U_s\text{, latents in } U_{sToT}. \tag{A.111}$$

Note that $\mathrm{an}^{ex}(Z_s)$ could, in general, contain variables in U_{sToT} which are not in U_s. However, this case is precluded by the condition (ii) above. Then, note that we have

$$Y \perp\!\!\!\perp U_s^x \mid X, U_s, \tag{A.112}$$

since any path from U_s^x to Y must be intercepted by X. Hence, we can write

$$\begin{aligned} P(y \mid x^{U_s}) &= \sum_{u_s, u_s^x} P(u_s, u_s^x) P(y \mid x, u_s, u_s^x) \quad \text{using Eq.} A.112 && \text{(A.113)} \\ &= \sum_{u_s, u_s^x, u_s^y} P(u_s, u_s^x) P(y \mid x, u_s, u_s^x, u_s^y) P(u_s^y \mid x, u_s, u_s^x) && \text{(A.114)} \end{aligned}$$

Now, note that we have

$$U_s^y \perp\!\!\!\perp X, U_s, U_s^x, \tag{A.115}$$

since U_s^y has no path to X, U_s, U_s^x. Denote by $u_{\overline{s}}$ the values u_s, u_s^x, u_s^y. Thus, we can re-write Eq. A.114 as

$$\begin{aligned} P(y \mid x^{U_s}) &= \sum_{u_{\overline{s}}} P(u_{\overline{s}}) P(y \mid x, u_{\overline{s}}) && \text{(A.116)} \\ &= \sum_{u_{\overline{s}}} P(u_{\overline{s}}) P(y \mid x, u_{\overline{s}}, z_s(u_{\overline{s}})) && \text{(A.117)} \\ &= \sum_{z_s} \sum_{u_{\overline{s}}} \mathbb{1}(Z_s(u_{\overline{s}}) = z_s) P(u_{\overline{s}}) P(y \mid x, u_{\overline{s}}, z_s(u_{\overline{s}})). && \text{(A.118)} \end{aligned}$$

Now, note that

$$Y \perp\!\!\!\perp U_{\overline{s}} \mid X, Z_s, \tag{A.119}$$

since X, Z_s close all paths from $U_{\overline{s}}$ to Y, which is also due to the fact that $Y \notin \mathrm{AS}(U_s)$. Without the latter condition, there could be an element in U_s^y which points directly to Y, and cannot be separated from Y. Therefore, using Eq. A.119, we finally have that

$$P(y \mid x^{U_s}) = \sum_{z_s} \sum_{u_{\overline{s}}} \mathbb{1}(Z_s(u_{\overline{s}}) = z_s) P(u_{\overline{s}}) P(y \mid x, z_s) \tag{A.120}$$

$$= \sum_{z_s} P(y \mid x, z_s) \sum_{u_s} \mathbb{1}(Z_s(u_{\overline{s}}) = z_s) P(u_{\overline{s}}) \tag{A.121}$$

$$= \sum_{z_s} P(y \mid x, z_s) P(z_s), \tag{A.122}$$

which completes the proof by giving an expression for $P(y \mid x^{U_s})$ based on the observational distribution $P(v)$. ■

B

Practical aspects of fairness measures

B.1 Identification of measures

The structure of the measures used in Causal Fairness Analysis was given by the Fairness Map in Thm. 4.8 (see also Fig. 4.5). Moreover, in Thm. 4.11 in Appendix A.2 we have shown that many of the measures in the map are identifiable from observational data in the standard fairness model (SFM) and we provided explicit expressions for their identification.

The natural question is whether these measures remain identifiable when some assumptions of the SFM are relaxed. To answer this question, we consider what happens to identifiability of different measures when we add bidirected edges to the $\mathcal{G}_{\text{SFM}}$.

B.1.1 Identification under Extended Fairness Model

There are five possible bidirected edges that could be added to the $\mathcal{G}_{\text{SFM}}$ (since the bidirected edge $X \dashleftarrow\dashrightarrow Z$ is assumed to be present already). The other five possibilities include the $Z \dashleftarrow\dashrightarrow Y$ (confounder-outcome), $W \dashleftarrow\dashrightarrow Y$ (mediator-outcome), $X \dashleftarrow\dashrightarrow W$ (attribute-mediator), $Z \dashleftarrow\dashrightarrow W$ (confounder-mediator) and $X \dashleftarrow\dashrightarrow Y$ (attribute-outcome). We analyze these cases in the respective order.

Bidirected edge $Z \dashleftarrow\dashrightarrow Y$. Consider the case of confounder-outcome confounding, represented by the $Z \dashleftarrow\dashrightarrow Y$ edge. An example of such a model is

	Measure	ID expression
general	$\text{TE}_{x_0,x_1}(y)$	$\sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)]P(z)$
	$\text{Exp-SE}_x(y)$	$\sum_z P(y \mid x, z)[P(z) - P(z \mid x)]$
	$\text{NDE}_{x_0,x_1}(y)$	$\sum_{z,w} [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_0, z)P(z)$
	$\text{NIE}_{x_0,x_1}(y)$	$\sum_{z,w} P(y \mid x_0, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z)$
x-specific	$\text{ETT}_{x_0,x_1}(y \mid x)$	$\sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)]P(z \mid x)$
	$\text{Ctf-SE}_{x_0,x_1}(y)$	$\sum_z P(y \mid x_0, z)[P(z \mid x_0) - P(z \mid x_1)]$
	$\text{Ctf-DE}_{x_0,x_1}(y \mid x)$	$\sum_{z,w} [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_0, z)P(z \mid x)$
	$\text{Ctf-IE}_{x_0,x_1}(y \mid x)$	$\sum_{z,w} P(y \mid x_0, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z \mid x)$
z-specific	$z\text{-TE}_{x_0,x_1}(y \mid x)$	$P(y \mid x_1, z) - P(y \mid x_0, z)$
	$z\text{-DE}_{x_0,x_1}(y \mid x)$	$\sum_w [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_0, z)$
	$z\text{-IE}_{x_0,x_1}(y \mid x)$	$\sum_w P(y \mid x_0, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]$

Table B.1: Population level and x-specific causal measures of fairness in the TV-family, and their identification expressions under the standard fairness model $\mathcal{G}_{SFM}$.

given on the r.h.s. of Table B.2. In this case, without expanding the Z set, none of the fairness measures are identifiable (due to the set Z not satisfying the back-door criterion with respect to variables X and Y). However, this does not necessarily mean there is no hope for identifying our fairness measures. What we do next is refine the Z set, in the hope that the additional assumptions obtained in this process will help us identify our quantities of interest. In some sense, the assumptions encoded in the clustered diagram are not sufficient for identification. However, spelling out the variable relations within a cluster may help with identification. Consider the example on the r.h.s. of Table B.2, where the full causal graph is given, after refining the previously clustered Z set. Interestingly, in this case the set $\{Z_1, Z_2\}$ can be shown as back-door admissible for the effect of X on Y. Furthermore, the identification expression for all the quantities remains the same as in the standard fairness model, given by the expressions in Table B.1.

Bidirected edge $\boldsymbol{W} \dashleftarrow\dashrightarrow \boldsymbol{Y}$. Next consider the case where there is a bidirected edge between the group of variables W and the outcome Y. Firstly, we note that the identification of causal (TE/ETT) and spurious measures (Exp-SE/Ctf-SE) is unaffected by the $W \dashleftarrow\dashrightarrow Y$ edge, and that these quantities are identified by the same expressions as in Table B.1. The quantities

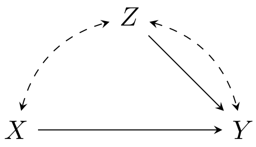
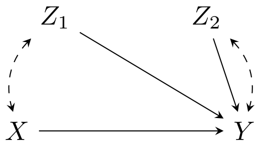
cluster model	refined model
	

Table B.2: An example of the extended fairness model with a bidirected $Z \leftarrow\!\!\dashrightarrow Y$ edge (left side), in which refining the set of variables Z yields a graph (right side) in which all fairness measures are identifiable.

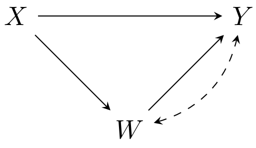
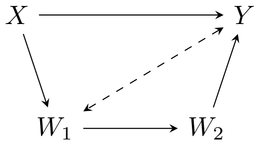
cluster model	refined model
	

Table B.3: An example of the extended fairness model with a bidirected $W \leftarrow\!\!\dashrightarrow Y$ edge (left side), in which refining the set of variables W yields a graph (right side) in which all fairness measures are identifiable.

measuring direct and indirect effects are not identifiable, at least not without further refining the W set. Consider the example given in Table B.3. In the l.h.s. of the table we have a model in which W is clustered and NDE or NIE quantities are not identifiable. On the r.h.s., after expanding the previously clustered W set, the natural direct (and indirect) effects can be identified, by the virtue of the *front-door criterion* (Pearl, 2000). However, note that in this case, the identification expression for the natural direct effect *is different from the identification expression for the natural direct effect in the standard fairness model.* Whenever front-door identification is used, we expect the expression to change, compared to the baseline SFM case.

Bidirected edge $\boldsymbol{X \leftarrow\!\!\dashrightarrow W}$. The case of the $X \leftarrow\!\!\dashrightarrow W$ edge is similar to that of $W \leftarrow\!\!\dashrightarrow Y$, yet slightly different. None of the measures discussed are identifiable in this case, before refining the W set. However, similarly as in the $W \leftarrow\!\!\dashrightarrow Y$ example in Table B.3, when refining the W set, we might find that in fact the effect of X on Y is identifiable via the front-door. Again, the identification expression in this case will change. For the sake of brevity we skip an explicit example.

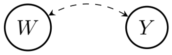
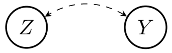
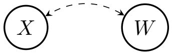
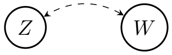
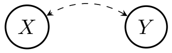
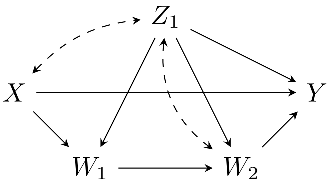
	✔	✔	Refine W	Refine W
	Refine Z	Refine Z	Refine Z	Refine Z
	Refine W	Refine W	Refine W	Refine W
	Refine Z, W	Refine Z, W	Refine Z, W	Refine Z, W
	✘	✘	✘	✘

Table B.4: Identification of causal fairness measures under latent confounding.

Bidirected edge $Z \dashleftarrow\dashrightarrow W$. In the case of the $Z \dashleftarrow\dashrightarrow W$ edge, none of the measures are identifiable. However, refining the Z and W sets may help. To see an example, consider the following graph



.

In this case, all of the measures of fairness in Table B.1 are identifiable, but again with different expressions than those presented in the table.

Bidirected edge $X \dashleftarrow\dashrightarrow Y$. The attribute-outcome confounding represented by the $X \dashleftarrow\dashrightarrow Y$ edge is the most difficult case. When this edge is present, none of the fairness quantities can be identified. The reason why this case is hard is that the $X \dashleftarrow\dashrightarrow Y$ introduces a bidirected edge between X and its child Y. This causes the effect of X on Y to be non-identifiable (Tian and Pearl, 2002). For more general identification strategies for when a combination of observational and experimental data is available, we refer the reader to (Lee *et al.*, 2019) and (Correa *et al.*, 2021a), and for partial identification ones, see (Zhang *et al.*, 2022).

The summary of the discussion of the five cases of bidirected edges in the extended fairness model, and what can be done under their presence, is given in Table B.4. We end with an example (see Fig. B.1) that fits the extended fairness model with all bidirected edges apart from the $X \dashleftarrow\dashrightarrow Y$, but in which case all the fairness measures in Table B.1 are identifiable (albeit not with the same expression as in the table), showing that refining Z and W sets

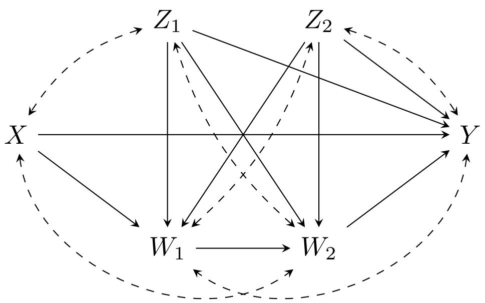


Figure B.1: Causal diagram compatible with the SFM with all bidirected arrows apart from $X \dashleftarrow\dashrightarrow Y$, in which all effects are identifiable.

sometimes may help. We leave the derivation of the identification expressions in this instance as an exercise for the curious reader.

B.2 Estimation of measures

Suppose we found that a target causal measure of fairness is identifiable from observational data (after possibly refining the SFM). The next question is then how to estimate the causal measure in practice. There is a large body of literature on the estimation of causal quantities, based on which our own implementation is built. We focus on describing how to estimate $\mathbb{E}(y_x)$ and $\mathbb{E}(y_{x_1,W_{x_0}})$. Most fairness measures can then be derived from taking (conditional) differences of these two estimands.

Doubly Robust Estimation

In the SFM, a standard way of computing the quantity $\mathbb{E}(y_x)$ would be using inverse propensity weighting. The mediator W can be marginalized out and the estimator

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}(X_i = x)Y_i}{\widehat{p}(X_i \mid Z_i)}, \tag{B.1}$$

where $\widehat{p}(X_i \mid Z_i)$ is the estimate of the conditional probability $\mathbb{P}(X_i = x \mid Z_i)$, can be used. There is an additional assumption necessary for such an approach:

Definition B.1 (Positivity assumption)**.** The positivity assumption holds if $\forall\, x, z$, $\mathbb{P}(X = x \mid Z = z)$ is bounded away from 0, that is

$$\delta < \mathbb{P}(X = x \mid Z = z) < 1 - \delta,$$

for some $\delta > 0$.

Such an assumption is needed for the estimation of causal quantities we discuss (together with the assumptions encoded in the SFM that are used for identification).

However, more powerful estimation techniques have been developed and applied very broadly. In particular, *doubly robust* estimators have been proposed for the estimation of causal quantities (Robins *et al.*, 1994; Robins and Rotnitzky, 1995; Bang and Robins, 2005). In context of the estimator in Eq. B.1, a doubly robust estimator would be

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}(X_i = x)(Y_i - \widehat{\mu}(Y_i \mid Z_i, X_i))}{\widehat{p}(X_i \mid Z_i)} + \widehat{\mu}(Y_i \mid Z_i, X_i), \tag{B.2}$$

where $\widehat{\mu}$ denotes the estimator of the conditional mean $\mathbb{E}[Y \mid Z = z, X = x]$. In fact, only one of the two estimators $\widehat{\mu}(Y_i \mid Z_i, X_i)$ and $\widehat{p}(X_i \mid Z_i)$ needs to be consistent, for the entire estimator in Eq.B.2 to be consistent. Such robustness to model misspecification is a rather desirable property.

Estimating $\mathbb{E}(y_{x_1,W_{x_0}})$ in a robust fashion is somewhat more involved. This problem has been studied under the rubric of causal mediation analysis (Robins and Greenland, 1992; Pearl, 2001; Robins, 2003). Tchetgen and Shpitser, 2012 proposed a multiply robust estimator of the expected potential outcome $\mathbb{E}[Y_{x_1,W_{x_0}}]$ defined via:

$$\begin{aligned}\phi_{x_0,x_1}(X,W,Z) =& \frac{\mathbb{1}(X = x_1)f(W \mid x_0, Z)}{p_{x_1}(Z)f(W \mid x_1, Z)}[Y - \mu(x_1, W, Z)] \\ &+ \frac{\mathbb{1}(X = x_0)}{p_{x_0}(Z)}\left[\mu(x_1, W, Z) - \int_{\mathcal{W}} \mu(x_1, w, Z)f(w \mid x_0, Z)\, dw\right] \\ &+ \int_{\mathcal{W}} \mu(x_1, w, Z)f(w \mid x_0, Z)\, dw.\end{aligned} \tag{B.3}$$

The estimator is given by $\frac{1}{n}\sum_{i=1}^{n}\widehat{\phi}_{x_0,x_1}(X_i, W_i, Z_i)$, where in $\widehat{\phi}$ the quantities $p_x(Z)$, $\mu(X, W, Z)$ and $f(W \mid X, Z)$ are replaced by respective estimates. Such an estimator is multiply robust (one of the three models can be misspecified). However, the estimator also requires the estimation of the conditional density $f(W \mid X, Z)$. In case of continuous or high-dimensional W, estimating the conditional density could be very hard and the estimator could therefore suffer in performance. We revisit the estimation of $\mathbb{E}[y_{x_1,W_{x_0}}]$ shortly.

Double Machine Learning

Doubly (and multiply) robust estimation allows for model misspecification of one of the models, while retaining consistency of the estimator. However, we have not discussed the convergence rates of these estimators yet. In some cases fast, $O(n^{-\frac{1}{2}})$ rates are attainable for doubly robust estimators, under certain conditions. For example, one such condition is that $p_x(Z), \mu(X, W, Z)$ and their estimates belong to the Donsker class of functions (Benkeser *et al.*, 2017). For a review, refer to (Kennedy, 2016). However, modern ML methods do not belong to the Donsker class.

In a recent advance, Chernozhukov *et al.*, 2018 showed that the Donsker class condition can, in many cases (including modern ML methods), be relaxed by using a cross-fitting approach. This method was named *double machine learning* (DML). For estimating $\mathbb{E}[Y_x]$ we make use of the estimator in Eq. B.2 and proceed as follows:

1. Split the data $\mathcal{D}$ into K disjoint folds $\mathcal{D}_1$, $\mathcal{D}_2$, ..., $\mathcal{D}_K$,
2. Using the complement of fold $\mathcal{D}_k$ (labeled $\mathcal{D}_k^C$) compute the estimates $\widehat{p_x}^{-(k)}(Z)$, $\widehat{\mu}^{-(k)}(X, Z)$ of $P(X = x \mid Z = z)$ and $\mathbb{E}[Y \mid Z = z]$,
3. Compute
$$\frac{\mathbb{1}(X_i = x)(Y_i - \widehat{\mu}(Y_i \mid Z_i, X_i))}{\widehat{p}(X_i \mid Z_i)} + \widehat{\mu}(Y_i \mid Z_i, X_i), \tag{B.4}$$
for each observation (X_i, Z_i, Y_i) in $\mathcal{D}_k$ by plugging in estimators $\widehat{p_x}^{-(k)}(Z)$, $\widehat{\mu}^{-(k)}(X, Z)$ obtained on the complement $\mathcal{D}_k^C$,
4. Taking the mean of the terms in Eq. B.4 across all observations.

For estimating $\mathbb{E}[y_{x_1, W_{x_0}}]$ we follow the approach of Farbmacher *et al.*, 2020. The authors propose a slightly different estimator than that based on Eq. B.3, where they replace $\phi_{x_0,x_1}(X, W, Z)$ by

$$\begin{aligned}\psi_{x_0,x_1}(X, W, Z) = &\frac{\mathbb{1}(X = x_1) p_{x_0}(Z, W)}{p_{x_1}(Z, W) p_{x_0}(Z)} [Y - \mu(x_1, W, Z)] \\ &+ \frac{\mathbb{1}(X = x_0)}{p_{x_0}(Z)} [\mu(x_1, W, Z) - \mathbb{E}[\mu(x_1, W, Z) \mid X = x_0, Z]] \\ &+ \mathbb{E}[\mu(x_1, W, Z) \mid X = x_0, Z],\end{aligned} \tag{B.5}$$

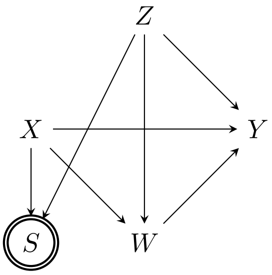
which avoids the computation of densities in a possibly high-dimensional case. The terms $\psi_{x_0,x_1}(X, W, Z)$ are estimated in a cross-fitting procedure as described above, with the slight extension in Step 2, where we further split the complement $\mathcal{D}_k^C$ into two parts, to estimate the conditional mean $\mu(X, W, Z)$ and the nested conditional mean $\mathbb{E}[\mu(x_1, W, Z) \mid X = x_0, Z]$ on disjoint subsets of the data. This approach is used in the `faircause` R-package.

C

Selection Bias Interpretation

The majority of the manuscript was concerned with the standard fairness model (SFM) from Def. 2.7. In the SFM, there is a bidirected edge $X \dashleftarrow\dashrightarrow Z$, which represents some latent (possibly historical) context which is a source of common variation between the protected attribute X and confounders Z. In particular, we now discuss the version of the SFM which considers a selection bias process based on X, Z, instead of latent confounding. In particular, consider the following definition:

Definition C.1 (SFM with Selection Bias). The standard fairness model with selection bias (SFM-SB) is the causal diagram $\mathcal{G}_{\text{SFM-SB}}$ over endogenous variables $\{X, Z, W, Y\}$ and given by



.

In the above causal model, we are considering a selection process $S(x, z)$ based on which individuals are included in the dataset. If $S(x, z) = 1$, the individual is included in our dataset, and $S(x, z) = 0$ otherwise. As there are no open

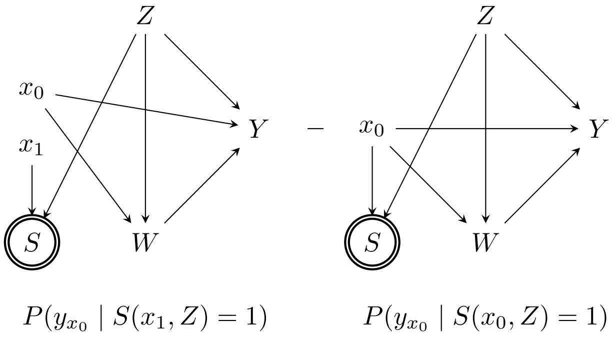

Z
x_0
x_1
Y
S
W
−
Z
x_0
Y
S
W
$P(y_{x_0} \mid S(x_1, Z) = 1)$
$P(y_{x_0} \mid S(x_0, Z) = 1)$

Figure C.1: Quantity Ctf-SBE$_{x_0,x_1}(y)$ represented graphically as a contrast.

back-door paths between X and Y, we know that the spurious effect between X and Y is 0, so we can ignore it. However, the TV measure $P(y \mid x_1) - P(y \mid x_0)$ does include variations originating from the selection process at node S. In particular, we can define an effect associated with the selection process at S:

Definition C.2 (Counterfactual Selection Bias Effect). The counterfactual selection bias effect (Ctf-SBE) is defined as:

$$\text{Ctf-SBE}_{x_0,x_1}(y) = P(y_{x_0} \mid S(x_1, Z) = 1) - P(y_{x_0} \mid S(x_0, Z) = 1). \qquad \text{(C.1)}$$

We also write S_x as an abbreviation for $S(x, Z) = 1$.

The definition is shown graphically in Fig. C.1. On the r.h.s. we have the baseline in which the variables W, Y respond to the value $X = x_0$, and the selection process on individuals at S also takes the value $X = x_0$. This setting is compared to the setting on the l.h.s., in which W, Y still respond to the value of the $X = x_0$, but the individuals are subject to the selection process of $X = x_1$. Intuitively, due to a different selection process for value x_0, x_1, the observed conditional distributions

$$Z \mid X = x_0 \text{ and } Z \mid X = x_1$$

are different, even though there are no common causes of X and Z. The contrast in Eq. C.1 and its graphical representation in Fig. C.1 capture precisely the difference in outcome Y arising from this difference in the selection processes $S(x_0, \cdot)$ and $S(x_1, \cdot)$. Importantly, the model SFM-SB allows us to decompose the total variation measure. For doing so, we need the notions of direct and indirect effects, which are defined as follows:

$$\text{Ctf-DE}_{x_0,x_1}(y \mid S_{x_0}) = P(y_{x_1,W_{x_0}} \mid S_{x_0}) - P(y_{x_0} \mid S_{x_0}) \qquad \text{(C.2)}$$

$$\text{Ctf-IE}_{x_0,x_1}(y \mid S_{x_0}) = P(y_{x_0,W_{x_1}} \mid S_{x_0}) - P(y_{x_0} \mid S_{x_0}). \qquad \text{(C.3)}$$

The notions are entirely analogous to the notions of direct and indirect effects from Def. 4.5, apart from the fact that the conditioning on $X = x$ is replaced by conditioning on the selection process S_x. Armed with such analogues of the direct and indirect effects for the SFM-SB model, we decompose the TV as follows:

Proposition C.1 (Decomposition of TV for SFM-SB)**.** The total variation measure can be decomposed into the selection bias effect, indirect effect, and direct effect as follows:

$$\mathrm{TV}_{x_0,x_1}(y) = \text{Ctf-DE}_{x_0,x_1}(y \mid S_{x_0}) - \text{Ctf-IE}_{x_1,x_0}(y \mid S_{x_0}) - \text{Ctf-SBE}_{x_1,x_0}(y) \tag{C.4}$$

$$= \text{Ctf-SBE}_{x_0,x_1}(y) - \text{Ctf-DE}_{x_1,x_0}(y \mid S_{x_1}) + \text{Ctf-IE}_{x_0,x_1}(y \mid S_{x_1}) \tag{C.5}$$

Importantly, the decomposition in Prop. C.1 can be identified from observational data in the following way:

Proposition C.2. The quantities appearing in the TV decomposition in Eq. C.4 are identifiable from observational data under selection bias, and have the following identification expressions:

$$\text{Ctf-DE}_{x_0,x_1}(y \mid S_{x_0}) = \sum_{z,w} [P^*(y \mid x_1, z, w) - P^*(y \mid x_0, z, w)] \cdot P^*(w \mid x_0, z) P^*(z \mid x_0) \tag{C.6}$$

$$\text{Ctf-IE}_{x_1,x_0}(y \mid S_{x_0}) = \sum_{z,w} P^*(y \mid x_1, z, w) \cdot [P^*(w \mid x_0, z) - P^*(w \mid x_1, z)] P^*(z \mid x_0) \tag{C.7}$$

$$\text{Ctf-SBE}_{x_1,x_0}(y) = \sum_{z} P^*(y \mid x_1, z)[P^*(z \mid x_0) - P^*(z \mid x_1)], \tag{C.8}$$

where P^* is the observational distribution under selection bias, defined by

$$P^*(v) = P(v \mid S = 1). \tag{C.9}$$

Proof. We prove the identification expression for the Ctf-SBE term, and the other two expressions follow from a similar argument. Note that:

$$P(y_{x_1} \mid S_x = 1) = \sum_{z} P(y_{x_1} \mid z, S_x = 1) P(z \mid S_x = 1). \tag{C.10}$$

The first term within the sum can be expanded as:

$$\begin{aligned} P(y_{x_1} \mid z, S_x = 1) &= P(y_{x_1} \mid z, x, S_x = 1) \quad Y_{x_1} \perp\!\!\!\perp X \mid Z, S_x && \text{(C.11)} \\ &= P(y_{x_1} \mid z, x, S = 1) \quad \text{Consistency Axiom} && \text{(C.12)} \\ &= P(y_{x_1} \mid z, x_1, S = 1) \quad Y_{x_1} \perp\!\!\!\perp X \mid Z, S && \text{(C.13)} \\ &= P(y \mid z, x_1, S = 1) \quad \text{Consistency Axiom} && \text{(C.14)} \\ &= P^*(y \mid z, x_1) \quad \text{by definition.} && \text{(C.15)} \end{aligned}$$

For the second term within the sum, we have that

$$\begin{aligned} P(z \mid S_x = 1) &= P(z \mid S_x = 1, x) \quad Z \perp\!\!\!\perp X \mid S_x && \text{(C.16)} \\ &= P(z \mid S = 1, x) \quad \text{Consistency Axiom} && \text{(C.17)} \\ &= P^*(z \mid x) \quad \text{by definition.} && \text{(C.18)} \end{aligned}$$

Putting together with the first term, the derivation yields the identification expression in Eq. C.8. ■

The crucial takeaway from the above proposition is that the identification expressions we obtain are identical to those obtained when decomposing the TV based on the SFM. In particular, this implies that *even if we work with the SFM, but SFM-SB is the true underlying model, the decomposition we obtain is valid*, but has a slightly different interpretation. This result can be seen formally in the following corollary:

Corollary C.1 (SFM and SFM-SB decomposition ID equivalence)**.** Let $\mathcal{M}_1$ be an SCM compatible with the SFM, and let $P_1(V)$ denote its observational distribution. Let $\mathcal{M}_2$ be an SCM compatible with the SFM-SB, and let $P_2(V)$ denote its observational distribution. Suppose moreover that

$$P_1(V) = P_2(V) = P(V), \tag{C.19}$$

that is, the observational distributions of $\mathcal{M}_1$ and $\mathcal{M}_2$ are the same. Then it follows that

$$\begin{aligned} \text{Ctf-DE}^{\mathcal{M}_1}_{x_0,x_1}(y \mid x_0) &= \text{Ctf-DE}^{\mathcal{M}_2}_{x_0,x_1}(y \mid S_{x_0}) && \text{(C.20)} \\ \text{Ctf-IE}^{\mathcal{M}_1}_{x_1,x_0}(y \mid x_0) &= \text{Ctf-IE}^{\mathcal{M}_2}_{x_1,x_0}(y \mid S_{x_0}) && \text{(C.21)} \\ \text{Ctf-SE}^{\mathcal{M}_1}_{x_1,x_0}(y) &= \text{Ctf-SBE}^{\mathcal{M}_2}_{x_1,x_0}(y), && \text{(C.22)} \end{aligned}$$

that is, the decomposition of the TV measure for the two SCMs has the same terms.

Proof. We leverage the identification expressions from Prop. C.2 and check they are equal to the identification expressions for Ctf-SE, Ctf-DE, and Ctf-IE shown in Tab. B.1. ■

In words, the terms appearing in the TV decomposition of the SFM are the same as the terms appearing in the TV decomposition when using the SFM-SB, if two SCMs have the same observational distribution. What this shows is that we are agnostic to the choice of the model, between the SFM and SFM-SB, when decomposing the TV - the only difference in the decomposition arises in the *interpretation of the effects*. In particular, the if the SFM model is the true model, then the Ctf-SE$_{x_1,x_0}(y)$ measures the change in outcome between conditioning on $X = x_0$ and $X = x_1$, while keeping $X = x_1$ along all causal pathways. If the SFM-SB model is the true model, then the Ctf-SBE$_{x_1,x_0}(y)$ measures the change in outcome induced by the selection process S_{x_0} compared to S_{x_1}, while keeping $X = x_1$ along all causal pathways. The qualitative interpretation of the two terms differs, but the quantitative value is the same regardless of the model. This shows a fundamental analogy between the bidirected arrow $X \dashleftarrow\dashrightarrow Z$ in the SFM and the selection process at the node S governed by X, Z in the SFM-SB.

D

Multi-valued and Continuous Protected Attributes

In this appendix, we discuss how to extend the main results of the manuscript to a setting with multi-valued or continuous protected attributes. We also quickly discuss how we may address the setting with multiple protected attributes.

Throughout, let $\mathcal{X}$ denote the domain of the protected attribute X. In the multi-valued, discrete case, we consider $|\mathcal{X}|$ to be an integer, whereas for X continuous, we assume that $\mathcal{X}$ in a subset of the reals, $\mathcal{X} \subseteq \mathbb{R}$. We next explain how some of the key results may be extended to the case of a multi-valued X.

(1) The definition of the total variation (TV) measure is updated, and the new criterion we consider is

$$\mathbb{E}[Y \mid X = x] = \mathbb{E}[Y] \, \forall x. \tag{D.1}$$

Suppose we select a fixed baseline value of X, say $x_0 \in \mathcal{X}$. Then, we could consider a collection of measures $\mathbb{E}[Y \mid X = x] - \mathbb{E}[Y \mid X = x_0]$, for each $x \in \mathcal{X}$. Alternatively, a single measure over the entire domain could be considered, e.g.,

$$\text{iTV}_{x_0,X}(y) = \mathbb{E}_{X \sim P(X)} \left[\mathbb{E}[Y \mid X] - \mathbb{E}[Y \mid X = x_0] \right], \tag{D.2}$$

where iTV stands for integrated TV measure.

(2) Notions of direct, indirect, and spurious effects also need to be updated accordingly. For instance, given a baseline value of $X = x_0$, we may

consider the following measures of the direct, indirect, and spurious effects

$$\text{NDE}_{x_0,x}(y) = P(y_{W_{x_0},x}) - P(y_{x_0}) \tag{D.3}$$

$$\text{NIE}_{x_0,x}(y) = P(y_{W_x,x_0}) - P(y_{x_0}) \tag{D.4}$$

$$\text{Exp-SE}_x(y) = P(y \mid x) - P(y_x), \tag{D.5}$$

and further analogues can be written for x, z, or v'-specific measures of direct / indirect effects. In case a single measure[1] is of interest instead of a collection of measures, we may consider measures such as

$$\text{iNDE}_{x_0,X}(y) = \mathbb{E}_{X\sim P(X)}[\text{NDE}_{x_0,X}(y)] \tag{D.6}$$

that integrates the NDE value over the entire domain of X.

(3) The Fundamental Problem of Causal Fairness Analysis (FPCFA, Def. 3.6) requires a decomposability property. If one considers measures such as $\text{NDE}_{x_0,x}(y)$ for each x separately, then the property of decomposability will be satisfied for each value of x separately. For the integrated measures, iTV measure can be decomposed as

$$\text{iTV}_{x_0,X}(y) = \text{iNDE}_{x_0,X}(y) - \text{iNIE}_{X,x_0}(y) \tag{D.7}$$

$$+ \text{iExp-SE}_X(y) - \text{Exp-SE}_{x_0}(y). \tag{D.8}$$

Other decomposition results, such as in Thms. 4.3, 4.4, and 4.5 can be adapted similarly. Further, the integrated measures are still admissible to the structural measures, i.e.,

$$\text{Str-DE} = 0 \implies \text{iNDE}_{x_0,X}(y) = 0 \tag{D.9}$$

$$\text{Str-IE} = 0 \implies \text{iNIE}_{x_0,X}(y) = 0 \tag{D.10}$$

$$\text{Str-SE} = 0 \implies \text{iExp-SE}_X(y) = 0. \tag{D.11}$$

For each $x \in \mathcal{X}$, the NDE, NIE, and Exp-SE measures are also admissible with respect to structural criteria.

(4) The Fairness Map (Thm. 4.8, Fig. 4.5) was defined as having two separate axes, corresponding to different units of the population, and different mechanisms. In the continuous case, there is an additional, third axis, which indicates which value of $x \in \mathcal{X}$ is being compared against the baseline value $X = x_0$.

[1] One may also attempt to detect discrimination by using measures such as $\sup_{x\in\mathcal{X}} |\text{NDE}_{x,x_0}(y)|$ which would also be a valid choice, but the property of decomposability as in Eq. D.7 would not hold true.

(5) The decomposition of the predictive parity measure (PPM) from Thm. 4.13 can still be applied, but now again there is a unique measure for each $x \in \mathcal{X}$, $\text{PPM}_{x_0,x}(y) = P(y \mid x, \widehat{y}) - P(y \mid x_0, \widehat{y})$. Furthermore, the principles of Causal Predictive Parity (Def. 4.14) can also be extended to the continuous case, by adding a quantifier $\forall x \in \mathcal{X}$, e.g., causal predictive parity along the direct pathway could be written as

$$\mathbb{E}[y_{x,W_{x_0}} \mid E] - \mathbb{E}[y_{x_0} \mid E] = \mathbb{E}[\widehat{y}_{x,W_{x_0}} \mid E] - \mathbb{E}[\widehat{y}_{x_0} \mid E] \; \forall x \in \mathcal{X}, E. \tag{D.12}$$

(6) In the context of decision-making, the Benefit Fairness criterion (Def. 5.10) can be adapted to require that

$$P(d \mid x, \Delta = \delta) = P(d \mid x, \Delta = \delta) \; \forall x \in \mathcal{X}, \delta. \tag{D.13}$$

The definition of Causal Benefit Fairness (Def. 5.11) could be adapted to the continuous case by adding a quantifier over $x \in \mathcal{X}$, for instance, Causal BF along the direct pathway would be defined as

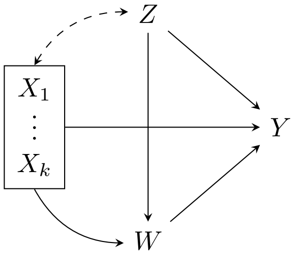
$$\mathbb{E}(y_{x,W_{x_0},d_1} - y_{x,W_{x_0},d_0} \mid x, z, w) = \mathbb{E}(y_{x_0,d_1} - y_{x_0,d_0} \mid x, z, w) \; \forall x, z, w \tag{D.14}$$

$$P(d \mid \Delta, x_0) = P(d \mid \Delta, x_1) \; \forall x, \delta. \tag{D.15}$$

As the above reasoning shows, extending the results of the manuscript to multi-valued and continuous protected attributes X would be conceptually possible. However, we note that continuous protected attributes may complicate the estimation of some of the quantities described above, and we do not consider these challenges in this manuscript.

Multiple Protected Attributes. Finally, we mention how one may wish to handle multiple protected attributes $X_1, \ldots, X_k$. Firstly, we will only consider the case in which the attributes $X_1, \ldots, X_k$ satisfy the assumptions of the standard fairness model (SFM), defined as follows:

Definition D.1 (Multi-Attribute Standard Fairness Model)**.** The multi-attribute standard fairness model (MA-SFM) is the cluster causal diagram $\mathcal{G}_{\text{SFM}}$ over endogenous variables $\{X_1, \ldots, X_k, Z, W, Y\}$ and given by



.

The cluster $\{X_1, \ldots, X_k\}$ allows for arbitrary causal or confounding relationships between the variables $X_1, \ldots, X_k$.

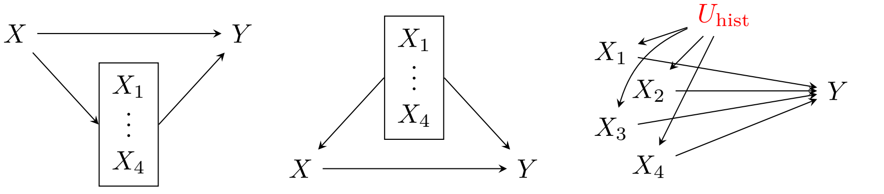
Now, if we are dealing with a setting of multiple protected attributes that satisfy the MA-SFM model, we proceed as follows. Let $\mathcal{X}_1, \ldots, \mathcal{X}_k$ be the domains of $X_1, \ldots, X_k$, respectively. Then, we define the product protected attribute as $X^p = (X_1, \ldots, X_k)$ taking values in $\mathcal{X}^p = \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$. Then, based on the product attribute X^p and the values it takes, we reduce the problem to a setting with a single multi-valued (or continuous) protected attribute that can be handled as discussed above.

In general, the protected attributes $X_1, \ldots, X_k$ may not necessarily satisfy the assumptions of the MA-SFM. If this is the case, a suggested route for considering fairness with respect to $X_1, \ldots, X_k$ would be to consider $X_1, \ldots, X_k$ one-by-one, and perform the analyses described in the manuscript for a single $X = X_i$ at a time.

D.1 On the Semantics of Manipulating the Protected Attribute

In this section, we discuss various questions related to the meaning of manipulating the protected attribute X. In particular, commonly considered protected attributes such as race, gender, or religion are not subject to a real-world "intervention" of setting the attribute to a fixed value. In other words, we cannot simply design an experiment in which we randomize the allocation of individuals to males and females, or to majority and minority group applicants. Furthermore, some works have argued that the meaning of the counterfactual Y_x may not be well-defined (Hu and Kohler-Hausmann, 2020), with some even arguing that counterfactual reasoning may be inappropriate for capturing discrimination (Kohler-Hausmann, 2018; Dembroff and Kohler-Hausmann, 2022). All of these works seek more precision in the semantics around the concept of "manipulating race", which is certainly a worthwhile question to ask. More broadly, in the causal inference literature, many have argued for the mantra "no causation without manipulation" (Rubin, 1986; Hernán, 2005; Gelman and Hill, 2006), and here we wish to alleviate most of these concerns, by discussing the semantics of manipulating attribute such as race, gender, or religion.

In our discussion, we focus on the arguments put forth by Hu and Kohler-Hausmann, 2020, as these arguments are articulated in the language of graphical causal models. We analyze a number of claims made by the authors, and propose specific tools for addressing their concerns. Crucially, we phrase some of the elusive philosophical concepts in a formal mathematical language, thereby adding to the existing discussion about the validity of hypothetical manipulations of the protected attribute. In particular, we address the following three arguments of Hu and Kohler-Hausmann, 2020:



(a) Mediator model. **(b)** Confounder model. **(c)** Historical process confounding.

Figure D.1: Modeling options for religion as a bundle of sticks.

(A) Protected attributes are a *"bundle of sticks"* (Sen and Wasow, 2016), formed from multiple constitutive, and not defining features,

(B) The effects of interventions on attributes such as sex, race, and religion thus cannot be reasoned about in the framework of structural causality and graphical causal models, since such effects are not well-defined,

(C) Explanations originating from counterfactual worlds where the protected attribute is manipulated are not meaningful for explaining discrimination in the current world.

D.1.1 Issue A: Attributes as a bundle of sticks.

The example put forward by Hu and Kohler-Hausmann, 2020 takes religion as the protected attribute, with $X \in \{0, 1\}$ representing whether an individual is or is not Catholic. A number of constitutive features of X are then mentioned, namely the following beliefs and practices: Resurrection of Christ (X_1), Papal Infallibility (X_2), Saints (X_3), and Sunday Mass (X_4), to name a few. The authors then argue that, for a given outcome Y, one of the two causal models is possible, shown in Figs. D.1a, D.1b. Their conclusion is that either (i) $X_1, \dots, X_4$ are causal descendants of X as in Fig. D.1a; or (ii) $X_1, \dots, X_4$ causally precede X as in Fig. D.1b. The very concept of Catholic surely depends on all of the mentioned constitutive features, and hence (Hu and Kohler-Hausmann, 2020) conclude that the setting (i) seems unlikely. Similarly, one may notice that reasoning about the concept of Catholic itself seems to be meaningless without $X_1, \dots, X_4$. Therefore, the Fig. D.1b also seems inappropriate. From this, the authors conclude that causal diagrams may be insufficient for representing concepts that are formed from constitutive features, such as religion, race, or gender.

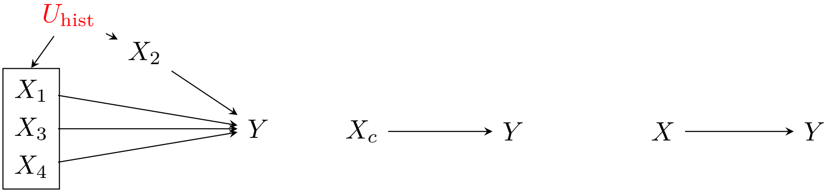
However, not all modeling options are exhausted after considering diagrams in Fig. D.1a and D.1b. In fact, the standard fairness model (SFM) introduced in Def. 2.7 was partially motivated by such ambiguities in specifying diagrams in the context of fairness analysis – and in particular, there is a *bidirected* *arrow* $X \dashleftarrow\dashrightarrow Z$ between the protected attribute X and the set of confounders Z. The reason for this modeling choice is that one may not be able to commit to the complex historical processes that introduce co-variations between the protected attribute, and the usually observed demographics. Importantly, the very same modeling choice can be used for the bundle of sticks representation of religion – clearly, belief in the Resurrection of Christ and Papal Infallibility are correlated, yet there is no clear causal relation between them. Instead, we may say that a set of historical process and practices confounds these two variables, indicated by the latent, unobserved U_{hist} in Fig. D.1c. In the analysis of the second issue, we discuss how the causal diagram in Fig. D.1c can be used for a meaningful analysis.

D.1.2 Issue B: Effects of interventions on race, sex, or religion are not well-defined through structural causality.

Hu and Kohler-Hausmann, 2020 argue that, partly for reasons outlined above, one cannot reason about the causal effects of attributes such as race, sex, or religion. Even though the question of manipulating protected attributes is subtle, and clarity on the semantics of such manipulations is a worthy endeavor, we disagree with the conclusions of (Hu and Kohler-Hausmann, 2020). We next discuss a number of methodological options that ground the semantics of such manipulations, and allow one to reason about fairness through structural causality.

In particular, we cover three different approaches for defining how the manipulations of the protected attribute can be defined in light of considering constitutive features. The described approach is related to the reasoning presented in (Weinberger, 2022), based on the notion of *signal manipulation*. The approaches we discuss are twofold, based on whether the constitutive features of the protected attribute (features $X_1, \ldots, X_4$ in our running example) are observed and available in the data. We thus discuss an approach for the case of observed features, and an interpretation for the case of unobserved features.

Observed Constitutive Features and Multi-valued Attributes. Consider now the case of the causal diagram in Fig. D.1c, with $X_1, \ldots, X_4$, and Y observed. The first modeling step required is to draw a boundary that determines what are the constitutive features of the protected attribute. For instance, should the protected attributes be constituted from all of the features $X_1, \ldots, X_4$? Or, alternatively, should one choose only a subset of them as constitutive of the protected group? For instance, one may consider X_1, X_3, X_4 only as constitutive of the protected group. The choice of constitutive features may be application-specific, and should be performed by the data analyst,



(a) Cluster model. **(b)** Continuous attribute. **(c)** Binary attribute.

Figure D.2: Modeling options for religion as a bundle of sticks.

while also taking into account domain knowledge. Once the constitutive features have been grouped into a cluster[2], the remaining features become a confounder, as displayed in Fig. D.2a. Once the cluster diagram after grouping the variables has been established, there are two ways we can proceed, which are discussed next.

The first option is to treat all of the constitutive features separately. Consider the multi-valued vectors that represent all the possible combinations of (X_1, X_3, X_4). There are 2^3 possible values that are attained, and we set the value $(0, 0, 0)$ as the baseline value (corresponding to an individual for whom no constitutive characteristics are present). Then, we can compare each vector (x_1, x_3, x_4) against $(0, 0, 0)$, and measure the effect of *manipulating the* $x_i \neq 0$ *to* 0[3]. Interesting structure may be uncovered in this way, namely, perhaps $\mathbb{E}[Y_{(0,0,1)} - Y_{(0,0,0)}]$ is much larger (in absolute value) than $\mathbb{E}[Y_{(1,0,0)} - Y_{(0,0,0)}]$, possibly implying that the feature x_1 plays a more important role in explaining the phenomenon than the feature x_4. In fact, this argument can be made formal under the assumption of no interactions in the f_Y mechanism but we do not go into its detail here.

Another option would be construct a mapping $f_X : (X_1, X_3, X_4) \mapsto X$, that assigns a value to the entire cluster. One such possible function is just setting $X = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} X_i$, where $\mathcal{I}$ is the index set of all constitutive features. Naturally, other possible mappings exist, and the mapping could also be stochastic. Once a cluster value X has been defined, we can again use the methods proposed for multi-valued attributes in Appendix D, and compare different values of $X = x$ against the baseline $X = 0$. In the extreme case, the mapping f_X may create a binary label for X. We next explain why this simplification step can still be meaningful.

[2] This clustering process is similar to the clustering of Z or W variables when constructing the Standard Fairness Model (Def. 2.7). For more details, we refer the reader to (Anand *et al.*, 2021).

[3] For instance, such manipulations can be conceptualized as a person "writing a different value on their application".

Unobserved Constitutive Features and Soft Interventions. Consider now the case of the causal diagram in Fig. D.2a, with $X_1, \ldots, X_4$, not observed, but instead we are given an imperfect value of the cluster, labeled X. That is, we are only given the output of the $f_X : (X_1, X_3, X_4) \mapsto X$ mapping described in the previous paragraph. The key question we answer next is the following: If we hypothesize interventions on the variable X, do such operations have a valid syntactic interpretation?

To give a positive answer to this question, we describe an interpretation via soft-interventions (Correa and Bareinboim, 2020). Soft interventions are an extension of atomic interventions, which were considered throughout this manuscript. Atomic interventions set X to a specific, fixed value, say $X = x_0$. Soft interventions, on the other hand, may set the value of X to a *policy*, e.g., we may consider a policy intervention that sets the value of X to x_0 with probability 0.6, whereas it sets it to x_1 with probability 0.4.

We continue illustrating our point by means example. Consider a hypothetical setting in which we have a continuous variable $X_c \sim \text{Unif}[0, 1]$ that represents the protected attribute, and the true causal diagram is given in Fig. D.2b. The variable is chosen as continuous to indicate a possible complexity in determining the protected attribute (as described in previous paragraphs). Instead of having access to X_c, we only have access to an imperfect version of it, say $X \in \{0, 1\}$, and we posit the diagram in Fig. D.2c. For simplicity, suppose that $X = \mathbb{1}(X_c \geq \frac{1}{2})$ but we are not given this information.

A possible issue may lie in the fact that the mechanism f_Y in fact responds to X_c, while we are trying to conceptualize interventions on X, and the f_Y mechanism responds to X_c, and not its abstraction X. However, as it turns out, an atomic intervention in the model in Fig. D.2c corresponds to a soft-intervention in the model in Fig. D.2b. In particular, in this case, we may write

$$
\begin{aligned}
P(Y_{X=x_0} = 1) &= P(Y = 1 \mid X = x_0) && \text{(D.16)}\\
&= \int_{[0,\frac{1}{2}]} P(Y = 1 \mid X_c = x_c, X = x_0) f_{X_c \mid X = x_0}(x_c) dx_c && \text{(D.17)}\\
&= \int_{[0,\frac{1}{2}]} 2P(Y = 1 \mid X_c = x_c, X = x_0) dx_c && \text{(D.18)}\\
&= P(Y = 1 \mid X_c \sim \text{Unif}[0, \frac{1}{2}]) && \text{(D.19)}\\
&= P(Y = 1 \mid do(X_c \sim \text{Unif}[0, \frac{1}{2}])) && \text{(D.20)}\\
&= P(Y = 1; \sigma_{X_c}) && \text{(D.21)}
\end{aligned}
$$

where σ_{X_c} indicates a policy intervention that sets X_c uniformly to the $[0, \frac{1}{2}]$ interval. Through this analysis, the meaning of, say, the total effect of X on

Y, written $P(y_{x_1}) - P(y_{x_0})$, becomes more apparent:

$$\text{TE}_{x_0,x_1}(y) = P(y \mid do(X_c \sim \text{Unif}[\frac{1}{2}, 1])) - P(y \mid do(X_c \sim \text{Unif}[0, \frac{1}{2}])). \tag{D.22}$$

That is, the total effect compares the outcome of a policy that sets X_c uniformly to $[0, \frac{1}{2}]$, against a policy that sets X_c uniformly to $[\frac{1}{2}, 1]$, given a clear semantical interpretation to the quantity $\text{TE}_{x_0,x_1}(y)$ in terms of the true underlying, though unobserved, quality X_c.

In fact, this construction generalizes to arbitrary mappings satisfying minor assumptions. Suppose that $X_c \sim F_{X_c}$ according to some probability distribution F_{X_c} that admits a density. Then, suppose that $f_X : X_c \mapsto X$ is an arbitrary mapping from the domain of X_c into $\{0, 1\}$. We can then write

$$\begin{aligned}
P(Y_{X=x_0} = 1) &= P(Y = 1 \mid X = x_0) && \text{(D.23)}\\
&= \int_{f_X^{-1}(x_0)} P(Y = 1 \mid X = x_0, X_c = x_c) f_{X_c|X=x_0}(x_c) dx_c && \text{(D.24)}\\
&= \int_{f_X^{-1}(x_0)} P(Y = 1 \mid X_c = x_c) f_{X_c|X=x_0}(x_c) dx_c && \text{(D.25)}\\
&= P(Y = 1 \mid X_c \sim F_{X_c|X=x_0}) && \text{(D.26)}\\
&= P(Y = 1 \mid X_c \sim do(F_{X_c|X=x_0})) && \text{(D.27)}\\
&= P(Y = 1; \sigma_{X_c}), && \text{(D.28)}
\end{aligned}$$

where σ_{X_c} now indicates a stochastic intervention that sets X_c to its conditional distribution given $X = x_0$. In other words, the interpretation given to the total effect in our first example with a uniformly distribution and a threshold mapping was not an idiosyncrasy. Instead, it follows from a more general approach in Eqs. D.23-D.28.

We now recap the importance of the above result. Crucially, in the real world, the f_Y mechanism responds to a continuous random variable X_c. The mechanism is unaware of the value of the "binarized" attribute X, and does not respond to it. Nonetheless, in a simplified causal diagram with X taken as the treatment instead of X_c, the total effect still has a meaningful interpretation with respect to the underlying true structural causal model, in which f_Y responds to changes in X_c, and not to X.

D.1.3 Issue C: Counterfactual Worlds Do Not Explain Social Phenomena in the Current World

The final point we address concerns the validity of counterfactual causal reasoning for explaining discrimination in the current real world. Here, we leverage the Berkeley admissions example introduced in Ex. 2.1. As a quick recap, the protected attribute X represents gender, a mediator D represents the choice of department to which the student applies, and Y represents the admission outcome. In particular, Hu and Kohler-Hausmann, 2020 write: "Modular counterfactuals of the type, 'What would the effect of sex on admissions be in a world when men and women apply at the same rates to math departments?' – do not necessarily tell us anything empirically relevant to the normative question about whether a current practice is discriminatory in our current world where those premises are counter to fact".

Some clarification is in order regarding what causal modeling is attempting to answer in such instances. The dataset under analysis was generated from a specific structural causal model that represents the decision-making mechanism that was used by the university's committee, labeled f_Y. One can perform a thought experiment, in which the committee spends infinite time deliberating admissions, and produces an output decision for any input and possible value of the noise variables. Any causal analysis undertaken is strictly concerned with this generative model of reality, and does not attempt to answer anything about how the committee would have acted *on a different occasion*, on which the correlation between department of application and gender vanished. Instead, the type of question we are asking is, for the committee *fixed in time and place*, how would they have evaluated students had they been given applications of students in which, for instance, the gender was randomized? That is, causal modeling is relative to the underlying model of reality, and does not purport to answer questions on how downstream mechanisms (evaluation of applications) would change over time had an upstream mechanism (choosing department of application) been affected.

We address one final point of Hu and Kohler-Hausmann, 2020. The authors write that "more people sexcoded 'male' than 'female' apply to math departments and that means, cognitively, that decision-makers associate male and math more than they associate female and math. That is, after all, the problem. It is not clear why knowing how people sexcoded 'female' would be treated in a counterfactual world where equal numbers of people sexed female and male applied to math departments is helpful for sorting out whether in our world, where math is a male-y thing, the current admission practices constitute discrimination". Some key methodological developments in causal inference are entirely ignored in the considerations of authors, similarly as in (Kohler-Hausmann, 2018). In fact, as we discuss next, causal methodology allows us to: (i) determine whether math is seen as a male-y thing by the committee, or if females are treated unfairly for other reasons; (ii) quantify the contribution of math being a male-y thing compared to other forms of discrimination.

The issue at hand hand, best illustrated through an example, has to do with *interactions* among variables. Consider the following example:

Example D.1 (Berkeley Admissions – continued). Consider the Berkeley admissions setting from Ex. 2.1. Let X be gender (x_0 female, x_1 male), D department choice (d_0 non-math, d_1 math), and Y admission outcome (y_1 for admission). Consider the following SCM:

$$X \leftarrow \text{Bernoulli}(0.5) \tag{D.29}$$

$$D \leftarrow \text{Bernoulli}(0.5 + \alpha X) \tag{D.30}$$

$$Y \leftarrow (0.1 + \beta X + \gamma D + \delta X D). \tag{D.31}$$

Now, notice that there is an interaction term in the f_Y mechanism, namely δXD. Due to this term, the probability of admission increases for individuals *who are male, and apply to the math department*. This term, therefore, in words of Hu and Kohler-Hausmann, 2020 measures how much math is male-y thing, as perceived by the committee. The other part of this story about how much math is male-y thing is the difference in the rate of application to math departments, given by the parameter α.

Importantly, other forms of discrimination also exist. For instance, if $\beta > 0$, male applicants are given advantage over female candidates, in way that has nothing to do with math being a male-y thing.

A technical question, in this scenario, is the following. Can we test for the existence of the interaction term? And secondly, if the interaction term exists, can we obtain a quantity that captures it? To answer affirmatively to both questions, we first compute the NDE for both $x_0 \to x_1$ and $x_1 \to x_0$ transitions:

$$\text{NDE}_{x_0,x_1}(y) = \beta + \frac{\delta}{2} \tag{D.32}$$

$$\text{NDE}_{x_1,x_0}(y) = \beta + \frac{\delta}{2} + \alpha\delta. \tag{D.33}$$

Notice that if either $\alpha = 0$, or $\delta = 0$, the two NDEs are the same. In fact, a hypothesis test

$$H_0 : \text{NDE}_{x_0,x_1}(y) = \text{NDE}_{x_1,x_0}(y) \tag{D.34}$$

is a test for the existence of an interaction between direct and indirect pathways. In fact, the difference between the two NDEs

$$\text{NDE}_{x_1,x_0}(y) - \text{NDE}_{x_0,x_1}(y) = \alpha\delta \tag{D.35}$$

quantifies the strength of the interaction of direct and indirect pathways, e.g., the impact of the entire phenomenon of math being a male-y thing (males are more likely to apply to math departments, in conjunction with the committee perceiving males as more qualified) on the disparity observed in outcome. □

The discussion of the above example does not only address the simple parametric instance in Eqs. D.29-D.31, but can also be generalized to more complex settings and interactions, that is, to arbitrary SCM mechanisms. Therefore, when diagnosing issues with the causal methodology for detecting discrimination, one also needs to carefully consider the methodological capabilities at hand to the data analyst.

E

Process Fairness

In this appendix, we discuss the connection of causal fairness analysis with the notion of process fairness (Grgic-Hlaca *et al.*, 2016). Process fairness offers a different normative view on fairness when compared to the legal doctrines of disparate treatment and disparate impact, around which most of the discussion in this manuscript revolved. The discussion in this appendix builds on the tools developed in Sec. 3 and Sec. 4.

The disparate treatment and impact doctrines are usually discussed in the context of outcome fairness, focusing on disparities in the outcome itself. Complementary to this, the notion of process fairness is focused on how decisions come about, and, in particular, which variables are used in the decision-making process. In this context, the causal approach to fairness discussed earlier also plays an important role. The crucial point is that considerations about outcome fairness, when paired with appropriate causal assumptions, may also give insights about process fairness. We formalize this statement in the sequel.

The disparate treatment doctrine is concerned with differential outcomes for similarly situated individuals who differ in the protected characteristic. If $Z = z, W = w$ denote the values of the confounders and mediators, respectively, such as disparity can be written as

$$P(y \mid x_1, z, w) - P(y \mid x_0, z, w) \neq 0. \tag{E.1}$$

However, a statistical claim, such as in Eq. E.1, in itself does not make any claims about the decision-making process, unless paired with causal assumptions. To produce a causal claim, we can consider the quantity

$$(x, z, w)\text{-DE}_{x_0,x_1}(y \mid x_0, z, w) = P(y_{x_1,W_{x_0}} \mid x_0, z, w) - P(y_{x_0} \mid x_0, z, w), \tag{E.2}$$

which measures the direct effect of a $x_0 \to x_1$ transition for the group of units with covariate values x_0, z, w (see Sec. 4 for details). Crucially, this quantity may have causal implications since it is admissible (Def. 3.4) with respect to the *structural direct effect* (Def. 3.2). This implies that

$$(x, z, w)\text{-DE}_{x_0,x_1}(y \mid x_0, z, w) \neq 0 \implies \text{Str-DE} \neq 0. \tag{E.3}$$

In words, if the causal quantity is different from 0, then the protected attribute X is known to be used as an input to the decision-making mechanism f_Y that determines the values of the outcome. Put differently, this allows one to establish a qualitative claim about the *process itself*, as discussed in (Grgic-Hlaca *et al.*, 2016). Now, the key piece of the puzzle is how to move from the statistical claim in Eq. E.1 to a counterfactual claim about $(x, z, w)\text{-DE}_{x_0,x_1}(y \mid x_0, z, w)$. As it turns out, the latter quantity is *identifiable* under the SFM, and in fact equals exactly the expression in Eq. E.1. The main point here is that, in absence of appropriate causal assumptions, the quantity $(x, z, w)\text{-DE}_{x_0,x_1}(y \mid x_0, z, w)$ need not equal the expression in Eq. E.1, and observing a disparity in outcome does not imply anything about the process of decision-making in general. However, based on this disparity, one may be able to produce claims about the process of decision-making with the help of appropriate causal assumptions.

A similar line of reasoning, although somewhat more involved, applies for the doctrine of disparate impact, and the indirect and spurious effects. For instance, based on the admissibility of measures such as natural indirect effect (Def. 4.2) and experimental spurious effect (Def. 4.1) with respect to structural indirect and spurious effects, respectively, we know that

$$\text{NIE}_{x_0,x_1}(y) \neq 0 \implies \text{Str-IE} \neq 0, \tag{E.4}$$

$$\text{Exp-SE}_x(y) \neq 0 \implies \text{Str-SE} \neq 0. \tag{E.5}$$

Once again, this allows one to make qualitative claims about the decision-making process (in particular, Str-IE $\neq 0$ implies mediators are used as an input to the mechanism f_Y, and that the mediators are affected by the protected attribute X; Str-SE $\neq 0$ implies that confounders are used as an input to f_Y, and that there are common variations of the confounders and the attribute X).

Finally, we mention another fundamental connection of process and outcome fairness that follows from the causal approach. Based on the decomposition of the TV measure in Thm. 4.3, we have that

$$\text{TV}_{x_0,x_1}(y) = x\text{-DE}_{x_0,x_1}(y \mid x_0) - x\text{-IE}_{x_1,x_0}(y \mid x_0) - x\text{-SE}_{x_1,x_0}(y). \tag{E.6}$$

The TV measure captures the entire observed disparity, related to outcome fairness. However, each of the terms on the r.h.s. of Eq. E.6 is related to a specific part of the decision-making process – whether the attribute is used directly (term x-DE); whether the attribute influences the mediators, which are then used in decision-making (term x-IE); and whether the attribute has common variations with the confounders, which are used in decision-making (term x-SE). Crucially, once we compute each of the terms on the r.h.s. of Eq. E.6, it allows us to quantify how much *each part of the decision process* contributes to the overall *disparity in the outcome* that was observed in an aggregate measure such as TV. Therefore, the causal analysis allows the data scientist to attribute outcome disparities found in the data to the causal mechanisms that generate them, and therefore permit simultaneous reasoning about both disparities in outcome and how they came about – thereby considering outcome and process fairness within a unified framework.

References

Act, C. R. (1964). “Civil rights act of 1964”. *Title VII, Equal Employment Opportunities.*

Agarwal, A., A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. (2018). “A reductions approach to fair classification”. In: *International Conference on Machine Learning.* PMLR. 60–69.

Anand, T., A. Ribeiro, J. Tian, and E. Bareinboim. (2021). “Effect Identification in Causal Diagrams with Clustered Variables”.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. (2016). “Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.” *ProPublica.* May. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Avin, C., I. Shpitser, and J. Pearl. (2005). “Identifiability of path-specific effects”. In: *Proceedings of the 19th international joint conference on Artificial intelligence (IJCAI’05).* 357–363.

Balke, A. and J. Pearl. (1994). “Counterfactual probabilities: Computational methods, bounds and applications”. In: *Uncertainty Proceedings 1994.* Elsevier. 46–54.

Bang, H. and J. M. Robins. (2005). “Doubly robust estimation in missing data and causal inference models”. *Biometrics.* 61(4): 962–973.

Bareinboim, E. and J. Pearl. (2016). “Causal Inference and The Data-Fusion Problem”. In: *Proceedings of the National Academy of Sciences.* Ed. by R. M. Shiffrin. Vol. 113. National Academy of Sciences. 7345–7352.

Bareinboim, E., J. D. Correa, D. Ibeling, and T. Icard. (2022). “On Pearl’s Hierarchy and the Foundations of Causal Inference”. In: *Probabilistic and Causal Inference: The Works of Judea Pearl.* 1st. New York, NY, USA: Association for Computing Machinery. 507–556.

Barocas, S., M. Hardt, and A. Narayanan. (2017). “Fairness in machine learning”. *Nips tutorial.* 1: 2017.

Barocas, S. and A. D. Selbst. (2016). “Big data’s disparate impact”. *Calif. L. Rev.* 104: 671.

Ben-Michael, E., K. Imai, and Z. Jiang. (2022). “Policy learning with asymmetric utilities”. *arXiv preprint arXiv:2206.10479.*

Benkeser, D., M. Carone, M. V. D. Laan, and P. Gilbert. (2017). “Doubly robust nonparametric inference on the average treatment effect”. *Biometrika.* 104(4): 863–880.

Bickel, P. J., E. A. Hammel, and J. W. O’Connell. (1975). “Sex bias in graduate admissions: Data from Berkeley”. *Science.* 187(4175): 398–404.

Breiman, L. (2001). “Random forests”. *Machine learning.* 45: 5–32.

Brimicombe, A. J. (2007). “Ethnicity, religion, and residential segregation in London: evidence from a computational typology of minority communities”. *Environment and Planning B: Planning and Design.* 34(5): 884–904.

Buolamwini, J. and T. Gebru. (2018). “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency.* Ed. by S. A. Friedler and C. Wilson. Vol. 81. *Proceedings of Machine Learning Research.* NY, USA. 77–91.

Calders, T. and S. Verwer. (2010). “Three Naive Bayes Approaches for Discrimination-Free Classification”. *Data Mining journal.*

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. (2018). “Double/debiased machine learning for treatment and structural parameters”.

Chiappa, S. (2019). “Path-specific counterfactual fairness”. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 33. No. 01. 7801–7808.

Chouldechova, A. (2017). “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. *Tech. rep.* No. arXiv:1703.00056. arXiv.org.

Cinelli, C. and C. Hazlett. (2020). “Making sense of sensitivity: Extending omitted variable bias”. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 82(1): 39–67.

Cinelli, C., D. Kumor, B. Chen, J. Pearl, and E. Bareinboim. (2019). “Sensitivity analysis of linear structural causal models”. In: *International conference on machine learning.* PMLR. 1252–1261.

Commission, E. (2021). “EU Artificial Intelligence Act”. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%5C%3A52021PC0206.

Corbett-Davies, S. and S. Goel. (2018). “The measure and mismeasure of fairness: A critical review of fair machine learning”. *arXiv preprint arXiv:1808.00023.*

Correa, J. and E. Bareinboim. (2020). “A calculus for stochastic interventions: Causal effect identification and surrogate experiments”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 06. 10093–10100.

Correa, J., S. Lee, and E. Bareinboim. (2021a). “Nested Counterfactual Identification from Arbitrary Surrogate Experiments”. In: *Advances in Neural Information Processing Systems*. Vol. 34.

Correa, J., S. Lee, and E. Bareinboim. (2021b). “Nested counterfactual identification from arbitrary surrogate experiments”. *Advances in Neural Information Processing Systems*. 34: 6856–6867.

Coston, A., A. Mishler, E. H. Kennedy, and A. Chouldechova. (2020). “Counterfactual risk assessments, evaluation, and fairness”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 582–593.

Dembroff, R. and I. Kohler-Hausmann. (2022). “Supreme confusion about causality at the Supreme Court”. *CUNY L. Rev.* 25: 57.

Detrixhe, J. and J. B. Merrill. (2019). “The fight against financial advertisers using Facebook for digital redlining”.

Ding, P. and T. J. VanderWeele. (2016). “Sensitivity analysis without assumptions”. *Epidemiology (Cambridge, Mass.)* 27(3): 368.

Ding, Q. J. and T. Hesketh. (2006). “Family size, fertility preferences, and sex ratio in China in the era of the one child family policy: results from national family planning and reproductive health survey”.

Dutta, S., D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney. (2020). “Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing”. In: *International conference on machine learning*. PMLR. 2803–2813.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. (2012). “Fairness through awareness”. In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

Farbmacher, H., M. Huber, L. Lafférs, H. Langen, and M. Spindler. (2020). “Causal mediation analysis with double machine learning”. *arXiv preprint arXiv:2002.12710*.

Frangakis, C. E. and D. B. Rubin. (2002). “Principal stratification in causal inference”. *Biometrics*. 58(1): 21–29.

Friedler, S. A., C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. (2019). “A comparative study of fairness-enhancing interventions in machine learning”. In: *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

Friedler, S. A., C. Scheidegger, and S. Venkatasubramanian. (2016). “On the (im)possibility of fairness”. *Tech. rep.* No. 1609.07236. arxiv.org. URL: http://arxiv.org/abs/1609.07236.

Gelman, A. and J. Hill. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Grgic-Hlaca, N., M. B. Zafar, K. P. Gummadi, and A. Weller. (2016). "The case for process fairness in learning: Feature selection for fair decision making". In: *NIPS symposium on machine learning and the law.* Vol. 1. No. 2. Barcelona, Spain. 11.

Grimmelmann, J. and D. Westreich. (2016). "Incomprehensible discrimination". *Calif. L. Rev. Circuit.* 7: 164.

Guth, L. (2009). "Minimax problems related to cup powers and Steenrod squares". *Geometric And Functional Analysis.* 18: 1917–1987.

Hajian, S. and J. Domingo-Ferrer. (2012). "A Study on the Impact of Data Anonymization on Anti-discrimination". In: *ICDM International Workshop on Discrimination and Privacy-Aware Data Mining.* Ed. by T. Calders and I. Zliobaite. IEEE.

Halpern, J. Y. (2016). *Actual causality.* MiT Press.

Hamburg, M. A. and F. S. Collins. (2010). "The path to personalized medicine". *New England Journal of Medicine.* 363(4): 301–304.

Hardt, M., E. Price, and N. Srebro. (2016). "Equality of opportunity in supervised learning". *Advances in neural information processing systems.* 29: 3315–3323.

Harwell, D. (2019). "Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use". https://www.washingtonpost.com/technol
study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/.

Heckman, J. J., H. Ichimura, and P. Todd. (1998). "Matching as an econometric evaluation estimator". *The review of economic studies.* 65(2): 261–294.

Hernán, M. A. (2005). "Invited commentary: hypothetical interventions to define causal effects—afterthought or prerequisite?" *American journal of epidemiology.* 162(7): 618–620.

Hernandez, J. (2009). "Redlining revisited: mortgage lending patterns in Sacramento 1930–2004". *International Journal of Urban and Regional Research.* 33(2): 291–313.

Hesketh, T., L. Lu, and Z. W. Xing. (2005). "The effect of China's one-child family policy after 25 years".

Hu, L. and I. Kohler-Hausmann. (2020). "What's sex got to do with fair machine learning?" *arXiv preprint arXiv:2006.01770.*

Imai, K. and Z. Jiang. (2020). "Principal fairness for human and algorithmic decision-making". *arXiv preprint arXiv:2005.10400.*

Insel, T. R. (2009). "Translating scientific opportunity into public health impact: a strategic plan for research on mental illness". *Archives of general psychiatry.* 66(2): 128–133.

Ji, S., J. Kollár, and B. Shiffman. (1992). "A global Łojasiewicz inequality for algebraic varieties". *Transactions of the American Mathematical Society.* 329(2): 813–818.

Kamiran, F. and T. Calders. (2009). "Classifying without Discriminating". In: *Proc. IC4 09*. IEEE.

Kamiran, F. and T. Calders. (2012). "Data preprocessing techniques for classification without discrimination". *Knowledge and Information Systems*. 33(1): 1–33.

Kamiran, F., T. Calders, and M. Pechenizkiy. (2010). "Discrimination Aware Decision Tree Learning". In: *International Conference on Data Mining*. IEEE.

Kamiran, F., A. Karim, and X. Zhang. (2012). "Decision theory for discrimination-aware classification". In: *2012 IEEE 12th International Conference on Data Mining*. IEEE. 924–929.

Kamishima, T., S. Akaho, H. Asoh, and J. Sakuma. (2012). "Fairness-aware classifier with prejudice remover regularizer". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 35–50.

Kennedy, E. H. (2016). "Semiparametric theory and empirical processes in causal inference". In: *Statistical causal inferences and their applications in public health research*. Springer. 141–167.

Kilbertus, N., M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. (2017). "Avoiding discrimination through causal reasoning". *arXiv preprint arXiv:1706.02744*.

Kohler-Hausmann, I. (2018). "Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination". *Nw. UL Rev.* 113: 1163.

Kotz, N. (2005). *Judgment Days: Lyndon Baines Johnson, Martin Luther King, Jr., and the Laws That Changed America*. HMH.

Kusner, M. J., J. Loftus, C. Russell, and R. Silva. (2017). "Counterfactual fairness". *Advances in neural information processing systems*. 30.

Larson, J., S. Mattu, L. Kirchner, and J. Angwin. (2016). "How we analyzed the COMPAS recidivism algorithm". *ProPublica (5 2016)*. 9.

Lee, S., J. Correa, and E. Bareinboim. (2019). "General Identifiability with Arbitrary Surrogate Experiments". In: *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. Tel Aviv, Israel: AUAI Press.

Luong, B. T., S. Ruggieri, and F. Turini. (2011). "k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention". In: *17th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2011)*. ACM.

Mancuhan, K. and C. Clifton. (2014). "Decision Tree Classification on Outsourced Data". In: *Workshop on Data Ethics held in conjunction with KDD 2014*. New York, NY.

Moore, M. (2019). "Causation in the Law". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University.

Nabi, R. and I. Shpitser. (2018). "Fair inference on outcomes". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1.

Nilforoshan, H., J. D. Gaebler, R. Shroff, and S. Goel. (2022). "Causal conceptions of fairness and their consequences". In: *International Conference on Machine Learning*. PMLR. 16848–16887.

Oppenheimer, D. B. (1994). "Kennedy, King, Shuttlesworth and Walker: The Events Leading to the Introduction of the Civil Rights Act of 1964". *USFL Rev.* 29: 645.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Pearl, J. (2001). "Direct and Indirect Effects". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 411–420.

Pearl, J. and D. Mackenzie. (2018). *The Book of Why: The New Science of Cause and Effect*. 1st. New York, NY, USA: Basic Books, Inc.

Pearson, K. (1899). "IV. Mathematical contributions to the theory of evolution.—V. On the reconstruction of the stature of prehistoric races". *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*. 1(192): 169–244.

Pedreschi, D., S. Ruggieri, and F. Turini. (2008). "Discrimination-aware data mining". In: *14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008)*. ACM.

Pedreschi, D., S. Ruggieri, and F. Turini. (2009). "Measuring Discrimination in Socially-Sensitive Decision Records". In: *9th SIAM Conference on Data Mining (SDM 2009)*. 581–592.

Plečko, D., N. Bennett, and N. Meinshausen. (2021). "fairadapt: Causal Reasoning for Fair Data Pre-processing". *arXiv preprint arXiv:2110.10200*.

Plečko, D. and N. Meinshausen. (2020). "Fair data adaptation with quantile preservation". *Journal of Machine Learning Research*. 21: 242.

Pleiss, G., M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. (2017). "On Fairness and Calibration". In: *NIPS*. URL: https://arxiv.org/abs/1709.02012.

Robins, J. M. (2003). "Semantics of causal DAG models and the identification of direct and indirect effects". *Oxford Statistical Science Series*: 70–82.

Robins, J. M. and S. Greenland. (1992). "Identifiability and exchangeability for direct and indirect effects". *Epidemiology*: 143–155.

Robins, J. M. and A. Rotnitzky. (1995). "Semiparametric efficiency in multivariate regression models with missing data". *Journal of the American Statistical Association*. 90(429): 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao. (1994). "Estimation of regression coefficients when some regressors are not always observed". *Journal of the American statistical Association*. 89(427): 846–866.

Rodolfa, K. T., H. Lamba, and R. Ghani. (2021). "Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy". *Nature Machine Intelligence*. 3(10): 896–904.

Romei, A. and S. Ruggieri. (2014). "A multidisciplinary survey on discrimination analysis". *The Knowledge Engineering Review*. 29(5): 582–638.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*. 66(5): 688.

Rubin, D. B. (1986). "Statistics and causal inference: Comment: Which ifs have causal answers". *Journal of the American Statistical Association*. 81(396): 961–962.

Rubin, D. B. (2005). "Causal inference using potential outcomes: Design, modeling, decisions". *Journal of the American Statistical Association*. 100(469): 322–331.

Ruggieri, S., D. Pedreschi, and F. Turini. (2011). "DCUBE: Discrimination Discovery in Databases". In: *17th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2011)*. ACM.

Rutherglen, G. (1987). "Disparate impact under title VII: an objective theory of discrimination". *Va. L. Rev.* 73: 1297.

Sen, M. and O. Wasow. (2016). "Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics". *Annual Review of Political Science*. 19: 499–522.

Shapley, L. S. *et al.* (1953). "A value for n-person games".

Shpitser, I. and J. Pearl. (2007). "What Counterfactuals Can Be Tested". In: *Proceedings of the Twenty-third Conference on Uncertainty in Artificial Intelligence*. 352–359.

Shpitser, I. and E. T. Tchetgen. (2016). "Causal inference with a graphical hierarchy of interventions". *Annals of statistics*. 44(6): 2433.

Singal, R., G. Michailidis, and H. Ng. (2021). "Flow-based attribution in graphical models: A recursive shapley approach". In: *International Conference on Machine Learning*. PMLR. 9733–9743.

Tchetgen, E. J. T. and I. Shpitser. (2012). "Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis". *Annals of statistics*. 40(3): 1816.

Tian, J. and J. Pearl. (2000). "Probabilities of causation: Bounds and identification". *Annals of Mathematics and Artificial Intelligence*. 28(1): 287–313.

Tian, J. and J. Pearl. (2002). "A general identification condition for causal effects". In: *Aaai/iaai*. 567–573.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

Weinberger, N. (2022). "Signal manipulation and the causal analysis of racial discrimination".

Wright, M. N., S. Wager, and P. Probst. (2020). "Ranger: A fast implementation of random forests". *R package version 0.12*. 1.

Wu, Y., L. Zhang, X. Wu, and H. Tong. (2019). "Pc-fairness: A unified framework for measuring causality-based fairness". *Advances in neural information processing systems*. 32.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. (2013). "Learning Fair Representations". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S. Dasgupta and D. Mcallester. Vol. 28. No. 3. 325–333.

Zenou, Y. and N. Boccard. (2000). "Racial discrimination and redlining in cities". *Journal of Urban economics*. 48(2): 260–285.

Zhang, B. H., B. Lemoine, and M. Mitchell. (2018). "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

Zhang, J. and E. Bareinboim. (2018a). "Equality of Opportunity in Classification: A Causal Approach". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Montreal, Canada: Curran Associates, Inc. 3671–3681.

Zhang, J. and E. Bareinboim. (2018b). "Fairness in decision-making—the causal explanation formula". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. No. 1.

Zhang, J. and E. Bareinboim. (2018c). "Non-parametric path analysis in structural causal models". In: *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*.

Zhang, J., J. Tian, and E. Bareinboim. (2022). "Partial Counterfactual Identification from Observational and Experimental Data". In: *Proceedings of the 39th International Conference on Machine Learning*.

Zliobaite, I., F. Kamiran, and T. Calders. (2011). "Handling Conditional Discrimination". In: *International Conference on Data Mining*. IEEE.