

Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning

Other titles in Foundations and Trends® in Machine Learning

A Friendly Tutorial on Mean-Field Spin Glass Techniques for Non-Physicists

Andrea Montanari and Subhabrata Sen

ISBN: 978-1-63828-212-9

Reinforcement Learning, Bit by Bit Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband and Zheng Wen

ISBN: 978-1-63828-254-9

Introduction to Riemannian Geometry and Geometric Statistics: From Basic Theory to Implementation with Geomstats

Nicolas Guigui, Nina Miolane and Xavier Pennec

ISBN: 978-1-63828-154-2

Graph Neural Networks for Natural Language Processing: A Survey

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei and Bo Long

ISBN: 978-1-63828-142-9

Model-based Reinforcement Learning: A Survey

Thomas M. Moerland, Joost Broekens, Aske Plaat and Catholijn M. Jonker

ISBN: 978-1-63828-056-9

Divided Differences, Falling Factorials, and Discrete Splines: Another Look at Trend Filtering and Related Problems

Ryan J. Tibshirani

ISBN: 978-1-63828-036-1

Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning

Drago Plečko

ETH Zürich

drago.plecko@stat.math.ethz.ch

Elias Bareinboim

Columbia University

eb@cs.columbia.edu

now

the essence of knowledge

Boston — Delft

Foundations and Trends[®] in Machine Learning

Published, sold and distributed by:

now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:

now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

D. Plečko and E. Bareinboim. *Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning*. Foundations and Trends[®] in Machine Learning, vol. 17, no. 3, pp. 304–589, 2024.

ISBN: 978-1-63828-331-7

© 2024 D. Plečko and E. Bareinboim

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanical, photocopying, recording or otherwise, without prior written permission of the publishers.

Photocopying. In the USA: This journal is registered at the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923. Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by now Publishers Inc for users registered with the Copyright Clearance Center (CCC). The 'services' for users can be found on the internet at: www.copyright.com

For those organizations that have been granted a photocopy license, a separate system of payment has been arranged. Authorization does not extend to other kinds of copying, such as that for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale. In the rest of the world: Permission to photocopy must be obtained from the copyright owner. Please apply to now Publishers Inc., PO Box 1024, Hanover, MA 02339, USA; Tel. +1 781 871 0245; www.nowpublishers.com; sales@nowpublishers.com

now Publishers Inc. has an exclusive license to publish this material worldwide. Permission to use this content must be obtained from the copyright license holder. Please apply to now Publishers, PO Box 179, 2600 AD Delft, The Netherlands, www.nowpublishers.com; e-mail: sales@nowpublishers.com

Foundations and Trends[®] in Machine Learning

Volume 17, Issue 3, 2024

Editorial Board

Editor-in-Chief

Michael Jordan

University of California, Berkeley
United States

Ryan Tibshirani

University of California, Berkeley
United States

Editors

Genevera Allen
Rice University

Peter Bartlett
UC Berkeley

Yoshua Bengio
Université de Montréal

Avrim Blum
*Toyota Technological
Institute*

Craig Boutilier
University of Toronto

Stephen Boyd
Stanford University

Carla Brodley
Northeastern University

Inderjit Dhillon
Texas at Austin

Jerome Friedman
Stanford University

Kenji Fukumizu
ISM

Zoubin Ghahramani
Cambridge University

David Heckerman
Amazon

Tom Heskes
Radboud University

Geoffrey Hinton
University of Toronto

Aapo Hyvarinen
Helsinki IIT

Nan Jiang
UIUC

Leslie Pack Kaelbling
MIT

Michael Kearns
UPenn

Daphne Koller
Stanford University

John Lafferty
Yale

Michael Littman
Brown University

Gabor Lugosi
Pompeu Fabra

David Madigan
Columbia University

Pascal Massart
Université de Paris-Sud

Andrew McCallum
*University of
Massachusetts Amherst*

Marina Meila
University of Washington

Andrew Moore
CMU

Vianney Perchet
CREST

John Platt
Google

Luc de Raedt
KU Leuven

Christian Robert
Paris-Dauphine

Sunita Sarawagi
IIT Bombay

Robert Schapire
Microsoft Research

Bernhard Schoelkopf
Max Planck Institute

Richard Sutton
University of Alberta

Csaba Szepesvari
University of Alberta

Larry Wasserman
CMU

Bin Yu
UC Berkeley

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2024, Volume 17, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Contents

1	Introduction	3
2	Foundations of Causal Inference	14
2.1	Structural Causal Models	15
2.2	Observational and Counterfactual Distributions	18
2.3	Encoding Structural Assumptions through Causal Diagrams	22
3	Foundations of Causal Fairness Analysis	29
3.1	Structural Fairness Criteria	31
3.2	Explaining Factual and Counterfactual Variations	37
4	Total Variation Family	44
4.1	Population-level Contrasts - $P(u)$	44
4.2	Summary of the TV-family and the Fairness Map	67
4.3	The Identification Problem and the FPCFA in Practice	72
4.4	Other Relations with the Literature	78
5	Fairness Tasks	105
5.1	Task 1: Bias Detection and Quantification	106
5.2	Task 2: Fair Prediction	117
5.3	Task 3: Fair Decision-Making	132

6	Disparate Impact and Business Necessity	165
6.1	Causal Inference Procedures	167
6.2	Refining Spurious Discrimination	175
6.3	Refining Indirect Effects	197
6.4	Extended Fairness Map	203
6.5	Extended Fairness Cookbook	214
7	Conclusions	219
	Acknowledgments	221
	Appendices	222
A	Proofs of Main Theorems and Derivations	223
B	Practical Aspects of Fairness Measures	246
C	Selection Bias Interpretation	254
D	Multi-valued and Continuous Protected Attributes	260
E	Process Fairness	274
	References	277

Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning

Drago Plečko¹ and Elias Bareinboim²

¹*Seminar für Statistik, ETH Zürich, Switzerland;*

drago.plecko@stat.math.ethz.ch

²*Department of Computer Science, Columbia University, USA;*

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 monograph, 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,

Drago Plečko and Elias Bareinboim (2024), “Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning”, Foundations and Trends® in Machine Learning: Vol. 17, No. 3, pp 304–589. DOI: 10.1561/2200000106.

©2024 D. Plečko and E. Bareinboim

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 target¹ 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

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

²One 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

In this monograph, 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 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).

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.

³This 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 monograph.

⁴Interestingly, 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.

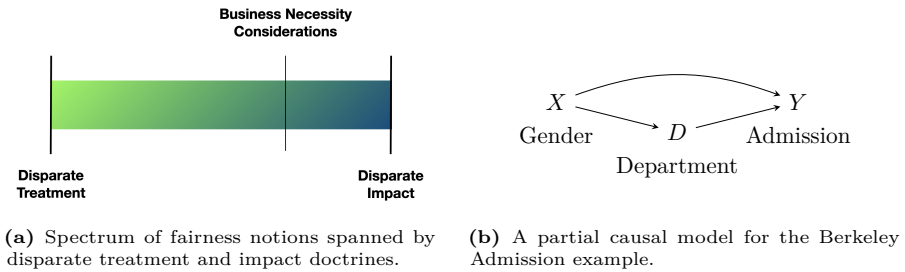


Figure 1.1: Spectrum of fairness notions and the Berkeley Admissions example.

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. 1.1a). 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 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.

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. 1.1b. 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 mechanism⁵ 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), 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

⁵As 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.

⁶Society 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.

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. 1.1a 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 monograph, 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 policymakers, 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.

⁷We 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.

Monograph Roadmap and 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. 1.2, which also serves as a roadmap of how this monograph 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 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

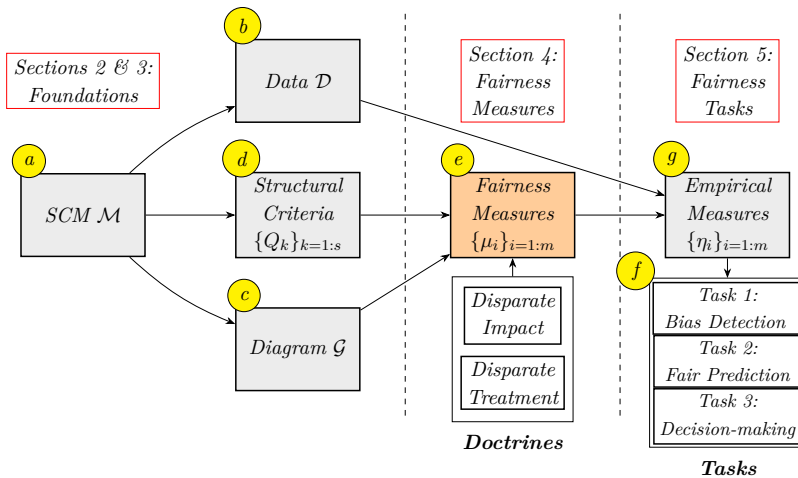


Figure 1.2: A mental map of the Causal Fairness Analysis framework.

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. 3.2).

2. We introduce the Fundamental Problem of Causal Fairness Analysis (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

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

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 settings (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 monograph.

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 and Derivations

In this section, we provide the proofs of the main theorems presented in the monograph. 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 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 $Z \subset V'$):

Table A.1: List of implications in the Fairness Map in Fig. 4.5.

	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

$$\text{unit-TE}_{x_0, x_1}(y(u)) = 0 \forall u \implies v'\text{-TE}_{x_0, x_1}(y | v') = 0 \forall v' \quad (\text{A.1})$$

$$\implies z\text{-TE}_{x_0, x_1}(y | z) = 0 \forall z \quad (\text{A.2})$$

$$\implies \text{ETT}_{x_0, x_1}(y | x) = 0 \forall x \quad (\text{A.3})$$

$$\implies \text{TE}_{x_0, x_1}(y) = 0, \quad (\text{A.4})$$

$$\text{unit-DE}_{x_0, x_1}(y(u)) = 0 \forall u \implies v'\text{-DE}_{x_0, x_1}(y | v') = 0 \forall v' \quad (\text{A.5})$$

$$\implies z\text{-DE}_{x_0, x_1}(y | z) = 0 \forall z \quad (\text{A.6})$$

$$\implies \text{Ctf-DE}_{x_0, x_1}(y | x) = 0 \forall x \quad (\text{A.7})$$

$$\implies \text{NDE}_{x_0, x_1}(y) = 0, \quad (\text{A.8})$$

$$\text{unit-IE}_{x_0, x_1}(y(u)) = 0 \forall u \implies v'\text{-IE}_{x_0, x_1}(y | v') = 0 \forall v' \quad (\text{A.9})$$

$$\implies z\text{-IE}_{x_0, x_1}(y | z) = 0 \forall z \quad (\text{A.10})$$

$$\implies \text{Ctf-IE}_{x_0, x_1}(y | x) = 0 \forall x \quad (\text{A.11})$$

$$\implies \text{NIE}_{x_0, x_1}(y) = 0. \quad (\text{A.12})$$

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.

$$\begin{aligned} \text{TE}_{x_0, x_1}(y) &= P(y_{x_1}) - P(y_{x_0}) \\ &= \sum_x [P(y_{x_1} | x) - P(y_{x_0} | x)]P(x) \\ &= \sum_x \text{ETT}_{x_0, x_1}(y | x)P(x). \end{aligned}$$

Therefore, if $\text{ETT}_{x_0, x_1}(y | x) = 0 \forall x$ then $\text{TE}_{x_0, x_1}(y) = 0$. Next, we can write

$$\begin{aligned} \text{ETT}_{x_0, x_1}(y | x) &= P(y_{x_1} | x) - P(y_{x_0} | x) \\ &= \sum_z [P(y_{x_1} | x, z) - P(y_{x_0} | x, z)]P(z | x) \\ &= \sum_z [P(y_{x_1} | z) - P(y_{x_0} | z)]P(z | x) \quad Y_x \perp\!\!\!\perp X | Z \text{ in SFM} \\ &= \sum_z z\text{-TE}_{x_0, x_1}(y | z)P(z | x). \end{aligned}$$

Therefore, if $z\text{-TE}_{x_0, x_1}(y | z) = 0 \forall z$ then $\text{ETT}_{x_0, x_1}(y | x) = 0 \forall x$. Next, for a set $V' \subseteq V$ such that $Z \subseteq V'$, we can write

$$\begin{aligned} z\text{-TE}_{x_0, x_1}(y) &= P(y_{x_1} | z) - P(y_{x_0} | z) \\ &= \sum_{v' \setminus z} [P(y_{x_1} | z, v' \setminus z) - P(y_{x_0} | z, v' \setminus z)]P(v' \setminus z | z) \\ &= \sum_{v' \setminus z} v'\text{-TE}_{x_0, x_1}(y | v')P(v' \setminus z | z). \end{aligned}$$

Therefore, if $v'\text{-TE}_{x_0, x_1}(y | v') = 0 \forall v'$ then $z\text{-TE}_{x_0, x_1}(y | z) = 0 \forall z$. Next, notice that

$$\begin{aligned} v'\text{-TE}_{x_0, x_1}(y) &= P(y_{x_1} | v') - P(y_{x_0} | v') \\ &= \sum_u [y_{x_1}(u) - y_{x_0}(u)]P(u | v') \\ &= \sum_u \text{unit-TE}_{x_0, x_1}(y(u))P(u | v'). \end{aligned}$$

Therefore, if $\text{unit-TE}_{x_0, x_1}(y(u)) = 0 \forall u$ then $v'\text{-TE}_{x_0, x_1}(y | v') = 0 \forall v'$. ■

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,

$$\text{Exp-SE}_x(y) = 0 \forall x \iff \text{Ctf-SE}_{x,x'}(y) = 0 \forall x \neq x'. \quad (\text{A.13})$$

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:

$$\text{S-DE} \implies \text{unit-DE}_{x_0,x_1}(y(u)) = 0 \forall u. \quad (\text{A.14})$$

Proof. Suppose that $X \notin \text{pa}(Y)$. Note that:

$$\begin{aligned} \text{unit-DE}_{x_0,x_1}(y(u)) &= y_{x_1,W_{x_0}}(u) - y_{x_0}(u) \\ &= f_Y(x_1, W_{x_0}(u), Z(u), u_Y) - f_Y(x_0, W_{x_0}(u), Z(u), u_Y) \\ &= f_Y(W_{x_0}(u), Z(u), u_Y) \\ &\quad - f_Y(W_{x_0}(u), Z(u), u_Y) \quad X \notin \text{pa}(Y) \\ &= 0. \end{aligned}$$

■

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:

$$\text{S-IE} \implies \text{unit-IE}_{x_1,x_0}(y(u)) = 0 \forall u. \quad (\text{A.15})$$

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:

$$\begin{aligned} \text{unit-IE}_{x_1, x_0}(y(u)) &= y_{x_1, W_{x_0}}(u) - y_{x_1}(u) \\ &= f_Y(x_1, (W_{de}^C)_{x_0}(u), Z(u), u_Y) \\ &\quad - f_Y(x_1, (W_{de}^C)_{x_1}(u), Z(u), u_Y) \\ &= f_Y(x_1, W_{de}^C(u), Z(u), u_Y) \\ &\quad - f_Y(x_1, W_{de}^C(u), Z(u), u_Y) \quad W_{de}^C \notin \text{de}(X) \\ &= 0. \end{aligned}$$

■

Lemma A.5 (Admissibility w.r.t. structural spurious). The structural spurious effect criterion ($U_X \cap \text{an}(Y) = \emptyset$ and $\text{an}(X) \cap \text{an}_{\underline{X}}(Y) = \emptyset$) implies counterfactual spurious effect is 0. Formally:

$$\text{S-SE} \implies \text{Ctf-SE}_{x_0, x_1}(y) = 0 \forall u. \quad (\text{A.16})$$

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:

$$\begin{aligned} \text{Ctf-SE}_{x_0, x_1}(y) &= P(y_{x_0} | x_1) - P(y | x_0) \\ &= P(y_{x_0}) - P(y | x_0) \quad \text{since } Y_x \perp\!\!\!\perp X \\ &= P(y_{x_0}) - P(y_{x_0}) \quad \text{Action/Observation Exchange} \\ &= 0. \end{aligned}$$

■

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:

$$\text{TE}_{x_0,x_1}(y) = \text{NDE}_{x_0,x_1}(y) - \text{NIE}_{x_1,x_0}(y) \tag{A.17}$$

$$\text{ETT}_{x_0,x_1}(y | x) = \text{Ctf-DE}_{x_0,x_1}(y | x) - \text{Ctf-IE}_{x_1,x_0}(y | x) \tag{A.18}$$

$$z\text{-TE}_{x_0,x_1}(y | z) = z\text{-DE}_{x_0,x_1}(y | z) - z\text{-IE}_{x_1,x_0}(y | z) \tag{A.19}$$

$$v'\text{-TE}_{x_0,x_1}(y | v') = v'\text{-DE}_{x_0,x_1}(y | v') - v'\text{-IE}_{x_1,x_0}(y | v') \tag{A.20}$$

$$\text{unit-TE}_{x_0,x_1}(y(u)) = \text{unit-DE}_{x_0,x_1}(y(u)) - \text{unit-IE}_{x_1,x_0}(y(u)). \tag{A.21}$$

Proof. The proof follows from the structural basis expansion from Eq. 3.24. In particular, note that

$$E\text{-TE}_{x_1,x_0}(y | E) = P(y_{x_1} | E) - P(y_{x_0} | E) \tag{A.22}$$

$$= P(y_{x_1} | E) - P(y_{x_1,W_{x_0}} | E) \tag{A.23}$$

$$+ P(y_{x_1,W_{x_0}} | E) - P(y_{x_0} | E)$$

$$= -E\text{-IE}_{x_1,x_0}(y | E) + E\text{-DE}_{x_1,x_0}(y | E). \tag{A.24}$$

By using different events E the claim follows. ■

Lemma A.7 (TV Decompositions). The total variation (TV) measure admits the following two decompositions

$$\text{TV}_{x_0,x_1}(y) = \text{Exp-SE}_{x_1}(y) + \text{TE}_{x_0,x_1}(y) - \text{Exp-SE}_{x_0}(y) \tag{A.25}$$

$$= \text{ETT}_{x_0,x_1}(y | x_0) - \text{Ctf-SE}_{x_1,x_0}. \tag{A.26}$$

Proof. We write

$$\begin{aligned} \text{TV}_{x_0,x_1}(y) &= P(y | x_1) - P(y | x_0) \\ &= P(y | x_1) - P(y_{x_1}) + P(y_{x_1}) - P(y_{x_0}) + P(y_{x_0}) - P(y | x_0) \\ &= \text{Exp-SE}_{x_1}(y) + \text{TE}_{x_0,x_1}(y) - \text{Exp-SE}_{x_0}(y). \end{aligned}$$

Alternatively, we can write

$$\begin{aligned} \text{TV}_{x_0,x_1}(y) &= P(y | x_1) - P(y | x_0) \\ &= P(y | x_1) - P(y_{x_1} | x_0) + P(y_{x_1} | x_0) - P(y | x_0) \\ &= \text{ETT}_{x_0,x_1}(y | x_0) - \text{Ctf-SE}_{x_1,x_0}(y), \end{aligned}$$

which completes the proof. ■

A.2 Soundness of the SFM: Proof of Thm. 4.10 and 4.11

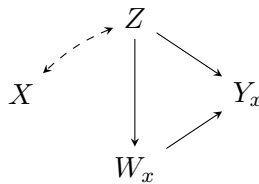
Proof. The proof consists of two parts. In the first part, we show that the quantities where the event E is either of \emptyset , $\{x\}$, $\{z\}$ (corresponding to the first three rows of the fairness map) are identifiable under the assumptions of the Standard Fairness Model. We in particular show that $\text{TE}_{x_0, x_1}(y)$, $\text{Exp-SE}_x(y)$, $\text{TE}_{x_0, x_1}(y | z)$, $\text{ETT}_{x_0, x_1}(y | x)$, and $\text{Ctf-DE}_{x_0, x_1}(y | x)$ are identifiable (it follows from very similar arguments that all other quantities are also identifiable). Additionally, we also show that $(x, w)\text{-DE}_{x_0, x_1}(y | x, w)$ and $(x, z, w)\text{-DE}_{x_0, x_1}(y | x, z, w)$ 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}_{SFM} , the quantities of interest are (i) identifiable; (ii) their identification expression is the same. This in turn shows that using \mathcal{G}_{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 $(x, w)\text{-DE}$ and $(x, z, w)\text{-DE}$) is not identifiable under some very mild conditions (namely the existence of a path $X \rightarrow W_{i_1} \rightarrow \dots \rightarrow W_{i_k} \rightarrow Y$). 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}_{SFM} contains “groups” of variables Z and W , making the standard identification machinery (Pearl, 2000) not directly applicable.

Part I: Note that for identifying $\text{TE}_{x_0, x_1}(y)$ we need to identify $P(y_x)$. We can write

$$\begin{aligned} P(y_x) &= P(y | do(x)) \\ &= \sum_z P(y | do(x), z)P(z | do(x)) \quad \text{Law of Total Probability} \\ &= \sum_z P(y | x, z)P(z) \quad (Y \perp\!\!\!\perp X | Z)_{\mathcal{G}_{\underline{X}}}, (X \perp\!\!\!\perp Z)_{\mathcal{G}_{\overline{X}}} \end{aligned}$$

from which it follows that $\text{TE}_{x_0, x_1}(y) = \sum_z [P(y | x_1, z) - P(y | x_0, z)]P(z)$. Note that the identifiability of $\text{TE}_{x_0, x_1}(y | z)$ also follows from the above derivation, namely $\text{TE}_{x_0, x_1}(y | z) = \sum_z [P(y |$

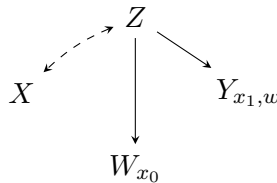
$x_1, z) - P(y \mid x_0, z)]$, and so does $\text{Exp-SE}_x(y) = \sum_z P(y \mid x, z)[P(z \mid x) - P(z)]$. We are now left with showing that $\text{ETT}_{x_0, x_1}(y \mid x)$ and $\text{Ctf-DE}_{x_0, x_1}(y \mid x)$ are also identifiable. These are Layer 3, counterfactual quantities and therefore rules 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,

$$\begin{aligned}
 \text{ETT}_{x_0, x_1}(y) &= P(y_{x_1} \mid x) - P(y_{x_0} \mid x) \\
 &= \sum_z [P(y_{x_1} \mid x, z) - P(y_{x_0} \mid x, z)]P(z \mid x) \quad \text{Law of Tot. Prob.} \\
 &= \sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)]P(z \mid x) \quad Y_x \perp\!\!\!\perp X \mid Z.
 \end{aligned}$$

Finally, for identifying $\text{Ctf-DE}_{x_0, x_1}(y \mid x)$ we use *make-cg* applied to \mathcal{G}_{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 $\text{Ctf-DE}_{x_0, x_1}(y \mid x)$ equals to:

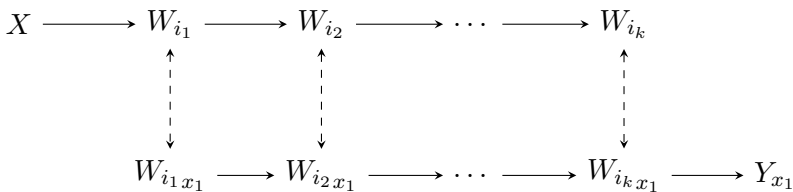
$$\begin{aligned}
 & P(y_{x_1, W_{x_0}} \mid x) - P(y_{x_0, W_{x_0}} \mid x) \\
 &= \sum_z [P(y_{x_1, W_{x_0}} \mid x, z) - P(y_{x_0, W_{x_0}} \mid x, z)]P(z \mid x) && \text{Law of Tot. Prob.} \\
 &= \sum_{z, w} [P(y_{x_1, w, w_{x_0}} \mid x, z) - P(y_{x_0, w, w_{x_0}} \mid x, z)]P(z \mid x) && \text{Ctf. unnesting} \\
 &= \sum_{z, w} [P(y_{x_1, w} \mid x, z) - P(y_{x_0, w} \mid x, z)]P(w_{x_0} \mid z)P(z \mid x) && Y_{x_1, w} \perp\!\!\!\perp W_{x_0} \mid Z \\
 &= \sum_{z, w} [P(y_{x_1, w} \mid x, z) - P(y_{x_0, w} \mid x, z)]P(w \mid x_0, z)P(z \mid x) && W_{x_0} \perp\!\!\!\perp X \mid Z \\
 &= \sum_{z, w} [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_0, z)P(z \mid x) && Y_{x, w} \perp\!\!\!\perp X \mid Z.
 \end{aligned}$$

From the above, one can also show that

$$\begin{aligned}
 (x, w)\text{-DE}_{x_0, x_1}(y \mid x, w) &= \sum_z [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)] \\
 &\quad \cdot P(w \mid z, x)P(z \mid x), \\
 (x, z, w)\text{-DE}_{x_0, x_1}(y \mid x, z, w) &= P(y \mid x_1, z, w) - P(y \mid x_0, z, w),
 \end{aligned}$$

completing the first part of the proof.

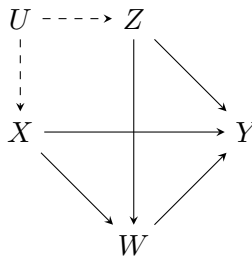
Part II: We next need to show that any contrast with either $W = w$ or $Y = y$ in the event E (excluding (x, w) -DE and (x, z, w) -DE) is not identifiable, even if using the full graph \mathcal{G} . We show this for the quantity $P(y_{x_1} \mid x_0, w)$, since other similar quantities work analogously. 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 $X \rightarrow W_{i_1} \rightarrow \dots \rightarrow W_{i_k} \rightarrow Y$ 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 $\{W_i, W_{ix_1}\}$ 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

$$Y = \sum_{V_i \in X, Z, W} a_{V_i Y} V_i + \epsilon_Y,$$

and the linear predictor of Y , labeled f can be written as

$$f(X, Z, W) = \sum_{V_i \in X, Z, W} \tilde{a}_{V_i Y} V_i.$$

The objective of the optimization (i.e., the MSE) can then be written as

$$\begin{aligned} \mathbb{E}[Y - f(X, Z, W)]^2 &= \mathbb{E}\left[\sum_{V_i \in X, Z, W} (a_{V_i Y} - \tilde{a}_{V_i Y}) V_i + \epsilon_Y \right]^2 \\ &= \mathbb{E}[\epsilon_Y^2] + \mathbb{E}\left[\sum_{V_i, V_j \in X, Z, W} (a_{V_i Y} - \tilde{a}_{V_i Y})(a_{V_j Y} - \tilde{a}_{V_j Y}) V_i V_j \right] \\ &= 1 + (a_{VY} - \tilde{a}_{VY})^T \mathbb{E}[VV^T](a_{VY} - \tilde{a}_{VY}), \end{aligned}$$

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 \tilde{a}_{VY} , since we have that

$$\text{TV}_{x_0, x_1}(f) = (\mathbb{E}[V | x_1] - \mathbb{E}[V | x_0])^T \tilde{a}_{VY}.$$

We write

$$c = \mathbb{E}[V | x_1] - \mathbb{E}[V | x_0], \tag{A.32}$$

$$\Sigma = \mathbb{E}[VV^T] \tag{A.33}$$

and note that our optimization problem can be written as

$$\arg \min_{\tilde{a}_{VY}} (a_{VY} - \tilde{a}_{VY})^T \Sigma (a_{VY} - \tilde{a}_{VY}) \tag{A.34}$$

$$\text{subject to } c^T \tilde{a}_{VY} = 0. \tag{A.35}$$

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 $t = \Sigma^{\frac{1}{2}}(a_{VY} - \tilde{a}_{VY})$), the solution can be derived explicitly as

$$\hat{a}_{VY} = a_{VY} - \frac{c^T a_{VY} \Sigma^{-1} c}{c^T \Sigma^{-1} c}.$$

We next analyze the constraints

$$\text{Ctf-DE}_{x_0, x_1}(\hat{f}_{\text{fair}} | x_0) = \text{Ctf-IE}_{x_1, x_0}(\hat{f}_{\text{fair}} | x_0) = \text{Ctf-SE}_{x_1, x_0}(\hat{f}_{\text{fair}}) = 0.$$

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 $c_1 = (1, 0, \dots, 0)^T$. 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 $U \leftarrow\!\!\rightarrow X$ 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

$$\mathbb{E}[V_i V_j] = \sum_{\substack{\text{treks } T_s \\ \text{from } V_i \text{ to } V_j}} \lambda(T_s) \prod_{\substack{\text{edges } V_k \rightarrow V_l \\ \in T_s}} a_{V_k V_l}$$

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

$$c_1^T \left(a_{VY} - \frac{c^T a_{VY} \Sigma^{-1} c}{c^T \Sigma^{-1} c} \right) = 0, \tag{A.36}$$

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

$$\text{deg}(c_1^T (a_{VY} - \frac{c^T a_{VY} \Sigma^{-1} c}{c^T \Sigma^{-1} c})) \leq \text{deg}(c_1) + \text{deg}(a_{VY}) + \text{deg}(\frac{c^T a_{VY} \Sigma^{-1} c}{c^T \Sigma^{-1} c}) \tag{A.37}$$

and also that

$$\text{deg}(\frac{c^T a_{VY} \Sigma^{-1} c}{c^T \Sigma^{-1} c}) \leq \text{deg}(c^T a_{VY} \Sigma^{-1} c) + \text{deg}(c^T \Sigma^{-1} c) \tag{A.38}$$

$$\leq 2\text{deg}(c) + \text{deg}(a_{VY}) + \text{deg}(\Sigma^{-1}) + 2\text{deg}(c) + \text{deg}(\Sigma^{-1}). \tag{A.39}$$

Now, one can observe the following bounds, where $p = |V|$:

$$\text{deg}(c) \leq p \text{ from Eq. A.32,} \tag{A.40}$$

$$\text{deg}(a_{VY}) = 1 \text{ by definition,} \tag{A.41}$$

$$\text{deg}(\Sigma^{-1}) \leq p^2 \cdot \max_{i,j} \text{deg}(\Sigma_{ij}) = p^4 \text{ from Eq. A.33.} \tag{A.42}$$

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

$$\inf_{z \in Z} \|x - z\|_2 \leq k_1 |f(x)|^{k_2} \quad \forall x \in K. \tag{A.43}$$

Therefore, there exist constants k_1, k_2 such that:

$$\text{vol}(\mathcal{H}^{DE}(\epsilon)) = \text{vol}\{a \in [-1, 1]^{|E|} \mid |c_1^T(a_{VY} - \frac{c^T a_{VY} \Sigma^{-1} c}{c^T \Sigma^{-1} c})| \leq \epsilon\} \tag{A.44}$$

$$= \text{vol}\{a \in [-1, 1]^{|E|} \mid d(a, \mathcal{H}^{DE}(0)) \leq k_1 \epsilon^{k_2}\}, \tag{A.45}$$

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

$$\text{vol}(\mathcal{H}^{DE}(\epsilon)) \leq k_1 \epsilon^{k_2} |E|^{|E|/2} C(|E|) \text{deg}(\mathcal{H}^{DE}(0)), \tag{A.47}$$

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 $\mathcal{S}_{n_Z, n_W}^{linear}$ 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

$$\epsilon = \left(\frac{2^{p(p+1)}}{8k_1 C(|E|)(p+1)^2(p(p+1))^{\frac{p(p+1)}{2}}} \right)^{1/k_2} \tag{A.49}$$

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 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:

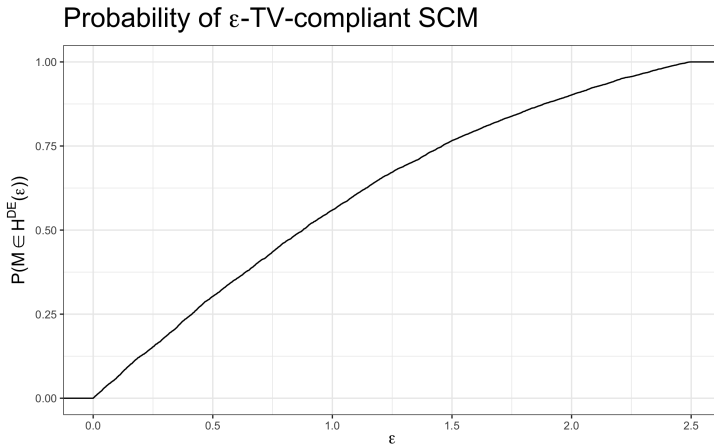


Figure A.1: Estimating empirically the probability that a random SCM in $\mathcal{S}_{n_Z, n_W}^{linear}$, for $n_Z = n_W = 5$, has a direct effect smaller than ϵ after ensuring that TV equals 0.

$$X \leftarrow f_X(u_x, u_z) \tag{A.54}$$

$$Z \leftarrow f_Z(u_x, u_z) \tag{A.55}$$

$$W \leftarrow f_W(X, Z, u_w) \tag{A.56}$$

$$Y \leftarrow f_Y(X, Z, W, u_y). \tag{A.57}$$

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 $Z | X = x_1$ is transported onto $Z | X = x_0$. Write τ^Z for the transport map. On the level of the SCM, this corresponds to extending the equations by an additional mechanism

$$\tilde{Z} \leftarrow \begin{cases} f_Z(u_x, u_z) & \text{if } f_X(u_x, u_z) = x_0 \\ f_Z(\pi^Z(u_x, u_z)) & \text{if } f_X(u_x, u_z) = x_1 \end{cases}. \tag{A.58}$$

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:

$$\begin{aligned} \pi^Z &:= \arg \min_{\pi} \int_{\mathcal{U}_X \times \mathcal{U}_Z} \|f_Z(\pi(u_z, u_x)) - f_Z(u_z, u_x)\|^2 du_{xz}^{X=x_1} \\ \text{s.t.} \quad & f_Z(\pi(u_z, u_x)) \stackrel{d}{=} f_Z(u_z, u_x) \quad u_x, u_z \sim U_{X, U_Z | X=x_1} \end{aligned} \tag{A.59}$$

The measure $du_{xz}^{X=x_1}$ in the objective is the probability measure associated with the distribution $P(u_x, u_z | X = x_1)$. The constraint ensures that after the transport, $\tilde{Z} | X = x_1$ is equal in distribution to $\tilde{Z} | X = x_0$. In the second step of the procedure, we are transporting the distribution of W . This results in adding the mechanism:

$$\tilde{W} \leftarrow \begin{cases} f_W(x_0, \tilde{Z}, u_w) & \text{if } X = x_0 \\ f_W(x_0, \tilde{Z}, \pi^W(u_w)) & \text{if } X = x_1 \end{cases} \tag{A.60}$$

Similarly as for π^Z , π^W is a mapping that solves following optimization problem:

$$\begin{aligned} \pi^W &:= \arg \min_{\pi} \int_{\mathcal{U}_W} \|f_W(x_0, \tilde{z}, \pi(u_w)) - f_W(x_1, \tilde{z}, u_w)\|^2 du_w \\ \text{s.t.} \quad & f_W(x_0, \tilde{z}, \pi(u_w)) \stackrel{d}{=} f_W(x_0, \tilde{z}, u_w). \end{aligned} \tag{A.61}$$

The above optimization problem is thought of being solved separately for each value of $\tilde{Z} = \tilde{z}$. Finally, in the last step, we are constructing the additional mechanism:

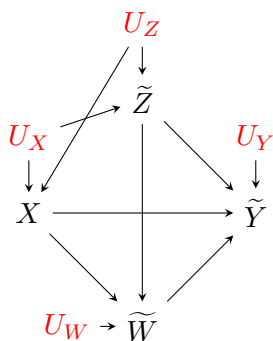
$$\tilde{Y} \leftarrow \begin{cases} f_Y(x_0, \tilde{Z}, \tilde{W}, u_y) & \text{if } X = x_0 \\ f_Y(x_0, \tilde{Z}, \tilde{W}, \pi^Y(u_y)) & \text{if } X = x_1 \end{cases} \tag{A.62}$$

Again, the implicit mapping π^Y is constructed so that it is the solution to

$$\begin{aligned} \pi^Y &:= \arg \min_{\pi} \int_{\mathcal{U}_Y} \|f_Y(x_0, \tilde{z}, \tilde{w}, \pi(u_y)) - f_Y(x_1, \tilde{z}, \tilde{w}, u_y)\|^2 du_y \\ \text{s.t.} \quad & f_Y(x_0, \tilde{z}, \tilde{w}, \pi(u_y)) \stackrel{d}{=} f_Y(x_0, \tilde{z}, \tilde{w}, u_y). \end{aligned} \tag{A.63}$$

where the problem is solved separately for each fixed choice of parents $\tilde{Z} = \tilde{z}$, $\tilde{W} = \tilde{w}$.

After constructing the additional mechanisms \tilde{Z} , \tilde{W} , and \tilde{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 $X, \tilde{Z}, \tilde{W}, \tilde{Y}$. Therefore, it follows that the identification expressions for the spurious, indirect, and direct effects are known, and given by:

$$x\text{-DE}_{x_0, x_1}(\tilde{y} \mid x_0) = \sum_{\tilde{z}, \tilde{w}} [P(\tilde{y} \mid x_1, \tilde{z}, \tilde{w}) - P(\tilde{y} \mid x_0, \tilde{z}, \tilde{w})] P(\tilde{w} \mid x_0, \tilde{z}) P(\tilde{z} \mid x) \tag{A.64}$$

$$x\text{-IE}_{x_0, x_1}(\tilde{y} \mid x_0) = \sum_{\tilde{z}, \tilde{w}} P(\tilde{y} \mid x_0, \tilde{z}, \tilde{w}) [P(\tilde{w} \mid x_1, \tilde{z}) - P(\tilde{w} \mid x_0, \tilde{z})] P(\tilde{z} \mid x) \tag{A.65}$$

$$x\text{-SE}_{x_1, x_0}(\tilde{y}) = \sum_{\tilde{z}} P(\tilde{y} \mid x_1, \tilde{z}) [P(\tilde{z} \mid x_0) - P(\tilde{z} \mid x_1)]. \tag{A.66}$$

To finish the proof, notice that by construction (the matching of distributions via optimal transport), we have that

$$P(\tilde{y} \mid x_1, \tilde{z}, \tilde{w}) = P(\tilde{y} \mid x_0, \tilde{z}, \tilde{w}) \tag{A.67}$$

$$P(\tilde{w} \mid x_1, \tilde{z}) = P(\tilde{w} \mid x_0, \tilde{z}) \tag{A.68}$$

$$P(\tilde{z} \mid x_0) = P(\tilde{z} \mid x_1), \tag{A.69}$$

implying that all three effects in Eq. A.64-A.66 are equal to 0 (the argument for showing that $x\text{-DE}_{x_1, x_0}(\tilde{y} \mid x_0)$ and $x\text{-DE}_{x_1, x_0}(\tilde{y} \mid x_0)$ 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 $\mu(\hat{y}) = 0$ implies that

$$\sum_u [\hat{y}_{C_1}(u) - \hat{y}_{C_0}(u)]P(u | E) = 0. \tag{A.70}$$

For part (a), assume that the policy D is a linear function of \hat{Y} , i.e., $f_D(\hat{y}) = a\hat{y} + b$. Then we simply have that:

$$\mu(d) = \sum_u [d_{C_1}(u) - d_{C_0}(u)]P(u | E) \tag{A.71}$$

$$= a \cdot \sum_u [\hat{y}_{C_1}(u) - \hat{y}_{C_0}(u)]P(u | E) \tag{A.72}$$

$$= a\mu(\hat{y}) = 0. \tag{A.73}$$

For part (b), assume that the measure μ is a unit-level measure (the event $E = \{U = u\}$). Then, the fairness condition implies that $\hat{y}_{C_1}(u) = \hat{y}_{C_0}(u) \forall u$, from which it follows that

$$d_{C_1}(u) = f_D(\hat{y}_{C_1}(u)) = f_D(\hat{y}_{C_0}(u)) = d_{C_0}(u) \forall u. \tag{A.74}$$



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:

$$P(y_{d_0} = 0, y_{d_1} = 1 | w, x) = P(U_Y - \frac{w}{5} < 0.5, U_Y + \frac{w}{3} - \frac{w}{5} > 0.5) \tag{A.75}$$

$$= P(U_Y < 0.5 + \frac{w}{5}, U_Y > 0.5 + \frac{w}{5} - \frac{w}{3}) \tag{A.76}$$

$$= P(0.5 + \frac{w}{5} - \frac{w}{3} < U_Y < 0.5 + \frac{w}{5}) \tag{A.77}$$

$$= \frac{w}{3} \quad (\text{using } U_Y \sim \text{Unif}[0, 1]), \tag{A.78}$$

from which Eq. 5.143 follows.

A.7 Proof of Thm. 4.13 and Cor. 4.14

Proof. For the theorem proof, consider that:

$$\mathbb{E}(y \mid x_1, \hat{y}) - \mathbb{E}(y \mid x_0, \hat{y}) = \mathbb{E}(y_{x_1} \mid x_1, \hat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_0, \hat{y}_{x_0}) \quad (\text{A.79})$$

$$= \underbrace{\mathbb{E}(y_{x_1} \mid x_1, \hat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_1, \hat{y}_{x_1})}_{\text{Term (I)}} \quad (\text{A.80})$$

$$+ \underbrace{\mathbb{E}(y_{x_0} \mid x_1, \hat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_1, \hat{y}_{x_0})}_{\text{Term (II)}} \quad (\text{A.81})$$

$$+ \underbrace{\mathbb{E}(y_{x_0} \mid x_1, \hat{y}_{x_0}) - \mathbb{E}(y_{x_0} \mid x_0, \hat{y}_{x_0})}_{\text{Term (III)}}. \quad (\text{A.82})$$

Since by assumption no backdoor paths between X and Y, \hat{Y} exist, Term (III) vanishes. By noting that $\mathbb{E}(y_x \mid x_1, \hat{y}_{x_1}) = \mathbb{E}(y_x \mid x_1, \hat{y}) \ \forall x$ 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 \hat{Y} is efficient, i.e., $\hat{Y}(x, w) = \mathbb{E}[Y \mid x, w]$. In the linear case, the efficiency simply translates to the fact that

$$\alpha_{W\hat{Y}} = \alpha_{WY}, \quad (\text{A.83})$$

$$\alpha_{X\hat{Y}} = \alpha_{XY}. \quad (\text{A.84})$$

Due to linearity, for every unit u , we have that

$$y_{x_1}(u) - y_{x_0}(u) = \alpha_{XW}\alpha_{WY} + \alpha_{XY}, \quad (\text{A.85})$$

and since Term (I) can be written as $\sum_u [y_{x_1}(u) - y_{x_0}(u)]P(u \mid x_1, \hat{y})$, Eq. 4.222 follows. We next look at Term (II), which can be expanded as

$$\sum_u \hat{y}_{x_0}(u) [P(u \mid \hat{y}_{x_1}) - P(u \mid \hat{y}_{x_0})]. \quad (\text{A.86})$$

We now look at units u which are compatible with $\hat{Y}_{x_1}(u) = \hat{y}$ and $\hat{Y}_{x_0}(u) = \hat{y}$. We can expand $\hat{Y}_{x_1}(u)$ as

$$\hat{Y}_{x_1}(u) = \alpha_{X\hat{Y}} + \alpha_{XW}\alpha_{W\hat{Y}} + \alpha_{W\hat{Y}}u_W. \quad (\text{A.87})$$

Thus, we have that

$$\hat{Y}_{x_1}(u) = \hat{y} \implies \alpha_{W\hat{Y}}u_W = \hat{y} - \alpha_{X\hat{Y}} + \alpha_{XW}\alpha_{W\hat{Y}}. \quad (\text{A.88})$$

Similarly, we also obtain that

$$\hat{Y}_{x_0}(u) = \hat{y} \implies \alpha_{W\hat{Y}} u_W = \hat{y}. \quad (\text{A.89})$$

Due to the efficiency of learning which implies that $\alpha_{W\hat{Y}} = \alpha_{WY}$ and $\alpha_{X\hat{Y}} = \alpha_{XY}$, Eq. A.88 and A.89 imply

$$y_{x_0}(u) = \hat{y} - (\alpha_{XY} + \alpha_{XW}\alpha_{WY}) \quad \forall u \text{ s.t. } \hat{Y}_{x_1}(u) = \hat{y}, \quad (\text{A.90})$$

$$y_{x_0}(u) = \hat{y} \quad \forall u \text{ s.t. } \hat{Y}_{x_0}(u) = \hat{y}, \quad (\text{A.91})$$

which in turn shows that

$$\mathbb{E}(y_{x_0} \mid \hat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid \hat{y}_{x_0}) = -\alpha_{XY} - \alpha_{XW}\alpha_{WY}. \quad (\text{A.92})$$

■

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 \tilde{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 \bar{D}^{UT} with a higher or equal utility that takes the form

$$\bar{D}^{UT} = \begin{cases} 1 & \text{if } \Delta(x_1, z, w) > \delta^{(x_1)'}, \\ 1 & \text{if } \Delta(x_0, z, w) > \delta^{(x_0)'}, \\ 0 & \text{otherwise.} \end{cases} \tag{A.93}$$

with $\delta^{(x_0)'}$, $\delta^{(x_1)'}$ non-negative (otherwise, the policy can be trivially improved). In words, for any policy \bar{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 \bar{D}^{UT} is using a larger budget than D^{UT} . However, D^{UT} uses a budget of b , making \bar{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 \bar{D}^{UT} introduces a larger disparity than D^{UT} . Since the disparity of D^{UT} is at least M , the disparity of \bar{D}^{UT} is strictly greater than M . Further, note that $\delta^{(x_0)'} > \delta^{(x_0)} \geq 0$, showing that \bar{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 \bar{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

$$\mathbb{E}[\Delta \mathbb{1}(\Delta \in [\delta^{(x_1)}, \delta^{(x_1)'})] \mathbb{1}(X = x_1)] \tag{A.99}$$

$$- \mathbb{E}[\Delta \mathbb{1}(\Delta \in [\delta^{(x_0)'}, \delta^{(x_0)}]) \mathbb{1}(X = x_0)] \tag{A.100}$$

$$\geq \delta^{(x_1)} \mathbb{E}[\mathbb{1}(\Delta \in [\delta^{(x_1)}, \delta^{(x_1)'})] \mathbb{1}(X = x_1)] \tag{A.101}$$

$$- \delta^{(x_0)} \mathbb{E}[\mathbb{1}(\Delta \in [\delta^{(x_0)'}, \delta^{(x_0)}]) \mathbb{1}(X = x_0)] \tag{A.102}$$

$$\geq \delta^{(x_0)}(\mathbb{E}[\mathbb{1}(\Delta \in [\delta^{(x_1)}, \delta^{(x_1)'}])\mathbb{1}(X = x_1)]) \quad (\text{A.103})$$

$$- \mathbb{E}[\mathbb{1}(\Delta \in [\delta^{(x_0)'}, \delta^{(x_0)}])\mathbb{1}(X = x_0)]) \quad (\text{A.104})$$

$$= \delta^{(x_0)}(P(\Delta \in [\delta^{(x_1)}, \delta^{(x_1)'}], x_1)) \quad (\text{A.105})$$

$$- P(\Delta \in [\delta^{(x_0)'}, \delta^{(x_0)}], x_0)) \quad (\text{A.106})$$

$$\geq 0, \quad (\text{A.107})$$

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 $|P(d | x_1) - P(d | x_0)| \leq M$ or treat everyone with a positive benefit in the disadvantaged group. ■

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}_{SFM} .

B.1.1 Identification under Extended Fairness Model

There are five possible bidirected edges that could be added to the \mathcal{G}_{SFM} (since the bidirected edge $X \leftrightarrow Z$ is assumed to be present already). The other five possibilities include the $Z \leftrightarrow Y$ (confounder-outcome), $W \leftrightarrow Y$ (mediator-outcome), $X \leftrightarrow W$ (attribute-mediator), $Z \leftrightarrow W$ (confounder-mediator) and $X \leftrightarrow Y$ (attribute-outcome). We analyze these cases in the respective order.

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} .

	Measure	ID expression
general	$TE_{x_0, x_1}(y)$	$\sum_z [P(y x_1, z) - P(y x_0, z)]P(z)$
	$Exp-SE_x(y)$	$\sum_z P(y x, z)[P(z) - P(z x)]$
	$NDE_{x_0, x_1}(y)$	$\sum_{z, w} [P(y x_1, z, w) - P(y x_0, z, w)]P(w x_0, z)P(z)$
	$NIE_{x_0, x_1}(y)$	$\sum_{z, w} P(y x_0, z, w)[P(w x_1, z) - P(w x_0, z)]P(z)$
x -specific	$ETT_{x_0, x_1}(y x)$	$\sum_z [P(y x_1, z) - P(y x_0, z)]P(z x)$
	$Ctf-SE_{x_0, x_1}(y)$	$\sum_z P(y x_0, z)[P(z x_0) - P(z x_1)]$
	$Ctf-DE_{x_0, x_1}(y x)$	$\sum_{z, w} [P(y x_1, z, w) - P(y x_0, z, w)]P(w x_0, z)P(z x)$
	$Ctf-IE_{x_0, x_1}(y x)$	$\sum_{z, w} P(y x_0, z, w)[P(w x_1, z) - P(w x_0, z)]P(z x)$
z -specific	$z-TE_{x_0, x_1}(y x)$	$P(y x_1, z) - P(y x_0, z)$
	$z-DE_{x_0, x_1}(y x)$	$\sum_w [P(y x_1, z, w) - P(y x_0, z, w)]P(w x_0, z)$
	$z-IE_{x_0, x_1}(y x)$	$\sum_w P(y x_0, z, w)[P(w x_1, z) - P(w x_0, z)]$

Bidirected edge $Z \leftrightarrow Y$. Consider the case of confounder-outcome confounding, represented by the $Z \leftrightarrow Y$ edge. An example of such a model is 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.

Table B.2: An example of the extended fairness model with a bidirected $Z \leftrightarrow Y$ edge (left side), in which refining the set of variables Z yields a graph (right side) in which all fairness measures are identifiable.

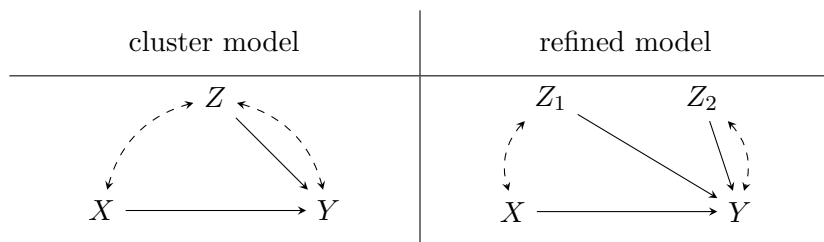
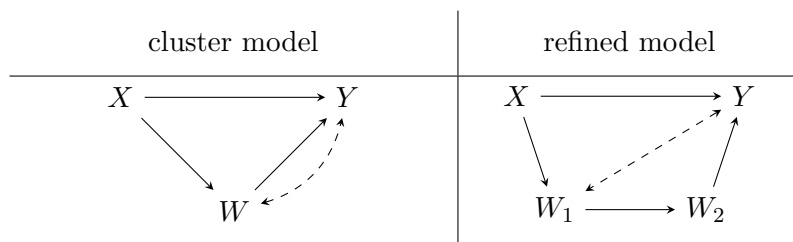


Table B.3: An example of the extended fairness model with a bidirected $W \leftrightarrow Y$ edge (left side), in which refining the set of variables W yields a graph (right side) in which all fairness measures are identifiable.



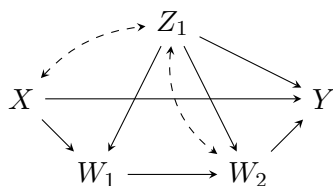
Bidirected edge $W \leftrightarrow Y$. Next consider the case where there is a bidirected edge between the group of variables W and the outcome Y . Firstly, we note that the identification of causal (TE/ETT) and spurious measures (Exp-SE/Ctf-SE) is unaffected by the $W \leftrightarrow Y$ edge, and that these quantities are identified by the same expressions as in Table B.1. The quantities 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 $X \leftrightarrow Y$. The case of the $X \leftrightarrow Y$ edge is similar to that of $W \leftrightarrow Y$, yet slightly different. None of the measures discussed are identifiable in this case, before refining the W set. However, similarly as in the $W \leftrightarrow Y$ 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.

Bidirected edge $Z \leftrightarrow W$. In the case of the $Z \leftrightarrow W$ edge, none of the measures are identifiable. However, refining the Z and W sets may help. To see an example, consider the following graph

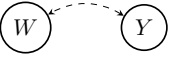
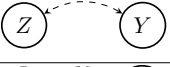





In this case, all of the measures of fairness in Table B.1 are identifiable, but again with different expressions than those presented in the table.

Bidirected edge $X \leftrightarrow Y$. The attribute-outcome confounding represented by the $X \leftrightarrow Y$ 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 $X \leftrightarrow Y$ 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.

Table B.4: Identification of causal fairness measures under latent confounding.

	✓	✓	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
	✗	✗	✗	✗

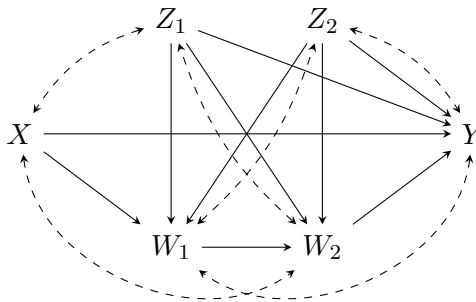


Figure B.1: Causal diagram compatible with the SFM with all bidirected arrows apart from $X \leftrightarrow Y$, in which all effects are identifiable.

We end with an example (see Fig. B.1) that fits the extended fairness model with all bidirected edges apart from the $X \leftrightarrow Y$, 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 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

$$\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}(X_i = x) Y_i}{\hat{p}(X_i | Z_i)}, \quad (\text{B.1})$$

where $\hat{p}(X_i | Z_i)$ is the estimate of the conditional probability $\mathbb{P}(X_i = x | Z_i)$, can be used. There is an additional assumption necessary for such an approach:

Definition B.1 (Positivity assumption). The positivity assumption holds if $\forall x, z, \mathbb{P}(X = x | Z = z)$ is bounded away from 0, that is

$$\delta < \mathbb{P}(X = x | Z = z) < 1 - \delta,$$

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

$$\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}(X_i = x)(Y_i - \hat{\mu}(Y_i | Z_i, X_i))}{\hat{p}(X_i | Z_i)} + \hat{\mu}(Y_i | Z_i, X_i), \quad (\text{B.2})$$

where $\hat{\mu}$ denotes the estimator of the conditional mean $\mathbb{E}[Y | Z = z, X = x]$. In fact, only one of the two estimators $\hat{\mu}(Y_i | Z_i, X_i)$ and $\hat{p}(X_i | Z_i)$ 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:

$$\begin{aligned} \phi_{x_0, x_1}(X, W, Z) &= \frac{\mathbb{1}(X = x_1)f(W | x_0, Z)}{p_{x_1}(Z)f(W | x_1, Z)}[Y - \mu(x_1, W, Z)] \\ &\quad + \frac{\mathbb{1}(X = x_0)}{p_{x_0}(Z)}[\mu(x_1, W, Z) - \int_{\mathcal{W}} \mu(x_1, w, Z)f(w | x_0, Z) dw] \\ &\quad + \int_{\mathcal{W}} \mu(x_1, w, Z)f(w | x_0, Z) dw. \end{aligned} \tag{B.3}$$

The estimator is given by $\frac{1}{n} \sum_{i=1}^n \hat{\phi}_{x_0, x_1}(X_i, W_i, Z_i)$, where in $\hat{\phi}$ the quantities $p_x(Z)$, $\mu(X, W, Z)$ and $f(W | X, Z)$ 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 $f(W | X, Z)$. 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 $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K$,
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

$$\frac{\mathbb{1}(X_i = x)(Y_i - \widehat{\mu}(Y_i \mid Z_i, X_i))}{\widehat{p}(X_i \mid Z_i)} + \widehat{\mu}(Y_i \mid Z_i, X_i), \quad (\text{B.4})$$

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

$$\begin{aligned} \psi_{x_0, x_1}(X, W, Z) &= \frac{\mathbb{1}(X = x_1)p_{x_0}(Z, W)}{p_{x_1}(Z, W)p_{x_0}(Z)} [Y - \mu(x_1, W, Z)] \\ &\quad + \frac{\mathbb{1}(X = x_0)}{p_{x_0}(Z)} [\mu(x_1, W, Z) - \mathbb{E}[\mu(x_1, W, Z) \mid X = x_0, Z]] \\ &\quad + \mathbb{E}[\mu(x_1, W, Z) \mid X = x_0, Z], \end{aligned} \quad (\text{B.5})$$

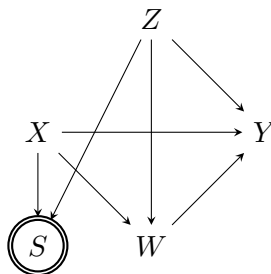
which avoids the computation of densities in a possibly high-dimensional case. The terms $\psi_{x_0, x_1}(X, W, Z)$ are estimated in a cross-fitting procedure as described above, with the slight extension in Step 2, where we further split the complement \mathcal{D}_k^C into two parts, to estimate the 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 monograph was concerned with the standard fairness model (SFM) from Def. 2.7. In the SFM, there is a bidirected edge $X \leftrightarrow Z$, 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



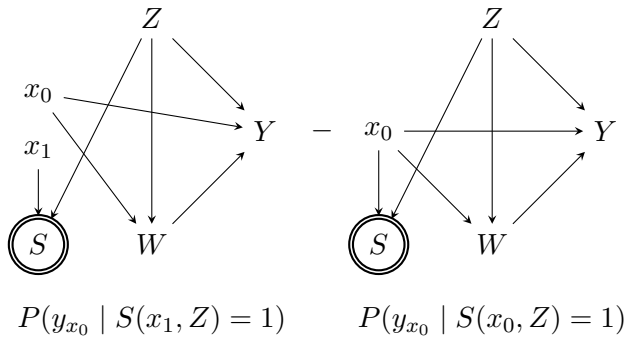


Figure C.1: Quantity $\text{Ctf-SBE}_{x_0, x_1}(y)$ represented graphically as a contrast.

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 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 $P(y | x_1) - P(y | x_0)$ 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:

$$\text{Ctf-SBE}_{x_0, x_1}(y) = P(y_{x_0} | S(x_1, Z) = 1) - P(y_{x_0} | S(x_0, Z) = 1). \tag{C.1}$$

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

$$Z | X = x_0 \text{ and } Z | X = x_1$$

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:

$$\text{Ctf-DE}_{x_0, x_1}(y | S_{x_0}) = P(y_{x_1, W_{x_0}} | S_{x_0}) - P(y_{x_0} | S_{x_0}) \quad (\text{C.2})$$

$$\text{Ctf-IE}_{x_0, x_1}(y | S_{x_0}) = P(y_{x_0, W_{x_1}} | S_{x_0}) - P(y_{x_0} | S_{x_0}). \quad (\text{C.3})$$

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:

$$\text{TV}_{x_0, x_1}(y) = \text{Ctf-DE}_{x_0, x_1}(y | S_{x_0}) - \text{Ctf-IE}_{x_1, x_0}(y | S_{x_0}) - \text{Ctf-SBE}_{x_1, x_0}(y) \quad (\text{C.4})$$

$$= \text{Ctf-SBE}_{x_0, x_1}(y) - \text{Ctf-DE}_{x_1, x_0}(y | S_{x_1}) + \text{Ctf-IE}_{x_0, x_1}(y | S_{x_1}) \quad (\text{C.5})$$

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:

$$\text{Ctf-DE}_{x_0, x_1}(y | S_{x_0}) = \sum_{z, w} [P^*(y | x_1, z, w) - P^*(y | x_0, z, w)] \quad (\text{C.6})$$

$$\cdot P^*(w | x_0, z) P^*(z | x_0)$$

$$\text{Ctf-IE}_{x_1, x_0}(y | S_{x_0}) = \sum_{z, w} P^*(y | x_1, z, w) \quad (\text{C.7})$$

$$\cdot [P^*(w | x_0, z) - P^*(w | x_1, z)] P^*(z | x_0)$$

$$\text{Ctf-SBE}_{x_1, x_0}(y) = \sum_z P^*(y | x_1, z) [P^*(z | x_0) - P^*(z | x_1)], \quad (\text{C.8})$$

where P^* is the observational distribution under selection bias, defined by

$$P^*(v) = P(v | S = 1). \quad (\text{C.9})$$

Proof. We prove the identification expression for the Ctf-SBE term, and the other two expressions follow from a similar argument. Note that:

$$P(y_{x_1} | S_x = 1) = \sum_z P(y_{x_1} | z, S_x = 1) P(z | S_x = 1). \quad (\text{C.10})$$

The first term within the sum can be expanded as:

$$P(y_{x_1} | z, S_x = 1) = P(y_{x_1} | z, x, S_x = 1) \quad Y_{x_1} \perp\!\!\!\perp X | Z, S_x \quad (\text{C.11})$$

$$= P(y_{x_1} | z, x, S = 1) \quad \text{Consistency Axiom} \quad (\text{C.12})$$

$$= P(y_{x_1} | z, x_1, S = 1) \quad Y_{x_1} \perp\!\!\!\perp X | Z, S \quad (\text{C.13})$$

$$= P(y | z, x_1, S = 1) \quad \text{Consistency Axiom} \quad (\text{C.14})$$

$$= P^*(y | z, x_1) \quad \text{by definition.} \quad (\text{C.15})$$

For the second term within the sum, we have that

$$P(z | S_x = 1) = P(z | S_x = 1, x) \quad Z \perp\!\!\!\perp X | S_x \quad (\text{C.16})$$

$$= P(z | S = 1, x) \quad \text{Consistency Axiom} \quad (\text{C.17})$$

$$= P^*(z | x) \quad \text{by definition.} \quad (\text{C.18})$$

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

$$P_1(V) = P_2(V) = P(V), \quad (\text{C.19})$$

that is, the observational distributions of \mathcal{M}_1 and \mathcal{M}_2 are the same. Then it follows that

$$\text{Ctf-DE}_{x_0, x_1}^{\mathcal{M}_1}(y | x_0) = \text{Ctf-DE}_{x_0, x_1}^{\mathcal{M}_2}(y | S_{x_0}) \quad (\text{C.20})$$

$$\text{Ctf-IE}_{x_1, x_0}^{\mathcal{M}_1}(y | x_0) = \text{Ctf-IE}_{x_1, x_0}^{\mathcal{M}_2}(y | S_{x_0}) \quad (\text{C.21})$$

$$\text{Ctf-SE}_{x_1, x_0}^{\mathcal{M}_1}(y) = \text{Ctf-SBE}_{x_1, x_0}^{\mathcal{M}_2}(y), \quad (\text{C.22})$$

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 $\text{Ctf-SE}_{x_1, x_0}(y)$ 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 $\text{Ctf-SBE}_{x_1, x_0}(y)$ 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 $X \leftrightarrow Z$ 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 monograph 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} is 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

$$\mathbb{E}[Y \mid X = x] = \mathbb{E}[Y] \quad \forall x. \tag{D.1}$$

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.,

$$\text{iTV}_{x_0, X}(y) = \mathbb{E}_{X \sim P(X)} [\mathbb{E}[Y \mid X] - \mathbb{E}[Y \mid X = x_0]], \tag{D.2}$$

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

$$\text{NDE}_{x_0,x}(y) = P(y_{W_{x_0,x}}) - P(y_{x_0}) \quad (\text{D.3})$$

$$\text{NIE}_{x_0,x}(y) = P(y_{W_{x,x_0}}) - P(y_{x_0}) \quad (\text{D.4})$$

$$\text{Exp-SE}_x(y) = P(y \mid x) - P(y_x), \quad (\text{D.5})$$

and further analogues can be written for x, z , or v' -specific measures of direct / indirect effects. In case a single measure¹ is of interest instead of a collection of measures, we may consider measures such as

$$\text{iNDE}_{x_0,X}(y) = \mathbb{E}_{X \sim P(X)}[\text{NDE}_{x_0,X}(y)] \quad (\text{D.6})$$

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

$$\text{iTV}_{x_0,X}(y) = \text{iNDE}_{x_0,X}(y) - \text{iNIE}_{X,x_0}(y) \quad (\text{D.7})$$

$$+ \text{iExp-SE}_X(y) - \text{Exp-SE}_{x_0}(y). \quad (\text{D.8})$$

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.,

$$\text{Str-DE} = 0 \implies \text{iNDE}_{x_0,X}(y) = 0 \quad (\text{D.9})$$

$$\text{Str-IE} = 0 \implies \text{iNIE}_{x_0,X}(y) = 0 \quad (\text{D.10})$$

$$\text{Str-SE} = 0 \implies \text{iExp-SE}_X(y) = 0. \quad (\text{D.11})$$

¹One 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.

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 $X = x_0$.
- (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 $x \in \mathcal{X}$, $\text{PPM}_{x_0,x}(y) = P(y | x, \hat{y}) - P(y | x_0, \hat{y})$. 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

$$\mathbb{E}[y_{x,W_{x_0}} | E] - \mathbb{E}[y_{x_0} | E] = \mathbb{E}[\hat{y}_{x,W_{x_0}} | E] - \mathbb{E}[\hat{y}_{x_0} | E] \quad \forall x \in \mathcal{X}, E. \quad (\text{D.12})$$

- (6) In the context of decision-making, the Benefit Fairness criterion (Def. 5.10) can be adapted to require that

$$P(d | x, \Delta = \delta) = P(d | x_0, \Delta = \delta) \quad \forall x \in \mathcal{X}, \delta. \quad (\text{D.13})$$

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

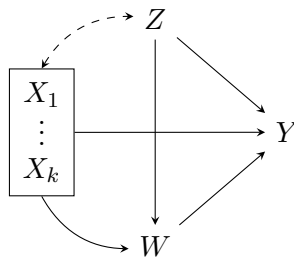
$$\mathbb{E}(y_{x,W_{x_0},d_1} - y_{x,W_{x_0},d_0} | x, z, w) = \mathbb{E}(y_{x_0,d_1} - y_{x_0,d_0} | x, z, w) \quad \forall x, z, w \quad (\text{D.14})$$

$$P(d | \Delta, x_0) = P(d | \Delta, x_1) \quad \forall x, \delta. \quad (\text{D.15})$$

As the above reasoning shows, extending the results of the monograph 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 monograph.

Multiple Protected Attributes. Finally, we mention how one may wish to handle multiple protected attributes X_1, \dots, X_k . Firstly, we will only consider the case in which the attributes X_1, \dots, 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}_{SFM} over endogenous variables $\{X_1, \dots, X_k, Z, W, Y\}$ and given by



The cluster $\{X_1, \dots, X_k\}$ allows for arbitrary causal or confounding relationships between the variables X_1, \dots, 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 $\mathcal{X}_1, \dots, \mathcal{X}_k$ be the domains of X_1, \dots, X_k , respectively. Then, we define the product protected attribute as $X^p = (X_1, \dots, X_k)$ taking values in $\mathcal{X}^p = \mathcal{X}_1 \times \dots \times \mathcal{X}_k$. 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, \dots, 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, \dots, X_k would be to consider X_1, \dots, X_k one-by-one, and perform the analyses described in the monograph 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) 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.

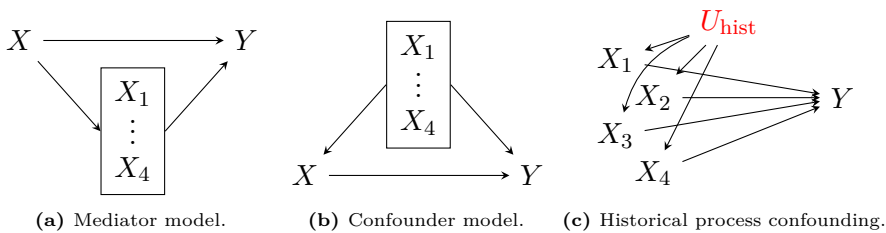


Figure D.1: Modeling options for religion as a bundle of sticks.

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* $X \leftrightarrow Z$ 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, \dots, 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, \dots, 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, \dots, X_4 ? Or, alternatively, should one choose

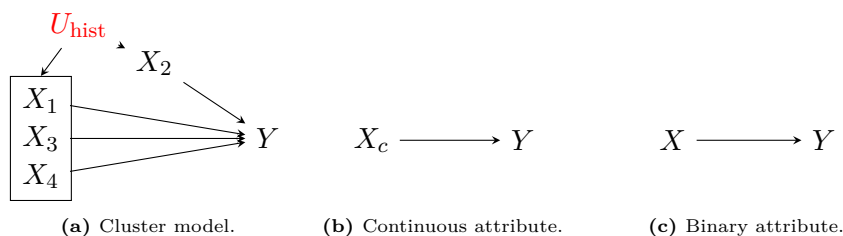


Figure D.2: Modeling options for religion as a bundle of sticks.

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, 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.

²This 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).

³For instance, such manipulations can be conceptualized as a person “writing a different value on their application”.

Another option would be to 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.

Unobserved Constitutive Features and Soft Interventions. Consider now the case of the causal diagram in Fig. D.2a, with X_1, \dots, 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 monograph. 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

$$P(Y_{X=x_0} = 1) = P(Y = 1 \mid X = x_0) \quad (\text{D.16})$$

$$= \int_{[0, \frac{1}{2}]} P(Y = 1 \mid X_c = x_c, X = x_0) f_{X_c \mid X=x_0}(x_c) dx_c \quad (\text{D.17})$$

$$= \int_{[0, \frac{1}{2}]} 2P(Y = 1 \mid X_c = x_c, X = x_0) dx_c \quad (\text{D.18})$$

$$= P(Y = 1 \mid X_c \sim \text{Unif}[0, \frac{1}{2}]) \quad (\text{D.19})$$

$$= P(Y = 1 \mid do(X_c \sim \text{Unif}[0, \frac{1}{2}])) \quad (\text{D.20})$$

$$= P(Y = 1; \sigma_{X_c}) \quad (\text{D.21})$$

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:

$$\text{TE}_{x_0, x_1}(y) = P(y \mid do(X_c \sim \text{Unif}[\frac{1}{2}, 1])) - P(y \mid do(X_c \sim \text{Unif}[0, \frac{1}{2}])). \quad (\text{D.22})$$

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

$$P(Y_{X=x_0} = 1) = P(Y = 1 \mid X = x_0) \quad (\text{D.23})$$

$$= \int_{f_X^{-1}(x_0)} P(Y = 1 \mid X = x_0, X_c = x_c) f_{X_c \mid X=x_0}(x_c) dx_c \quad (\text{D.24})$$

$$= \int_{f_X^{-1}(x_0)} P(Y = 1 \mid X_c = x_c) f_{X_c \mid X=x_0}(x_c) dx_c \quad (\text{D.25})$$

$$= P(Y = 1 \mid X_c \sim F_{X_c \mid X=x_0}) \quad (\text{D.26})$$

$$= P(Y = 1 \mid X_c \sim do(F_{X_c \mid X=x_0})) \quad (\text{D.27})$$

$$= P(Y = 1; \sigma_{X_c}), \quad (\text{D.28})$$

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 uniform 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

D.1. *On the Semantics of Manipulating the Protected Attribute* 271

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, 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:

$$X \leftarrow \text{Bernoulli}(0.5) \quad (\text{D.29})$$

$$D \leftarrow \text{Bernoulli}(0.5 + \alpha X) \quad (\text{D.30})$$

$$Y \leftarrow (0.1 + \beta X + \gamma D + \delta XD). \quad (\text{D.31})$$

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 $x_0 \rightarrow x_1$ and $x_1 \rightarrow x_0$ transitions:

$$\text{NDE}_{x_0, x_1}(y) = \beta + \frac{\delta}{2} \quad (\text{D.32})$$

$$\text{NDE}_{x_1, x_0}(y) = \beta + \frac{\delta}{2} + \alpha\delta. \quad (\text{D.33})$$

Notice that if either $\alpha = 0$, or $\delta = 0$, the two NDEs are the same. In fact, a hypothesis test

$$H_0 : \text{NDE}_{x_0, x_1}(y) = \text{NDE}_{x_1, x_0}(y) \quad (\text{D.34})$$

is a test for the existence of an interaction between direct and indirect pathways. In fact, the difference between the two NDEs

$$\text{NDE}_{x_1, x_0}(y) - \text{NDE}_{x_0, x_1}(y) = \alpha\delta \quad (\text{D.35})$$

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. \square

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 monograph 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

$$P(y \mid x_1, z, w) - P(y \mid x_0, z, w) \neq 0. \quad (\text{E.1})$$

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

$$(x, z, w)\text{-DE}_{x_0, x_1}(y \mid x_0, z, w) = P(y_{x_1, W_{x_0}} \mid x_0, z, w) - P(y_{x_0} \mid x_0, z, w), \quad (\text{E.2})$$

which measures the direct effect of a $x_0 \rightarrow x_1$ 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

$$(x, z, w)\text{-DE}_{x_0, x_1}(y \mid x_0, z, w) \neq 0 \implies \text{Str-DE} \neq 0. \quad (\text{E.3})$$

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

$$\text{NIE}_{x_0, x_1}(y) \neq 0 \implies \text{Str-IE} \neq 0, \quad (\text{E.4})$$

$$\text{Exp-SE}_x(y) \neq 0 \implies \text{Str-SE} \neq 0. \quad (\text{E.5})$$

Once again, this allows one to make qualitative claims about the decision-making process (in particular, $\text{Str-IE} \neq 0$ implies mediators are used as an input to the mechanism f_Y , and that the mediators are affected by the protected attribute X ; $\text{Str-SE} \neq 0$ 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

$$\text{TV}_{x_0, x_1}(y) = x\text{-DE}_{x_0, x_1}(y | x_0) - x\text{-IE}_{x_1, x_0}(y | x_0) - x\text{-SE}_{x_1, x_0}(y). \quad (\text{E.6})$$

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 $x\text{-DE}$); whether the attribute influences the mediators, which are then used in decision-making (term $x\text{-IE}$); and whether the attribute has common variations with the confounders, which are used in decision-making (term $x\text{-SE}$). 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. 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”. URL: <https://www.washingtonpost.com/technology/2019/12/19/federal-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.
- 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. 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. Boccoard. (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.