
Class of Science

Ph.D. in Data Science

XXXVI cycle

# Causality for Fair Machine Learning: Selected Topics and Applications

Scientific Disciplinary Area **INF/01**

**Candidate**
José M. Álvarez

**Supervisor**
Prof. Salvatore Ruggieri

Academic year 2023–2024

Thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy in Data Science

*Para mi nonna Sophia Gambaro,*
*mis avós María Milagros Vázquez y Manuel Álvarez,*
*y nuestras patrias portátiles.*

But you said we have a situation.

I didn't say it. The computer did. The whole system says it. [...] This doesn't mean anything is going to happen to you as such, at least not today or tomorrow. It just means you are the sum of your data. No man escapes that.

---

*White Noise* by Don Dellilo

# Summary

This thesis is about how we can use causality, in particular, in the form of structural causal models (SCM), to address fair machine learning (Fair ML) problems. We use SCM as auxiliary, declarative knowledge to contextualize and, in turn, enhance the formulation of such problems. We focus on automated decision-making (ADM) scenarios, in which a learned ML model, trained on historical data, is tasked with predicting the outcomes of incoming data. We address the following topics and applications.

*How can we use causal reasoning to better test for discrimination?* Based on the comparative nature of discrimination testing, the contribution to this question is twofold. First, we revisit the comparator used for testing the complainant's discrimination claim. Defining the comparator is at the center of all modeling tools for testing discrimination. We define two classes of comparators: the *ceteris paribus* (cp) comparator that represents an idealized comparison, and the *mutatis mutandis* (mm) comparator that represents a "fairness given the difference" comparison. Second, we propose counterfactual situation testing (CST), a new algorithmic tool for testing discrimination that uses the mm-comparator. Using a k-NN implementation, we compare CST to its standard counterpart that uses the cp-comparator.

*How can we use causal reasoning to operationalize subjective fairness?* The contribution to this question is the causal perception (CP) framework, in which we use SCM to represent how two individual agents interpret the same information differently. Perception is overlooked in Fair ML since we often consider a single, objective problem formulation. Further, many Fair ML applications disregard the risk of perception by assuming that all agents use these applications in the same way. Instead, with CP we propose a partial, subjective formulation of Fair ML problems in which decision-makers reason and decide differently on the same fairness problem or when using the same Fair ML application.

*How can we use causal reasoning to mitigate the bias from unrepresentative training data?* We use SCM to formalize the problem of unrepresentative data, both as a sample selection bias and domain adaptation problem, and motivate the use of weights to correct for the bias. The contribution to this question is twofold with a focus on data science applications. First, we revisit partial dependence plots (PDP) and modify this visualization tool proposing the weighted PDP (WPDP) as a solution. Under WPDP, the weights are used to correct for the contribution of each instance according to the underlying population distribution when drawing the plots. Second, we revisit the decision tree learning problem proposing a modification to the information gain split criterion and leading to what we define as domain adaptive decision trees (DADT). Under DADT, the entropy contribution for each instance when deciding the next tree split is weighted according to the target population distribution.

# Acknowledgements

This thesis started back home in Caracas and after jumping around between Pedder Bay, Gainesville, Toulouse, and Brussels, among other places, somehow ends in Pisa. Many people have made this thesis possible. It really takes a village.

I would like to start by thanking my advisor, Prof. Salvatore Ruggieri, for his kindness, encouragement, and guidance throughout the PhD. Grazie mille, Salvatore, for betting on me. I look forward to more future collaborations, now as colleagues. Similarly, I would like to thank my PhD Panel—Prof. Bettina Berendt, Prof. Carlos Castillo, Prof. Giovanni Comande, and Prof. Stan Matwin—as well as Prof. Franco Turini for supporting my work from its early stages. I would also like to thank Prof. Mykola Pechenizkiy and Prof. Joshua Loftus for reviewing this thesis and their feedback.

Speaking of mentorship, my path here has been far from linear. I would like to thank all of those mentors that brought me here. Thanks, in particular, to Dr. Nikos Skantzos, Prof. Eric Gautier, Dr. Shruti Sinha, Louise Strachan, the late Prof. Lawrence Kenny, Prof. Howard Louthan, Prof. Renata Serra, the late Bill Kolb, Samuel Pérez, Lee Fisher, José Correa, Reinaldo Solórzano, and Lily Bailer.

I am grateful to the NoBIAS – Artificial Intelligence without Bias Marie Skłodowska-Curie Innovative Training Network for funding my research. Special thanks to Antonio Bencini and Dr. Klaus Broelemann for hosting me. Thanks to all the ESRs, especially, to my co-authors Kristen Scott and Carlos Mougan.

During my PhD, I split my time between Pisa, Florence, and whatever city the next conference/workshop dictated. A big thank you to Laura State and Andrea Pugnana for four years of travels, discussions, and friendship; the colleagues from the 36th cycle for all the "un caffè?" after lunch; and Martina Cinquini for putting up with my broken Italian. You all made my commute from Florence to Pisa worth it. Thanks to the EUI Squadra for the times on and off the football pitch and to the EUI Economists for treating me like one of their own and making Florence my home. Thanks also to all of the colleagues I have met during conferences/workshops—especially to the EWAF 2023 crew: Christoph, Michele, Alessandro, Meike, and Corinna.

This thesis would not have been possible without my friends and family scattered between Venezuela, North America, and Europe: se les quiere y extraña mucho. Above all, I want to thank my mother, Tibaire, and my sister, Gabriela. This thesis is theirs as much as it is mine. Finally, I want to thank my partner, Marina. As with many other things in my life, I could not have done any of this without her.

Gracias totales,
José Manuel

# Contents

# Chapter 1

# Introduction

## 1.1 Prologue

Why do butchers wear white? The answer to this question has fascinated me since I first heard it in Prof. Jeffrey S. Adler's *US Urban History* (AMH 3460).[1] During his course, we studied the growth of American cities and, through them, the development of urban civilization as we know it today, at least, in the West. Cities, through their own social and physical structures, acted both as melting pots (e.g., through ports and multifamily housing) and cutting boards (e.g., though zoning laws and racial covenants), causing the fabric of the industrialized world. Suburbs, the anti-city response, did the same through their own social and physical structures. This course taught me that many of the aspects we take for granted in our daily, post-industrial, urban lives often summarize uncomfortable historical processes. It also taught me how imposed structure, from physical barriers to systematic policy interventions, leads to self-fulfilling prophecies.

The use of Machine Learning (ML) for automating, often critical, human decision-making processes has brought back attention to many of these taken-for-granted aspects. In the past decade, while we waited for the promises of automated decision-making (ADM) (e.g., [160, 197]), we encountered instead a range of disappointing algorithmic decision-makers (e.g., [17, 77, 138]). At first, we labeled these ML models as biased. Amazon's recruitment algorithm [77], for instance, was sexist because it penalized female CVs over similar male CVs. But then, when looking closer at these biased ML models, we could not separate them from the society that had trained them. Amazon's recruitment algorithm was trained on the company's male-dominated workforce. The question was not just "why is the algorithm unfair to female applicants?" but also "why is Amazon's workforce predominately male?" These biased ML models got more uncomfortable when we prohibit them from using sensitive information, like race and gender, as they managed still to infer and use it. Amazon's recruitment algorithm, even after training it on gender-neutral CVs, continued to be sexist by using other information associated to males and females. Now the question was not just "why is the algorithm still unfair to female applicants?" but also "why are the CVs used by the algorithm unable to provide neutral information?"

These ML models embody the omnipresence of social issues often believed outdated, like sexism, by showing that the downstream effects of these social issues, like inequality,

[1] https://history.ufl.edu/directory/jeffrey-adler/

are prevalent enough to motivate patterns of unfair decision-making. Clear examples to this argument include the "realization" that ML models can infer race from a zip code or gender from a resume [174]. Humans can infer that too [157]; we are just not as consistent nor scalable as a ML model. Further, humans can always outcast a shameful decision-maker. Doing the same with a ML model is much more difficult. For an object that summarizes thousands of past decisions, a ML model represents more of a social mirror than a possible scapegoat.

So why do butchers wear white, and why should I care? Regarding the first question, it was a uniform imposed by a newly created and growing middle class to signal discipline, restrain, and respect. In the USA of the 1800s, with rapid industrialization came the appearance of a new social class that wanted to join the wealthy class while distinguishing itself from the working class. One way of doing so, motivated in part by the Second Great Awakening, was to create a lifestyle centered around hard work. Starting and ending a work shift with a pristine white shirt was a sign of someone who belonged, like a manager, or of someone who aspired to belong, like a worker, to the rising middle class. This symbolism, as counterintuitive as it was, e.g., for butchers, permeated most jobs. A dirty white shirt for a butcher meant that he lacked the traits needed to someday be running the shop and, relative to his boss's clean white shirt, meant that he deserved to be at the bottom of the social hierarchy.[2]

Regarding the second question, in principle, you should not care as this thesis is not about symbolism in 1800s USA; however, I hope the reader can appreciate the parallels between a 19th century US manager and a 21st century ML model eager to draw meaning from seemingly neutral traits (like wearing a clean white shirt) about an individual's character (like work worthiness). These symbols, views, and practices from the past have created the patterns that the ML models are detecting today from data and using for ADM. A lot of them are patterns we take for granted, like butchers wearing white, but that reappear when we have to explain why the ML model has used the cleanliness of a worker's shirt to determine his or her next promotion. The ML model does not need to know what the cleanliness of the white shirt represents to act upon it, which makes it dangerous. Maybe the sole fact that the ML model is acting upon this pattern is what makes us aware of the meaning behind the cleanliness of the white shirt. In any case, once aware, meaning that the pattern can no longer be taken for granted, the question is then "how do we make the ML model as aware as us?"

This thesis is about fair machine learning and how causality as auxiliary knowledge, particularly in the form of structural causal models, can help us represent additional information about the data instances the potentially biased ML model classifies. It is about using causality to declare what we know or, at least, to acknowledge it in a way that is useful to the ML model. It is about confronting today these past patterns, their causes and their effects, as the drivers behind ADM.

[2]This explanation is based on my own recollection of Prof. Adler's course; however, several history books illustrate this pattern of cultural and class formation through symbols and practices intended for sorting individuals. See, e.g., Anbinder [15]'s *Five Points: The 19th-century New York City neighborhood that invented tap dance, stole elections, and became the world's most notorious slum*, or Prof. Adler's own *Murder in New Orleans: the creation of Jim Crow policing* [2].

## 1.2 Motivation, Challenges, and Contributions

Machine Learning (ML) models are increasingly used for automated decision-making (ADM) in, often, consequential settings such as resume selection [77], pre-trial bail [17], and government-aid allocation [138]. These ML models either guide or replace the original human decision-maker. Further, these ML models, having been trained on past data, often exhibit biased (as in unfair) decision-making toward individuals protected by non-discrimination law. All of this has motivated the study of fairness in ML, or Fair ML, as a field [13, 26, 244] and the increasing concerns by institutions and other stakeholders to regulate ADM under the prospect of algorithmic discrimination [92, 298].

Understanding what drives these biased ML models and the algorithms that power them has become central to both researchers and regulators. With many of these ML models illustrating past (or, arguably, ongoing) known biased patterns in our societies, such as a tech company not hiring enough women [77] or a conservative government being hostile to immigrant families [138], researchers and regulators have been facing retrospective questions to understand why these patterns are still relevant. In doing so, multiple fields, such as Computer Science, Sociology, and Law, have joined forces to better understand these socially loaded questions. The multidisciplinary approach to bias in ML models is illustrative of the need to have a more holistic and complex approach to fairness in ML models, often meaning the contextualization of the ML problem within its underlying social, economic, and historical forces [39, 128, 149, 150, 245]. To the ML research community such contextualization, in practice, has meant developing new ML frameworks able to represent and convey additional information to the ML problem formulation that is not necessarily captured by the data. Causal ML, or, broadly, causal reasoning for ML has emerged as one of these ML frameworks [254, 255].

Causality deals with formalizing *because* answers to *why* questions. Causality has a long tradition from Ancient Greece [303], an established reputation within Economics [16] and Law [176], and, recently, an increasing acceptance within ML researchers due to the works of Pearl [218] on structural causal models (SCM). These models allow us to represent cause-effect pairs that are understandable to ML models and intuitive to ML researchers. Causal ML, in particular through SCM, has become useful for understating, formalizing, and mitigating Fair ML problems [38, 181, 191]. Although it requires strong assumptions, modeling fairness in terms of causes and effects helps to contextualize the ML problem in ways in which we are able to address, e.g., concerns around discriminatory decision-making, knowledge representation, and data generation.

The broader focus of this thesis is on how we can use causality to better tackle Fair ML problems. In particular, we focus on the problems of algorithmic discrimination; representation of additional auxiliary information; and unrepresentative data due to non-random sampling. All of these topics, and their corresponding applications, share the use SCM to formalize context-specific knowledge. The following research questions are addressed in this thesis:

**Q1:** *How can we use causal reasoning to test for discrimination so that we capture the role of protected attributes, such as race and gender, on the other seemingly neutral attributes that are used for the decision-making process?*

**Q2:** *How can we use causal reasoning to formalize scenarios where fairness is, essentially, subjective as in dependent on who is making the decision?*

**Q3:** *How can we use causal reasoning to mitigate the potential bias in a learned model from using an unrepresentative sample as training data?*

We start with Chapter 2, in which we present the background knowledge for causal Fair ML. In this chapter, in particular, we take a critical view on key concepts, such as bias, fairness, and discrimination; discuss SCM and how they embody an interventionist view on causality; and make the case for using causality as auxiliary knowledge for guiding Fair ML. The background knowledge, as with the majority of this thesis, is specific to the European Union (EU) context as prescribed under EU non-discrimination law.

Chapters 3 and 4 address Q1. In Chapter 3, we revisit the discrimination comparator used for testing the discrimination claim made by a complainant. The comparator represents an individual profile sufficiently close to the individual profile of the complainant, in practice, meaning an individual that shares the same profile except membership to the group protected by non-discrimination law. Following Kohler-Hausmann [176]'s work on discrimination testing, we argue that the comparator is a counterfactual representation of the factual complainant. We further argue for the popularity of counterfactual reasoning in tools for testing discrimination. We then revisit the comparator by defining two kinds based on the type of counterfactual representation implemented.

The two kinds of comparators are the cp and mm comparators. The cp-comparator, or the *ceteris paribus* comparator, which is the standard comparator, represents an idealized counterfactual representation of the complainant: all non-protected attributes are the same, only the protected attribute changes. The mm-comparator, or the *mutatis mutandis* comparator, instead, represents a more flexible counterfactual representation of the complainant: the non-protected attributes, if needed, are adjusted according to the downstream effects produced by changing the protected attribute. Through the mm-comparator we present a causal critique against the standard way we test for discrimination [176]. We also present a comparator that embodies the EU non-discrimination law's goal of substantive equality [287].

In Chapter 4, we introduce counterfactual situation testing (CST), which is an extension to Thanh et al. [277]'s k-NN situation testing (ST) under a mm-comparator. In CST, we use the generated counterfactual distribution of all complainants to derive our comparators. Such distribution represents the "what would have been if" of the complainant under a non-protected status, including adjustments to the complainant's neutral attributes. Given that the mm-comparator is more flexible than the cp-comparator, we detect a higher number of discrimination cases under CST than ST using the same k-NN implementation. These results show the impact on testing for discrimination when changing how we view similarity between individuals through the comparators. We also explore the problem of multiple and intersectional discrimination [308], and show that a counterfactually fair [177] decision-maker can be discriminatory.

We address Q2 in Chapter 5, in which we formulate the problem of causal perception (CP). Perception refers to the situation in which to individuals interpret the same information differently. Although widely studied by cognitive psychologist [157, 158, 279, 280, 281], precisely as a source of bias in human decision-making, perception has been largely overlooked in Fair ML. Using SCM, we formalize the problem of perception under causal reasoning. We present the general problem and then argue for two kinds of CP, unfaithful and inconsistent, based on the causal properties of faithfulness [224] and consistency [241]. Finally, we make the case that CP will play an important

role in human-centered AI since we can expect for different individuals to make sense differently of the information provided by the same ADM model.

Finally, Chapter 6 addresses Q3. This chapter argues that the problem of sample selection bias, meaning the situation in which the training data has not been chosen randomly, can be understood using SCM. We then discuss two applications, weighted partial dependence plots (WPDP) (Section 6.2) and domain adaptive decision tress (DADT) (Section 6.3), which are solutions, respectively, to the problem of interpreting a ML model and learning a decision tree under unrepresentative training data.

Chapter 7 concludes this thesis with a discussion on future research directions; contributions; and limitations. Beyond that chapter, the appendices provide additional information corresponding to each chapter. Additionally, where needed, we provide the link to the code (in the form of a GitHub repository)[3] for each chapter.

Throughout the thesis, we use the first-person plural, which is the standard in scientific writing. In certain situations, though, in which I wish, for instance, to address the reader directly or express a personal opinion, I switch to the first-person singular.

## 1.3 Publications

I list below the journal and conference papers published by the time of writing:

- [13] J. M. Álvarez, A. Bringas-Colmenarejo, A. Elobaid, S. Fabbrizzi, M. Fahimi, A. Ferrara, S. Ghodsi, C. Mougan, I. Papageorgiou, P. Reyero, et al. Policy advice and best practices on bias and fairness in ai. *Ethics and Information Technology*, 26 (2):31, 2024.

- [8] J. M. Álvarez and S. Ruggieri. Counterfactual situation testing: Uncovering discrimination under fairness given the difference. In *EAAMO*, pages 2:1–2:11. ACM, 2023.

- [203] C. Mougan, J. M. Álvarez, S. Ruggieri, and S. Staab. Fairness implications of encoding protected categorical attributes. In *AIES*, pages 454–465. ACM, 2023.

- [12] J. M. Álvarez, K. M. Scott, B. Berendt, and S. Ruggieri. Domain adaptive decision trees: Implications for accuracy and fairness. In *FAccT*, pages 423–433. ACM, 2023.

- [245] S. Ruggieri, J. M. Álvarez, A. Pugnana, L. State, and F. Turini. Can we trust fair-AI? In *AAAI*, pages 15421–15430. AAAI Press, 2023.

- [178] M. Lazzari, J. M. Álvarez, and S. Ruggieri. Predicting and explaining employee turnover intention. *Int. J. Data Sci. Anal.*, 14(3):279–292, 2022.

I also include a list of papers under submission at the time of writing:

- [9] J. M. Álvarez and S. Ruggieri. Causal perception. *CoRR*, abs/2401.13408, 2024.

- [14] J. M. Álvarez, A. Mastropietro, and S. Ruggieri. The initial screening order problem. *CoRR*, abs/2307.15398, 2024.

[3] All codes are available here: `https://github.com/cc-jalvarez`.

- [10] J. M. Álvarez and S. Ruggieri. Uncovering algorithmic discrimination: An opportunity to revisit the comparator. *CoRR*, abs/2405.13693, 2024.
- [215] F. Palomba, A. Pugnana, J. M. Alvarez, and S. Ruggieri. A causal framework for evaluating deferring systems. *CoRR*, abs/2405.18902, 2024.

Additionally, I include other relevant publications:

- [11] J. M. Alvarez, A. Fabris, C. Heitz, C. Hertweck, M. Loi, and M. Zehlike, editors. *Proceedings of the 2nd European Workshop on Algorithmic Fairness, Winterthur, Switzerland, June 7th to 9th, 2023*, volume 3442 of *CEUR Workshop Proceedings*, 2023. CEUR-WS.org.

Not all of the publications listed above are included in this thesis. I will highlight when a chapter or (sub)section of a chapter is based on a published work. The excluded publications are: Mougan et al. [203], Álvarez et al. [13], and Alvarez et al. [11].

Mougan et al. [203] studies the fairness effects of encoding categorical protected attributes, while Álvarez et al. [13] is an interdisciplinary survey of the Fair ML literature based on the NoBIAS ITN. Both of these works address Fair ML concerns, but do not require the use of causal reasoning. This is why I exclude both of them. Alvarez et al. [11] is just the workshop proceedings for EWAF 2023 of which I was an organizer.[4]

[4]For more information, visit: `https://sites.google.com/view/ewaf23/`.

# Chapter 2

# Causality and Fairness

In this chapter, we present the overall background knowledge for the thesis. We consider as a general setup, otherwise specified, the supervised learning setting with the set of $j$ *predictive variables* $\mathbf{X}$ and the *outcome variable* $Y$. Given the sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i = \{x_{i,1}, \ldots, x_{i,j}\}$, we wish to learn the *predictive model* $\hat{f}$ using empirical risk minimization such that

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell\big(f(\mathbf{x}_i), y_i\big) \tag{2.1}$$

with some loss function $\ell$ and pre-specified function class $\mathcal{F}$. The sample is *i.i.d.*, or independent and identically distributed, from a distribution $\mathcal{D}$ with domain $\mathbf{X} \times Y$. The learned model represents the model-specification that minimizes the expected loss function over the sample distribution given by $\mathbb{E}_{\mathbf{X},Y\sim\mathcal{D}}\big[\ell\big(f(\mathbf{X}), Y\big)\big]$.

Through the learned model, $\hat{f}$, we obtain the *predicted outcome variable* $\hat{Y} = \hat{f}(\mathbf{X})$. The outcome variable $Y$ can be either discrete (like a category) or continuous (like a score). When $Y$ is discrete, $\hat{f}$ solves for a classification problem; when $Y$ is continuous, $\hat{f}$ solves for a regression problem. The set of predictive variables $\mathbf{X}$ can contain both discrete and continuous variables.

Given our focus on fairness, we also consider the set of *protected predictive (or sensitive) variables* $\mathbf{A}$. For simplicity, we define $\mathbf{A} \subset \mathbf{X}$ and distinguish between protected and non-protected predictive variables only when necessary. For instance, as we will show later in this chapter, there are different fairness (and even legal) implications to using $\mathbf{X}$ versus $\mathbf{X} \setminus \mathbf{A}$ as inputs for learning $\hat{f}$. In general, we emphasize the role of $\mathbf{A}$ only when dealing with fairness problems; otherwise, we refer to the generic $\mathbf{X}$.

Throughout the thesis we use uppercase letters when referring to generic aspects of a variable and lowercase letters when referring to the realizations of such variable. Hence, $\mathbf{x}_i$ refers to the *ith* vector of values for all $j$ predictive variables, while $x_{i,j}$ refers to the *ith* value of the *jth* predictive variable. Also, for $\mathbf{X}$ (and, thus, $\mathbf{A}$) we use *attributes* or *features* interchangeably with predictive variable. Similarly, for $Y$ (and, thus, $\hat{Y}$) we simply use *outcome* interchangeably with outcome variable.

As it is standard in Machine Learning (ML), the sample is split into training and testing data. As the names suggest, the *training data* is used for learning the ML model $\hat{f}$ while the *test data* is used for evaluating the ML model $\hat{f}$. Under a random split, both datasets follow the same distribution $\mathcal{D}$. Further, under an *i.i.d.* sample, we assume that the *incoming data* (or new samples) to be used by the learned ML model $\hat{f}$ for making

predictions also follow the distribution $\mathcal{D}$.

We will also consider the setting when this *i.i.d.* assumption is not true. Consequently, the ML model $\hat{f}$ is learned on data that is not representative of the population on which it is being deployed. Under this scenario, often the training data represents a *source domain* with distribution $\mathcal{D}_S$ and the test data represents a *target domain* with distribution $\mathcal{D}_T$ such that $\mathcal{D}_S \neq \mathcal{D}_T$. We assume, otherwise specified, the standard setting in which source and target domains align under the same distribution $\mathcal{D}$ with domain $\mathbf{X} \times Y$.

The rest of this chapter is organized as follows. We start by introducing structural causal models in Section 2.1, which is the causality framework of choice. We then discuss bias in automated decision-making and present the interdisciplinary field of Fair ML and its links to non-discrimination law within the European context in Section 2.2. We conclude the chapter by discussing popular fairness definitions used within Fair ML in Section 2.3. The later chapters include specific background knowledge sections that expand on what is presented here.

## 2.1 Structural Causal Models

The underlying conceptual framework of the thesis is causality. Discovering, formulating, and evaluating causal claims of the form $X \rightarrow Y$, or "$X$ causes $Y$", has played a central role in explaining our world since (at least) Plato [240]. There seems to be a shared view among researchers on what is not causation. Such view is often summarized with "correlation does not imply causation", a mantra known to all students of Statistics 101. What is causation, though, is less clear. From philosophers to economists, causality carries different implications. When talking about causation, thus, we must be precise on what view of causation.

Causality implies structure. In a causal world everything that happens is determined by that world's laws and initial conditions. There is an inherent order to what happens, a sequence of events driven by cause-effect mechanisms representing how information flows in the world. Studying causality, thus, requires an implicit suspension of disbelief (with varying degrees)[1] on whether said structures exist at all. Nonetheless, despite the epistemic debates around it, causality remains a useful framework across multiple fields, unsurprisingly making its way in recent years into the field of Machine Learning [254, 295] and, thus, its sub-field of Fair Machine Learning [181, 191].

In this thesis we use structural causal models as popularized by Pearl [218] to formulate causality.[2] Structural causal models are probabilistic graphical models that combine Bayesian networks and structural equation models. Structural causal models allow us to describe the *data generating model* (DGM) behind a (joint) probability distribution of interest. Further, as we will see later in this section, structural causal models allow us to manipulate the DGM such that we are able to generate new probability distributions that answer to observational (*what is*), interventional (*what if*), and counterfactual (*what would have been if*) queries [221]. It is this ability of structural causal models to generate new representations in the form of probability distributions what, among

[1] Consider, e.g., the views by someone like Cartwright [58] against someone like Schölkopf [254].

[2] Pearl himself, however, has always credited the early works by Wright [305] and Haavelmo [121] as the basis for structural causal models. See, e.g., Pearl [219].

other properties, makes structural causal models the causality framework of choice in Machine Learning over other frameworks like, e.g., potential outcomes [52].

The ability to generate counterfactual distributions is central to this work, which strongly motivated the choice of structural causal models as our causality framework. Overall, the choice to use structural causal models was motivated by three reasons. First, as previously stated, it is the dominant causality framework in Machine Learning. For instance, most causal fairness definitions (e.g., [63, 167, 177]) are based on structural causal models. Second, structural causal models are based on how humans reason about the world [220], which permits us to tackle human notions that are often causally formulated like discrimination [176]. Third, structural causal models are a convenient way for organizing assumptions about the world, which facilitates stakeholder participation when reasoning about contested concepts like fairness [206].

**Definition 2.1.1.** (Structural Causal Model) A *structural causal model* (SCM) [218] is a tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$ describing the data-generating process that transforms a set of $p$ exogenous latent random variables $\mathbf{U} \sim P_{\mathbf{U}}$ into a set of $p$ endogenous observed random variables $\mathbf{V}$ according to a set of structural equations $\mathbf{F}$ such that:

$$P_{\mathbf{U}} = P(U_1, \ldots, U_p) \qquad V_j := f_j(V_{pa(j)}, U_j) \qquad \text{for } j \text{ in } 1, \ldots, p \tag{2.2}$$

where $U_j \in \mathbf{U}$, $V_j \in \mathbf{V}$, and $f_j \in \mathbf{F}$. Each *jth* function $f_j$ maps the *jth* exogenous variable $U_j$ to the *jth* endogenous variable $V_j$ based on the subset of endogenous variables that directly cause $V_j$, or the causal parents $V_{pa(j)}$.

Notice that we use the operator $:=$ instead of $=$ in (2.2). It implies an *assignment operation*, and is equivalent to writing $\leftarrow$. Contrary to the standard equality operator $=$, the assignment operator *implies a flow of information*, denoting cause(s)-effect or parent(s)-child pairs rather than equals. Intuitively, Definition 2.1.1 describes the data-generating model of a given dataset (or context) in terms of cause-effect relations. For instance, we can use it to represent causally the relations between $\mathbf{X}$ and $Y$ in (2.1), allowing us to reason causally about these variables and the problem they address.

We assume *causal sufficiency* for (2.2). It means that there are no hidden common causes or confounders in model. This assumption implies the independence among the exogenous latent random variables in $\mathbf{U}$:

$$P(U_1, \ldots, U_j) = P(U_1) \times \cdots \times P(U_j) \tag{2.3}$$

which allows us to factorize $P_{\mathbf{U}}$ into its individual components. It is both difficult to assume and test for causal sufficiency. However, it is a common (though not necessary) assumption as it allows to generate counterfactuals more easily.

The SCM $\mathcal{M}$ induces a corresponding *causal graph* $\mathcal{G}$ in which each node represents a random variable and each edge a causal relation. For instance, the edge $V_i \rightarrow V_j$ is in the causal graph if $i \in pa(j)$. Namely, the causal graph describes visually the causal dependencies of the SCM $\mathcal{M}$. The causal graph $\mathcal{G}$ is by definition a *directed graph*. We assume it to be acyclical (hence, why $\mathcal{G}$ is also referred to as a *causal directed acyclical graph,* or causal DAG), meaning there are no feedback loops.

In the remainder of this section, we introduce the preliminary knowledge to better understand the foundations of SCM (Section 2.1.1); we present the implicit manipulationist (or interventionist) account of causality behind SCM (Section 2.1.2); and finalize

with a discussion on counterfactual generation under SCM (Section 2.1). For further reading, Pearl [218] remains the work of reference.

### 2.1.1 Preliminaries

Given a set of random variables, we often want to say something meaningful, statistically speaking, about one variable given the other variables. For this section, let us consider the set of four random variables $\{X_1, X_2, X_3, X_4\}$ with probability distribution $P$.

**Markov chains.** Suppose that we are interested in the variable $X_4$. Using the *chain rule*, based on Bayes' Theorem, we can factorize the set's joint probability distribution as a sequence of conditional probabilities with respect to the other three variables. The factorization allows us to relate $X_4$ with $X_3$, $X_2$, and $X_1$ through $P$:

$$\begin{aligned} P(X_4, X_3, X_2, X_1) &= P(X_4|X_3, X_2, X_1)P(X_3, X_2, X_1) \\ &= P(X_4|X_3, X_2, X_1)P(X_3|X_2, X_1)P(X_2, X_1) \\ &= P(X_4|X_3, X_2, X_1)P(X_3|X_2, X_1)P(X_2|X_1)P(X_1) \end{aligned}$$

In the above factorization, order is important but unclear. We can derive other equally valid factorizations for this set by focusing on other random variables at each factorization. The choice to start with $X_4$ and continue with $X_3$ over $X_2$ or $X_1$ is not inherent to $\{X_1, X_2, X_3, X_4\}$. A way to impose order explicitly is to treat the set as a Markov process, or Markov chain, which is a type of stochastic process that describes a sequence of random variables based on the Markov property.

The *Markov property* states that the next state of a sequence of random variables is conditionally independent of all past states given the current state. Formally, the Markov property states that for a sequence $X = (X_n)_{n\geq 0}$ of $n$ steps, with $X_0$ as the initial state, the future state $\{n + 1\}$ is conditionally independent of the past states $\{n - 1, n - 2, \ldots, 0\}$ given the present state $\{n\}$:

$$X_{n+1} \perp\!\!\!\perp X_{n-1}, \ldots, X_1, X_0 \mid X_n \tag{2.4}$$

Revisiting under (2.4) the factorization previously obtained using the chain rule, i.e. treating the set $\{X_1, X_2, X_3, X_4\}$ as a Markov chain with each subscript denoting the order of the sequence, we obtain a much more simple factorization:

$$\begin{aligned} P(X_4, X_3, X_2, X_1) &= P(X_4|X_3, X_2, X_1)P(X_3|X_2, X_1)P(X_2|X_1)P(X_1) \\ &= P(X_4|X_3)P(X_3|X_2)P(X_2|X_1)P(X_1). \end{aligned}$$

Readers familiar with Markov processes will note that the above factorization induces a corresponding Markov chain $\mathcal{G}$, which is similar to the causal graph $\mathcal{G}$ introduced earlier in Section 2.1 for the structural causal model $\mathcal{M}$. The key difference between a Markov chain and a causal graph lies in the semantics of each graphical object. With a Markov chain, we wish to represent changes in states. For instance, the probability $P(X_4|X_3)$ denotes both the existence of a path from $X_3$ to $X_4$ and the probability associated to crossing that path. With a structural causal model, we instead wish to represent cause-effect pairs. The probability $P(X_4|X_3)$ denotes the cause $X_3$ and its effect $X_4$.

For our purposes, what Markov processes and structural causal models share is a focus on representing order and how it influences the flow (or dependence) of information among a set of random variables as captured by the joint probability distribution $P$. This point will become clearer when we present the notion of *Markovian parents*, which is a building block of structural causal models.

**Graph terminology.** Recall that the causal graph $\mathcal{G}$ is a directed (assumed) acyclical graph, or DAG, where each node represents a random variable and each directed edge a cause-effect pair. Let us define the following additional terms for $\mathcal{G}$:

- A *directed path* between nodes $i$ and $j$ is the sequence of distinct, successive, and adjacent directed edges that point toward $j$ from $i$.
- A *directed cycle* consists of a directed path $(i, \dots, j, k)$ plus an edge $k \to i$. Under a DAG, cycles are not allowed.
- If there is a directed edge between $i$ and $j$, i.e. $i \to j$, then $i$ is a *parent* of $j$ and $j$ is a *child* of $i$.
- If there is a directed path from $i$ to $j$, then $i$ is an *ancestor* of $j$ and $j$ is a *descendant* of $j$. Each node is an ancestor and descendant of itself.

Consider, for instance, the Markov chain in Figure 2.1 for the previous factorization of the set $\{X_1, X_2, X_3, X_4\}$, which is also a DAG. For node $X_3$ we write its parents as $\text{pa}(X_3) = \cup_{k \in X_3} \text{pa}(X_3) = \{X_2\}$; its ancestors as $\text{an}(X_3) = \{X_3, X_2, X_1\}$; its children as $\text{ch}(X_3) = \{X_4\}$; and its descendants as $\text{desc}(X_3) = \{X_3, X_4\}$. Finally, the non-descendants of $X_3$, by definition $\text{nondesc}(X_3) := V \setminus \text{desc}(X_3)$, is the set $\{X_2, X_1\}$.

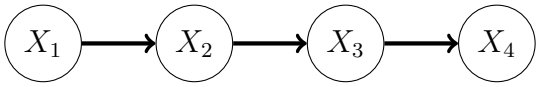



Figure 2.1: An example of a Markov chain and DAG.

**Markovian parents.** There is an implicit ordering (of states) in causality: for the effect (or current state) to occur, the cause must have occurred first (or past states). The Markov property (2.4) provides a formalization of this dynamic. It is useful to rewrite it using the previous graph terminology:

$$X_S \perp\!\!\!\perp X_{\text{nondesc}(S) \setminus \text{pa}(S)} \mid X_{\text{pa}(S)} \tag{2.5}$$

for any collection of nodes $S$. Back to $X_3$ in the Markov chain in Figure 2.1, e.g., when applying (2.5) we get $X_3 \perp\!\!\!\perp X_1 \mid X_2$. It follows from (2.5) that we can factorize the joint probability distribution $P$ for any $n$ random variables as parent-child relationships:

$$P(X_1, X_2, \dots, X_n) = \prod_{j=1}^{n} P(X_j | X_{\text{pa}(j)}) \tag{2.6}$$

which allows to "draw" the DAG $\mathcal{G}$ from the observed dependencies between the random variables in the joint probability $P$, meaning $P$ implies $\mathcal{G}$ or $P \Rightarrow \mathcal{G}$. For instance, the factorization of $P(X_1, X_2, X_3, X_4) = P(X_4|X_3)P(X_3|X_2)P(X_2|X_1)P(x_1)$ implies the DAG in Figure 2.1.

**d-separation.** Similarly, the inverse of $P \Rightarrow \mathcal{G}$ is also of interest: i.e., how to factorize the joint probability $P$ given the DAG $\mathcal{G}$? Essentially, we want to read-off variable independencies from a graph. This is possible using the concept of directional-separation, or *d-separation*, where we define a set of nodes $S$ such that it "blocks" the path, or induces conditional independence, between nodes $i$ and $j$.

To illustrate the concept of d-separation, consider three random variables $X$, $Y$, and $Z$. We want to define the set $S$ of nodes to block $X$ and $Y$. Mathematically, we mean $X \perp\!\!\!\perp Y|S$. Graphically, we want $S$ to d-separate these two nodes by blocking all paths between them. The Figure 2.2 shows the three cases of dependence between $X$ and $Y$. We, thus, need to know how to use the node $Z$ in each of these cases to induce independence between $X$ and $Y$, where the options are $S = \{Z\}$ or $S = \{\}$.

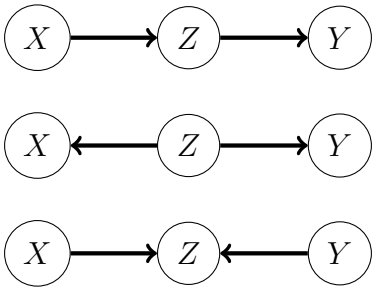



Figure 2.2: From top to bottom DAG: $Z$ acting as a mediator, fork, and collider.

From top to bottom DAG in Figure 2.2, $Z$ acts structurally as, respectively, a mediator, fork, and chain collider $X$ and $Y$. More specifically:

- Under $Z$ as a *mediator*, we define $S = \{Z\}$. This is because the flow of information that goes from $X$ to $Y$ passes through and it is, thus, mediated by $Z$. Controlling or conditioning for $Z$, blocks the information coming out of $X$ and into $Y$.
- Under $Z$ as a *fork*, we also define $S = \{Z\}$. This is because the flow of information starts from $Z$ and goes into $X$ and $Y$ simultaneously. With $Z$ being the source of information, controlling for it blocks the information going into $X$ and $Y$ and stop any association between these two variables. Here, $Z$ is a confounder or "common cause" between $X$ and $Y$.
- Under $Z$ as a *collider*, we define $S = \emptyset$. This case is the least intuitive of the three. Essentially, $X$ and $Y$ are two variables that would otherwise have no association between them but "collide" into $Z$ and are thus linked to each other. Ignoring $Z$, meaning not controlling for it, avoids the association through collision. The set $S$ should also not include descendants of $Z$ as this would also activate the collider.

**Global Markov and faithfulness.** We summarize this section with the two properties that motivate any structural causal model $\mathcal{M}$: the global Markov property and the faithfulness property. To illustrate each of these properties, consider the nodes $A$, $B$, and $S$ in $\mathcal{G}$ and corresponding random variables $X_A$, $X_B$, and $X_S$ in $P$, or vice-versa.

*Global Markov Property.* A (joint) probability distribution $P$ is *global Markov* with respect to a DAG $\mathcal{G}$ (i.e., $\mathcal{G} \Rightarrow P$) if:

$$A \text{ and } B \text{ are d-separated by } S \text{ in } \mathcal{G} \Rightarrow X_A \perp\!\!\!\perp X_B \mid X_S \text{ in } P$$

*Faithfulness Property.* A (joint) probability distribution $P$ is *faithful* with respect to a DAG $\mathcal{G}$ (i.e., $\mathcal{G} \Leftarrow P$) if all pairwise disjoint subsets $A$, $B$ and $S$ of nodes:

$$X_A \perp\!\!\!\perp X_B \mid X_S \text{ in } P \Rightarrow A \text{ and } B \text{ are d-separated by } S \text{ in } \mathcal{G}$$

Under these two properties, we are able to express the factorization of a joint probability distribution (i.e., a chain of conditionals) into a DAG or, equivalently, read conditional dependencies off from a DAG and write them as a chain of conditionals. Both properties are essential assumptions made on either $P$ or $\mathcal{G}$ that condition the overall problem of *learning causality* (see, e.g., [120, 224]). Learning causality can be split into two camps: causal discovery and causal inference.

In *causal discovery*, we wish to learn the causal structures using data. We, thus, assume that the causal dependencies in $P$ are indicative of some graph $\mathcal{G}$. Since several graphs explain the same probability distribution, making the link from $P$ to a specific $\mathcal{G}$, thus, requires "faith" in the discovered graph.

In *causal inference*, we wish to model the causal effects among variables using data. We, thus, assume that the causal structures in $\mathcal{G}$ contain all valuable information to infer $P$. The graph simplifies this task by allowing us to focus only on the parent-child relationships among the variables. Making the link from $\mathcal{G}$ to $P$, thus, requires for the existence of Markovian parents at a global level. Therefore, for any structural causal model $\mathcal{M}$ there is an implicit assumption being made on the global Markov and faithfulness properties of the system.

**Structural equations.** To conclude the preliminaries, we address the set of structural equations $\mathbf{F}$ that powers any structural causal model $\mathcal{M}$.

Behind the probability distribution and DAG of $\mathcal{M}$, there is the set of structural equations $\mathbf{F}$ functionally modeling the parent-child or cause-effect relationships. Recall from (2.2) that $\mathbf{F}$ denotes a set of equations such that $V_j := f_j(V_{pa(j)}, U_j)$ for the *jth* variable or node in $\mathbf{V}$. Each structural equation $f_j$ links the child to its parents along with its background information. As the name suggest, these equations model the (causal) structure. They represent how are parent-child relationships causally related.

In practice, these structural equations (unless known) need to be learned using data: i.e., we must find the $\hat{f}_j$ that approximates each $f_j$. This procedure is done under empirical risk minimization (2.1) for each structural equation in $\mathbf{F}$. Hence, even under a known DAG $\mathcal{G}$ and fully observed distribution $P$, we still need to define the equations in $\mathbf{F}$. This step is known as *model specification*. When done improperly, as with any other modeling approach, it can lead to model specification bias [304].

### 2.1.2 A Manipulationist Account

There are multiple accounts of causality. Structural causal models are based on the *manipulationist* or *interventionist* account of causality. In this section, we present briefly this account and how it shapes Pearl's work on causality (in particular, through the do-operator). For further reading, we recommend Woodward [303].

What does it mean for a variable to cause another variable? Pearl's view and, thus, that of structural causal models is based on the notion of intervention or manipulation. We say $X$ causes $Y$ if when we intervene or manipulate $X$ we observe changes in $Y$.

This idea, with marginal modifications over the years [123, 125, 126], is at the center of structural causal models [218]. Structural causal models are equipped with interventional capabilities in the form of the *do-operator*. This operator, which sets any random variable to a constant value, allows to envision hypothetical scenarios given a structural causal model $\mathcal{M}$. The do-operator is, thus, an interventionist (or manipulationist) account of causality as it assumes that the system can be intervened (or manipulated), suspending its internal laws without jeopardizing the stability of the system itself in terms of its structure [286, 303].

**The do-operator.** Pearl [218] defines the do-operator as a localized intervention or "micro-surgery" that sets a random variable $X$ uniformly to the value $x'$ or, formally, $do(X := x')$. In other words, $X$ is purposely assigned the value $x'$ across the population.[3] As we will see in the next section, this act, in turn, generates a new probability distribution as though we were re-setting the data generating model described by $\mathcal{M}$.

To illustrate the do-operator, let us consider the following structural causal model $\mathcal{M}_1$ (as in model one) in Figure 2.3 with a set of endogenous variables $\{Y, X, Z, W_1, W_2\}$, corresponding exogenous variables $\{U_1, U_2, U_3, U_4, U_5\}$, and corresponding structural equations $\{f_1, f_2, f_3, f_4, f_5\}$. For the structural equations, we assume additive noise and linear inputs with real coefficient weights such that:

$$
\begin{aligned}
Z &:= U_5 \\
X &:= \alpha_3 Z + \alpha_4 W_1 + U_2 \\
W_1 &:= U_3 \\
W_2 &:= \beta_2 X + U_4 \\
Y &:= \beta_1 X + \alpha_1 W_1 + \alpha_2 W_2 + U_1
\end{aligned}
$$

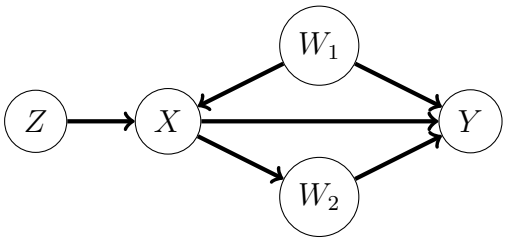



Figure 2.3

Figure 2.3, which shows $\mathcal{M}_1$'s DAG and corresponding set of structural equations, represents the probability distribution $P(Z, X, W_1, W_2, Y)$. We avoid drawing the noise terms (or latent variables/initial conditions) in the DAG.

Suppose now that we carry out the intervention $x'$ on the variable $X$ in $\mathcal{M}_1$, or $do(X := x')$. Figure 2.4, which shows the intervened $\mathcal{M}_1$'s DAG and corresponding set of structural equations (both marked in red), represents the post-intervention distribution $P(Z, X, W_1, W_2, Y \mid X := x')$. Under such intervention to $\mathcal{M}_1$, we have that:

$$
\begin{aligned}
Z &:= U_5 \\
X &:= x' \\
W_1 &:= U_3 \\
W_2 &:= \beta_2 x' + U_4 \\
Y &:= \beta_1 x' + \alpha_1 W_1 + \alpha_2 W_2 + U_1
\end{aligned}
$$

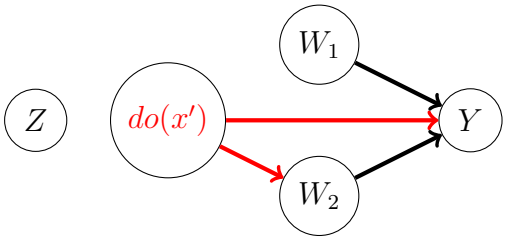



Figure 2.4

[3] Today, this operation is also referred to as a *hard intervention*. Recent works (see, e.g., Massidda et al. [193]) explore situation where the random variable is set to a non-constant value, or a *soft intervention*.

Figure 2.4 presents essentially a new model: a hypothetical scenario based on the original model $\mathcal{M}_1$ in which we have suspended the laws around the node $X$, i.e., a sort of "small miracle" [286]. Under this intervention, we have replaced completely the right-hand side of $X$ for the value $x'$ in the system of equations that, in turn, implies that all functions where $X$ is an input are now assigned the value $x'$ (in red). Similarly, we have removed the incoming directed edges to $X$ and have set the outgoing directed edges (in red) from $X$ as this node equals the value $x'$.

The remaining parts of $\mathcal{M}_1$ not directly affected by $X$ are unchanged. Notice, though, that the effects of the intervention are experienced for all nodes that share a path with descendants of $X$. In short, we *wipe out* the equations and *break up* the edges of $X$ [303, p. 47-48], and let the information spread. Figures 2.3 and 2.4 clearly illustrate the orderly nature of causality and how, by having a view on the cause-effects pairs (read, structure), we can reason about the implications from manipulating elements of the system onto other elements (i.e., the rest) of the system.

**Invariance.** The notion of invariance is what allows the do-operator to function and makes possible the manipulationist or interventionist account of causality. The idea is that a causal relationship between two variables implies an *invariant relationship* between the two variables. If $A$ causes $B$, whatever we decide to do with either $A$ or $B$ does not change the fact that $A$ is a cause of $B$ and $B$ an effect of $A$. Hence, the notion of invariance motivates structure and vice versa.

Invariance allows to conceive scenarios in which we are able to manipulate one or several relationships locally and contemplate the remaining relationships to hold. The existence of structure, in turn, motivates interventions. To quote Woodward [303, p. 33], "[o]ne may learn, through passive observation, that two variables $A$ and $B$ are correlated. However, this fact by itself tells one nothing about whether one can, by acting so as to change or manipulate $A$, also changes $B$." The presence or, at least, assumption of structure is crucial for manipulation claims.

A consequence of invariance, is the difference between *conditioning* (or *looking*) and *intervening* (or *doing*) or, more formally, the difference between $\mathbb{E}[Y|W_1, W_2, Z, X = x']$ and $\mathbb{E}[Y|W_1, W_2, Z, do(X := x')]$. Here, both expectations are claims on the joint probability distribution with a focus on $Y$ and the effect of $X$ on it. Ordinary conditioning is observed from the data. When we condition on $X$ to see how $Y$ changes, we are changing the whole system at the same time, and it is not possible to attribute fully to $X$ the changes in $Y$. With the *do-operator*, however, we intervene on the desired node(s) and the changes spread accordingly, meaning that the other conditional distributions remain unchanged. The only case in which conditioning and intervening coincide is when there are no edges coming into $X$, i.e. $\text{pa}(X) = \emptyset$. This situation occurs, for instance, in randomized control trails.

This distinction between conditioning and intervening is rooted in the notion of *entangled and disentangled factorizations* of the joint distribution [254]. The structural causal model's implicit structure (read, invariance) allows to control for the flow of information in a local sense, ensuring causal claims. To observe this point, we write the

joint distribution of all nodes, $P(\mathbf{V})$, in model $\mathcal{M}_1$ pre-intervention on $X$ as:

$$P(\mathbf{V}) = \left\{ \prod_{i \in \mathbf{V} \setminus X} P(V_i | V_{pa(i)}) \right\} P(X | X_{pa(X)})$$

and post-intervention on $X$:

$$P(\mathbf{V} | do(X := x')) = \left\{ \prod_{i \in \mathbf{V} \setminus X} P(V_i | V_{pa(i)}) \right\} \mathbb{1}\{X = x'\}$$

where the joint distribution is truncated by using the do-operator. Here, the notion of *modularity*, a consequence of assuming invariance and structure, becomes apparent. Each child-parent factorizations, motivated by (2.6), acts as an *independent modulus*. As shown in the pre- and post-intervention $P(\mathbf{V})$ for $\mathcal{M}_1$, intervening $X$ only affects the modulus affecting/affected by it; the other modulus remain unaffected.

Modularity is, clearly, a strong premise to structural causal models. It has been criticized, mainly by philosophers, as too simplistic [58, 59, 60]. Cartwright [59], for instance, refers to modularity as an "epistemic convenience." Our view here is that modularity, invariance, and the whole of structural causal models are useful concepts to make sense of the world. Claims for and against these concepts, yet crucial and valid, are outside the scope of this thesis.

**Causal identification.** Making any causal claim requires identification. For instance, to claim the causal effect of $X$ on $Y$ in model $\mathcal{M}_1$ in Figure 2.3 requires precision on what sort of causal effect we wish to claim. Is it a *total (causal) effect* as in all information leaving from $X$ and reaching $Y$? Or is it a *direct (causal) effect* as in the information leaving from $X$ and directly affecting $Y$? Or is it an *indirect (causal) effect* as in all the information leaving from $X$ and affecting $Y$ through $W_2$?

To identify any causal claim we resort to *covariate adjustment*. Using the notions of d-separation (Section 2.1.1), we define the adjustment set $\mathbf{S} \subset \mathbf{V}$ that blocks all active paths besides the path(s) of interest. This step needs to be done before intervening: i.e., *identification before intervention*. We define the adjustment formula as:

$$P(Y | do(X)) = \int_{\mathbf{S}} P(Y | X, \mathbf{S}) P(\mathbf{S}) d\mathbf{S} \tag{2.7}$$

for identifying the causal effect of $X$ on $Y$ such that $Y, X, \mathbf{S} \in \mathbf{V}$ for a given structural causal model $\mathcal{M}$.

There are three graphical criterion (or identification strategies) used for defining $\mathbf{S}$: back-door criterion, adjustment criterion, and front-door criterion [218]. Here, we mostly focus on the back-door criterion. It states that $\mathbf{S}$ cannot contain the nodes $X$ and $Y$; cannot include descendants from $X$; and should block all "back-door paths" between $X$ and $Y$ that start with a direct edge going into $X$. In practice, it consists of controlling for all confounders between $X$ and $Y$.

Intuitively, the back-door criterion ensures that the only source of variation between $X$ and $Y$ comes from changes directly on and from $X$. Consider, for instance, $\mathcal{M}_1$ in Figure 2.3. Notice that $X$ causes $Y$ directly through $X \to Y$ and indirectly through

$X \to W_2 \to Y$. These are the two paths from which the causal effect of $X$ reaches $Y$. Also notice that both $X$ and $Y$ are caused by $W_1$, making it a confounder of $X$ and $Y$. Suppose we want to answer whether $X$ causes $Y$. To do so, we indeed need to intervene $X$ via the do-operator (as shown in Figure 2.4) and observe whether $Y$ changes; however, we first need to ensure the identification of this causal effect. In particular, we want to make sure that the changes observed in $Y$ after having intervened $X$ are only due to changing $X$ and not from the causal effect of $W_1$ on both $X$ and $Y$. Here, $X \leftarrow W_1 \to Y$ represents a back-door. Therefore, $\mathbf{S} = \{W_1\}$ in (2.7) under the back-door criterion.

**Potential outcomes framework.** A popular alternative to structural causal models, especially among economists and other social scientists, is the potential outcomes framework (PO) [151]. It is also known as the *Rubin causal model* due to the works of statistician Donald Rubin (see, e.g., [152]). Here, we briefly present PO; see Angrist and Pischke [16] for further reading.

This framework argues that individual $i$ units have a set of available potential outcomes $Y$ to be realized conditional on some intervention or treatment $T$. For instance, given the existence of a drug able to cure a disease, the *ith* patient will have one potential outcome with, $Y_1 = (Y|T = 1)$, based on receiving the treatment and another potential outcome without, $Y_0 = (Y|T = 0)$, based on not receiving the treatment. In practice, however, this setting means observing the *ith* patient twice. In this *single world* we live in, it is not possible to observe simultaneously $Y_1$ and $Y_0$ for any $i$. PO is used for designing experiments where we randomly allocate $T$ to different groups of patients and obtain *experimental data* to test causal claims around $T$ and its effect on $Y$. The gold standard here are randomized control trials (RCTs) in which we insure, through experimental design, that $T$ is randomly allocated. Other methods, though, exist for performing PO using non-RCTs data [16].

PO is largely used within the social sciences, though some recent lines of work in Machine Learning (e.g., risk assessment instruments [72, 198]) resort to it over structural causal models. These frameworks are not at odds with each; rather, each serves different communities with different research goals. Overall, PO is preferred when testing a given treatment or intervention, while structural causal models are preferred when learning causal representations [120].

**No causation without manipulation.** To conclude this section, we address the famous "no causation without manipulation" phrase coined by Holland [142]. This phrase is often used to question conceptually whether a given variable with immutable (i.e., non-manipulable) properties, like race or gender, should be considered at the center of a causal claim. See, e.g., Hu and Kohler-Hausmann [149], Kohler-Hausmann [176].

The standard workaround for such immutable variables, especially by the PO crowd, has been to argue for the manipulation of the perception of said attributes, not the attributes themselves. This line of argument is, in fact, very similar to the one adopted by a broader manipulationist crowd in which we are just concerned with generating or imagining hypothetical scenarios [303]. I argue that it aligns well in meaning with how interventions are carried out in structural causal models.

Our view on immutable variables is context-based. If, for instance, it is useful to picture race as a mutable variable, then we allow interventions on the variable race.

That is not to say that we assume that a given individual can change his or her race in a specific context, but that considering such a hypothetical scenario has value for understanding said context. Whether we wish to frame it as the perception of race or as race itself is not important, though we carefully discuss it within each context.

### 2.1.3 Counterfactuals

For a given structural causal model (SCM) $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$, according to Pearl and Mackenzie [220]'s *causal hierarchy*, it is possible to answer three levels of causal queries. Each query induces a corresponding *generated* (or *learned*) *distribution* that represents the query. These levels are, in increasing order of complexity: associative, interventional, and counterfactual queries. The corresponding distributions are the observed, interventional, and counterfactual distributions. In this work, we are mainly interested in the latter type of queries that are often used for modeling causal fairness.

**Pearl's causal hierarchy.** Consider the endogenous variables $X, Y \in \mathbf{V}$, and suppose we are interested in causal claims between these two variables.

At the *associative or first level*, which is the standard correlation-based modeling, we form predictions based on observations of $X$ and $Y$. We consider the observed distribution $P(Y|X)$. There is no causal meaning given to the relationship between $X$ and $Y$; in fact, there is no need for knowledge of $\mathcal{M}$. This level answers to *what is* questions.

At the *interventional or second level*, we intervene $X$ to measure what changes in $Y$ as a consequence of the intervention. We consider the interventional distribution $P(Y|do(X := x)$. Causal knowledge in the form of $\mathcal{M}$ is needed at this level. This level answers to *what if* questions.

At the *counterfactual or third level*, we also intervene $X$ to measure what changes in $Y$ but while accounting for individual or unit-level variation. Such variation comes from the fact that each structural equation is at the population level and, thus, they might over or under estimate causal effects at the individual level. We consider the counterfactual distribution $P(Y_{X \leftarrow x}|Y', X')$, which reads as the probability of event $Y$ had $X$ been intervened to $x$ given that $Y', X'$ are observed. This final level answers to *what would have been if* questions.

Few papers have tried to formalize the causal hierarchy; an exception is Bareinboim et al. [24]. Structural causal models are used across a range of applications in which the three types of learned (or generated) distributions may have different implications. Further, the introduction of deep learning models for causal modeling has blurred (in some settings) the distinction between the second and third level (see, e.g., Javaloy et al. [154]). We note that this is still an ongoing area of research in causal ML.

**Generating counterfactual distributions.** Based on Pearl et al. [221], three steps are necessary for generating a counterfactual distribution given an SCM $\mathcal{M}$: abduction, action, and prediction. To generate counterfactuals, we assume the SCM $\mathcal{M}$ (2.2) to be an *additive noise model* (ADM) [147], meaning $V_j := f_j(V_p a(j)) + U_j$.[4] The goal with assuming ADM is to be able to separate a variable's parents from its noise.

[4]It can be also be "imposed" under, e.g., multiplicative noise by applying a logarithmic transformation.

Let us consider the model $\mathcal{M}_1$ from Figure 2.3. Suppose we wish to answer what would have been of $Y$ if $W_1$ equaled $w'$. Formally, we wish to compute the distribution $P(Y_{W_1 \leftarrow w'}(\mathbf{U})|Y, W_1, W_2, X, Z)$ following these three steps:

- *Abduction:* Using the observations of all variables, or evidence, we compute the values for the unobserved latent space $\mathbf{U}$ based on the set of structural equations $\mathbf{F}$ and the induced causal graph $\mathcal{G}$. Under causal sufficiency (2.3), this step amounts to estimating the *ith* individual error terms of each *jth* structural equation: $\hat{U}_{i,j}$. Otherwise, such as under the presence of confounders, we compute the posterior $P(U|Y, W_1, W_2, X, Z)$ and draw from it the *ith* individual "units": $\hat{U}_i$. This is by far the most complex and controversial step when generating counterfactuals.

- *Action:* Once we draw an approximation to the distribution of the latent space, $P_{\hat{\mathbf{U}}}$, we intervene the model, $do(W_1 := w')$, updating the structural equations and causal graph accordingly.

- *Prediction:* Given the intervened structural causal model and $P_{\hat{\mathbf{U}}}$, we re-estimate (or let the information flow) and observe, in this case, the distribution of $Y$.

We stress that estimating the abduction step depends on the assumption made about the structural causal model, in particular, on $\mathbf{F}$ and $P_{\mathbf{U}}$. Is it common, as it is much simpler, to work with an ADM $f_j \in \mathbf{F}$. Given ADM and causal sufficiency, it is relatively straightforward to estimate $\hat{\mathbf{U}}$ as it reduces to calculating the individual error terms. See, e.g., the implementations of Karimi et al. [163] and Álvarez and Ruggieri [8].

Given ADM and causal insufficiency, the estimation requires more computational extensive methods based on Bayesian statistics: essentially, how can we update the prior $P_{\mathbf{U}}$ given the evidence and the causal graph. In this case, some works use Monte Carlo Markov Chains (MCMC) [268] to draw the posterior. See, e.g., Kusner et al. [177]. More recent works have focused on using deep learning models, such as variational auto encoders (VAE) [32, 172, 248, 251] and normalizing flows [154], to learn the posterior.

**Naming convention: counterfactuals.** Throughout this work, when using the term "counterfactuals" we are exclusively referring to instances, distributions, or objects generated via a structural causal model as prescribed by Pearl et al. [221].

Counterfactuals are not to be confused with *counterfactual explanations* [119] as first defined by Wachter et al. [287]. These are two distinct fields built around counterfactual reasoning, though counterfactual explanations are not, in principle, causally based. Some researchers (including myself) argue that the correct term should be *contrastive explanations* instead of counterfactual explanations to avoid further confusion.

Similarly, this kind of counterfactuals, which are based on a manipulationist or interventionist view of causality (Section 2.1.2), are also known as *non-backtracking counterfactuals* or *interventionist counterfactuals* [286]. This distinction comes from the fact that the counterfactuals are generated through interventions while keeping the initial conditions (i.e., the latent space) intact and, thus, do not allow for backtracking. Backtracking counterfactuals, however, are generated by not intervening the structural causal model and instead finding a new set of initial conditions that generate the desired counterfactual distribution when down-streaming its effects. See von Kügelgen et al. [286] for more details. In this work, counterfactuals refer to the non-backtracking, interventionist kind.

## 2.2 Bias in Automated Decision-Making

Machine learning (ML) is increasingly used today for automated decision-making (ADM) in hiring, lending, and other high-risk scenarios. These models–often trained on historical and, thus, likely biased data–are good at perpetuating and even finding new unfair and potentially discriminatory patterns. Famous examples include the COMPAS [17], Amazon recruiting [77], and Dutch Government [138] scandals in which, respectively, potentially racist, sexist, and xenophobic ML models were used in high-risk decision-making scenarios for ADM.

The growing interdisciplinary field of Fair ML studies how to detect, mitigate, and improve biased models under fairness constraints (mostly) based on notions of equality. The field has increased considerably in the past decade, mainly in the form of tailor-made conferences like ACM FAccT, ACM AIES, and ACM EAAMO and special tracks and workshops in larger, more traditional ML conferences like ECML, ICML, and NeurIPS. Other umbrella terms, such as *Fair AI* and *algorithmic fairness*, are used as well to refer to the field. Here, we mostly use Fair ML.

We note that, often, the term *ethical AI* is used to refer to algorithmic fairness and explainable AI (xAI), which are the two largest interdisciplinary communities working on this topic. These two communities are not mutually exclusive, with many researchers working across both. As the name suggests, however, xAI focuses more on developing methods for making ML models interpretable to humans. This is because many ML models are black-boxes. We do not draw much from the xAI literature in this thesis. See Molnar [200] for a recent book on the topic.

In this section, we discuss the foundational notions of bias, fairness, and discrimination behind Fair ML (Section 2.2.1) as well as introduce the characteristics of algorithmic discrimination that are specific to the European Union (EU) context (Section 2.2.2). For further reading, we recommend Ntoutsi et al. [214] for an interdisciplinary introduction to bias in ML; Barocas et al. [26] for an exhaustive survey of Fair ML; and Ruggieri et al. [245] for a recent critical take on Fair ML. We also recommend Álvarez et al. [13] for an interdisciplinary, EU-based, comprehensive introduction to Fair ML that includes policy implications and key takeaways for practitioners.

### 2.2.1 From Bias to Fairness to Discrimination

Central to Fair ML are the notions of bias, fairness, and discrimination. Although all three notions are interlinked and are often used interchangeably, they carry different meanings. Here, we (briefly) define each of these concepts. Further, in doing so, I wish to illustrate the ordering among these foundational concepts. I argue that all decisions deemed discriminatory are unfair, but not all unfair decisions are discriminatory; similarly, all unfair decisions are biased, but not all biased decisions are unfair.

To illustrate these three concepts we consider the Amazon recruiting scandal [77], where the company trained a ML model on current employees to filter out potential applicants based on their CVs. The company had to shutdown the ML model after discovering that it was ranking male applicants above female applicants with similar qualifications. The algorithm apparently associated the gender of an applicant to the probability of success within the company.

**Bias.** The starting point for Fair ML is bias. The problem with bias is that it means different things to different people, which is problematic in an interdisciplinary field like Fair ML. In practice, researchers often take for granted the potential subjectivity of this concept and move on without addressing it. Barocas et al. [26], for instance, use bias to refer "to demographic disparities in algorithmic systems that are objectionable for societal reasons." It is a definition clearly open to interpretation as what is objectionable depends on the society we assume or envision. This is not problematic as long as we do not assume some inherent universality for bias. Even a single, given society changes over time and what was once objectionable may no longer be so today.

The issue persists in more technical definitions of bias, such as statistical bias, commonly used in other quantitative fields like Econometrics [16, 111, 304]. Here, a statistical estimator (e.g., the average) is considered biased if it differs in expectation from the population parameter it wants to estimate (the mean). While the estimator in question is sample-specific and, thus, observed, the population parameter of interest is purely theoretical.[5] The question of subjectivity still remains as the bias depends on assuming the existence of a population parameter and its distribution. In fact, by assuming the representation of societal values in the from of a distribution, the statistical bias formulation captures the broader bias definition from the previous paragraph.

We view *bias* as a deviation (often, an undesirable one) from an expected reference point. What this deviation represents and how this deviation is represented are based on some (implicit) agreed context, view, or interpretation on how the world is and/or should be. In other words, defining bias means making normative statements about the world. Again, there is nothing, in principle, problematic about this practice as long as we acknowledge it.

In the Amazon recruiting scandal, we speak of a biased ML model toward female applicants as we *expect* applicants with the same qualifications (regardless of gender) to be classified the same by the ML model. This is because, given our current societal views and understanding of Amazon's business, gender has no informational value in determining an applicant's potential. We can both identify and condemn this deviation by stressing that the model goes against societal expectations, or, similarly, by measuring the distance between the observed and expected distribution of the model predictions.

This definition of bias is on purpose subjective. I want to emphasize the role of the researcher and the research community when addressing bias in Fair ML since we are always at risk of being biased ourselves when we talk about bias. It is something very human. As writer David Foster Wallace [290] once so accurately put it:

> Here is just one example of the total wrongness of something I tend to be automatically sure of: everything in my own immediate experience supports my deep belief that I am the absolute center of the universe; the realest, most vivid and important person in existence. We rarely think about this sort of natural, basic self-centeredness because it's so socially repulsive. But it's pretty much

[5] Formally, suppose we want to estimate the population parameter $\theta$ and we obtain the estimate $\hat{\theta}$ from a random sample using a known statistical estimator for $\theta$, then we define *bias* as:

$$bias(\theta) = \mathbb{E}[\hat{\theta}] - \theta.$$

> the same for all of us. It is our default setting, hard-wired into our boards at birth. Think about it: there is no experience you have had that you are not the absolute center of.

This view of bias, I would argue, falls along the lines of what feminist philosopher Donna Haraway defined as *situated knowledge* [129]. To Haraway, our objectivism is a function of what we choose to see (*situation*), how we choose to see it (*location*), and from where we choose to see it (*position*)–all of which are telling of privilege and existing power relations. Haraway advocates for embracing partial views of the world that are based on the situation at hand. "The moral is simple," she writes, "only partial perspective promises objective vision." Such conception of bias implies that whatever issues we are addressing through the ML model, we are limited to some implicit context.

**Fairness.** It follows that not all deviations from an expected reference point are unfair acts, but all unfair acts are deviations from an expected reference point. In Fair ML, fairness has been defined in terms of equality of opportunities and/or treatment. These definitions, which we cover in the next section, have sufficed the ML community while triggered moral philosophers and legal scholars alike: the mere idea of equality, for instance, is contested among scholars of these fields (see, e.g., Westen [297]).

For our purposes, we view *(un)fairness* as the act of acknowledging that some biases and inequalities in outcome and/or treatment due to the deviation are unacceptable and must be addressed. From a ML perspective, the most approachable treatment of (un)fairness I have come across is John Rawls's *Justice as Fairness* [231]. Rawls argues that societies are built on the principle of justice, defined as "simply the acknowledgement of certain principles of judgment, fulfilling certain general conditions, to be used in criticizing the arrangement of their [as in the parties involved] common affairs."

According to Rawls [231], when we talk about something being just it is with respect to some common agreement on how practices between the parties are to be conducted, which, in turn, relies on the conception of fairness as it is based on *the choice of practices*. Importantly, to talk about justice we first have to agree on the practices by which justice is conceived. Rawls [231] draws on two principles:

> first, each person participating in practice, or affected by it, has an equal right to the most extensive liberty compatible with a like liberty for all [*the principle of equal liberty*], and second, inequalities are arbitrary unless it is reasonable to expect that they will work out for everyone's advantage [*the principle of permissible inequalities*].

Departure without justification from these principles implies unfair practices and, thus, an unjust arrangement for dealing with common affairs among the parties. The first principle holds, all else equal, that a departure from an "initial position" is possible as long as it is properly justified (with the burden of proof on the individual who departs from it). The second principle holds that the resulting inequalities (with respect to the initial position) are permissible if there is a reason that the practices resulting in the inequalities "work for the advantage of *every* party engaging in it" [231, p.165-7]. These principles are apparent in the fairness definitions. And, as with bias, it seems that fairness is a product of its context.

In the Amazon recruiting scandal, the ML model's bias toward female applicant is considered unfair as it is treating unequally (or, also, providing unequal opportunities) to similar male and female applicants. There is no justification to make the inequalities in treatment permissible in this context as we recognize female and male applicants as equal parties. We, in principle, would find as permissible a discrepancy in the number of successful male applicants versus female applicants as long as it could be justified by, say, male applicants having on average a better background that their female counterparts and, as long as, both parties had access to the same set of opportunities.

**Discrimination.** The concept of *discrimination* refers to an unjustified difference in treatment toward an individual or group of individuals based on (perceived) membership to a protected by non-discrimination laws group. We view, thus, discrimination as unfairness and, thus, bias sanctioned by law. In turn, this means that discrimination has to be proven in court, meaning not all unfairness can or will be considered discriminatory while all what is considered discriminatory will by default be viewed as unfair.

Essential to discrimination are the *protected groups*. As the name suggests, these are groups of individuals deemed vulnerable by society that are currently protected by non-discrimination law. The list of protected groups varies per country, though these usually include gender, race, and religion [234]. The choice of defining a group of individuals as protected requires (some) acknowledgement of (past) wrong doing as a society. Recognition under (non-discrimination) law is an ultimate goal for many of the issues tackled by Fair ML. While notions like unfairness (and bias) can be contested as they require an agreement on what unfair (and biased) decisions are, discrimination, in principle, already establishes an agreed starting point for all parties involved. Focus is then given on (dis)proving the discrimination claim.

Back to the Amazon recruiting scandal, because the disparity in treatment by the ML model appears to affect female applicants and benefit male applicants, the protected attribute gender is of interest. Although no charges were filled against Amazon, the unfairness of their ML model could have presented grounds for a potential discrimination case. This, however, needed to be determined by a court relevant to the case.

**Remark 2.2.1.** Throughout the thesis, we will use the phrase "knowing what we know" when discussing these three notions of Fair ML. This phrase allows to convey tacit, relevant, and shared background knowledge for a given ADM context.

Following up on the previous remark, notice, e.g., that at no point it was needed for me to explain why the Amazon recruiting scandal was a relevant Fair ML scandal to begin with. Knowing what we know about the treatment of women in our modern societies, their presence in STEM fields, and their treatment in tech industries, it goes without saying that the decisions made by Amazon's ADM system were concerning. Overall, it is important to recognize the historical processes behind all three terms, including the designation of protected groups, which can only be understood, in my view, through the lenses of history.

### 2.2.2 The EU Context

The European Union (EU) is taking the regulation of ADM systems very seriously. The latest example of this trend is the ongoing AI Act [92], which is intended to regulate

the use of Artificial Intelligence (AI) under a scale of risk scenarios. It remains ahead, in terms of scope and actual steps, from its US counterpart, the AI Bill of Rights [298]. The AI Act precedes another EU effort to regulate digital platforms, the GDPR [91], which focuses on ensuring the proper processing of user data under privacy concerns.

ADM systems pose a risk to multiple areas regulated by the EU. In this work, we focus mostly on non-discrimination law.[6] Here, we briefly address the general EU context and current challenges as the topic of algorithmic discrimination is recurrent throughout the thesis. In traditional discrimination (i.e., under a human decision-maker), the legal challenge is to determine how and why the decision was made and whether the protected attribute played role in it [174]. Proving discrimination is not easy, and ADM systems further complicate it as current laws were written for a human, not an algorithmic decision-maker. Existing EU and US anti-discrimination laws, e.g., do not provide an easy fit for ADM systems [25, 122].

This thesis is clearly not a legal work, though we draw considerably from the legal field. For the curious reader, we recommend the following recent papers tackling Fair ML and non-discrimination law under the EU: Hacker [122], Xenidis [308], Wachter et al. [288], Calvi and Kotzinos [55], Weerts et al. [296], and Panigutti et al. [217], among others. These are some of the EU-based works we draw from. Below we highlight characteristics specific to the EU context to consider when moving forward.

**Direct and indirect discrimination.** Under EU non-discrimination law, discrimination is classified as either direct or indirect. Under direct discrimination, the decision maker uses information on the protected attribute to make the decision, which is by default illegal. Under indirect discrimination, the decision maker uses non-protected and, thus, neutral attributes that in fact act (almost) as a proxy of the protected attribute. Often, direct discrimination is viewed as intentional (or premeditated) while indirect discrimination is viewed as unintentional. Indirect discrimination can be ruled out as long as the decision maker shows that the neutral attributes are used for a legitimate (business) purpose.

This split on discrimination types is similar to the US distinction between *disparate treatment* (for intentional discrimination) and *disparate impact* (for unintentional discrimination) [25]. Direct discrimination and disparate treatment are essentially the same. This is not the case for indirect discrimination and disparate impact. The key difference is that indirect discrimination still finds the decision maker liable despite lack of premeditation, which is not the case for disparate impact [122].

The dominant legal interpretation of direct and indirect algorithmic discrimination has been centered on whether the ADM system uses the protected attribute as an input [122]. If the protected attribute is an input to the model, we treat the scenario under direct discrimination; otherwise, we treat the scenario under indirect discrimination. See Hacker [122] for further details. Under this interpretation, the (expected) form of algorithmic discrimination (given our current legal conceptions) is indirect discrimination. This interpretation does not represent a general consensus, e.g., Xenidis [308] and Adams-Prassl et al. [1] have criticized it as insufficient.

The overall critique seems not to be aimed at a specific interpretation of algorithmic

[6] Another important area, e.g., is competition law. Calvano et al. [54], for instance, explore how algorithms can reach a state of collusion, which is considered illegal, without being trained to do so.

discrimination under current laws, but at the need to develop a new legal doctrine to properly address it as current laws are not well-equipped to do so. For instance, Adams-Prassl et al. [1] argues that the distinction between direct and indirect discrimination based on whether the protected attribute is a model input or not is pointless when ML models are able to infer the protected attribute using only neutral attributes. Hence, the potentially indirect discriminatory algorithm can be used to directly discriminate.

**Substantive and formal equality.** Legal scholars, according to Wachter et al. [288], interpret the equality objectives of EU non-discrimination law as substantive rather than formal. Under substantive equality, unlike formal equality, the status quo is not considered neutral. The implementation of EU non-discrimination law is seen, thus, as a tool not only for achieving equality as we observe it today but also for achieving the kind of equality we wish to experience in the future. In other words, it is seen as a corrective tool used to amend past injustices. This is not the case for US anti-discrimination law, which aims for formal equality [288].

How this difference materializes in practice from a ML point of view remains unclear and is an ongoing goal of Fair ML. Wachter et al. [288], based on this distinction on equality, classify ML methods as *bias preserving* and *bias transforming*. ML that is bias preserving assumes nothing about the status quo and, thus, preserves whatever existing biases are present in society. ML that is bias transforming, instead, sees the status quo itself as an issue and, thus, aims to change the biases it contains. Defining the current status quo and defining a preferred version of it are both up for interpretation.

From a ML perspective, the closest critique on these two equality objectives is the one raised by Dwork et al. [88] and Hardt et al. [131] against the demographic parity fairness definition (see Section 2.3). Forcing, for instance, a company to hire 50% male and female candidates might comply regarding formal equality, but it might mean little regarding substantive equality if within a year the majority of new employees that remain in the company are male. The focus on substantive equality is specific to the EU context and aligns well with the goals of Fair ML.

**Multi-dimensional discrimination.** Raised by the US legal scholar Crenshaw [74], intersectionality refers to individuals that cover multiple protected groups, like a black female individual. Current EU non-discrimination law does not consider intersectional discrimination. It instead only recognizes multiple discrimination, which imposes the individual to prove separate cases of discrimination across the protected groups while ignoring how these identities intersect [308].

The issue is that multiple discrimination tends to downplay intersectional discrimination, meaning an individual cannot be discriminated in the multiple sense while still be discriminated in the intersectional sense. Under current law, Xenidis [308] argues that, following Crenshaw [74]'s logic, a black female would have to prove multiple discrimination separately: as a female individual (based on gender) and as a black individual (based on race), with both claims having to hold simultaneously. The black female, though, even if not discriminated multiple times under gender and race, can still be discriminated at the intersection of gender and race.

Intersectional fairness is a pressing matter in Fair ML. Many of the ADM systems use multiple protected attributes and sometimes even combine them for better performance

[203]. Intersectional fairness remains largely understudied (with some exceptions, e.g., [203, 291, 310]). Recently, Roy et al. [239] provide a recent joint Computer Science and Law perspective on intersectionality and multiple discrimination; Romei and Ruggieri [234] also give a similar though briefer discussion.

## 2.3 Popular Fairness Definitions

To conclude this chapter, we present the main fairness definitions for this work. To do so, we focus on the joint probability distribution $P(Y, \hat{Y}, X, A)$ (and variants of it) based on the ML model $\hat{f}$. For simplicity, we assume single neutral and protected attributes. Further, let $Y = 1$ denote the desirable outcome (e.g., receiving a loan) and $A = 1$ membership to the protected group (e.g., female). These definitions extend beyond this simple setting. For illustrative purposes, we use once again the Amazon recruiting scandal, where $A$ denotes applicants' gender, $X$ applicants' university grades, $\hat{Y}$ the model's recommendation to interview an applicant, and $Y$ some measure of success within the company, like reaching the five-year mark or becoming a manager.

This section is not meant to be exhaustive as new fairness definitions continue to appear; we recommend Verma and Rubin [284] for a concise survey on the leading fairness definitions. We also recommend works like Binns [39] and Hutchinson and Mitchell [150] that position the Fair ML definitions relative to other fields' treatment of fairness. Further, the definitions covered here are aimed aimed at *classification* problems, which represent the majority of ADM settings. Still, these definitions can and have been extended to other ML problems. See, e.g., Zehlike et al. [315, 316] for a survey on fairness definitions for ranking problems.

In this section, we introduce the main correlation-based fairness definitions in Section 2.3.1 and the main causality-based fairness in Section 2.3.2. We conclude in Section 2.3.3 with a discussion on the Yule Effect that highlights the importance of causal knowledge when using these fairness definitions. Before moving forward, though, below we highlight relevant characteristics of these fairness definitions.

**Learning to decide versus learning to predict.** It is worth stressing that the role of most, it not all, ML models used in ADM is to calculate a $\hat{Y}$ that approximates a $Y$ of interest. This claim is trivial under the supervised learning setting in (2.1): we use $Y$ to train the model $\hat{f}$. The perfect classifier (or, overall, the perfect model) would mean $\hat{Y} = Y$. Under this supervised setting, studying the differences between these two variables has important fairness implications. However, this setting has further implications when we consider that the purpose of the model $\hat{f}$ is to be used on incoming, unlabeled samples of individuals that have yet to experience the outcome $Y$.

Kilbertus et al. [169] define this distinction in terms of *learning to predict* (i.e., training $\hat{f}$ over the "historical" labeled data) and *learning to decide* (i.e., using $\hat{f}$ over the new unlabeled data and using $\hat{Y}$ to infer $Y$). Hence, the existence of $Y$ for the incoming sample is debatable.[7] Under random sampling, which is rarely the case for a deployed

[7]This view, e.g., has motivated the works on *performative predictions* in which the ML model rather than inferring the outcome ends up inducing it through its predictions. It is based on early works by Grunberg and Modigliani [116] on the predictability of social events. We do not cover this line of work here; see Perdomo et al. [223] for details.

model, this distinction is not important as the learned model $\hat{f}$ is able to infer $Y$ via $\hat{Y}$.

**Pre-, in-, and post-processing.** Fair ML methods apply along the ML pipeline. Techniques are classified as pre-, in-, and post-processing. At the *pre-processing* stage, we are concerned with having a fair (often meaning a representative) training data for the model. It often involves over- and under-sampling, synthetic data generation, data augmentation, and data sampling techniques. This stage is also referred to as *bias prevention*. At the *in-processing* stage, we are concerned with learning a fair model. It often involves optimization techniques under fairness constraints. This stage is also referred to as *bias mitigation*. Finally, at the *post-processing* stage, we are concerned with ensuring the fairness of a learned model when implemented. It ranges from relabeling the model outcomes to monitoring the deployed model's behavior under flows of incoming data. This stage is also referred to as *bias detection*. See Ntoutsi et al. [214] for details.

**Fairness through awareness.** Earlier works on fairness varied along the lines of learning a *unaware model* $\hat{f}$ (meaning, one that does not require access to the protected attribute $A$) and learning an *aware model* $\hat{f}$ (meaning, one that does require access to the protected attribute $A$). In practice, *fairness through unawareness* consists of excluding $A$ when learning the model $\hat{f}$ and of making no attempts to adjust for any links between it and the other attributes used by $\hat{f}$.

Such approach is considered ineffective when $A$ is correlated with $X$ as removing $A$ shifts its probability mass onto $X$. We still obtain a biased model for $X$, but under the false promise that it is fair just because it does not require $A$ as an input (see, e.g., Mougan et al. [203]). Using an Econometrics term, such approach introduces a *missing variable bias* into the model [111]. The issue, however, is that often models are prohibited from using $A$ [122]. It remains an open discussion between ML and legal scholars.

Most of Fair ML works today consist of *fairness through awareness* methods that require (some) information on $A$ to be implemented. In the case in which it is not possible to use $A$ as input to a model, e.g., there are techniques for obtaining a fair representation of $X$, or $\tilde{X}$, that contains as little information as possible from $A$. See Zemel et al. [317] for the formulation of the *fair representation learning* problem. Also see Dwork et al. [88] for the first explicit formulation of the *fairness through awareness* problem.

Bringing these characteristics together, we revisit the empirical risk minimization problem for learning the model $\hat{f}$ (2.1) to achieve fairness. Let $\psi$ denote the fairness definition (or goal) to be used as the constraint. We have:

$$\begin{aligned} \hat{f} = \underset{f \in \mathcal{F}}{\arg\min} & \sum_{i=1}^{n} \ell\big(f(\mathbf{x}_i), y_i\big) \\ \text{subject to} \quad & \psi \end{aligned} \tag{2.8}$$

The above formulation is specific for the optimization problem of learning a fair $\hat{f}$, though similar formulations apply, e.g., to learning a fair representation for the set of neutral attributes $\mathbf{X}$ or relabeling the model outcomes $\hat{Y}$. That is, we always want to minimize the loss (of information) as long as it meets a fairness constraint. In principle, the constraint imposes an *accuracy-fairness trade-off*, though it remains unclear if such trade-off is always the case [233].

### 2.3.1 Correlation Based

Here, we present the correlation-based (or data-driven) fairness definitions. These definitions are split into *individual* and *group* definitions based on whether the definition aims at achieving fairness for the individual level (e.g., a female applicant) or group level (e.g., the female applicants). We express these definitions in terms of probabilities. This is because, implicitly, the learned model $\hat{f}$ is an approximation to the conditional distribution of interest, $P(Y|X, A)$, representing the variation in outcomes as explained by the available attributes.

**Definition 2.3.1.** (Demographic Parity) Under demographic parity (DP), the values of $A$ should not determine the predicted outcome provided by the learned model. Formally, for a binary $A$, it is equivalent to:

$$P(\hat{Y} = y \mid A = 1) = P(\hat{Y} = y \mid A = 0) \tag{2.9}$$

It implies $\hat{Y} \perp\!\!\!\perp A$, known as the *independence criterion* for non-discrimination [26]. DP is also referred to as *statistical parity*.

DP does not exclude and might require positive (or reverse) discrimination, which is considered illegal in some cases and varies per country [234]. DP and its various equivalent notions appears in several papers, such as Kamiran and Calders [161] and Feldman et al. [97]. It is based on the legal notion of *disparate impact* (or, its EU equivalent, *direct discrimination*) described under anti-discrimination law.

Notice that (2.9) completely disregards the true outcome $Y$. On one hand, this favors the use of DP in settings where the model is applied to unlabeled data, which is a common setting in ML. On the other hand, this ignores how the fairness goals might hinder the model's ability to infer the ground truth. Regardless, DP remains a popular and widely used fairness definition.

**Definition 2.3.2.** (Equalized Odds) Under equalized odds (EO), the model should predict the desirable outcome $Y = 1$ at the same true positive and false positive rates across the values of $A$. Formally:

$$P(\hat{Y} = 1 \mid A = 1, Y = y) = P(\hat{Y} = 1 \mid A = 0, Y = y) \tag{2.10}$$

It implies that the model prediction is conditionally independent of $A$ given $Y$, or $\hat{Y} \perp\!\!\!\perp A \mid Y$, known as the *separation criterion* for non-discrimination [26].

EO was defined by Hardt et al. [131], in part, as a response to DP's shortcomings in accounting for fairness in terms of $\hat{Y}$ and $Y$. On one hand, EO requires the availability of $Y$, which is often only the case when training the model and thus its implementation is, in principle, limited without making strong assumptions about the training and incoming data. On the other hand, EO allows to view fairness in terms of predictions and decisions as discussed in Kilbertus et al. [169].

An extension to EO, also proposed by Hardt et al. [131], is *equal opportunity* in which (2.10) only considers the true positive rate:

$$P(\hat{Y} = 1 \mid A = 1, Y = 1) = P(\hat{Y} = 1 \mid A = 0, Y = 1) \tag{2.11}$$

which is considered a weaker notion of non-discrimination.

We note that earlier works by Pedreschi et al. [222] and Ruggieri et al. [244] provided the first formalization to equal opportunity (and, implicitly, to equalized odds) before Hardt et al. [131]. This was during time in which Fair ML was treated as *discrimination discovery*. These works, however, focus more on a data mining setting and rely on logic-based reasoning rather than probabilistic ML as their modeling framework.

**Definition 2.3.3.** (Calibration) Under calibration (CA), the model is considered to be calibrated if when it predicts that an applicant has the label $y$, the probability of the applicant actually having this label is the same for all values of $A$, Formally:

$$P(Y = y \mid A = 1, \hat{Y} = y) = P(Y = y \mid A = 0, \hat{Y} = y) \tag{2.12}$$

It implies $Y \perp\!\!\!\perp A \mid \hat{Y}$, known as the *sufficiency criterion* for non-discrimination [26].

Introduced by Chouldechova [64], CA is essentially the reverse of EO. Although these two definitions are similar, they have been shown to be incompatible. For these two to be compatible, the prediction error of the model has to be zero or $Y$ has to be independent from $A$, which are unrealistic conditions in practice [173].

The estimation of EO and CA is based on calculating the *confusion matrix* for the learned model. Recall that the confusion matrix consists in estimating:

- the true positives (TP), or the total number of cases where $P(\hat{Y} = 1|Y = 1)$;
- the true negatives (TN), or the total number of cases where $P(\hat{Y} = 0|Y = 0)$;
- the false positives (FP), also known as the type-I error, or the total number of cases where $P(\hat{Y} = 1|Y = 0)$; and
- the false negatives (FN), also known as type-II error, or the total number of cases where $P(\hat{Y} = 0|Y = 1)$.

where from these four quantities we estimate the true positive rate (TPR), or $TRP = TP\,/\,(TP + FN)$, and the true negative rate (TNR), or $TNR = TN\,/\,(TN + FP)$, among other measures. See Verma and Rubin [284, Section 3] for further details.

The estimation of DP is more flexible but it essentially boils down to a measure of equal representation in $\hat{Y}$. For instance, it can be measured by looking at the distance between the two distributions $P(\hat{Y}|A = 1)$ and $P(\hat{Y}|A = 0)$. All definitions can be extended conditionally by controlling for (some of the) neutral attributes.

**Definition 2.3.4.** (Individual Fairness) Introduced by Dwork et al. [88], individual fairness (IF) formalizes the notion that similar individuals should be treated similarly. Formally, in terms of probabilities, for two distinct but similar profiles $i$ and $j$:

$$P(\hat{Y} = y_i \mid X = x_i, A = 1) \approx P(\hat{Y} = y_j \mid X = x_j, A = 0) \tag{2.13}$$

with similarity defined by the metric $d$ such that $d(x_i, x_j) \approx 0$. We can also re-write it in terms of some acceptable $\epsilon$-deviation, meaning the model is considered individually fair as long as:

$$|P(\hat{Y} = y_i \mid X = x_i, A = 1) - P(\hat{Y} = y_j \mid X = x_j, A = 0)| \leq \epsilon \tag{2.14}$$

which allows to consider a range on IF notions.

The formalization (2.13) is one of many for IF as it depends on how and on what we define similarity. Already learning or assuming any given $d$ is a difficult task as defining two individuals as similar is never entirely objective. Such focus is what positions IF at the individual level, separating it from the previous three fairness definitions. Although IF has no specific non-discrimination criterion, it is the basis of non-discrimination law: treating similar individuals similarly, as argued since the time of Aristotle, guides the West's views of non-discrimination [288].

Today Fair ML currently has a dominant narrative focused on always classifying fairness definitions into individual and group level fairness. In principle, the tension is obvious if we consider that, e.g., under DP, while group fairness between male and female applicants is met, it is possible for the model to select a poorly qualified female but never her similar male counterpart, violating IF. However, this discussion is not as clear cut given the strong similarity statement required when implementing individual fairness. It can be the case, say, that one person views the poorly qualified female applicant and her male counterpart as similar while another person does not. That is because *similarity is a normative statement*, which not only poses a challenge in Fair ML but also in proving discrimination claims [296]. Binns [40], e.g., questions this apparent tension between individual and group level fairness using legal and political philosophy arguments along these same lines.

**Remark 2.3.1.** Overall, I find this discussion troublesome as the Fair ML field pays too much attention on whether a definition is individual or group level instead of what fairness notions the definitions are meant to operationalize and why. It is worth stressing that Dwork et al. [88] never introduce IF (2.13) as individual fairness: the paper simply operationalizes philosophical works by people like Rawls [231] while criticizing how DP (2.9) fails to consider that membership to $A$ is not the only attribute that defines an individual. We need to be more open to new definitions that move between the present notions of individual- and group-level fairness definitions.

### 2.3.2 Causality Based

Here, we present the fairness definitions that require auxiliary knowledge in the form of a structural causal model $\mathcal{M}$ (2.2), making them causality-based definitions. This is a growing area within Fair ML as causality, mainly in the form of causal inference, has been previously used for testing discrimination cases by social scientists [136, 176]. This is because discrimination focuses on whether the outcome is caused by, directly or indirectly, the protected attribute. We focus on the most impactful definitions for this work. For a broader view on these definitions, see the causal fairness surveys by Loftus et al. [181] and Makhlouf et al. [191]. General surveys on causality for ML are also helpful. We recommend Nogueira et al. [212].

**Definition 2.3.5.** (Counterfactual Fairness) A predictor $\hat{Y}$ of $Y$ is counterfactually fair given the protected attribute $A = a$, an unobserved (latent) variable $U$, and any observed variables $X$ if:

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a) \tag{2.15}$$

for all $y$ and $a' \neq a$. It was first introduced by Kusner et al. [177].

The interpretation of (2.15) is simple: a decision is counterfactually fair if it would have been the same had an applicant been from a different group in $A$. Notation-wise, $\hat{Y}_{A\leftarrow a'}(U)$ in (2.15) reads as "what would have been of $\hat{Y}$, under the latent variable $U$, had $A$ equalled $a'$. Given a SCM $\mathcal{M}$ and using the individual values for $X$ and $A$, each individual counterfactual quantity is generated via the abduction, action, and prediction steps introduced in Section 2.1.3.

Counterfactual fairness remains the most important causal fairness definition. Since its introduction, it has been extended and studied further several times: e.g., Chiappa [63] considers counterfactual fairness under specific paths of the SCM $\mathcal{M}$, while Kilbertus et al. [168] explores the robustness of counterfactual claims in the presence of hidden confounders in the SCM $\mathcal{M}$.[8]

Russell et al. [246], the companion paper to Kusner et al. [177], consider counterfactual fairness under a set of causal graphs (i.e., worldviews) for the same problem. It introduces *approximate counterfactual fairness*. The main idea is that the model is approximately counterfactually fair up until some $\epsilon$ difference across worlds, meaning the model can claim to be fair under multiple causal graphs.

**Remark 2.3.2.** Overall, the main challenge for causal definitions, especially under the kind of counterfactual queries involved for Fair ML, is that we always observe the factual *what is* $Y$ and never the counterfactual *what would have been if* $Y^{CF}$. This obvious challenge overall reduces to either a *matching problem*, which is the dominant approach outside ML, or a *representation learning problem*, which is the dominant approach within ML. Intuitively, counterfactual reasoning aims at answering questions relative to a unit or individual of interest in which, by construction, we only observe one outcome (what has happened) and resort to imagining an alternative outcome (what could have happened). The best possible setting would be to have access to the unit or individual of interest twice: its factual version and its counterfactual version. Beyond an exercise of imagination, this setting is not possible in practice. The second best setting is to find another unit or individual that resembles or to generate a unit or individual that represents the counterfactual version we wish to conceive.

**Remark 2.3.3.** The distinction between individual and group level fairness is blurry within causal fairness definitions. As most of these notions are based on comparing individual profiles, there is a more explicit link to individual fairness. This link is further stressed as, for instance, counterfactual reasoning is specific to the individual (e.g., Kusner et al. [177]). However, causality also implies a structure that applies to all individuals (e.g., Kilbertus et al. [167]). If the cause holds for one individual of a given group, why can it not hold for all other individuals that are members of the same group? This distinction is rarely discussed, but it is important as, for instance, discrimination claims are made at an individual level but argued at a group level.

We now turn to Kilbertus et al. [167], which uses causal reasoning in the form of a SCM $\mathcal{M}$ to formalize notions of discrimination. This work, as shown in the definitions below, relies heavily on the causal graph and its structure. The starting point is that all paths, direct or indirect, from the protected attribute $A$ are problematic. We then

[8] It is worth noting that some of these extensions, such as path-dependent counterfactual fairness, were already introduced by Kusner et al. [177] in the supplement material.

relax this "skeptic" point of view by acknowledging that some descendants of $A$ are less concerning to others.

Kilbertus et al. [167] defines resolving and proxy variables under the context of discrimination. The term *resolving variable* refer to any variable in the causal graph "that is influenced by $A$ in a manner that we accept as nondiscriminatory." The term *proxy variable* refers to any descendent of $A$ "that is significantly correlated with $A$, yet in our view should not affect the prediction."

**Definition 2.3.6.** (Unresolved Discrimination) A variable $X$ in a causal graph exhibits unresolved discrimination if there exists a directed path from the protected attribute $A$ to $X$ that is not blocked by a resolving variable and $X$ itself is non-resolving.

**Definition 2.3.7.** (Potential Proxy Discrimination) A variable $X$ in a causal graph exhibits potential proxy discrimination if there exists a directed path from $A$ to $X$ that is blocked by a proxy variable and $X$ itself is not a proxy.

**Definition 2.3.8.** (Proxy Discrimination) A predictor $\hat{Y}$ exhibits no proxy discrimination based on a proxy $X$ if for $x$ and $x'$:

$$P(\hat{Y}|do(X := x)) = P(\hat{Y}|do(X := x')) \tag{2.16}$$

All three definitions are subjective, illustrating the difficulty behind agreeing on discrimination and its problematic paths. We either claim all paths from $A$ are problematic or agree on what variables influenced by $A$ are resolving and, thus, which descendants are correlated enough with $A$ to be considered proxies. These are statements that might lie beyond data-driven methods and require the engagement of different stakeholders as emphasized in, e.g., Álvarez and Ruggieri [8]. We come back to the link between causality and discrimination in the next chapters.

**Remark 2.3.4.** The reason I include Kilbertus et al. [167] in this section is because it highlights how simple it is to define what discrimination is (and thus unfairness and bias) formally through a causal graph, but also how equally difficult it is to be precise about what makes something discriminatory through the same formalisms.

The definitions presented above are nothing new to legal scholars that study direct and indirect discrimination; Kilbertus et al. [167] simply re-introduce them under causal reasoning. These definitions are still subjective and context-specific: defining a resolving variable, e.g., is based on the problem formulation. In this work, I do not use Kilbertus et al. [167] beyond this section. However, works like Kilbertus et al. [167] help us to position how causal researchers view discrimination problems: $A \rightarrow Y$ represents potential direct discrimination while $A \rightarrow X \rightarrow Y$ represents potential indirect discrimination. In the first case, we need to find a method to block the direct path; in the second case, we first have to consider how acceptable is $X$ in its current form and then, if necessary, update it to account for the influence from $A$.

We conclude with a mention to causal reasoning for *harm*, a recent line of work by Beckers et al. [29] that I expect to grow in importance within Fair ML. The EU's AI ACT [92], e.g., addresses the issue of harm, not unfairness. The general idea is that a harmful decision is one that reduces the utility (below an agreed lower bound) derived by an individual from a decision. A decision can be unfair but, if it does not leave the

individual worst off (according to, say, a regulator) in terms of "utility", then it cannot be considered harmful. There is no reference to unfairness in this case. What the AI Act considers harmful is, for now, unclear. Further, Beckers et al. [29]'s causal definition of harm, which falls along the lines of counterfactual fairness but includes an individual-specific utility function, has no clear application either. It remains to be seen how harm evolves as a key notion for Fair ML.

### 2.3.3 Can Fair ML Be Unfair? The Yule Effect

> [W]e cannot infer independence of a pair of attributes within a sub-universe from the fact of independence within the universe at large.
>
> *G. Udny Yule [312, page 132]*

This section is partly based on the conference paper S. Ruggieri, J. M. Álvarez, A. Pugnana, L. State, and F. Turini. Can we trust fair-AI? In *AAAI*, pages 15421–15430. AAAI Press, 2023.

To conclude the discussion on fairness definitions, we study the Yule Effect [312] to highlight the importance of using auxiliary causal knowledge for Fair ML. We argue that the Yule Effect is introduced *by* the incorrect use of Fair ML methods. First, we present a causal reasoning approach for correcting the unfairness of the decision procedure behind $\hat{Y}$. Next, we describe a common approach to the problem that adopts group-level fairness correction. Finally, we discuss the consequences of such common approach with an example, highlighting the Yule Effect due to blindly correcting decision procedures.

Let us assume a scenario where we observe from historical data that $\hat{Y} \not\perp\!\!\!\perp A$, substantiated by a large risk difference. The risk difference, also called *total variation* in the causality literature, embeds direct, indirect, and spurious effects of $A$ on $\hat{Y}$ [226]. Spurious effects are introduced by confounding variables that cause both $A$ and $\hat{Y}$. We formalize the causal relations among $A$, $\hat{Y}$ and other observed variables using a DAG in Figure 2.5. Plecko and Bareinboim [226] call it the *standard fairness models*, as it summarizes most possible scenarios encountered.

From a causal perspective, we are interested in measuring the direct and indirect effects only, whose sum is the *average causal effect* (ACE):

$$P(\hat{Y} = 1|do(A = 1)) - P(\hat{Y} = 1|do(A = 0)) \tag{2.17}$$

which we explore using Figure 2.5. Let us consider now a third observed feature, called $Z$, which is the only input, together with $A$, to the decision procedure $\hat{Y}$. Let us develop a case-based reasoning around $Z$.

As a first case, consider the situation in which $Z$ is a mechanism through which the causal effect of $A$ propagates to $\hat{Y}$, acting as a mediator. Relative to Figure 2.5, this setting means that $W = Z$ and $V$ is removed. Examples of mediators include legitimate business requirements, such as level of education or prior working experience. In this case, the ACE is equal to the risk difference metric, and since $\hat{Y} \not\perp\!\!\!\perp A$, it is non-zero. Therefore, the decision procedure leading to $\hat{Y}$ is unfair, and it *should* be corrected.

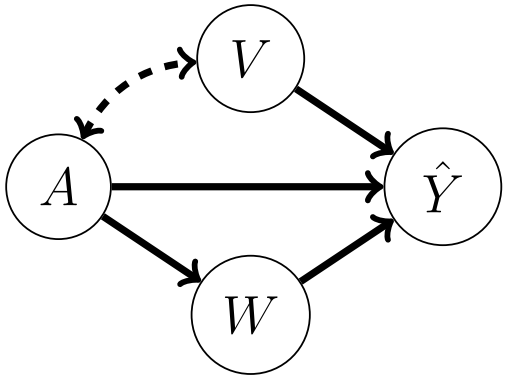



Figure 2.5: The standard fairness model according to Plecko and Bareinboim [226]. Direct edges model possible causal dependencies. The dashed edge models spurious correlation induced by unobserved features. $V$ is a confounder. $W$ is a mediator.

As a second case, consider the situation in which $Z$ is correlated with $A$, acting as a confounder. Relative to Figure 2.5, this setting means that $V = Z$ and $W$ is removed. Examples of confounders include demographic and geographic features. In such a case, the ACE can be calculated by averaging the stratified risk difference on $Z$ using the *adjustment formula* (2.7). We define the adjustment set in terms $Z$. Formally:

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

Given $Z$ as a confounder between the protected attribute and the decision procedure, let us distinguish two sub-cases. The first sub-case assumes $\hat{Y} \perp\!\!\!\perp A \mid Z$, and it is known as *Simpons' Paradox*:[9]

$$\hat{Y} \not\perp\!\!\!\perp R \quad \wedge \quad \hat{Y} \perp\!\!\!\perp A \mid Z \tag{2.18}$$

and it occurs when vanishing correlations in separate distributions do not produce a vanishing mixture. In such a case, each term in the previous sum centered on $Z$ above is zero, and, a fortiori, the ACE is zero. We *should not* correct the decision procedure leading to $\hat{Y}$. This reasoning extends to collapsible association measures, such as the selective risk ratio, for which the value in the mixture is a weighted average of the values in the separate distributions [218]. For non-collapsible metrics, the value at the mixture can be outside of the range of the values in the separate distributions. Hence, for non-collapsible metrics, the decision procedure *should* or *should not* be corrected based on the value at the mixture, which can be computed from the adjusted formula.

The second sub-case assumes $\hat{Y} \not\perp\!\!\!\perp A \mid Z$. At least one term of the previous sum centered on $Z$ is non-zero. Also, terms can be of opposite sign, which means that the overall sum can be zero or non-zero. The decision procedure *should not* or *should* be, respectively, corrected based on the result of the sum.

These cases highlight how the role of $Z$ with respect to $A$ and $\hat{Y}$, defined by the structure of the causal relationships, conditions the required correction. It can be difficult in practice to determine whether $Z$ is a mediator or a confounder [26, Chapter 5]. This confusion may lead to the wrong action with regard to the correction of the

[9]The term has been improperly extended to include the Yule Effect, see [270, Sect. 3.5.2].

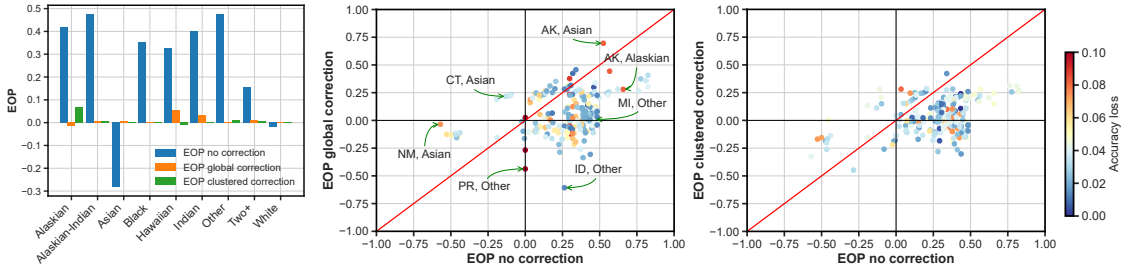



Figure 2.6: Left: EOPs for each racial group for classifiers with no correction, global correction, and clustered correction. Center: EOPs for each state and race, with color denoting the loss in accuracy after global correction. Right: EOPs for each state and race, with color denoting the loss in accuracy after clustered correction.

decision procedure. Non-causal approaches dismiss the above case-based reasoning altogether (recall, Section 2.3.1). They test independence only or separation only (recall Definitions 2.9 and 2.10). Hence, a typical approach after observing $\hat{Y} \perp\!\!\!\perp A$ consists of blindly correcting the decision procedure leading to $\hat{Y}$. With a few exceptions that will be recalled next, research papers adopting non-causal approaches fall back to this.

What are the consequences of failing to understand (or disregarding) the causal structure around $Z$? Let us assume that the decision procedure is corrected and deployed. We would then observe (close to) zero risk difference, which would support the conclusion $\hat{Y} \perp\!\!\!\perp A$. Is everything all right? According to the case-based reasoning above, the correction of the decision procedure may have mitigated or may have worsened fairness of the procedure.[10]

Let us consider an example based on the *ACSIncome* dataset, an excerpt of the U.S. Census data recently curated by Ding et al. [85].[11] With reference to Figure 2.5, we set $A$ to be the race of individuals, $\hat{Y}$ the predicted income (above 50K USD or not), $W$ the number of working hours per week, and $V$ the state of residence. Moreover, let $Y$ be the true income. We split the available data into 67% for training a classifier, and 33% for testing its predictive performances and fairness metrics. An initial classifier is built using LightGBM [165], a state-of-the-art gradient boosting approach. We adopt the separation metric of the quality of opportunity (EOP), or Definition 2.11, by Hardt et al. [131]:

$$P(\hat{Y} = 1|Y = 1) - P(\hat{Y} = 1|Y = 1, A = i)$$

which is the difference between the recall of positives (i.e., classified as having and actually having an income above 50K USD) at population-level and at the level of the $ith$ racial group present in the data.

The larger the EOP, the worse is the ability of the classifier to recall positives of the group compared to the average recall. The EOPs observed over the test set are reported in Figure 2.6 (left) in blue, from which we conclude $A \not\perp\!\!\!\perp \hat{Y} \mid Y$. Let us now correct the decision procedure by a post-processing method that specializes the decision threshold

[10]Interestingly, the graph in Figure 2.5 is likely no longer faithful to the new data. Since faithfulness is required by many approaches for causal discovery, reconstructing the causal structure of the new data (e.g., in an external audit study) may become problematic.

[11]The notebook is available at `https://github.com/ruggieris/DD/blob/main/notebooks/dd_ACSIncome_Yule.ipynb`.

for each racial group of $A$ [131]. The EOPs observed after this global correction are shown in Figure 2.6 (left) in orange. They are closer to the optimal value of zero.

In this context, we would expect the corrected classifier to be fair not only at the country level, but also at the state level. However, here the state is a confounder that the correction of the classifier has not accounted for. Figure 2.6 (center) shows that the EOPs of racial groups at each state have been affected by the correction in different ways. For instance, “Other races" in Michigan (MI) have a considerably lower EOP after the correction. Asians in Alaska (AK), instead, have a higher EOP metric after the correction. “Other races" under ID moved from being disfavored to being favored considerably: i.e., they moved from a recall much lower than average to a recall much higher than average. “Other races" in Puerto Rico (PR), which were not disadvantaged (close to zero EOP), after the correction result now to be advantaged (large negative EOP). Conversely, Asians in Connecticut (CT), which were favored, become disfavored after correction. Finally, notice that the loss in accuracy at state level after the correction, denoted by the color of dots, can be as high as 10% and it is not uniform across states, nor is there a clear pattern for how it is distributed.

**Definition 2.3.9.** (Yule's Effect) The Yule Effect occurs when vanishing correlation in the mixture of a few distributions does not produce vanishing correlation in separate distributions. Formally:

$$\hat{Y} \perp\!\!\!\perp A \quad \wedge \quad \hat{Y} \not\perp\!\!\!\perp A \mid Z \tag{2.19}$$

It can occur when positive and negative associations between the predicted outcome $\hat{Y}$ and the protected attribute $A$ when conditioning on a third attribute $Z$ cancel out.

**Remark 2.3.5.** Yule's Effect is precisely what has been pointed out in the *ACIncome* example above. Whenever we aim at group-level fairness, such as independence $\hat{Y} \perp\!\!\!\perp A$, but we wrongly disregard to control for $Z$ based on auxiliary causal knowledge, Fair ML algorithms may result in disparate effects on separate distributions, with some impacted positively (higher fairness) and other impacted negatively (lower fairness). Further, we can *combine Simpson's Paradox and Yule's Effect into a well-known general statement about conditional independence.* Relative to Figure 2.5,

*For* $\mathbf{W} \subset \mathbf{Z}$*, then* $\hat{Y} \perp\!\!\!\perp A \mid \mathbf{W}$ *neither implies nor is implied by* $\hat{Y} \perp\!\!\!\perp A \mid \mathbf{Z}$*.*

As a consequence of the above general statement, independence fairness ($\hat{Y} \perp\!\!\!\perp A$) does not imply nor is implied by conditional independence fairness ($\hat{Y} \perp\!\!\!\perp A \mid Z$). Here, recall that the independence criterion refers to DP (2.9). Similarly, separation fairness ($\hat{Y} \perp\!\!\!\perp A \mid Y$) does not imply nor is implied by conditional separation fairness ($\hat{Y} \perp\!\!\!\perp A \mid Y, Z$). Here, recall that the separation criterion refers to EO (2.10). Moreover, when multiple confounders are present, conditional independence/separation with respect to all of them does not imply nor is implied by conditional independence/separation with respect to *a subset* of them. Hence, if we want to properly implement (conditional) independence and separation metrics, we should be aware of all the confounders to control for all of them or, alternatively, we should be aware of a more detailed structural causal model that allows for finer reasoning about all the confounders. Either case would allow to better implement (conditional) independence and separation metrics.

Not adequately controlling for potential confounders may limit the implementation of (conditional) independence and separation metrics. Based on the above reasoning, in

practice, we could either ignore the potential confounders or account for all the potential confounders. Neither approach is desirable. If we ignore the potential confounders and, in turn, make no distinction between a confounder and a mediator, then notice that by wrongly controlling for a mediator, through the adjustment formula, we only measure the direct causal effect of $A$ on $\hat{Y}$ and are ignoring the indirect effect of $A$ on $\hat{Y}$. The indirect effect can be positive if the mediator positively affects the advantageous decision, and the social groups have disproportionate distributions over the mediator. This is the case, for instance, for education level in job candidate selection, since social groups have disproportionate access to education. The indirect effect can also be negative if the mediator results from the implementation of positive actions, such as quotas in favor of disabled individuals.

Similarly, if we account for all potential confounders, then notice that the number of strata to control for under these conditional definitions can be very high. In the above *ACIncome* example there are 51 states (counting Puerto Rico). In general, the number of strata is equal to the product of the cardinalities of the domains of the features to control for. As a partial solution, Kamiran et al. [162] proposes to cluster the strata into a few groups to control for. Figure 2.6 (right) reports the result of separately correcting the classifier for each of the five groups of states. These groups are obtained by clustering states based on the probability distribution of races within them using the k-means algorithm. Compared to the global correction, the clustered one is beneficial with respect to both EOP and accuracy loss. The mean absolute EOP is 0.258 for the uncorrected classifier, 0.119 for the globally corrected one, and 0.105 for the clustered corrected classifier. Still, such a step implies a loss of granularity that, depending on the task, makes it unappealing to use by a practitioner.

In either approach, auxiliary to causal knowledge would be helpful as it would helps us identify the confounders from the mediators or, alternatively, reduce the number of clusters to consider based on this distinction. Overall, the Yule Effect highlights a shortcoming of correlation-based Fair ML. Because these fairness definitions (recall Section 2.3.1) are oblivious to the causal structures underlying an ADM process, correcting for fairness may sometimes be worst than not correcting at all. This is because structure conditions the flow of information in a system and, thus, determines how we should correct for anything within that system. Such a focus on structure is clear in the causal-based Fair ML (recall Section 2.3.2).

Importantly, such sort of causal reasoning requires, first, accepting the role of auxiliary causal knowledge in reasoning about fairness problems and, second, full (or sufficient) access to such auxiliary causal knowledge for it to be useful. This is the case, at least, for this thesis as this entire chapter illustrates. Causal reasoning then allows us to avoid issues like Yule's Effect and, in that sense, to enhance popular fairness definitions, ensuring that Fair ML remains fair when implemented.

# Chapter 3

# Revisiting the Comparator

This chapter is based on the working paper: J. M. Álvarez and S. Ruggieri. Uncovering algorithmic discrimination: An opportunity to revisit the comparator. *CoRR*, abs/2405.13693, 2024. It is currently under submission.

> [I]t is far from clear to whom we owe a gesture of epistemic solidarity.
>
> *The Right to Sex* by Srinivasan [271, p. 11]

The discrimination *comparator*, henceforth comparator, is the individual profile used for testing the discrimination claim of the *complainant*. As the name suggests, the comparator serves as a comparison to the complainant, often varying only in terms of membership to the protected attribute on which the discrimination claim is based on but being similar (or comparable) on all other relevant attributes. For instance, suppose we are auditing a hiring process after a female candidate, Martina, complained that she was discriminated based on her gender. Now Martina becomes the complainant, and the first step would be to find (or generate) her comparator, Martin, meaning finding (or generating) a male candidate with a profile that approximates that of Martina. The next step would be to compare Martina and Martin under the same hiring process to test the discrimination claim. As both profiles are similar except for gender, the only source of variation that could explain a difference in outcomes would be the fact that Martina is female and Martin is male. To be certain, an intermediate step would be to find other "Martinas" and "Martins" (i.e., similar female profiles to Martina and similar male profiles to Martin) and compare the average difference in outcomes. This is because the literal comparison of Martina-vs-Martin is not enough evidence, in some cases, to rule out randomness from the decision process, requiring a many-to-many comparison centered around the individual discrimination claim of the complainant.

The comparator is present in all methods for testing discrimination,[1] illustrating the comparative element intrinsic to discrimination itself both as a societal concept and as a modeling problem. We can only claim that Martina was discriminated by comparing her to Martin and, if needed, to other female and male profiles like her. Further, the com-

[1] Unless we take an *individualized* view on justice [40]. This is a non-standard view on justice that we will discuss later in Chapter 7.

parator represents the counterfactual reasoning that underpins most, if not all, methods for testing discrimination. What is Martin other than an answer to what would have happened to Martina had she been male? Unable to compare confidently all the possible worlds of Martina, which are, by definition, hypothetical and based on our societal imagination, we rely on what is observed in this world by finding the closest possible Martin to represent (with some confidence) what could have been of Martina.

The concept of the comparator is simple and intuitive, but its implementation can be neither simple nor intuitive. What exactly does it mean for Martina and Martin to be similar? For instance, as argued in other works like Heckman [136], knowing what we know about gender in our society, is comparing these two profiles enough to "isolate" the effect of gender? Are we able, or should we attempt at all, to approximate Martina's counterfactual world(s) through Martin's experience? All of these questions translate into concerns about how we test for discrimination as well as modeling problems to be addressed when testing for discrimination.

In this chapter, we revisit the comparator under a causal perspective. Under the premise that the comparator and, thus, the methods for testing discrimination follow counterfactual reasoning, what Kohler-Hausmann [176] refers to as the *counterfactual causal model of discrimination*,[2] we propose two formulations of the comparator. We also propose a causal desiderata for testing discrimination based on our revision of the comparator as well as survey of the representative literature.

## 3.1 Establishing Discrimination

Moving forward, let us consider the tenure Example 3.1.1. This example is based on Morgan et al. [202]'s work on the unequal impact of parenthood in academia. It will helps us illustrate later on our two definitions for the comparator.

Morgan et al. [202] find that mothers in academia, on average, experience a negative impact on their careers relative to comparable fathers in academia. One example of this trend is the effect of parental leave on the number of publications. While male academics show an increase in the number of publications during parental leave, female academics show a decrease or no publications at all compared to their pre-parental leave level of output. A key reason for this difference is the different roles played by males and females when taking care of the newborn. Mothers take care of the newborn during the first months, which tend to be the more stressful ones, in order to, e.g., breastfeed the newborn, while fathers have a more prominent role in the later months, which tend to be less demanding. As a consequence of this timing, on average, female academics have time only for the newborn during parental leave while male academics are able to research during parental leave. Social expectations and household dynamics, as Morgan et al. [202] argue, are overall key factors in creating this difference. In short, based on these empirical results, what we know is that academic mothers, as the primary caregivers, tend to carry the burden of childbearing within their households, which consumes potential time that could be spent on advancing ones academic career

[2] I wrote *Counterfactual Situation Testing* [8], which we will cover in Chapter 4, in part, as a response to Kohler-Hausmann [176], which I consider seminal work for questioning our current understanding on testing for algorithmic discrimination.

by, e.g., writing papers at home or networking at conferences. Academic fathers tend to share less of this burden.

**Example 3.1.1.** (Tenure at the University of Pisa) For illustrative purposes, let us assume that the University of Pisa only looks at the number of publications to grant tenure. Suppose Clara, who was denied tenure for having published only 12 papers, files a discrimination claim against the university based on gender, becoming the complainant. To test her claim we must consider the relevant information used for the tenure decision (the number of publications) along with the information linked to the protected group (gender) and find a comparator, meaning a male individual that went through the same decision process. Suppose that we find Mike, who also published 12 papers and was also denied tenure by the university. Do we deny Clara's discrimination claim? Further, given what we know about parenthood's impact on academic performance (recall, Morgan et al. [202]), would our answer change knowing that Clara is a mother? In that case, assuming that Mike is also a father, would we still deny Clara's discrimination claim? Or, furthermore, assuming a measurable penalty in terms of number of publications for being a mother in academia, would we be willing to accept another comparator, Vincent, with 18 publications and recently tenured, to base our assessment of Clara's discrimination claim?

Regarding Example 3.1.1, in other words, what are willing to assume about gender and its effect on parenthood and academic performance given what we know from works like Morgan et al. [202]? *Clearly, whatever we assume will determine who we define as the comparator for the complainant Clara and whoever we define as the comparator will determine the validity of Clara's discrimination claim.* In shorty, who is more similar or comparable to Clara: Mike or Vincent? It is not an easy question to answer nor to model. It highlights the complexity behind the comparator.

For Example 3.1.1, let us denote gender as $A$, number of publications as $X$, and the tenure decision as $Y$ in Example 3.1.1. We can describe the decision process of the university using a SCM as $X \rightarrow Y$, while we can similarly describe Clara's claim as $A \rightarrow X \rightarrow Y$. We will come back to these causal graphs shortly.

**Establishing Discrimination in the EU.** In Section 2.2.1, Chapter 2, we introduced the two forms of discrimination conceived by EU non-discrimination law: *direct discrimination* and *indirect discrimination.* Following Weerts et al. [296], there are four main elements in a discrimination case, be it direct or indirect discrimination, that we present below. We add in italics the implications of each to automated decision making (ADM) where necessary.

- **"On grounds of"...** We need to determine whether the claim falls under direct or indirect discrimination, which will depend on whether the decision is taken based on (i.e., "on grounds off") the protected attribute or not. *For ADM, this element has been interpreted in terms of whether the protected attribute is (direct) or not (indirect) an input of the model [122]. Recent works have extended this debate by pointing out that if the model does not use the protected attribute but is able to infer it through other attributes, which, in turn, means that it affects its decision-making, (i.e., proxy discrimination [278]) then it falls under direct discrimination [1].*

- **..."a protected characteristic."** The discrimination claim must be based on an attribute (or characteristic) protected by non-discrimination EU law. *For ADM, this element, on top of the current discussions on intersectional versus multiple discrimination [308], is ongoing as there is fear that the models are able to create new protected attributes. For instance, Weerts et al. [296] stresses that the current scope of the law fails to protect for people's income or socioeconomic background, which can be used to inform the models.*

- **... where there is evidence for "less favorable treatment" or "particular disadvantage"...** The complainant, to establish a case of discrimination, needs to show *prima facie* evidence, meaning "sufficient evidence for a rebuttable presumption of discrimination to be established by the judge" [296]. This is where the comparator comes in as a form of evidence. It possible to provide a hypothetical comparator [296]. If *prima facie* discrimination is established, the burden of disproving discrimination falls to the defendant. *For ADM, by the nature of the model, evidence itself becomes standardized. In principle, for the evidence we must focus only on the input and corresponding output of the model.*

- **...unless there is an "objective justification."** Direct discrimination cannot be justified in principle. Indirect discrimination, instead, can be justifiable as long as it has a legitimate goal and passes the proportionality test. As Weerts et al. [296] point out, neither the law provides concrete guidelines for the proportionality test nor can it be settled in advance. Further, the "objective justifications" are settled on a case-by-case basis. *For ADM, this element raises the point that we can train a model that, in principle, could be used across multiple decision-making cases but to judge its fairness (for the purpose of testing discrimination) will depend on each case [289].*

Of the above elements, the third element, which refers to the *prima facie* discrimination evidence, is the most relevant one to the comparator. It highlights the role played by the comparator in establishing discrimination: it is the basis for the evidence. *What we define as the comparator inevitably determines what we consider and test for as discriminatory.* It is under this third element where the tools for testing discrimination enter. These are, after all, tools used for providing *prima facie* evidence.

Let us revisit Example 3.1.1 under these four elements. First, we would be facing an instance of indirect discrimination as the university only uses number of publications for the tenure decision. Second, the protected attribute is clearly gender, though the complainant would have to make the case that parenthood (as a consequence of gender) merits the same considerations. Third, we would need to provide evidence for Clara's claim by finding a comparator and, potentially, other individual profiles that went through the same process and suffered a different outcome. Fourth, assuming a potential indirect link between gender (through parenthood) and the tenure decision, the university can still make a legitimate case for why it uses number of publications for its decision. It would be up to the judge to decide if it is valid.[3]

**On controlling for chance.** We extend the third element for establishing discrimination of Weerts et al. [296] by emphasizing the role of *many-to-many comparisons versus*

[3]I am clearly not a lawyer. Again, this is a hypothetical example written for illustrative purposes.

*the literal one-to-one comparison between the complainant and its comparator.* What is interesting, though unsurprising, is that most tools for testing discrimination provide confidence intervals and, overall, measures of certainty around their claims (see, e.g., the multi-disciplinary survey by Romei and Ruggieri [234]). We view it as unsurprising because most of these tools come from a long and established modeling culture built on inferential statistics. Similarly, there is, again unsurprisingly, a lack of inferential statistics within the more recent tools, such as the FlipTest [41], precisely because they are built on a culture of predictive modeling that characterizes machine learning. Future work should look at this argument more systematically and jointly with legal scholars. We also recommend Breiman [49]'s position paper, *Statistical modeling: The two cultures.*

What is interesting is that this focus on quantifying certainty seems to have been adopted (or, at least, expected) by lawyers as literal comparisons may be considered as insufficient evidence [100]. Although we have yet to witness fair machine learning tools as evidence for *prima facie* discrimination [296], we believe there might be some tension between how these tools reports their findings and how lawyers expect these findings to be reported in terms of certainty. This too is unsurprising to us. If it takes more than a throw to check whether a coin is fair, why should it not be the same for a serious accusation like discrimination? Intuitively, we wish to control for chance from (or rule out out uncertainty in) the decision process in question to have (some) certainty on the pattern we are testing for. In principle, this is not possible under a literal comparison.

## 3.2 The Counterfactual Model of Discrimination

Discrimination is often conceived as a causal claim on the (in)direct effect of the protected attribute $A$ on the decision outcome of interest $Y$ [136]. For instance, was the candidate hired *because of* (or, equivalently, *as a cause of*) his or her race? Its causal underpinning, which is motivated by non-discrimination law's own definition of discrimination [296], has long motivated a range of methods for testing discrimination based on *counterfactual reasoning*. For instance, would the candidate have been hired had he or she been of another race? In practice, such counterfactual reasoning has led these methods to operationalize the scenario in which we are able to manipulate the protected attribute of the individual(s) making the claim, imagine the "what would have been if" (or the counterfactual) outcome, and compare it to the "what is" (or the factual) outcome to isolate the causal effect of $A$ on $Y$. Kohler-Hausmann [176] refers to this practice as the *counterfactual causal model of discrimination.*

As we will discuss in Section 3.4, the counterfactual model of discrimination[4] motivates the traditional methods, such as correspondence studies [35] and natural experiments [109], as well as the recent algorithmic methods, such as discrimination discovery [244] and individual fairness [88]. Overall, the counterfactual model of discrimination, as argued by Kohler-Hausmann [176] with whom we agree, dominates the methods used to test for discrimination.

[4]Henceforth, I drop the "causal" from the name as counterfactuals are causal by definition.

### 3.2.1 Why Counterfactuals?

It is a question, to the best of our knowledge, often taken for granted by the literature. We believe that this is the case because, given the current conception we have as a society on discrimination as a social phenomenon and a modeling problem, there is no better alternative to the counterfactual model of discrimination.

As a social phenomenon, quoting from Weerts et al. [296], who base their definition from Lippert-Rasmussen [180], "discrimination can generally be characterized by the morally objectionable practice of subjecting a person (or group of persons) to a treatment in some social dimension that, for no good reason, is disadvantageous compared to the treatment awarded to other persons who are in a similar situation, but who belong to another socially salient group." Lippert-Rasmussen [180] considers a group to be socially salient "if perceived membership of it is important to the structure of social interactions across a wide range of social contexts." Discrimination is a social phenomenon since, e.g., defining the socially salient group or the morally objectionable practice can only be done by first defining a specific, shared social context. Naturally, the social context is (meant to be) captured or, at least, (partially) established by the non-discrimination laws governing the society. The establishment of a protected (by non-discrimination law) attribute, for instance, illustrates this process.

At the core of this definition is the *comparative element*. Here, *we are not interested in the difference in treatment between any two individuals from different socially salient groups, but in the difference in treatment between two similarly situated individuals from different socially salient groups.* The notion of *similarity*, be it in terms of similarly situated individuals [180] or similar individuals [297], is, thus, central to discrimination. In other words, discrimination occurs when similarly situated (or similar) individuals that differ on membership to a socially salient group are treated differently. Conversely, discrimination does not occur when similarly situated (or similar) individuals that differ on membership to a socially salient group are treated equally. As Weerts et al. [296] point out, under this setting, discrimination becomes the opposite of equality.

Discrimination, under this setting, appears simple and even intuitive, capturing an argument that dates back to Socrates [297]: treat similar individuals similarly. This simple argument too presents a path to follow when testing for discrimination: show that similar individuals are treated similarly. The simplicity of this argument, however, breaks down when defining similarity between individuals. In *The Empty Idea of Equality*, Westen [297] points out at this limitation of non-discrimination law by arguing that equality is a circular concept: to argue for equality we first must define what it means to be equal. It follows, thus, that similarity is also a circular concept.

If conceiving discrimination is a social phenomenon, then testing for discrimination is a modeling problem. To test for discrimination, we must first formally define what it means for two individuals to be similar. It is possible to argue for similarity between individuals without resorting to mathematical formalism in the form of, e.g., defining a distance function $d$ for similarity (see, e.g., Loi et al. [183]), but these arguments still require the backing of evidence from data. For instance, in Europe it is encouraged to provide evidence beyond the literal comparison [100], meaning that the argument around the discrimination claim must be backed by other cases or data. The need for evidence, in turn, poses the discrimination claim as a modeling problem. We need to find or generate other similar individuals to build our evidence, and, for that, using $d$ is

more convenient than arguing case by case.

Resorting to mathematical formalisms to define similarity, be it because it gives a (false) sense of objectivism or simply because it is more scalable, is the common practice [176, 234]. The fact that the methods in Section 3.4, e.g., exist and are used (sometimes even required) for proving discrimination illustrates the role played by modeling in testing for discrimination. Defining (implicitly or explicitly) similarity under $d$ becomes a formal objective when testing for discrimination, and what we define as similar under $d$ determines what we test for as discriminatory or unequal.

Still, among other modeling frameworks, why the preference for counterfactual reasoning when testing for discrimination? We argue that it is due to two factors. First, it is due to the framing of the protected attribute $A$ as a cause of the outcome $Y$, making the testing for discrimination a causal problem. There is a long tradition in causality, especially within SCM (see Chapter 2.1), to formulate *what constitutes a cause in terms of counterfactual reasoning* (see, e.g., [124, 125, 126, 218]): essentially, if intervening $A$ induces a counterfactual outcome $Y^{CF}$ different from the factual outcome $Y^F$, then $A$ must be a cause of the outcome $Y$.[5] It is a form of reasoning rooted in the manipulationist view of causality [303], in which only through the manipulation $A$ can we test its causal effect on $Y$.

Overall, although not discussed enough [8], this first factor is rooted in the idea of discrimination as a pattern, which, in turn, implies structure. Intuitively, a discriminatory decision-maker has (un)consciously set up a decision process that links causally $A$ with $Y$, leading to the formation of a discriminatory decision-making pattern. Under this setting, to test for the existence of such pattern would imply to examine the impact of $A$ on the decision process by changing $A$ itself. If there is such a pattern, then observing a change in outcome due to having changed only $A$ (i.e., by controlling for everything else) would, in theory, first, confirm the existence of the pattern, and, second, confirm the role of $A$ within it.

Second, the popularity of counterfactual reasoning for testing discrimination, we also argue, is due to its (apparent) intuitiveness. This argument would also explain the wide acceptance of the counterfactual model of discrimination by modelers, lawyers, and other stakeholders. Although it is trivial to point out, no individual $i'$ can be more similar to another individual $i$ than that same individual $i$. Conceptually, when testing for discrimination we are, in turn, trying to imagine how the same individual would have been like in, essentially, another life or possible world. This is the sort of mental exercise we carry out when asking "what would have been of a female candidate had she been male?" It is a hypothetical question that, regardless, we aim to answer in practice.

We cannot observe, for any individual $i$, his or her "what is," or factual, and "what would have been if," or counterfactual, outcomes; we are only certain of the former, factual outcome. Given this issue, the best we can do is to look for another individual $i'$ that represents (or approximates) the latter, counterfactual outcome of $i$. Defining similarity between individuals through $d$ is an attempt at controlling for all factors that might influence an outcome in order to be able to answer whether one of those factors, i.e., $A$, is a cause of $Y$. We stress that this line of reasoning is implicit to all of the methods built over the counterfactual model of discrimination. Under the counterfactual model

[5] These definitions are, of course, much more formal (e.g., in terms of probability distributions) but this phrase captures the main idea behind all of them.

of discrimination, similarity is not only a statement between two individuals $i$ and $i'$, but it is also a statement on the alternative paths attributed to an individual $i$ through the lived experience of another individual $i'$. It is what makes counterfactual reasoning as a framework both intuitive to accept and complex to implement because we cannot escape the need to define similarity between individuals.

As with discrimination and its link to equality, it is fair to assume that we, as a society, might all agree on the principle behind counterfactual reasoning for testing discrimination but might also have serious disagreements when answering the question "similar to what?" This critique against counterfactual reasoning for discrimination is not new. Heckman [136], e.g., argues that causal methods that test for discrimination fail to understand that no $d$ will render two individuals similar enough to silence all disagreements. To be fair, though, in practice it is unlikely that anybody actually believes that such a $d$ exists; the problem is that we still need to define one to test for discrimination. Kohler-Hausmann [176], however, extended past criticism by pointing out that, in our search for similar individuals when testing for discrimination, we are ignoring the role of the protected attribute in shaping the set of available paths of any individual in question. We will comes back to this point in the next two subsections.

It seems unlikely that we move away from the counterfactual model of discrimination as the comparative element of discrimination is difficult to escape both when defining and testing discrimination. The causal wording behind discrimination makes it even harder. Hence, moving forward, we must focus on the limitations of the counterfactual model of discrimination, the main one being its circular nature due to the need to always needing to define what similarity between individuals means. In other words, we must revisit the comparator.

### 3.2.2 The Comparator

The comparator, as the name suggests, refers to the individual (profile) $i'$ used for testing the discrimination claim of the individual (profile) $i$. We also refer to the individual (profile) $i$ as the complainant. *The comparator captures the essence of what we mean by similarly situated or similar individuals that belong to different socially salient groups.* Hence, how we (choose to) define, identify, or generate the comparator says a lot about our view on similarity between individuals through $d$ and, consequently, determines how we test for discrimination.

**Definition 3.2.1.** (The Comparator) For the individual profile $i$, $\langle \mathbf{x}_i, a_i, y_i \rangle$, we define its comparator as the individual profile $i'$, $\langle \mathbf{x}_{i'}, a_{i'}, y_{i'} \rangle$, where $a_i \neq a_{i'}$, such that

$$d\big(\mathbf{x}_i, \mathbf{x}_{i'}\big) \leq \epsilon$$

with $d$ denoting the similarity measure (or distance function) and $\epsilon \in \mathbb{R}^+$ the similarity threshold allowed.

Let us stress three central characteristics of the above definition. First, we stress, as also argued by Weerts et al. [296], that Definition 3.2.1 is a *normative statement* on not only what similarity between individuals means but also on what equality (or lack thereof) means. Intuitively, the comparator is a sufficiently close $i'$ individual profile to $i$ individual profile according to $d$ and what is observed. What Definition 3.2.1 implies is

that the complainant $i$ and its comparator $i'$ should be treated equally in terms of their respective outcomes $y_i$ and $y_{i'}$ as long as they are equal in terms of their respective sets of non-protected attributes $\mathbf{x}_i$ and $\mathbf{x}_{i'}$; deviations from this would point at potential *prima facie* discrimination.

Second, we stress that in Definition 3.2.1 *similarity comes down to a comparison on the non-protected attributes* $\mathbf{X}$ between $i'$ and $i$. What this means is that the notion of similarly situated (or similar) individuals rests solely on what we observe (or have measured) as $\mathbf{X}$ for $i$ and $i'$. Notice that we have yet to make any claims between $\mathbf{X}$ and $A$. As argued by Kohler-Hausmann [176], this is an important and often overlooked link that we will address in the next subsection.

Third, we stress that the comparator $i'$ in Definition 3.2.1 *represents a sufficiently close counterfactual representation of* $i$, thus, embodying the counterfactual model of discrimination criticized by Kohler-Hausmann [176]. As discussed previously, finding that similarly situated (or similar) individual $i'$ to the individual $i$ is not only a statement on overall similarity between $i$ and any $i'$, but also a statement on the counterfactual life, in theory, available to $i$ had he or she been a member of $i'$'s group. In other words, the comparator $i'$ is a representation of the counterfactual world of $i$ according to $d$ [8, 176, 245]. Implicitly, by finding (or generating) the individual profile of $i'$, we are answering the counterfactual question "what would have been of the individual profile $i$ had the individual been of a different category within the protected attribute?"

Back to Example 3.1.1, we observe all three characteristics when defining Clara's comparator. First, by defining Mike or Vincent as her comparator, we implicitly make a normative statement on what two similar male and female academics, in that context, should look like. Second, regardless of the comparator we choose, similarity manifests through comparing the number of publications between Clara and her chosen comparator. Third, what we aim to answer with either Mike or Vincent (or any other male profile that we deem comparable to Clara) is where would Clara be had she been born male and undertaken academia as a career path.

### 3.2.3 Fairness Given the Difference

The phrase *fairness given the difference*, or FGD, we argue, embodies Kohler-Hausmann [176]'s main criticism of the counterfactual model of discrimination. We use FGD interchangeably with the KHC, or the *Kohler-Hausmann Critique*.

What is the Kohler-Hausmann Critique? Under the premise that the comparator $i'$ in Definition 3.2.1 represents a counterfactual representation of the complainant $i$, Kohler-Hausmann [176] criticizes the counterfactual model of discrimination for ignoring how membership to a protected group conditions the material outcomes of individuals and, thus, limits any meaningful counterfactual analysis between protected and non-protected individuals. Kohler-Hausmann [176] argues that protected attributes, like race, are social constructs, which is a view widely held by social scientists [42, 128, 261]. What it means is that these social categories were once (or still are) used not only to classify individuals but to also delimit the opportunities available to them [192].

According to Kohler-Hausmann [176], at least within the US context, it is inconceivable to picture a counterfactual white version of a black individual as defined in Definition 3.2.1. Under such definition, the ideal comparator $i'$ will have a set of non-protected

attributes identical to those of $i$, or $d(\mathbf{x}_i, \mathbf{x}_{i'}) = 0$. By envisioning a counterfactual world where $i$ can be exactly as he or she is today, in terms of non-protected attributes, while belonging to the non-protect group, this mental exercise requires from us to accept that the protected individual $i$ would have arrived at the same point had he or she been a non-protected individual $i'$, which reduces the pervasive influence of $A$.

Formally, we argue, under Definition 3.2.1 envisions a scenario where we can manipulate $A$ and expect $\mathbf{X}$ to remain unchanged. As Kohler-Hausmann [176] writes in the case of the protected attribute race:

> The problem with identifying discrimination with the treatment effect of race is that it misrepresents what race is and how it produces effects in the world, and concomitantly, what makes discrimination of race a moral wrong. [...] [I]f the signifiers of racial categories fundamentally structure the interpretation and relevance of other characteristics or traits of the unit, then it is a mistake to talk about identical units that differ only by raced statues.

Kohler-Hausmann [176] later extends the above argument in Hu and Kohler-Hausmann [149] for the protected attribute gender. In short, the Kohler-Hausmann Critique implies that the comparator $i'$ as conceived in Definition 3.2.1 is wrong and, thus, tools for testing discrimination built using such comparator are also wrong.

In their article, *The Trouble with Disparity*,[6] Walter Benn Michaels and Adolph Reed Jr. make a similar critique to Kohler-Hausmann [176] regarding non-discrimination law. Michaels and Reed, however, focus on the tension between race and class and use the coverage given to black communities during the COVID-19 pandemic to illustrate their points.[7] During the pandemic in the USA, blacks were dying from the virus at a higher rate than any other social group, which prompted "scientific" questions on whether there was a genetic pre-disposition that made this group more vulnerable to the virus relative to other groups. No such evidence was found. Once we controlled for other factors to isolate the so-called treatment effect of race, it became clear that "race" had nothing to with the likelihood of dying from the virus.

Drawing parallels to non-discrimination law and its focus on parity-based representation, Michaels and Reed point at the obvious or, in their words, *deeper cause* that nobody wanted to address regarding the disparate death rates among social groups from COVID-19: inequality. Blacks were not more likely to die from COVID-19 because of the color of their skin but because of their socioeconomic background. Being poor is what made an individual in the USA more vulnerable to COVID-19, and the majority of poor individuals in the USA are black.

We mention this article because it highlights the complexity of testing discrimination. The answer to "why are most low-income individuals in the USA today black?" is the same answer to "why is race considered a protected attribute?": because there is a history

[6] `https://nonsite.org/the-trouble-with-disparity/`

[7] Keep in mind that this article takes a Marxist-like view on discrimination, which is, by definition, based on the constant class struggle that drives human history. I think their remarks on race in the USA and non-discrimination law are valid and strong enough to not rely on Marxist theory. Overall, I appreciate Marxist thinkers, like Silvia Federici [94, 95], as they tend to be highly critical of the status quo and provide interesting points worth reconsidering under other frameworks. I am also aware, though, that most of these thinkers are at the margins of their own fields and that, overall, Marxist theory has yet to show that class formation and class identity explains historical patterns.

of systematic policies against individuals perceived as black (or, in general, non-white). Now, because of that same answer, the question we aim to answer when testing for the discriminatory effects of race should not be limited to race alone. The same way we have established that race and other protected attributes exist and have acknowledge their effects in our society, we must also recognize how their effects have materialized in limiting the opportunities of multiple generations.

Back to the COVID-19 example, it might have been race what initially caused the exclusionary policies toward non-whites that, in turn, created the current social context, but, precisely because of this historical process, race alone cannot claim full responsibility of the present disparities. Larger forces, though, still associated to race, such as inequality and social mobility, are stronger causes to the present disparities. See, e.g., Chetty et al. [62]. The problem is that, ironically, as Michaels and Reed argue, it has become easier to speak of racial discrimination than income discrimination.

We emphasize that such arguments are, in principle, not new. Wachter et al. [288], in particular, argue that European non-discrimination law focuses on substantive rather than formal equality. Our understanding of this distinction on equality is that formal equality requires, e.g., that certain representational quotas are met for a given decision process. Instead, substantive equality requires that the process itself systematically reaches an envisioned level of equality. The former assumes a neutral status quo, while the latter sees the status quo as biased and it is non-discrimination law's goal to change it for the better. However, it is unclear how non-discrimination law should be enforced to reach substantive equality [288]. Arguments like those by Wachter et al. [288], in our view, align with the criticisms previously mentioned on non-discrimination law.

Given this complex link between a phenotype like race and other, in principle, neutral attributes like income and education, can we (or, should we) conceive a comparator as defined in Definition 3.2.1? This discussion comes back to clarifying the link between $A$ and $\mathbf{X}$. If we consider the legal European context, as discussed in Section 2.2.2 from Chapter 2, we have two discrimination scenarios: $A \rightarrow Y$, or *direct discrimination*; and $A \rightarrow \mathbf{X} \rightarrow Y$, or *indirect discrimination.* Let us focus on the latter, which is the common case within automated decision-making [174]. At a conceptual level, given the causal structure $A \rightarrow \mathbf{X} \rightarrow Y$, we argue that *the key question addressed by the Kohler-Hausmann Critique is: what happens to* $\mathbf{X}$ *once we manipulate* $A$ *to derive the comparator?* The standard approach, which drives the counterfactual model of discrimination, is captured by Definition 3.2.1, while Kohler-Hausmann [176] among other fewer works like [8, 149, 297] call for a departure from such approach by recognizing and accounting for the causal link between $A$ and $\mathbf{X}$ (in the form of modeling the downstream effects) when testing for discrimination. Therefore, we propose revisiting the comparator under *two types of causal interventions* for counterfactual reasoning: *ceteris paribus* (CP), or "all else equal;" and *mutatis mutandis* (MM), or "adjusting for what needs to be adjusted." These causal interventions, in turn, define two kinds of comparators $i'$ for the complainant $i$ that, respectively, capture the *idealized comparison* and the *fairness given the difference comparison.* We define them below.

**Definition 3.2.2.** (The CP-Comparator) Consider the individual profile $\langle \mathbf{x}_i, a_i, y_i \rangle$ for the complainant $i$. Given the distance function $d$, with similarity threshold $\epsilon$, the *ceteris paribus* (i.e., all else equal) comparator $i'$ with profile $\langle \mathbf{x}_{i'}, a_{i'}, y_{i'} \rangle$ satisfies:

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) \leq \epsilon$$

where $a_i \neq a_{i'}$. We define it as the cp-comparator as we ideally expect for the protected attribute $A$ to change but all other attributes $\mathbf{X}$ to remain the same when considering the counterfactual representation $i'$ of the factual $i$.

Notice that the distance function $d$ for the cp-comparator is oblivious to (or unaware of) the protected attribute $A$. In other words, given the complainant profile $i$, the cp-comparator profile $i'$ that is sufficiently close to $i$ under $d$ and $\epsilon$ is based only on the set of non-protected attributes $\mathbf{X}$. Let us define $\tilde{d}$ as a protected-aware distance function that, unlike $d$, considers both $A$ and $\mathbf{X}$ as inputs for defining the comparator.

**Definition 3.2.3.** (The MM-Comparator) Consider the individual profile $\langle \mathbf{x}_i, a_i, y_i \rangle$ for the complainant $i$. Given the protected-aware distance function $\tilde{d}$, with similarity threshold $\epsilon$, the *mutatis mutandis* (i.e., adjusting for what needs to be adjusted) comparator $i'$ with profile $\langle \mathbf{x}_{i'}, a_{i'}, y_{i'} \rangle$ satisfies:

$$\tilde{d}((\mathbf{x}_i, a_i), (\mathbf{x}_{i'}, a_{i'})) \leq \epsilon$$

where $a_i \neq a_{i'}$. We define it as the mm-comparator as we allow for the comparator to be dissimilar from the complainant in terms of $\mathbf{X}$.

Under the distance function $\tilde{d}$, two instances with different non-protected attributes, $\mathbf{x}_i \neq \mathbf{x}_{i'}$, can be considered similar given information, respectively, about $a_i$ and $a_{i'}$. In other words, given the complainant profile $i$, the mm-comparator profile $i'$ that is sufficiently close to $i$ under $d$ and $\epsilon$ is based on the set of non-protected attributes $\mathbf{X}$ and the link between $\mathbf{X}$ and protected attribute $A$.

**Remark 3.2.1.** When $a_i = a_{i'}$, the mm-comparator is the same as the cp-comparator. Intuitively, if we are interested in comparing the complainant $i$ to other profiles within the same membership under the protected attribute $A$, then we can use either $d$ or $\tilde{d}$. This is because, as we are moving within the same group, we assume that the influence of $A = a$ on $\mathbf{X}$ is shared among all $i$ individuals $a_i$. Implicitly, we assume that defining similarity between individuals of the same group is straightforward, requiring no additional considerations due to a shared experience under $A = a$.

**Remark 3.2.2.** If $\mathbf{X}$ is independent of $A$ (or, vice versa) then the mm-comparator is the same as the cp-comparator when $a_i \neq a_{a'}$.

Given the previous remarks, it is clear that, under the premise that we have a case of indirect discrimination, meaning $A \rightarrow \mathbf{X} \rightarrow Y$, the cp-comparator ignores the downstream effect of $A$ on $\mathbf{X}$ while the mm-comparator aims to account for it. To see this last point, we must ask how can we find a comparator under Definition 3.2.3 when, by definition, any $i'$ dissimilar to $i$ according to $\mathbf{X}$ will do? In other words, how can we account for the influence of $A$ on $\mathbf{X}$ (or the overall relationship between this two) to reach a notion of similarity that satisfies *fairness given the difference*?

A way to implement this, in particular, a causal way based on counterfactual generation (Section 2.1.3), is by learning $\tilde{\mathbf{X}}$, an adjusted representation of $\mathbf{X}$. The key idea here is that the mm-comparator is a statement on the observable profiles of the complaint $i$ and comparator $i'$. If we want to account for the downstream effects of $A$, then we need to update $\mathbf{X}$ as a function of $A$: i.e., there exist a mapping $g$ such that $\tilde{\mathbf{X}} = g(\mathbf{X}, A)$.

This "updating" is meant to account for the change experienced in $\mathbf{X}$ when manipulating $A$. As discussed in Section 2.1.3, $\tilde{\mathbf{X}}$ comes from the counterfactual distribution of intervening on $A$.

Formally, let $\tilde{\mathbf{x}}_i$ represent the counterfactual set of non-protected attributes of the complainant $i$ due to manipulating (i.e., intervening via the *do-operator*) $a_i$ to $a_{i'}$, then we derive the mm-comparator $i'$ such that:

$$d(\tilde{\mathbf{x}}_i, \mathbf{x}_{i'}) \leq \epsilon \tag{3.1}$$

where possibly $\tilde{\mathbf{x}}_i \neq \mathbf{x}_i$. Given the causal relation $A \rightarrow \mathbf{X}$, the comparator $i'$ is similar to the counterfactual profile of $i$ but dissimilar to its factual profile. Here, $d$ represents once again a distance function unaware of the protected attribute $A$.

We summarize the link between the Definition 3.2.3 and equation (3.1) for the complainant $i$ and its comparator $i'$ as:

$$\tilde{d}\big((\mathbf{x}_i, a_i), (\mathbf{x}_{i'}, a_{i'})\big) = d\big(g(\mathbf{x}_i, a_i), \mathbf{x}_{i'}\big) \tag{3.2}$$

where the left-hand-side of the equation represents the general definition of the mm-comparator while the right-hand-side represents an implementation under the function $g$. Intuitively, under $g$ we are able to position $i$, through its generated counterfactual representation, within the same "world" as $i'$.

**Remark 3.2.3.** Keep in mind that the purpose of both Definitions 3.2.2 and 3.2.3 is finding the comparator $i'$. Hence, regarding $\tilde{\mathbf{x}}_i$, the goal is to generate it to be able find $i'$ under $d$ and $\epsilon$ based $\mathbf{x}_{i'}$. We are not, in principle, interested in using the generated counterfactual of the complainant beyond a reference for the mm-comparator.

To illustrate both cp and mm comparators, let us consider the Example 3.1.1. Under the cp-comparator, the profile we would choose as comparator to Clara would be that of Mike who has the same number of publications. We would do so whether Clara or Mike are themselves, respectively, parents. What matters under the cp-comparator is to reach an idealized comparison to test the discrimination claim. Since the number of publications is what matters for the tenure decision, then we should focus on what is observed on $X$. This is the most conservative approach to similarity between the complainant and its comparator.

Under the mm-comparator, instead, the profile we would choose would be that of Vincent, who has 18 publications or, i.e., 6 more than Clara. Here, we would, e.g., base our comparator choice on Clara being a mother and the known unequal impact of parenthood on academia [202]. The idea behind comparing these two different profiles is based on Clara's counterfactual having a similar output to Vincent (3.1). In other words, assuming a penalty of $A$ on $X$, the world in which Clara is a male academic and a father is also the world in which she no longer has the same social expectations, e.g., to be the main caregiver of her house. It is also, in turn, the world where she can benefit from this additional time to focus more on her research. This is a more flexible approach to similarity between the complainant and its comparator, but one that aims to account for the effects of $A$ on $X$.

These comparators offer two views on similarity and, thus, two different views on what it means to test for discrimination. We do not, however, claim that one is preferred

over the other. Conceptually, it depends on what role the comparator is expected to play. For instance, the cp-comparator is intuitive but conservative, ignoring completely a constructivist view of $A$; the mm-comparator, instead, is more flexible but requires an additional step to envision a $\mathbf{X}$ that accounts for the effects of $A$. Obviously, the main challenge behind Definition 3.2.3 is modelling the downstream effects of $A$ on $X$. The mm-comparator further requires a view of non-discrimination law as intended for substantive equality, and an overall believe that the tools for testing discrimination should act as transformative societal interventions that aim for a better status quo [288]. The cp-comparator avoids such discussion, which could make it a more agreeable (or less controversial) option among a group of stakeholders.

In Chapter 4 we show how to implement and test for discrimination under each comparator. In particular, we compare the k-NN situation testing by Thanh et al. [277] to its counterfactual version [8]. Unsurprisingly, as the results of these experiments show, the number of discrimination cases increases under a mm-comparator (the more flexible choice) relative to a cp-comparator (the more conservative choice). In fact, as we later discuss in our analysis of these experiments, all cases detected under the cp-comparator are also detected under the mm-comparator. Intuitively, it can be that for some individuals going from the factual to the counterfactual world (as prescribed by our causal model) leaves them exactly where they are. After all, the idealized comparison still denotes a form of counterfactual reasoning.

## 3.3 Causal Desiderata

Under the premise that testing for discrimination is a modeling problem that rests on the counterfactual model for discrimination [176], as argued in the last section, we now present a causal desiderata for testing discrimination.

This desiderata is aimed at the more technical crowd working on algorithmic fairness and the overall development of auditing tools for ADM systems. As we will discuss in the next section when we cover some of the most representative tools, the properties we present here are rarely accounted for together by these auditing tools. This causal desiderata rests on top of the structural causal models (SCM) discussed in Section 2.1. We discuss three properties that discrimination testing tools should strive for.

**Participatory.** Based on the four elements required for establishing discrimination in the EU (Section 3.1), testing for discrimination is a participatory process. We argue that the tools for testing discrimination should also be participatory. For this, in particular, we view the role of the causal graph $\mathcal{G}$ as central. *We say that a tool for testing discrimination is participatory if it allows for different stakeholders (beyond the modeler) to engage in key aspects of the testing pipeline.*

Under SCM, how we choose to model the discrimination problem in terms of $\mathcal{G}$ will influence what we eventually test for as discriminatory. For instance, discussing whether the protected attribute is a cause or not of the relevant attribute(s) used for the decision process can be expressed via $A \rightarrow X$. It is a simple and intuitive representation that could help different stakeholders engage over a contested issue like discrimination.

In particular, a participatory tool should consider how the drawing of the causal graph affects the rest of its pipeline. If it cannot account for it, then it should recognize that

there is a possibility for multiple graphs to compete in formalizing the same problem.

**Reliable.** Regarding the third element for establishing discrimination in the EU, i.e., evidence, it is clear that tools for testing discrimination should be equipped with some measure of certainty around the results provided. *We say that a tool for testing discrimination is reliable if it provides a level of confidence for its results.*

This property goes back to the two modeling cultures [49] briefly discussed in Section 3.1. Indeed, standard methods that come from a statistical inference tradition, such as correspondence studies [35] and natural experiments [109], are reliable methods given their focus on reporting the statistical significance of their results. The earlier algorithmic works on discrimination, also referred to as discrimination discovery [222, 244, 277], which translated the standard methods under data mining techniques, also kept these practices. More recent algorithmic methods based on predictive modeling tend no to report the certainty of their results [41, 163, 177]. This trend is interesting as, given the role of the individual complainant and its discrimination claim, the algorithmic fairness literature likes to position discrimination under individual-level fairness [88]: the ADM system discriminates when similar individuals are not treated similarly. It is common for fairness papers to equate the unfairness of a model to discrimination, despite the fact that such results, e.g., may only amount to *prima facie* discrimination evidence.

Now, there is an interesting tension regarding the reliability of recent algorithmic tools for testing discrimination. We highlight two aspects. First, if we have access to the model then, in principle, why would we need to measure the certainty of our results? The model should be deterministic: if it rejects Clara but accepts Mike despite having the same profile, then clearly there is something not right with the model. This is true even with more complex cases like proxy discrimination, again, because what we are using for testing discrimination are the inputs to the model. This scenario changes when we consider discovering discrimination, as in we only have access to the data but no to the model. In other words, we cannot evaluate the model's decision-making through its inputs. In that case, reliability is clearly required.

A second aspect to consider for reliability, though, is how algorithmic tools need to account for the way evidence for *prima facie* discrimination is presented. To the best of our knowledge, this is unclear as we have yet to take an ADM system to court for discrimination (see Weerts et al. [296, Section 3]). The question here is whether the literal comparison, at the individual level, is enough? Further, the broader question is whether evidence for algorithmic discrimination will in the long run require some sort of statistical significance? For instance, is the ADM system considered discriminatory if it is (in a statistical sense) significantly better than its human equivalent? For us, this is an open question.

**Meaningful.** Similarly, regarding the third element for establishing discrimination in the EU, i.e., evidence, the choice of comparator is also important. Thus, the meaningfulness of a tool is a function of how it defines the comparator for any complainant. Defining the comparator is a normative statement [296] and, thus, such procedure should reflect a societal goal. Equivalently, meaningfulness refers to discrimination tools that aim for substantive equality, and, overall, bias transforming (over bias preserving) machine learning as defined by Wachter et al. [288]. *We say that a tool for testing discrimination*

*is meaningful if it aims for establishing substantive equality.*

Based on works like Weerts et al. [296], Wachter et al. [288], and Kohler-Hausmann [176], we view meaningfulness in terms of implementing a mm-comparator over a cp-comparator, meaning testing discrimination under "fairness given the difference". Meaningfulness is at the core of the discussion in Section 3.2.3.

In practice, it means using a method that meaningfully represents the hypothetical scenario of the complainant not being from the protected group. Having access to a causal graph $\mathcal{G}$ helps to achieve meaningfulness, via the generation of counterfactual distributions (abduction, action, and prediction; recall Section 2.1.3), but we do not view meaningfulness specific to causality. Overall, this property requires the implementation of $\tilde{X} = g(X, A)$.

## 3.4 On Discrimination Testing Tools

In this section, we survey a representative set of the methods (or tools) for testing discrimination with a focus on the causal desiderata discussed previously. We recommend Romei and Ruggieri [234] for an extensive and multidisciplinary survey.

Broadly, all these tools amount to gathering (*prima facie* discrimination) evidence against the a decision-maker, be it algorithmic or non-algorithmic. In practice, this means finding (or generating) a *comparator individual profile* $i'$ for a given *complainant individual profile* $i$ and comparing the decisions under the same decision-maker. We view all these tools as instances of the counterfactual model of discrimination [176].

### 3.4.1 Standard Methods

Here, by standard, we mean essentially non-algorithmic methods. These include natural experiments (e.g., [109]), field experiments (e.g., [34]), and audits (e.g., [99]), among others. These methods share a focus on either gathering or evaluating data that is, by design, able to capture the effect of the protected attribute on the decision outcome. In other words, data that is characterized by a before-and-after or a with-and-without mechanism around the protected attribute that ensures its identifiability. It will become clear to the reader that most of these methods, at least in their original form, are somewhat outdated for how decisions are taken today given our digital technologies.

Let us start with *audits*, which is a more holistic and human-dependent method, less focused on modeling. Audits, as the name suggests, consist in examining the decision process in question through human auditors. In practice, it means sending a group of experts to observe during a period of time and study the practices of the decision maker to corroborate the claims of the complainant [99]. Auditors, e.g., tend to interview key stakeholders. In the case of an ADM system, such a practice (in its intended form) only makes sense if the model is used by a human decision-maker or if the focus is to audit the procedure by which the model was trained, which is increasingly being viewed as a priority over, e.g., interpretable models, by the EU. For this discussion on transparency as a procedural rather interpretable process, see Panigutti et al. [217].

In any case, audits are time consuming, expensive, and human-driven that are not scalable nor accessible to all discrimination cases. Here, the role of the comparator is not necessarily central. Indeed, the auditors may focus on identifying cases comparable

to those of the complainant but most of the focus is on understanding the overall process and identifying procedural concerns.

Natural and field experiments, instead, tend to have more of a focus on discrimination as a modeling problem (i.e., identifying the causal effect of the protected attribute) and, thus, prioritize designing/gathering data suitable for this purpose. These methods are a product of their time. In particular, these methods are a product of the so-called "empirical revolution" experienced by the social sciences (mainly Economics) in which a mixture of data availability, better computers, and an intuitive causal framework (the Rubin-causal model or the potential outcomes framework) made it possible to test causal claims via experiments [16].

Let us first consider *natural experiments*. The overall idea is simple. A policy or shock occurs creating a before-and-after (or with-without) scenario that, in principle, allows us to test for the effect of the protected attribute. Hence, we find ourselves naturally in an experimental setting. Goldin and Rouse [109]'s *Orchestrating Impartiality: The Impact of"Blind" Auditions on Female Musicians* is an example of how a natural experiment was used to test for discrimination. US orchestras used to audition musicians without the use of screens, meaning the musician would perform in full sight to the evaluating committee. US orchestras back then were also male dominated and, some, through their conductors, openly sexist toward female musician, which motivated the suspicion that there was discrimination against women in the US orchestra system. Then, in the 1970s, orchestras started to introduce screens to carry out blind auditions to ensure an impartial process. Some orchestras even introduced carpets so that female musicians would not be identified by the sound of their hills when walking up the stage.

Female representation in orchestras increased after the introduction of the screen, though Goldin and Rouse [109], using data from the auditions, set out to study whether this amounted to evidence for discrimination within the US orchestra system. There was indeed a before-and-after mechanism with the introduction of the screen that coincided with an increase in female musicians playing in orchestras, but to claim a causal effect other factors needed to be controlled for. In practice, this meant controlling for such factors in the regression model: see Goldin and Rouse [109] for details. For instance, the introduction of the screen also coincided with an increase of females in the US work force, meaning that more women were also attending and graduating from music conservatories than ever before. Similarly, auditions were specific to an instrument and not all instruments within an orchestra had the same turnover. Also, not all instruments had the same split between male and female musicians. All of these factors needed to be accounted for by the regression model. The paper finds inconclusive evidence for discrimination against female musicians. It is a seminal paper[8] because it shows a meticulous discussion on the importance of controlling for all other factors that could explain a change in the decision: i.e., the key role of deriving a comparator good-enough to isolate the effect of the protected attribute.

Now *field experiments*, instead, focus on designing an experiment to create the proper data for testing the discrimination claim. Probably the most famous example is Bertrand and Mullainathan [35]'s *Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.* Also see Bertrand and Duflo [34]

[8]Claudia Goldin won the 2023 Nobel Prize in Economics for her research on gender and its influence on female participation in the labor market.

for a survey on field experiments for discrimination. Bertrand and Mullainathan [35] proposed the following experiment: let us create pairs of fictitious CVs tailored to job positions that are same except for the candidate's name, where the name is intended to reflect both gender and race within the US context. For instance, they would create the same CV for Mike and Jamal (respectively, a common white and black name according to census data at the time); send both CVs to the same job openings via mail (this was done in the early 2000s); and compare which CV received a callback. Here, intuitively, the CV with the black-sounding name was the (potential) complainant and the CV with the white-sounding name its comparator. All the CVs were drafted by the researchers. In the end, they found evidence for discrimination. This sort of experiment is known as *correspondence studies*. See Rooth [235] for a survey on correspondence studies.

Field experiments are a product of the empirical revolution. Consider, for instance, the method of *situation testing*, which is a form of field experiment that existed before the wide adoption of the more advanced modeling approaches we just discussed. Situation testing consists in recreating the situation of interest multiple times in order to identify the discriminatory pattern. In practice, situation testing is implemented by hiring group of actors that share key physical traits to the complainant and are equipped with similar profiles (or backstories) as well [236]. These actors are then sent through the same decision process, which enables us to compare the different outcomes. Similar to audits, it is a time consuming and human intensive task that is not scalable.

Standard methods have been at the center of most of the criticism against the counterfactual model of discrimination. This fact is not surprising, though. These methods have been present and used much longer than any other kinds of methods; they are, after all, the basis for the algorithmic methods that have followed. Heckman [136], e.g., is critical of the effectiveness of these methods in controlling for similarity and other factors and claims that we have been too confident in claiming to detect discrimination. More recently, Kohler-Hausmann [176] goes after these methods due to their experimental designs; she is particularly critical of Bertrand and Mullainathan [35]. For instance, if it is uncommon for a black male to, on average, attend Harvard and play Lacrosse, then what exactly are we testing for when we create two identical CVs with those characteristics that only differ in race?

Most of these standard methods rely on the cp-comparator. In fact, *ceteris paribus*, popularized by economists, is a key concept in most analysis involving inferential statistics. Our view on this debate is conditional on a case by case basis, meaning it depends on what type of discrimination we want to test as well as the overall setting. If direct discrimination is a concern, then the cp-comparator embodying an idealized comparison should be able to test for it. If, however, we are interested in indirect discrimination, then these experiments have to make a strong argument for the type of comparator used. In particular, this argument needs to be centered on the kind of signal (as in the relevant information) being used by the decision-maker and whether that signal has any link to the protected attribute.

Consider the case of the orchestra auditions. There, the decision-maker cares only for the playing quality of the musician, which should not be related to gender in any way.[9] Hence, the signal can be recovered once the screen is imposed as the sound from the in-

[9]We note that there is a reasonable case to be made here that instruments, by biased design, tend to favor male musicians [75].

strument can go through the screen but all other factors, like the musician's appearance and perceived gender, are blocked. Consider now the case of the CVs. There, even if we impose a "screen" by removing the names, there is still the possibility that the decision maker uses signals for candidate potential that are also linked to race or gender, such as extracurricular activities or the name of the high-school. These are valid signals, or at least signals that could be well argued by the decision-maker as necessary for measuring candidate potential. Given these signals and their link to race or gender, though, it also makes the use of the cp-comparator over the mm-comparator questionable. This is why Kohler-Hausmann [176] is critical of Bertrand and Mullainathan [35]: what exactly does it mean for Greg and Jamal to have identical CVs?

This concern on the usage of the cp-comparator for testing indirect discrimination further intensifies under ADM. As models are required not to use protected attributes as inputs, which, for now [1], places algorithmic decision-making within indirect discrimination law [122], it means that how we test for discrimination should account for this tension between *ceteris paribus* and *mutatis mutandis* comparators. If the focus were on direct discrimination, then using the cp-comparator for testing algorithmic discrimination would be fine; the issue, however, is that most ADM systems would be tested for indirect discrimination.

### 3.4.2 Discrimination Discovery

Discrimination discovery methods [162, 222, 244] represent the first attempt at what we now call algorithmic fairness. Mostly based in Europe, these works took the standard methods and combined them with data mining tools, shifting the focus from generating or finding the (experimental) data to exploring a data produced by a decision process. In part, this shift in focus was field-driven: within data mining and overall knowledge discovery, the focus of the problem formulation is on developing new and better algorithms for extracting information in a data-driven way. This is a different focus from the prevalent inferential statics approach in which the creation or gathering of the data is as important as the modeling approach. But this shift also reflected a change in testing for discrimination. With ADM systems becoming more prevalent (also fueled by the rise of big data, and faster and cheaper computing), the attention was now on a consistent, algorithmic decision-maker. To test whether the decision-maker discriminated, we could now look at the data it produced. It was a natural transition. Both the decision-maker in question and the tools for evaluating it were scalable. These works not only combined algorithms with discrimination, but they did so by borrowing directly from non-discrimination law. For instance, works by Pedreschi et al. [222] and Ruggieri et al. [244] provided the first formalization to equal opportunity (and, implicitly, to equalized odds) considerably before Hardt et al. [131]. Kamiran and Calders [161] did similarly for demographic parity.

Thanh et al. [277] is an illustrative example of the discrimination discovery methods. They propose a k-NN implementation of situation testing. Instead of hiring actors, the idea was to mine the dataset to find a suitable comparator for a complainant. Using these two instances as search centers, they build neighborhoods of similar protected and non-protected individual to compare and, thus, derive results with statistical significance. Overall, these tools, truthful to their standard methods roots, kept the many-to-many

comparisons to estimate the certainty around individual discrimination claims. These methods tend to be reliable as defined in our causal desiderata.

In terms of meaningfulness, overall, these methods focus on defining the best distance to reach the comparator for a given complainant. Some implementation are closer to the mm-comparator than to the cp-comparator. Thanh et al. [277], e.g., clearly use a cp-comparator as they minimize the distance between the complainant and potential comparators based only on the set of non-protected attributes. Zhang et al. [319], e.g., offer an extension to the k-NN situation testing [277] via causal weights, but their focus is on defining a distance function that prioritizes the important non-protected attributes when finding the comparator for a complainant. Qureshi et al. [229], .e.g., use a propensity score weight for the distance function with the protected attribute as the treatment. This last work approaches how we envision the mm-comparator.

### 3.4.3 Algorithmic Fairness for Discrimination

Eventually, discrimination discovery took a broader focus and became algorithmic fairness; recall the discussion in Section 2.3.1 and Section 2.3.2 in the previous chapter. Given the counterfactual model of discrimination and how we conceive discrimination, the main fairness works useful for discrimination are *counterfactual fairness* (2.15) and *individual fairness* (2.13). The former establishes the same treatment of the factual and its counterfactual instance under the same decision; the latter establishes the same treatment of similar instances under the same decision. Under the counterfactual model of discrimination, these two concepts intersect as we use the similarity between individuals to somehow approximate the comparison between an individual and its counterfactual-self. Other works like the *FlipTest* by Black et al. [41] or our own *Counterfactual Situation Testing* [8] (next chapter) build upon these two definitions.

Both definitions, recall the two modeling cultures [49], focus on a predictive modeling scenario. If the model is deterministic, then all the evidence we need is based on considering an input to the model and its output. Hence, the discussion on reliability requires further thought as well as a holistic approach based on how we use and plan to take to court ADM systems. It is important to define the source of uncertainty. With a human decision-maker, this point was clearer due to the inconsistent decision-making by the same decision-marker. But under ADM, in principle, this source of uncertainty is no longer a concern as the model may or may not be biased, but it should behave consistently. A valid question then is whether the literal comparison (between the complainant and its comparator) is enough evidence or if we need to consider the many-to-many comparison (between like-wise instances, respectively, to the complainant and its comparator). The answer to this question is unclear; it requires future and joint work with legal scholars and regulators.

In our view, the main concern regarding algorithmic fairness for discrimination is that of meaningfulness. At the moment, most ADM systems work with non-protected attributes that may or may not be (causally) linked to the protected attribute. Under this indirect discrimination setting, how we choose to define the comparator will affect the number of discrimination cases identified. This is where we see SCM and counterfactuals as useful tools: to envision hypothetical worlds under clear assumptions.

We acknowledge that Hu and Kohler-Hausmann [149] precisely goes after SCM as

a modeling framework, which led to works like Kasirzadeh and Smart [164], e.g., in criticizing counterfactual fairness and essentially classifying it as another cp-comparator implementation. We do not agree with these critiques, and would argue that they do not properly represent how counterfactual fairness works. Depending on the specification of the SCM, when generating the counterfactual distribution, we should be able to update $\mathbf{X}$ under the downstream effects of $A$. Of course, this whole point rests on strong modeling assumptions but it is still a move in the right direction.

For instance, if we were to revisit Bertrand and Mullainathan [35]'s famous CV experiment under counterfactual generation as formulated in Kusner et al. [177], assuming that the protected attribute is a root node causing other attributes that appear on the CV, then the generated CVs between a Greg and a Jamal would be different not only in terms of the name but also on all other perceived attributes linked to race. Indeed, a considerable modeling step would lie in between the factual data and the generation of the counterfactual data, but it could help the experiment to be more concrete about the hypothetical world we wish to picture. After all, is substantive equality not an exercise of counterfactual reasoning? To be able to discuss an alternative status quo we need to materialize first into some sort of representation that can be discussed between stakeholders. We should be, of course, careful with the promises behind counterfactual reasoning but we should also not dismiss its potential for developing better tools for testing discrimination.

## 3.5 Conclusion

In this chapter, building over Kohler-Hausmann [176], we revisited the discrimination comparator and presented two kinds: the *ceteris paribus* comparator and the *mutatis mutandis* comparator. We also presented a causal desiderata and briefly surveyed representative testing frameworks. Further, this chapter sets the tone for how we view causality as a tool for fairness and, in turn, discrimination.

Future work should include a more systematic review of a growing, multidisciplinary literature on testing for discrimination. It would benefit Fair ML to provide a table classifying each method in terms of the comparator being used and adherence to the desiderata. Such enterprises should be joint between Fair ML and Law as well as in line with current ADM regulations (e.g., the AI Act [92]).

We plan to formalize further the mm-comparator under causal reasoning. We believe that counterfactual generation will play a central role in creating more meaningful methods for testing discrimination. In that sense, generative ML tools could be repurposed for testing discrimination under causal reasoning. For instance, Black et al. [41] is an example of a non-causal discrimination testing tool that still operationalizes the counterfactual model of discrimination. Again, under the premise that the comparator represents a counterfactual representation of the complainant, the question is whether such a tool (and other tools similarly formulated) is indeed exempt from causal reasoning when testing for discrimination? We leave this discussion to future work.

# Chapter 4

# Counterfactual Situation Testing

This chapter is based on the conference paper: J. M. Álvarez and S. Ruggieri. Counterfactual situation testing: Uncovering discrimination under fairness given the difference. In *EAAMO*, pages 2:1–2:11. ACM, 2023.

We present counterfactual situation testing (CST), a causal data mining framework for detecting individual discrimination in a dataset of classifier decisions. CST answers the question "what would have been the model outcome had the individual, or complainant, been of a different protected status?" in an actionable and meaningful way. It extends the legally-grounded situation testing of Thanh et al. [277] by operationalizing the notion of *fairness given the difference* of Kohler-Hausmann [176] using counterfactual reasoning. In standard situation testing we find for each complainant similar protected and non-protected instances in the dataset; construct respectively a control and test group; and compare the groups such that a difference in decision outcomes implies a case of potential individual discrimination. In CST we avoid this idealized comparison by establishing the test group on the complainant's counterfactual generated via the steps of abduction, action, and prediction. The counterfactual reflects how the protected attribute, when changed, affects the other seemingly neutral attributes of the complainant. Under CST we, thus, test for discrimination by comparing similar individuals within each group but dissimilar individuals across both groups for each complainant. Evaluating it on two classification scenarios, CST uncovers a greater number of cases than ST, even when the classifier is counterfactually fair.

## 4.1 Introduction

Automated decision making (ADM) is becoming ubiquitous and its societal discontents clearer [17, 77, 138]. There is a shared urgency by regulators [92, 298] and researchers [174, 245] to develop frameworks that can asses these classifiers for potential discrimination based on protected attributes such as gender, race, or religion. Discrimination is often conceived as a causal claim on the effect of the protected attribute over an individual decision outcome [101, 136]. It is, in particular, a conception based on counterfactual reasoning—what would have been the model outcome if the individual, or *complainant*, were of a different protected status?—where we "manipulate" the protected attribute of the individual. Kohler-Hausmann [176] calls such conceptualization the *counterfactual causal model of discrimination* (CMD).

Several frameworks for proving ADM discrimination are based on CMD [176]. Central to these frameworks is defining "similar" instances to the complainant; arranging them based on their protected status into control and test groups; and comparing the decision outcomes of these groups to detect the effect of the protected attribute. Among the available tools [57, 191, 234], however, there is a need for one that is both *actionable* and *meaningful*. We consider a framework to be actionable if it can rule out random circumstances for the individual discrimination claim as often required by courts (e.g., [100, 101, 207]), and meaningful if it can account for known links between the protected attribute and all other attributes when manipulating the former as often demanded by social scientists (e.g., [42, 164, 261]). We regard actionability as an inferential concern to be handled by comparing multiple control-test instances around a complainant, while meaningfulness as an ontological concern to be handled by requiring causal domain-knowledge on the protected group of the complainant.

We present *counterfactual situation testing* (CST), a causal data mining framework for detecting instances of individual discrimination in the dataset used by a classifier. It combines (structural) counterfactuals [218, 221] with situation testing [277, 319]. *Counterfactuals* answer to counterfactual reasoning and are generated via structural causal models. Under the right causal knowledge, counterfactuals reflect at the individual level how changing the protected attribute affects other seemingly neutral attributes of a complainant. *Situation testing* is a data mining method, based on the homonymous legal tool [30, 236]. For each complainant, under some search algorithm and distance function for measuring similarity, it finds and compares a control and test group of similar protected and non-protected instances in the dataset, where a difference between the decision outcomes of the groups implies potential discrimination. CST follows the situation testing pipeline with the important exception that it constructs the test group around the complainant's counterfactual instead of the complainant.

**Example 4.1.1.** (An illustrative example) Let us consider the scenario in Figure 4.1 (used later in Section 4.4.1) in which a bank uses a classifier to accept or reject ($\hat{Y}$) individual loan applications based on annual salary ($X_1$) and account balance ($X_2$). Suppose a female applicant ($A = 1$) with $x_1 = 35000$ and $x_2 = 7048$ gets rejected and files for discrimination. The bank is using non-sensitive information to calculate $\hat{Y}$, but according to Figure 4.1 there is also a known link between $A$ and $\{X_1, X_2\}$ that questions the neutrality of such information.

Under standard situation testing, we would find a number of female (protected) and male (non-protected) instances with similar characteristics to the complainant. The resulting control and test groups to be compared would both have similar $X_1$ and $X_2$ to the complainant. On one hand, comparing multiple instances allows to check whether the complainant's claim is an isolated event or representative of an unfavorable pattern toward female applicants by the model (i.e., actionability). On the other hand, knowing what we know about $A$ and its influence, would it be fair to compare the similar female and male instances? As argued by previous works [149, 176], the answer is no. This *idealized comparison* underpinning standard situation testing takes for granted the effect of gender on annual salary and account balance.

Under counterfactual situation testing, instead, we would generate the complainant's counterfactual under the causal knowledge provided, creating a "male" applicant with a higher $x_1 = 50796$ and $x_2 = 13852$, and use it rather than the complainant to find

similar male instances. The resulting control and test groups would have different $X_1$ and $X_2$ between them. This disparate comparison embodies *fairness given the difference*, explicitly acknowledging the lack of neutrality when looking at $X_1$ and $X_2$ based on $A$ (i.e., meaningfulness). Here, the control group represents the observed factual world while the test group the hypothetical counterfactual world of the complainant.

In addition, with counterfactual situation testing we propose an actionable extension to *counterfactual fairness* by Kusner et al. [177], which remains the leading causal fairness framework [191]. A classifier is counterfactually fair when the complainant's and its counterfactual's decision outcomes are the same. These are the same two instances used by CST to construct, respectively, the control and test groups, which allows to equip this fairness definition with measures for uncertainty. Hence, CST links counterfactual fairness claims with notions of statistical significance.

Further, by looking at the control and test groups rather than the literal comparison of the factual versus counterfactual instances, CST evaluates whether the counterfactual claim is representative of similar instances. Hence, CST detects cases of individual discrimination that are also counterfactually fair, capturing the scenario where a deployed model discriminates when asked to evaluate a borderline instance multiple times.

Based on two case studies using synthetic and real data, we evaluate the CST framework using a *k*-nearest neighbor implementation, k-NN CST, and compare it to its situation testing counterpart, k-NN ST [277], as well as to counterfactual fairness [177]. Here, $k$ denotes the number of instances we wish to find for each control and test groups. The experiments show that CST detects a higher number of individual cases of discrimination across the different $k$ sizes. Further, the results also show that individual discrimination can occur even when the model is counterfactually fair. The results hold when dealing with multiple protected attributes as well as different implementation parameters.

**Summary of our contributions.** With CST we provide a new framework for detecting discrimination based on causal reasoning and popular data mining tools. Our main contributions are:

- With CST we present a meaningful framework (as in, it accounts for the downstream effects of the protected attribute on the attributes used for the decision) and actionable framework (as in, it uses many-to-many comparisons to account for uncertainty in the decision) for detecting individual discrimination.
- With CST we offer the first operationalization of *fairness given the difference* for discrimination analysis. In doing so, it introduces a new view on similarity that is more flexible than the common idealized comparison. Unsurprisingly, CST detects a considerably higher number of individual discrimination cases than standard ST and counterfactual fairness.
- With CST we introduce an actionable extension of counterfactual fairness equipped with confidence intervals. Under the many-to-many comparisons centered on the factual versus counterfactual pair, we provide evidence that a counterfactually fair algorithm can still be considered discriminatory.

In the reminder of this section, we present the related work. Moving forward, Section 4.2 explores the role of causal knowledge in CST. Section 4.3 presents the CST

framework and its k-NN implementation, while Section 4.4 showcases CST via two classification scenarios. Section 4.5 concludes the chapter. Finally, Appendix A contains additional supporting material.

### 4.1.1 Related Work

We position CST with current works along the goals of actionability and meaningfulness. Regarding actionability, when proving discrimination, it is important to insure that the framework accounts for sources of randomness in the decision process. Popular non-algorithmic frameworks—such as natural [109] and field [34] experiments, audit [99] and correspondence [35, 235] studies—address this issue by using multiple observations to build inferential statistics. Similar statistics are sometimes asked in court for proving discrimination (e.g., [100, Section 6.3]). Few algorithmic frameworks, instead, address this issue due to model complexity preventing formal inference [21]. An exception are data mining frameworks for discrimination discovery [222, 244] that operationalize the non-algorithmic notions, including situation testing [277, 319]. These frameworks (e.g., [3, 106, 229]) keep the focus on comparing multiple control-test instances for making individual claims, providing evidence similar to that produced by the quantitative tools used in court [176]. To the best of our knowledge, it remains unclear if the same can be said about existing causal fair machine learning methods [191] as these have yet to be used beyond academic circles.

Regarding meaningfulness, situation testing and the other methods have been criticized for their handling of the counterfactual question behind the causal model of discrimination [149, 164, 176]. In particular, these actionable methods take for granted the influence of the protected attribute on all other attributes. This can be seen, e.g., in how situation testing constructs the test group, which is equivalent to changing the protected attribute while keeping everything else equal. Such approach goes against how most social scientists interpret the protected attribute and its role as a social construct when proving discrimination [42, 128, 237, 261]. It is in that regard where structural causal models [218] and their ability for conceiving counterfactuals (e.g., [63, 310]), including counterfactual fairness [177], have an advantage. What the criticisms on counterfactuals [149, 164] overlook here is that generating counterfactuals, as long as the causal knowledge is properly specified, accounts modeling-wise for the effects of changing the protected attribute on all other observed attributes. A framework like counterfactual fairness, relative to situation testing and these other methods, is more meaningful in its handling of protected attributes. The novelty in CST is bridging these two lines of work, borrowing the actionability aspects from situation testing and the meaningful aspects from counterfactual fairness.

We highlight one recent individual discrimination framework, the FlipTest [41], that uses optimal transport instead of causal knowledge to obtain the control-test instances. Like algorithmic recourse [163] and counterfactual explanations [287], however, FlipTest requires the outcome of the machine learning model to flip, or cross the decision boundary. In CST we are not restricted by constructing a test group made from only individuals with positive outcomes, as a test group centered around a counterfactual with a negative outcome is also of interest for proving discrimination.

## 4.2 Causal Knowledge for Discrimination

Counterfactual situation testing requires access to the dataset of decision records of interest, $\mathcal{D}$, and the algorithmic decision-maker that produced it, $b()$. Let $\mathcal{D}$ contain the set of relevant attributes $X$, the set of protected attributes $A$, and the decision outcome $\hat{Y} = b(X)$. We describe $\mathcal{D}$ as a collection of $n$ tuples, each $(x_i, a_i, \widehat{y}_i)$ representing the $i^{th}$ individual profile, with $i \in [1, n]$. $\hat{Y}$ is binary with $\hat{Y} = 1$ denoting the positive outcome (e.g., loan granted). For illustrative purposes, we assume a single binary $A$ with $A = 1$ denoting the protected status (e.g., female), though we relax this assumption in the experiments of Section 4.4.2.

We also require causal knowledge in the form of a structural causal model (SCM) that describes the data generating model behind $\mathcal{D}$. We view this requirement as an input space for experts as these models are a convenient way for organizing assumptions on the source of the discrimination, facilitating stakeholder participation and supporting collaborative reasoning about contested concepts [206].

### 4.2.1 Structural Causal Models and Counterfactuals

We will use the SCM $\mathcal{M}$ (2.2) as presented in Definition 2.1.1. We will use the SCM $\mathcal{M}$ to generate counterfactual distributions. Refer to Section 2.1 for details.

Here, we assume causal sufficiency and an acyclical causal graph. These assumptions are necessary for generating counterfactuals. The causal sufficiency assumption is particularly deceitful as it is difficult to both test and account for a hidden confounder [76, 185, 195]. The risk of a hidden confounder is a general problem to modeling fairness. Here, the dataset $\mathcal{D}$ delimits our context. We expect it to contain all relevant information used by the decision-maker $b()$.

Input from several stakeholders is needed to derive (2.2). We see it as a necessary collaborative effort: *before we implement CST, we first need to agree over a worldview for the discrimination context.* Based on $\mathcal{D}$, a domain-expert motivates a causal graph $\mathcal{G}$. A modelling-expert then translates this graphical information into a SCM $\mathcal{M}$ . We do not cover this process here, but this is how we envision the initial implementation stage of counterfactual situation testing.

For a given SCM $\mathcal{M}$ we want to run *counterfactual queries* to build the test group for a complainant. Counterfactual queries answer to *what would have been if* questions. In CST, we wish to ask such questions around the protected attribute $A$, by setting $A$ to the non-protected status $\alpha$ using the *do-operator* $do(A := \alpha)$ [218] to capture the individual-level effects $A$ has on $X$ according to the SCM $\mathcal{M}$. Let $X^{CF}$ denote *the set of counterfactual variables* obtained via the three steps: abduction, action, and prediction. Further, let $P(X^{CF}_{A \leftarrow \alpha}(U) \mid X, A)$ denote *counterfactual distribution.*

We stress once again that generating counterfactuals and, thus, CST, unlike, e.g., counterfactual explanations [287] and discrimination frameworks like the FlipTest [41], does not require a change in the individual decision outcome. Hence, it is possible for $\widehat{Y} = \widehat{Y}^{CF}$ after manipulating $A$.

### 4.2.2 Conceiving Discrimination

The legal setting of interest is indirect discrimination under EU non-discrimination law. It occurs when an apparently neutral practice disadvantages individuals that belong to a protected group. Following [122], we focus on indirect discrimination for three reasons. First, unlike disparate impact under US law [25], the decision-maker can still be liable for it despite lack of premeditation and, thus, all practices need to consider potential indirect discrimination implications. Second, many ADM models are not allowed to use the protected attribute as input, making it difficult for regulators to use the direct discrimination setting. Third, we conceive discrimination as a product of a biased society where $b()$ continues to perpetuate the bias reflected in $\mathcal{D}$ because it cannot escape making a decision based on $X$ to derive $\hat{Y}$.

We view the indirect setting as the one that best describes how biased information can still be an issue for an ADM that never uses the protected attribute. Previous causal works [63, 167, 226] have focused more on whether the paths between $A$ and $\hat{Y}$ are direct or indirect. Here, the causal setting is much simpler. We know that $b()$ only uses $X$, and are more interested in how information from $A$ is carried by $X$ and how can we account for these links using causal knowledge. That said, this does not mean that CST cannot be implemented in other discrimination settings. We simply acknowledge that it was developed with the EU legal framework in mind. Proxy discrimination [278], e.g., is one setting that overlaps with the one we have considered.

Finally, we note that an open legal concern for CST is detecting algorithmic discrimination for various protected attributes, or $|A| > 1$. Two kinds of discrimination, *multiple* and *intersectional*, can occur. Consider, e.g., a black female as the complainant. On what protected attribute is she being potentially discriminated on? In multiple discrimination, we would need to detect separately whether the complainant was discriminated as a black and female individual. In intersectional discrimination, we would instead need to detect simultaneously if the complainant was discriminated as a black-female individual. Only multiple discrimination is currently recognized by EU law, which is an issue as an individual can be free from multiple discrimination but fall victim of intersectional discrimination [308]. CST can account operationally for both scenarios.

### 4.2.3 The Kohler-Hausmann Critique

Here, we make the case—very briefly—that the causal knowledge required for CST makes it a meaningful framework with respect to situation testing [277, 319] and other tools [234] for detecting discrimination. The reference work is Kohler-Hausmann [176]. We refer to the phrase *fairness given the difference*,[1] which best captures her overall critique toward the causal model of discrimination, as the *Kohler-Hausmann Critique* (KHC). CST aims to be meaningful by operationalizing the KHC. It builds the test group on the complainant's counterfactual, letting $X^{CF}$ reflect the effects of changing $A$ instead of assuming $X = X^{CF}$. This is because we view the test group as a representation of the hypothetical counterfactual world of the complainant. As the reader might notice, this line of argument is at the core of Chapter 3.

[1] A phrase by Kohler-Hausmann during a panel discussion at NeurIPS 2021 workshop on 'Algorithmic Fairness through the Lens of Causal Reasoning.

As argued by [176] and others before [42, 261], it is difficult to deny that most protected attributes, if not all of them, are *social constructs*. That is, these attributes were used to classify *and* divide groups of people in a systematic way that conditioned the material opportunities of multiple generations [192, 237]. Thus, *recognizing $A$ as a social constructs means recognizing that its effects can be reflected in seemingly neutral variables in $X$*. It is recognizing that $A$, the attribute, cannot capture alone the meaning of belonging to $A$ and that we might, as a minimum, have to link it with other attributes to better capture this, such as $A \rightarrow X$ where $A$ and $X$ change in unison. These attributes *summarize the historical processes that fairness researchers are trying to address* today and should not be treated lightly.[2]

The notion of *fairness given the difference* centers on how $A$ is treated in the counterfactual causal model of discrimination (CM). The critique goes beyond the standard manipulation concern [16] in which $A$ is an inmutable attribute. Instead, granted that we *can* or, more precisely, *have to* manipulate $A$ for running a discrimination analysis, the critique goes against how most discrimination frameworks operationalize such manipulation. The KCH emphasizes that *when $A$ changes, $X$ should change as well.*

Based on KHC, we consider two types of manipulations that summarize existing frameworks. The *ceteris paribus* (CP), or all else equal, manipulation in which $A$ changes but $X$ remains the same. Examples of it include situation testing [277, 319] but also, e.g., the famous correspondence study by Bertrand and Mullainathan [35]. The *mutatis mutandis* (MM), or changing what needs to be changed, manipulation in which $X$ changes when we manipulate $A$ based on some additional knowledge, like a structural causal model, that explicitly links $A$ to $X$. Counterfactual fairness [177], e.g, uses this manipulation. The MM is clearly preferred over the CP manipulation when we view $A$ as a social construct. Refer to Chapter 3, in particular, Definitions 3.2.2 and 3.2.3 for a formal discussion of the CP and MM manipulations for testing discrimination through the discrimination comparator.

## 4.3 Counterfactual Situation Testing

The objective of CST is to *construct* and *compare* a control and test group for each $c$ protected individual, or *complainant*, in $\mathcal{D}$ in a meaningful and actionable way. Let $(x_c, a_c, \widehat{y}_c) \in \mathcal{D}$ denote the *tuple of interest* on which the individual discrimination claim focuses on, where $c \in [1, n]$. We assume access to the ADM $b()$, the dataset $\mathcal{D}$, and a structural causal model $\mathcal{M}$ describing the discrimination context.

There are three key inputs to consider: the *number of instances per group*, $k$; the *similarity distance function of choice*, $d$; and the *strength of the evidence for rejecting the discrimination claim*, $\alpha$. A fourth key input that we fix in this paper is the *search algorithm of choice*, $\phi$, which we set as the *k-nearest neighbors algorithm* (k-NN) [133]. We do so as the k-NN is intuitive, easy to implement, and commonly used by existing situation testing frameworks. We discuss the CST implementation used in Section 4.3.4, including the choice of $d$.

[2] A clear example of this would be the use of race by US policy makers during the early post-WWII era. See, e.g., the historical evidence provided by Rothstein [238] (for housing), Schneider [253] (for narcotics), and Adler [2] (for policing).

### 4.3.1 Building Control and Test Groups

For complainant $c$, the control and test groups are built on the *search spaces* and *search centers* for each group. The search spaces are derived and, thus, delimited by $\mathcal{D}$: we are looking for individuals that have gone through the same decision process as the complainant. The search centers, however, are derived separately: the one for the control group comes from $\mathcal{D}$, while the one for the test group comes from the corresponding *counterfactual dataset* $\mathcal{D}^{CF}$. The test search center represents the *what would have been if* of the complainant under a *mutatis mutandis* (MM) manipulation of the protected attribute $A$ that motivates the discrimination claim.

**Definition 4.3.1** (Search Spaces)**.** Under a binary $A$, where $A = 1$ denotes the protected status, we partition $\mathcal{D}$ into the *control search space* $\mathcal{D}_c = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D} : a_i = 1\}$ and the *test search space* $\mathcal{D}_t = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D} : a_i = 0\}$.

**Definition 4.3.2** (Counterfactual Dataset)**.** The counterfactual dataset $\mathcal{D}^{CF}$ represents the counterfactual mapping of each instance in the dataset $\mathcal{D}$, with known decision maker $b()$ and SCM $\mathcal{M}$, via the abduction, action, and prediction steps [221] when setting a binary $A$ to the non-protected value, or *do*$(A := 0)$.

To obtain $\mathcal{D}^{CF}$, we consider an SCM $\mathcal{M}$ where $A$ has no causal parents, or is a root node, $A$ affects only the elements of $X$ considered by the expert(s), and $\hat{Y} = b(X)$. Therefore, when generating the counterfactuals on $A$ (Section 4.2.1), under the indirect discrimination setting (Section 4.2.2), the resulting $X^{CF}$ in $\mathcal{D}^{CF}$ should reflect an MM manipulation (Section 4.2.3). Under this structural representation, if $A$ changes then $X$ changes too. See, e.g., Figure 4.1 and Figure 4.3. The counterfactual dataset represents the world that the complainants would have experienced under $A = 0$ *given our worldview*. All three definitions extend to $|A| > 1$.

**Definition 4.3.3** (Search Centers)**.** For a complainant $c$, we use $x_c$ from the tuple of interest $(x_c, a_c, \hat{y}_c) \in \mathcal{D}$ as the control search center for exploring $\mathcal{D}_c \subset \mathcal{D}$, and use $x_c^{CF}$ from the tuple of interest's generated counterfactual $(x_c^{CF}, a_c^{CF}, \hat{y}_c^{CF}) \in \mathcal{D}^{CF}$ as the test search center for exploring $\mathcal{D}_t \subset \mathcal{D}$.

Given the factual $\mathcal{D}$ and counterfactual $\mathcal{D}^{CF}$ datasets, we construct the control and test groups for $c$ using the k-NN algorithm under some distance function $d(x, x')$ to measure similarity between two tuples $x$ and $x'$. We want each group or neighborhood to have a size $k$. For the *control group* (*k-ctr*) we use the (factual) tuple of interest $(x_c, a_c, \hat{y}_c) \in \mathcal{D}$ as search center to explore the protected search space $\mathcal{D}_c$:

$$\textit{k-ctr} = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D}_c : rank_d(x_c, x_i) \leq k\} \tag{4.1}$$

where $rank_d(x_c, x_i)$ is the rank position of $x_i$ among tuples in $\mathcal{D}_c$ with respect to the ascending distance $d$ from $x_c^{CF}$. For the *test group* (*k-tst*) we use the counterfactual tuple of interest $(x_c^{CF}, a_c^{CF}, \widehat{y}_c^{CF}) \in \mathcal{D}^{CF}$ as search center to explore the non-protected search space $\mathcal{D}_t$:

$$\textit{k-tst} = \{(x_i, a_i, \widehat{y}_i) \in \mathcal{D}_t : rank_d(x_c^{CF}, x_i) \leq k\} \tag{4.2}$$

where $rank_d(x_c^{CF}, x_i)$ is the rank position of $x_i$ among tuples in $\mathcal{D}_t$ with respect to the ascending distance $d$ from $x_c^{CF}$.

We use the same distance function $d$ for each group. Neither $A$ nor $\hat{Y}$ are used for constructing the groups. Both (4.1) and (4.2) can be expanded by including additional constraints, such as a maximum allowed distance $\epsilon > 0$. Formally, for instance, *k-ctr* $= \{x_i \in \mathcal{D}_c : rank_d(\mathbf{x}_c, x_i) \leq k \wedge d(x_c, x_i) \leq \epsilon\}$ and *k-tst* $= \{x_i \in \mathcal{D}_t : rank_d(x_c^{CF}, x_i) \leq k \wedge d(x_c^{CF}, x_i) \leq \epsilon\}$.

The choice of search centers (Definition 4.3.3) is what operationalizes *fairness given the difference* for counterfactual situation testing, making it a *meaningful framework* for testing individual discrimination. To build *k-ctr* and *k-tst* using, respectively, $x_c$ and $x_c^{CF}$ is a statement on how we perceive *within group ordering* as imposed by the protected attribute $A$. This is because the search centers must reflect the $A$-specific ordering of the search spaces that each center targets.

Let us consider our illustrative example from Section 4.1. If being a female ($A = 1$) in this society imposes certain systematic limitations that hinder $x_c$, then comparing $c$ to other female instances in the protected search space preserves the group ordering prescribed by $X|A = 1$ as all instances involved experience $A$ in the same way. Therefore, given our worldview, the generated counterfactual male instance for $c$ should then reflect the group ordering prescribed by $X|A = 0$. We expect $x_c \neq x_c^{CF}$ given what we know about the effects of $A$ on $X$. Using $x_c^{CF}$ as the test search center would allow us to compare $c$ to other male tuples in the non-protected search space without having to reduce $A$ to a phenotype.

One way to look at the previous statement is by considering the notion of effort. If being female requires a higher individual effort to achieve the same $x_c$, then it is fair to compare $c$ to other female instances. However, it is unfair to compare $c$ to other male instances without adjusting for the extra effort not incurred by the male instances for being males. The counterfactual $x_c^{CF}$ should reflect said adjustment. See [65, 66] on a similar, more formal critique on individual fairness [88] notions.

### 4.3.2 Detecting Discrimination

For a complainant $c$, we compare the control and test groups by looking at the *difference in proportion of negative decision outcomes*:

$$\Delta p = p_c - p_t \tag{4.3}$$

such that:

$$\begin{aligned} p_c &= \frac{|\{(x_i, a_i, \widehat{y_i}) \in \textit{k-ctr} : \hat{y}_i = 0\}|}{k} \\ p_t &= \frac{|\{(x_i, a_i, \widehat{y_i}) \in \textit{k-tst} : \hat{y}_i = 0\}|}{k} \end{aligned} \tag{4.4}$$

where $p_c$ and $p_t$ represents the count of tuples with a negative decision outcome ($\hat{Y} = 0$) in the control and test group. Only $\hat{Y}$ is used for deriving the proportions.

We compute $\Delta p$ for all protected tuples in $\mathcal{D}$ regardless of the their decision outcome $\hat{Y}$. CST has the option to include or exclude the search centers when calculating (4.4). If we exclude them, then $p_c$ and $p_t$ remain as is; if we include them, then $\hat{y}_c$ and $\hat{y}_c^{CF}$ are counted in $p_c$ and $p_t$, leading to a denominator in both of $k + 1$. We add this option to be able to compare CST against standard situation testing [277, 319], which excludes the search centers, and counterfactual fairness [177], which only uses the search centers.

Since $\Delta p$ is a proportion comparison, it is *asymptotically normally distributed*, which allows to build *Wald confidence intervals* (CI) around it [277]. Let $z_{\alpha/2}$ be the $1-\alpha/2$ quantile of the standard normal distribution $\mathcal{N}$ for a *significance level* of $\alpha$ (or, conversely, a *confidence level* $(1-\alpha)\cdot 100\%$). We write the two-sided CI for $\Delta p$ of $c$ as:

$$[\Delta p - w_\alpha, \Delta p + w_\alpha], \quad \text{with} \quad w_\alpha = z_{\alpha/2}\sqrt{\frac{p_c(1-p_c) - p_t(1-p_t)}{k}} \tag{4.5}$$

The confidence interval (4.5) responds to the hypothesis that there is individual discrimination, providing a measure of certainty on $\Delta p$ through a range of possible values. For a given claim, if the CI contains the *minimum accepted deviation* $\tau$, we cannot reject the hypothesis of no discrimination with $(1-\alpha)\cdot 100\%$ confidence. In other words, the *null hypothesis* $H_0 : \pi = \tau$ cannot be rejected in favor of the *alternative hypothesis* $H_1 : \pi > \tau$, where $\pi$ is the true difference in proportion of negative decision outcomes. $\tau$, with a default choice of $\tau = 0$, represents the minimum amount of difference between $p_c$ and $p_t$ that we need to observe to claim individual discrimination.

The overall choice of $\alpha$ and $\tau$ will depend on the context of the discrimination claim. It can be motivated, for instance, by legal requirements (set, e.g., by the court [277]), or technical requirements (set, e.g., via power analysis [68]), or both.

**Definition 4.3.4** (Individual Discrimination)**.** There is potential[3] individual discrimination toward the complainant $c$ if $\Delta p = p_c - p_t > \tau$, meaning the negative decision outcomes rate for the control group is greater than for the test group by some minimum deviation $\tau \in \mathbb{R}^+$.

We do not view Definition 4.3.4 as a matter of individual versus group fairness. When we test whether $b()$ discriminates against $c$, we inevitably pass judgement onto the classifier $b()$ in fear that this behaviour has happened before. In $\mathcal{D}$ we have more than one potential discrimination claim to consider under CST, allowing to draw individual-level conclusions while motivating group-level ones. If $b()$ discriminated against $c$, it also discriminated against what $c$ represents in terms of membership to $A$.

**Definition 4.3.5** (Confidence on the Individual Discrimination Claim)**.** The Wald confidence interval (4.5) gives a measure of certainty on $\Delta p$, which is (asymptotically) normally distributed. For a significance level $\alpha$, we are $(1-\alpha)\%$ confident on $\Delta p$. The claim is said to be statistically valid if the Wald confidence interval excludes $\tau$. This definition is a statistical inference extension of Definition 4.3.4.

The many-to-many comparison behind $\Delta p$ is what makes counterfactual situation testing an *actionable framework* for testing individual discrimination. Here, the notion of repetition and its relation to representativeness and certainty concerns is important. For proving individual discrimination a single comparison is not enough [100, Sec. 6.3]. This is because we want to ensure, one, that the individual claim is representative of the population, and two, be certain about the individual claim. Implicit to both concerns is finding a pattern of unfavorable decisions against the protected group to which the individual complainant belongs to, i.e., discrimination.

[3]Or *prima facie*. In practice, discrimination needs to be argued against/for. CST alone, as with any other discrimination analysis tool, cannot claim to prove discrimination. It can, however, provide evidence against/for a discrimination case [234].

Ideally, we would repeat the decision process multiple times for the discriminatory pattern to become apparent. This is not possible in practice. Back to our illustrative example from Section 4.1, we cannot ask the female complainant to apply multiple times to the same bank. We instead can look at other similar instances under the same process. This is what $p_c$ and $p_t$ (4.4) and Definition 4.3.4 represent. Similarly, if the bank's $b()$ is shown to discriminate against the female complainant, what rules out that it has not done it before or that this one time was an exception? Again, we cannot repeat the decision process until we are certain of the individual discrimination claim. We instead can assume a theoretical distribution of comparisons with $\pi$ to account for potential randomness in what we detect from the single point estimate that is $\Delta p$. This is what the CI (4.5) and Definition 4.3.5 represent.

### 4.3.3 Connection to Counterfactual Fairness

There is a clear link between CST and *counterfactual fairness* [177]. A decision maker is counterfactually fair if it outputs the same outcome for the factual tuple as for its counterfactual tuple, where the latter is generated based on the abduction, action, and prediction steps and the intervention on the protected attribute. Refer to (2.15) in Section 2.3.2. The factual $(x_c, a_c, \hat{y}_c)$ and counterfactual $(x_c^{CF}, a_c^{CF}, \hat{y}_c^{CF})$ tuples for $c$ used in CST are also the ones used for counterfactual fairness. We view CST, when including the search centers, as an actionable extension of counterfactual fairness.

**Proposition 4.3.1** (On Actionable Counterfactual Fairness)**.** Counterfactual fairness does not imply nor it is implied by Individual Discrimination (Definition 4.3.4).

We now present a sketch of proof to Proposition. 4.3.1. Consider the factual tuple $(x_c, a_c = 1, \widehat{y}_c = 0)$ and assume the generated counterfactual is $(x_c^{CF}, a_c^{CF} = 0, \widehat{y}_c^{CF} = 0)$. Since $\widehat{y}_c = \widehat{y}_c^{CF}$, this is a case where counterfactual fairness holds. However, the decision boundary of the model $b()$ can be purposely set such that the $k$-nearest neighbors of $x_c$ are all within the decision $\hat{Y} = 0$, and less than $1 - \tau$ fraction of the $k$-nearest neighbors of $x_c^{CF}$ are within the decision $\hat{Y} = 0$. This leads to a $\Delta p > 1 - (1 - \tau) = \tau$, showing that there is individual discrimination. The other way can be shown similarly by assuming $\widehat{y}_c \neq \widehat{y}_c^{CF}$ but the sets of $k$-nearest neighbors have rates of negative decisions whose difference is lower than $\tau$.

Proposition 4.3.1 alludes to the scenario where $b()$ is counterfactually fair yet discriminatory. Intuitively, it is possible to handle *borderline cases* where the tuple of interest and its counterfactual both get rejected by $b()$, though the latter is closer to the decision boundary. The model $b()$ would be considered counterfactually fair, but would that disprove the individual discrimination claim? CST, by constructing the control and test groups around this single comparison, accounts for this actionability concern.

CST further equips counterfactual fairness with confidence intervals. Previous works have addressed uncertainty in counterfactual fairness [168, 246], but with a focus on the structure of the SCM $\mathcal{M}$. We instead address certainty on the literal comparison that motivates the counterfactual fairness definition.

### 4.3.4 k-NN Implementation

Finally, we propose an implementation to our counterfactual situation testing framework. We already defined the search algorithm $\phi$ as the k-NN algorithm. We define as the similarity measure $d$ the same distance function between two tuples, $d(x, x')$, used in the k-NN situation testing implementation (k-NN ST) [277]. We do so because we want to compare our implementation, k-NN CST, against its standard counterpart, k-NN ST. We summarize the current algorithmic CST implementation in the next section.

Let us define the *distance between two tuples* as:

$$d(x, x') = \frac{\sum_{i=1}^{|X|} d_i(x_i - x'_i)}{|X|} \tag{4.6}$$

where (4.6) averages the sum of the per-attribute distances across $X$. Interpretation-wise, a lower (4.6) implies a higher similarity between the tuples $x$ and $x'$. CST can handle non-normalized attributes but, unless specified, we normalize them to insure comparable per-attribute distances.

In (4.6) equals the *overlap measurement* ($ol$) if the attribute $X_i$ is categorical; otherwise, it equals the *normalized Manhattan distance* ($md$) if the attribute $X_i$ is continuous, ordinal, or interval. We define $md$ as:

$$md(x_i, x_{i'}) = \frac{|x_i - x_{i'}|}{(\max(X) - \min(X))} \tag{4.7}$$

and define $ol$ as:

$$ol(x_i, x_{i'}) = \begin{cases} 1 & \text{if } x_i = x_{i'} \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

The choice of (4.6) is not restrictive. In subsequent works we hope to explore other distance options like , e.g., heterogeneous distance functions [301], as well as probability-based options, e.g., propensity score weighting [229].

CST is, above all, a framework for detecting discrimination. The choice of $d$ as well as $\phi$ are specific to the implementation of CST. What is important is that the test group is established around the complainant's counterfactual while the control group, like in other discrimination frameworks, is established around the complainant.

### 4.3.5 The Algorithms

We present the relevant algorithms for the k-NN CST implementation. The Algorithm 1 performs CST while Algorithm 2 returns the indices of the top-$k$ tuples with respect to the search centers based on the distance function $d$ (4.6).

The main difference in Algorithm 1 when creating the neighborhoods is that the search centers are drawn from the factual dataset for the control group $\mathcal{D}$ and the counterfactual dataset $\mathcal{D}^{CF}$ for the test group. Further, notice that we use the same $c$ or index for both as these two data-frames have the same structure by construction.

**Algorithm 1:** run_CST

**Input :** $\mathcal{D}, \mathcal{D}^{CF}, k$
**Output:** $[p_c - p_t]$

$prot\_condition \leftarrow \mathcal{D}[:,\ prot\_attribute] == prot\_value$
$\mathcal{D}_c \leftarrow \mathcal{D}[prot\_condition]$
$\mathcal{D}_t \leftarrow \mathcal{D}[\neg\ prot\_condition]$
$prot\_idx \leftarrow \mathcal{D}_c.index.to\_list(\ )\ diff\_list = [\ ]$
**for** $c,\ row \in prot\_idx$ **do**
  $res\_1 \leftarrow get\_top\_k(\mathcal{D}\quad [c,\ :], \mathcal{D}_c, k)\ res\_2 \leftarrow get\_top\_k(\mathcal{D}^{CF}[c,\ :], \mathcal{D}_t, k)$
  $p_c \leftarrow sum(\mathcal{D}[res_1,\ target\_attribute] == negative\_outcome)\ /\ len(res\_1)$
  $p_t \leftarrow sum(\mathcal{D}[res_2,\ target\_attribute] == negative\_outcome)\ /\ len(res\_2)$
  $diff\_list[c] \leftarrow p_c - p_t$
**end**
**return** $diff\_list$

**Algorithm 2:** get_top_k

**Input :** $t, t\_set, k$
**Output:** $[indices]$

$(idx, dist) \leftarrow k\_NN(t, t\_set, k+1)$ **if** *without search centers* **then**
  $remove(t, idx, dist)$
**end**
$idx' \leftarrow sort(idx, dist)$ **return** $idx'$

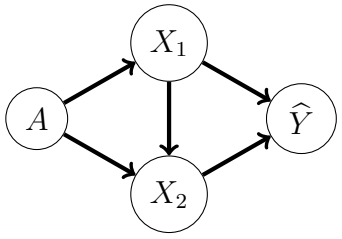
## 4.4 Experiments

We now showcase the counterfactual situation testing (CST) framework via its k-NN implementation using synthetic data in Section 4.4.1 and real data in Section 4.4.2. We contrast it to its situation testing counterpart (k-NN ST) [277], and to counterfactual fairness (CF) [177]. For the structural equations we assume additive noise. This is a convenient but not necessary assumption that simplifies the abduction step when generating the counterfactuals. Refer to Section 2.1.3 for details and the implications of additive noise model (ADM) for counterfactual generation.

We also consider the case of *positive discrimination* for both datasets in Section 4.4.3. For the real dataset, which contains multiple protected attributes, we consider $|A| = 2$ and study *multiple and intersectional discrimination* in Section 4.4.4. We extend the discrimination definitions presented in Section 4.3 accordingly.

Throughout the experiments, we use a significance level of $\alpha = 5\%$, a minimum deviation of $\tau = 0.0$, and a set of $k$ group sizes in $\{15, 30, 50, 100\}$ for CST runs that include and exclude the search centers. Also for comparison, we define individual discrimination as $\Delta p > \tau$ (Definition 4.3.4) for a single protected attribute. We still, though, demonstrate the use of confidence intervals (Definition 4.3.5) and how it would affect the final results. Finally, we assume $\mathcal{M}$ and $\mathcal{G}$ in both ADM scenarios.[4] Additional experiments in which we push this setting are presented in the Appendix A.

[4]The code and data are available at https://github.com/cc-jalvarez/counterfactual-situation-testing.




$$\mathcal{M} \begin{cases} A & \leftarrow U_A \\ X_1 & \leftarrow f_1(A) + U_1 \\ X_2 & \leftarrow f_2(X_1, A) + U_2 \end{cases}$$

$$\widehat{Y} = b(X_1, X_2)$$

Figure 4.1: The causal knowledge with corresponding SCM $\mathcal{M}$ and DAG $\mathcal{G}$ behind our (illustrative example) loan application dataset. Let $A$ denote an individual's gender, $X_1$ annual salary, $X_2$ bank balance, and $\widehat{Y}$ the loan decision based on the bank's ADM $b()$.

### 4.4.1 An Illustrative Example

We create a synthetic dataset $\mathcal{D}$ based on the scenario in Figure 4.1. It is a modified version of Karimi et al. [163, Figure 1], where we include the protected attribute gender $A$. Here, gender directly affects both an individual's annual salary $X_1$ and bank balance $X_2$, which are used by the bank's ADM $b()$ for approving ($\widehat{Y} = 1$) or rejecting ($\widehat{Y} = 0$) a loan application. We generate $\mathcal{D}$ for $n = 5000$ under $A \sim \text{Ber}(0.45)$ with $A = 1$ if the individual is female and $A = 0$ otherwise, and assume: $X_1 \leftarrow (-\$1500) \cdot \text{Poi}(10) \cdot A + U_1$; $X_2 \leftarrow (-\$300) \cdot \mathcal{X}^2(4) \cdot A + (3/10) \cdot X_1 + U_2$; and $\widehat{Y} = \mathbb{1}\{X_1 + 5 \cdot X_2 > 225000\}$ with $U_1 \sim \$10000 \cdot \text{Poi}(10)$ and $U_2 \sim \$2500 \cdot \mathcal{N}(0, 1)$. $\mathcal{D}$ represents *a known biased scenario*,[5] in which through $A$ we introduce a systematic bias onto the relevant decision attributes for female applicants.

To run CST we first generate the counterfactual dataset $\mathcal{D}^{CF}$ based on the intervention $do(A := 0)$, or *what would have happened had all loan applicants been male*? Comparing $\mathcal{D}$ to $\mathcal{D}^{CF}$ already highlights the unwanted systematic effects of $A$. This can be seen, for instance, in Figure 4.2 by the rightward shift experienced in $X_2$ for all female applicants when going from the factual to the counterfactual world. The loan rejection rate for females drops from 60.9% in $\mathcal{D}$ to 38.7% in $\mathcal{D}^{CF}$, which is now closer to the loan rejection rate of 39.2% experienced by males in both worlds. We run CST for all $k$ sizes. Results are shown in Table 4.1, where w/o refers to "without search centers" for CST.

Does $b()$ discriminate against female applicants? As Table 4.1 shows, all three methods detect a number of individual discrimination cases. On one hand, the bank clearly uses information that is neutral and needed for approving a loan request; on the other hand, this information is tainted by the effects of gender on such information and the bank, in turn, continues to perpetuate biases against women in this scenario. The results show how these neutral provisions are harmful toward female applicants, and uncover potential individual discrimination cases.

**CST relative to situation testing (ST).** Here, consider the CST version without the search centers as ST excludes them. What is clear from Table 4.1 is that CST finds more individual discrimination cases than ST for all $k$ sizes. For $k = 50$, e.g., CST (w/o) detects 20% while ST just 5%. These results highlight the impact of operationalizing *fairness*

[5]Such "penalties", e.g., capture the financial burdens female professionals face in the present after having been discouraged in the past from pursuing high-paying, male-oriented fields [75].

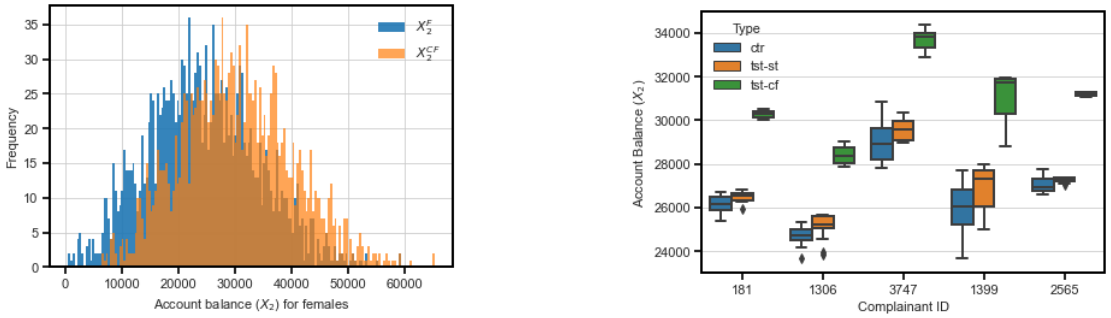



Figure 4.2: Left. Account balance ($X_2$) distribution for females in the factual $\mathcal{D}$ and counterfactual $\mathcal{D}^{CF}$ datasets. Right. A comparison on $X_2$ of the ST and CST (w/o) control group (ctr) versus the ST (tst-st) and CST (w/o) (tst-cf) test groups for five randomly chosen complainants detected by both methods, showing the *fairness given the difference* behind CST as tst-st is closer to ctr than tst-cf.

Table 4.1: Number (and %) of detected individual discrimination cases for the illustrative example based on gender.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 288 (16.8%) | 313 (18.3%) | 342 (20%) | 395 (23.1%) |
| ST [277] | 0 | 55 (3.2%) | 65 (3.8%) | 84 (5%) | 107 (6.3%) |
| CST | 0 | 420 (24.5%) | 434 (25.4%) | 453 (26.5%) | 480 (28%) |
| CF [177] | 376 (22%) | 376 (22%) | 376 (22%) | 376 (22%) | 376 (22%) |

*given the difference*, as the main difference between the two frameworks is how each individual test group is constructed based on the choice of search center. The control group is constructed the same way for both ST and CST.

The choice of the test search centers is what sets CST apart from ST. Note that ST performs an *idealized comparison*. Consider, e.g., the tuple $(x_1 = 35000, x_2 = 7948, a = 1)$ as the complainant $c$. With $c$ as the test search center, the most similar male profiles to the complainant under any distance $d$ would be tuples similar to $(x_1 = 35000, x_2 = 7948, a = 0)$. CST, conversely, performs a more *flexible comparison* under *fairness given the difference*. With the corresponding counterfactual tuple $(x_1^{CF} = 50796, x_2^{CF} = 13852, a^{CF} = 0)$ as the test search center, the most similar male profiles to the complainant under the same $d$ would be tuples similar to the counterfactual itself, not to the complainant $c$.

As a consequence of this idealized versus (more) flexible comparisons, the test groups we construct under CST is likely not to be similar to the equivalent ones we construct under ST nor to the same control groups we construct for both ST and CST. Figure 4.2 shows clearly this result for $k = 15$. We randomly chose five complainants that were discriminated by $b()$ according to both ST and CST and plot the distribution of $X_2$ for the control group (ctr), the ST test group (tst-st), and the CST test group (tst-cf). In this scenario, all 55 ST cases are also detected by CST.

**CST relative to counterfactual fairness (CF).** Here, consider the CST version including the search centers (though CST w/o is of interest also), as these represent the instances used by CF. We define CF discrimination as a case where the factual $\widehat{y}_c = 0$ becomes $\widehat{y}_c^{CF} = 1$ after the intervention of $A$. Under this definition, we detect 376 cases of CF discrimination in $\mathcal{D}$, or 22% of female applicants. CF is independent from $k$ as the framework applies only to the individual comparison of the factual and counterfactual tuples for complainant $c$. Table 4.1 show that CST detects a higher number of individual discrimination cases for each $k$ size (while CST w/o only passes CF at $k = 100$). In fact, in this scenario, all cases detected by CF are contained in CST.

What sets CST apart from CF is twofold. First, CST equips the CF comparison with certainty measures. This point is illustrated in Table 4.2 where we show individual cases of discrimination detected by both CF and CST along with confidence intervals (CI) (4.5) provided by the CST framework. Second, CST detects cases of individual discrimination that are counterfactually fair. This point is illustrated in Table 4.3 where we show individual cases that pass CF but still exhibit a discriminatory pattern when looking at $\Delta p$. Such results highlight why legal stakeholders require multiple comparisons to insure that $c$'s experience is representative of the discrimination claim.

Table 4.2: Subset of individual discrimination cases detected by both CST ($k = 15$) and CF with CI under $\alpha = 5\%$. The * denotes statistical significance.

| Comp. (ID) | $p_c$ | $p_t$ | $\Delta p$ | CI ($\alpha = 5\%$) |
|---|---|---|---|---|
| 44 | 1.00 | 0.00 | 1.00* | [1.00, 1.00] |
| 55 | 0.81 | 0.00 | 0.81* | [0.65, 0.97] |
| 150 | 1.00 | 0.94 | 0.06 | [-0.04, 0.16] |
| 203 | 1.00 | 0.88 | 0.13 | [-0.01, 0.26] |
| 218 | 0.56 | 0.00 | 0.56* | [0.36, 0.77] |

Table 4.3: Subset of individual discrimination cases detected by CST ($k = 15$) but not by CF with CI under $\alpha = 5\%$. The * denotes statistical significance.

| Comp. (ID) | $p_c$ | $p_t$ | $\Delta p$ | CI ($\alpha = 5\%$) |
|---|---|---|---|---|
| 5 | 0.06 | 0.0 | 0.06 | [-0.04, 0.16] |
| 147 | 0.50 | 0.0 | 0.5* | [0.29, 0.71] |
| 435 | 0.38 | 0.0 | 0.38* | [0.18, 0.58] |
| 1958 | 0.13 | 0.0 | 0.13 | [-0.01, 0.26] |
| 2926 | 0.75 | 0.0 | 0.75* | [0.57, 0.93] |

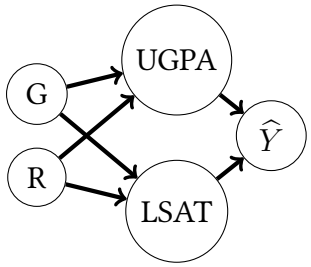
**Confidence in results.** Finally, notice that Tables 4.2 and 4.3 include cases where $\tau = 0$ falls within the individual CI. We detected these cases under $\Delta p > \tau$ (Definition 4.3.4). Under $\alpha = 5\%$, we would reject these cases as individual discrimination claims with confidence level of 95% since the minimum deviation is covered by the CIs (Definition 4.3.5). These are cases with small $\Delta p$'s that are too close to call. We denote those statistically significant cases with the asterisk on $\Delta p$.

The use of statistical significance, through the confidence intervals, shows the importance of considering uncertainty when making calls on the unfairness and, thus, potential discriminatory effects of an algorithmic decision-maker. As the asterisks in Tables 4.2 and 4.3 show, not all individual cases are representative of a larger, negative pattern. Therefore, extrapolating group-wide claims from single cases from such groups without considering this link, which is common within the fairness literature, is risky as well as detached from how these cases are argued for in court.

### 4.4.2 Law School Admissions

Based on the Law School Success example popularized by Kusner et al. [177, Figure 2] using US data from the Law School Admission Council survey [299], we create an admissions scenario to a top law school. We consider as protected attributes an applicant's gender, male/female, (*G*) and race, white/non-white, (*R*). We add an ADM $b()$ that considers the applicant's undergraduate grade-point average (*UGPA*) and law school admissions test scores (LSAT) for admission. If an applicant is successful, $\widehat{Y} = 1$; otherwise $\widehat{Y} = 0$. We summarize the scenario in Figure 4.3.

For the ADM $b()$, we use the median entry requirements for the top US law school to derive the cutoff $\psi$.[6] The cutoff is the weighted sum of 60% 3.93 over 4.00 in *UGPA* and 40% 46.1 over 48 *LSAT*, giving a total of 20.8; the maximum possible score under $b()$ is 22 for an applicant. The structural equations follow (2.2), as in [177], with $b_U$ and $b_L$ denoting the intercepts; $\beta_1$, $\beta_2$, $\lambda_1$, $\lambda_2$ the weights; and $UGPA \sim \mathcal{N}$ and $LSAT \sim$ Poi. the probability distributions.




$$\mathcal{M} \begin{cases} R & \leftarrow U_R \\ G & \leftarrow U_G \\ UGPA & \leftarrow b_U + \beta_1 \cdot R + \lambda_1 \cdot G + U_1, \\ LSAT & \leftarrow \exp\{b_L + \beta_2 \cdot R + \lambda_2 \cdot G + U_2\}, \end{cases}$$

$$\widehat{Y} = \mathbb{1}\{(0.6 \cdot UGPA + 0.4 \cdot LSAT) > \psi\}$$

Figure 4.3: The causal knowledge with corresponding SCM $\mathcal{M}$ and DAG $\mathcal{G}$ behind the law school admissions dataset, with $R$ denoting race ($R = 1$ for non-white) and $G$ denoting gender ($G = 1$ for female).

The dataset $\mathcal{D}$ contains $n = 21790$ applicants, 43.8% are females and 16.1% are non-whites. Despite the ADM $b()$ being externally imposed by us for the purpose of illustrating the CST framework, under $b()$ only 1.88% of the female applicants are successful compared to 2.65% of the male applicants; similarly, only 0.94% of the non-white applicants are successful compared to 2.58% of the white applicants. Therefore, is $b()$ discriminatory toward non-white and female applicants? We run CST along with ST and CF for each protected attribute. We generate the counterfactual dataset $\mathcal{D}^{CF}$ for each $R$–*what would have been the outcome had all applicants been white*?–and $G$–*what*

[6] At the time of writing, that being Yale University Law School—see `https://www.ilrg.com/rankings/law/index/1/asc/Accept`

*would have been the outcome had all applicants been male*?–and present the results, respectively, in Table 4.4 and Table 4.5. Both tables show CST detecting more individual cases of discrimination than ST and CF.

Table 4.4: Number (and %) of individual discrimination cases for the law school admissions scenario based on race.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 256 (7.3%) | 309 (8.81%) | 337 (9.61%) | 400 (11.41%) |
| ST [277] | 0 | 33 (0.94%) | 51 (1.45%) | 61 (1.74%) | 64 (1.83%) |
| CST | 0 | 286 (8.16%) | 309 (8.81%) | 337 (9.61%) | 400 (11.41%) |
| CF [177] | 231 (6.59%) | 231 (6.59%) | 231 (6.59%) | 231 (6.59%) | 231 (6.59%) |

Table 4.5: Number (and %) of individual discrimination cases in for the law school admissions scenario based on gender.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 78 (0.82%) | 120 (1.26%) | 253 (2.65%) | 296 (3.10%) |
| ST [277] | 0 | 77 (0.81%) | 101 (1.06%) | 229 (2.4%) | 258 (2.71%) |
| CST | 0 | 99 (1.04%) | 129 (1.35%) | 267 (2.80%) | 296 (3.10%) |
| CF [177] | 56 (0.59%) | 56 (0.59%) | 56 (0.59%) | 56 (0.59%) | 56 (0.59%) |

A similar analysis of Tables 4.4 and 4.5 for CST (w/o) versus ST and CST versus CF as in Section 4.4.1 follows. What the results here highlight, though, is how the two versions of CST compare to each other. In both tables, as $k$ increases, CST (w/o) catches on to CST in the number of cases. This is likely related to how the observations for female/male and non-white/white are distributed in the dataset; though, it also relates to the fact that the difference in size between the groups in each version is just one instance: $k$ versus $k + 1$. We should observe the same trend in Table 4.1 if we continued to increase $k$.

The results in Tables 4.4 and 4.5 show that the different runs of CST can eventually reach the same conclusions under a certain $k$ size. In practice, it means that we could implement one of the two versions of CST without compromising the number of detected individual discrimination cases, though further research is needed.

### 4.4.3 Positive Discrimination

Positive discrimination refers to cases where the protected individual (or complainant) is shown to be favored over the non-protected individual. It sounds similar to notions of *affirmative action*, however, for that to occur there needs to exist an explicit policy favoring the protected individuals, which is not the case for neither experiment. It also referred to as *tokenism* [234].

**Definition 4.4.1** (Positive Discrimination)**.** For a minimum deviation $\tau$ and a single protected attribute $A$, we define potential positive individual discrimination as $\Delta p < \tau$.

Cases of positive discrimination clearly do not fall under Definition 4.3.4, meaning they do not count as individual cases of discrimination. Regarding counterfactual fairness (CF), we define as positive discrimination when the factual has a positive decision outcome, $\hat{y}_c = 1$, but its counterfactual a negative one, $\hat{y}_c^{CF} = 0$.

Table 4.6: Number (and %) of positive individual discrimination cases in Section 4.4.1 based on gender.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| ST [277] | 0 | 45 (2.6%) | 50 (2.9%) | 77 (4.5%) | 118 (6.9%) |
| CST | 0 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| CF [177] | 0 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |

Table 4.7: Number (and %) of positive individual discrimination cases in Section 4.4.2 based on race.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 3 (0.0%) |
| ST [277] | 0 | 46 (1.3%) | 51 (1.5%) | 75 (2.1%) | 121 (3.5%) |
| CST | 0 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 3 (0.0%) |
| CF [177] | 0 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |

Table 4.8: Number (and %) of positive individual discrimination cases in Section 4.4.2 based on gender.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 57 (0.6%) | 128 (1.34%) | 73 (0.77%) | 104 (1.09%) |
| ST [277] | 0 | 44 (0.46%) | 144 (1.51%) | 108 (1.13%) | 159 (1.67%) |
| CST | 0 | 42 (0.4%) | 111 (1.16%) | 73 (0.77%) | 93 (0.98%) |
| CF [177] | 1 (0.0%) | 1 (0.0%) | 1 (0.0%) | 1 (0.0%) | 1 (0.0%) |

In all three tables ST detects more positive discrimination cases than either version of CST and CF. It does so, despite manipulating $A$ under a *ceteris paribus* type manipulation and, thus, taking for granted the effects of $A$ on $X$. Further research is needed as it is unclear why ST would, e.g., pick up on positive discrimination cases in Table 4.6 when we have introduced a systematic bias against female candidates in Figure 4.1.

We also find interesting that CST and CF align in detecting positive discrimination. Why this is the case, remains an open question to be formalized in subsequent works. It is possible that, under a *mutatis mutandis* type manipulation, this type of discrimination does not occur as non-protected individuals remain the same when moving from the factual to the counterfactual world.

### 4.4.4 Multiple and Intersectional Discrimination

This discrimination scenario only applies to the Law School Admissions scenario in Section 4.4.2, in which we have two protected attributes: gender and race, or $|A| = 2$. We draw mainly from Xenidis [308] for understanding and modeling the tension between multiple versus intersectional discrimination within the EU legal context. These two (as well as positive discrimination) are forms of discrimination that we consider understudied within algorithmic fairness. Hence, we want to show that our CST framework can handle them. We recognize, though, that further research, mainly in the normative sense, is needed.

**Definition 4.4.2** (Multiple Discrimination)**.** Under Definition 4.3.4 for potential individual discrimination and a set of $|A| = q > 1$ protected attributes, potential multiple individual discrimination occurs when the complainant $c$ is discriminated by the model for each $\{A_i\}_{i=1}^{q}$ protected attribute.

In the law school admission scenario (Section 4.4.2), Definition 4.4.2 means that a complainant must be discriminated separately as a female individual concerning the protected attribute gender *and* as a black individual concerning the protected attribute race. Here, only those individuals belonging to the protected classes of each protected attribute in the dataset $\mathcal{D}$ are eligible for multiple discrimination.

**Definition 4.4.3** (Intersectional Discrimination)**.** Under Definition 4.3.4 for potential individual discrimination and a set of $|A| = q > 1$ protected attributes, potential intersectional individual discrimination occurs when the complainant $c$ is discriminated by the model w.r.t. the protected attribute $A^* = \mathbb{1}\{A_1 = 1 \wedge A_2 = 1 \wedge \cdots \wedge A_q = 1\}$.

The Definition 4.4.3 also refers to those individuals belonging to the protected attributes gender and race in the law school admission scenario. However, it is a different definition than Definition 4.4.2 as we require the discrimination to be *simultaneous*.

The tension between Definition 4.4.2 and Definition 4.4.3 occurs because it is possible for multiple discrimination not to occur while intersectional discrimination occurs, which is troubling as only multiple discrimination is recognized under EU law [308]. We operationalize the two types of $|A| > 1$ discrimination under the CST framework in the following ways:

- Multiple discrimination: run CST separately for each $A_i$, including the generation of a counterfactual dataset for each $do(A_i := 0)$; and look for individual cases in which discrimination is detected across all runs.

- Intersectional discrimination: create the *intersectional protected attribute* $A^*$ as in Definition 4.4.3; generate the corresponding counterfactual dataset on $do(A^* := 0)$; and run a single CST as we would for $|A| = 1$.

Regarding counterfactual fairness (CF), for multiple discrimination we look at the factual-counterfactual tuples for each $A_i \in A$; for intersectional discrimination we looks at the factual-counterfactual tuples for the intersectional protected attribute $A^*$. Table 4.9 answers the question *what would have been the outcome had the complainant been male and had the complainant been white?* Table 4.10 answers the question what *would*

*have been the outcome had the complainant been male-white?* The last question implies that the non-protected individuals include the trivial male-white category but also the female-white and male-non-white categories.

Table 4.9: Number (and %) of multiple individual discrimination cases in Section 4.4.2 based on *gender and race*.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 8 (0.0%) | 10 (0.0%) | 20 (0.01%) | 20 (0.01%) |
| ST [277] | 0 | 5 (0.0%) | 5 (0.0%) | 12 (0.01%) | 19 (0.01%) |
| CST | 0 | 9 (0.0%) | 10 (0.00%) | 21 (0.01%) | 20 (0.01%) |
| CF [177] | 5 (0.0%) | 5 (0.0%) | 5 (0.0%) | 5 (0.0%) | 5 (0.0%) |

Table 4.10: Number (and %) of intersectional individual discrimination cases in Section 4.4.2 based on *gender-race*.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 130 (7.09%) | 138 (7.53%) | 148 (8.07%) | 160 (8.73%) |
| ST [277] | 0 | 14 (0.76%) | 14 (0.76%) | 17 (0.93%) | 24 (1.31%) |
| CST | 0 | 130 (7.09%) | 138 (7.53%) | 148 (8.07%) | 160 (8.73%) |
| CF [177] | 113 (6.16%) | 113 (6.16%) | 113 (6.16%) | 113 (6.16%) | 113 (6.16%) |

What is interesting between the results in Tables 4.9 and 4.10 is that we find less multiple individual discrimination cases than intersectional individual discrimination cases. It is interesting as these two types are linked to each other: both address the same subset within the protected attribute space. These, of course, are preliminary results but they seem to highlight the concerns by [308] that intersectional discrimination goes below the radar due to our current focus on multiple discrimination. Further, the numbers for intersectional discrimination in Table 4.10 are lower than for the respective individual discrimination cases in Tables 4.5 and 4.4. This is obvious as we would expect for the numbers to decrease when looking at the intersection of two attributes instead of one of its parts (i.e., the conjunction rule).

What is less obvious is why do we get more individual cases in the intersectional setting than in the multiple cases. We believe this might be because under multiple discrimination, we are looking at the strict intersection of the two protected attributes: we run CST separately for each and then look for simultaneously detected cases. Under intersectional discrimination, instead, we are looking at a new attribute, i.e., $A^*$, that defines a new, larger set of non-protected individuals. The higher numbers come from the fact of comparing a smaller but closer protected group to a larger but closer non-protected group. Further research is needed.

## 4.5 Conclusion

We presented counterfactual situation testing (CST), a new framework for detecting individual discrimination in a dataset of classifier decisions. Compared to other methods,

CST uncovers more cases even when the classifier is counterfactually fair. It also equips counterfactual fairness with uncertainty measures. CST acknowledges the pervasive effects of the protected attribute by comparing individual instances in the dataset that are observably different in the factual world but hypothetically similar in the counterfactual world. Thus, the results are not too surprising as CST operationalizes *fairness given the difference*, which is a more flexible take on similarity between individuals for testing discrimination than the standard, idealized comparison of two individuals that only differ on their protected status.

**Implementation.** CST is, above all, a framework for detecting discrimination that advocates for building the test group on the generated counterfactual of the complainant. How similarity is defined, e.g., obviously conditions the implementation. We presented a k-NN version with $d$ as (4.6); other implementations are possible and still loyal to CST as long as the construction of the control and test groups follows the *fairness given the difference* principle.

Similarly, detecting discrimination is a difficult, context-specific task. That is why for CST we emphasized the role of the expert in constructing the causal graph necessary for generating the counterfactual instances. Indeed, this step could be optimized using, e.g., causal discovery methods [224], but proving discrimination is time consuming and should remain as such given its sensitive role in our society.

Further, we are aware that, e.g., the experimental setting could be pushed further by considering higher dimensions or more complex causal structures. What is the point in doing so, though, if that is not the case with current ADM tools being deployed and audited in real life like the recent Dutch scandal [138]? Proving discrimination is not a problem exclusive to (causal) modeling. With CST we wanted to create a framework aware of the multiple angles to the problem of proving discrimination. The cases we have tackled here are intended to showcase what is possible implementation-wise.

**Limitations.** As future work, promising directions include extending the framework to cases where causal sufficiency does not hold, which is a common risk, and to cases where the decision maker $b()$ is non-binary or of a specific type (e.g., a decision tree). Here, we have also focused on tabular data. Future work should push CST further into more complex datasets to explore the scalability and robustness of the framework.

# Chapter 5

# Causal Perception

This chapter is based on the working paper: J. M. Álvarez and S. Ruggieri. Causal perception. *CoRR*, abs/2401.13408, 2024. It is currently under submission.

Perception occurs when two individuals interpret the same information differently. Despite being a known phenomenon with implications for bias in decision-making, as individual experience determines interpretation, perception remains largely overlooked in machine learning (ML) research. Modern decision flows, whether partially or fully automated, involve human experts interacting with ML applications. How might we then, e.g., account for two experts that interpret differently a deferred instance or an explanation from a ML model? To account for perception, we first need to formulate it. In this chapter, we define perception under causal reasoning using structural causal models (SCM). Our framework formalizes individual experience as additional causal knowledge that comes with and is used by a human expert (read, decision maker). We present two kinds of causal perception, unfaithful and inconsistent, based on the SCM properties of faithfulness and consistency. Further, we motivate the importance of perception within fairness problems. We illustrate our framework through a series of decision flow examples involving ML applications and human experts.

## 5.1 Introduction

The same information can be interpreted differently by two individuals. Consider, for instance, the rabbit-duck illusion made famous by Wittgenstein [302] in which individuals see either a rabbit or a duck when looking at a drawing for the first time. Psychologists refer to this phenomenon as *perception* [158]. It is a product of the mental shortcuts, or heuristics, used by humans that enable faster and potentially biased decision-making. These heuristics are shaped by each individual's experience, with an individual's socioeconomic background as a key contextual driver. Perception is harmless in cases such as the rabbit-duck illusion. In more sensitive cases, though, such as developing fair and transparent machine learning (ML) applications, the role of perception needs further consideration.

Ensuring fair ML applications, whether these are partially or fully automated, always involves some degree of human decision-making [245]. In learning to defer (LtD), e.g., the goal is to learn a model that abstains from predicting on instances it is not certain of and defers the decision to a human expert [188]. In explainable artificial intelligence

(xAI), e.g., the goal is to develop methods that explain to a human expert the predictions by a black-box model [118]. Clearly, perception can occur in both examples when multiple experts are involved. Notably, as both LtD and xAI account for the behavior of these experts, perception can impact the outputs of these ML applications. How might we learn to defer when two experts give different decisions on the same instance or design an explanation when two experts view differently the same statement? To answer these and similar questions, we first need to formalize perception in a way that is suitable for ML problems before treating it as a parameter of interest. That is the main objective of this chapter.

We present the *causal perception framework*. We consider the setting in which multiple individuals interpret the same information. Perception occurs as each individual's interpretation, shaped by their own experiences, is different. We formalize such setting by proposing a causal framework for perception based on structural causal models (SCM) by Pearl [218], allowing us to define what individual experience is and how it materializes through causal reasoning. In particular, we formalize perception as a difference in causal reasoning due to competing individual-specific SCMs. The proposed framework lays the basis for future work aimed at accounting for perception among individuals (read, decision makers) that interact with a ML application.

To illustrate the role of perception as we view it, consider Kleinberg et al. [174]'s college admissions example in which an admissions officer must decide between two applicants with similar SAT scores and high-school grades. The officer, using the applicants' addresses, "knows" that one lives in a wealthy neighborhood and the other one lives in a poor neighborhood. Which applicant should she choose? It is not a trivial question to ask. The answer may vary depending on who answers and what their views are on linking an applicant's socioeconomic background, observed performance, and unobserved potential. Choosing one applicant at random, e.g., may not be a decision shared by all officers. Example 5.2.1 extends this example under a decision flow involving the officer and her interaction with three common ML applications. We use it throughout the chapter to showcase the causal perception framework.

**Summary of our contributions.** Our main contributions are three.

- We provide a first definition of perception as a causal reasoning problem.
- In doing so, we present an intuitive framework suitable for ML applications for structuring and reasoning about the additional information (i.e., individual experience) that comes with the decision maker.
- We explore perception as a parameter for modeling fairness and introduce sensitive attributes, like gender, as loaded attributes due to their role in eliciting perception through stereotypes.

We emphasize that our work is conceptual. As we discuss in Sections 5.1.1 and 5.1.2, perception is important to model yet difficult to implement. As humans interact more with ML applications, though, which does not diminish the risk of perception occurring, there is a clear need to formulate perception in order to implement it within modern fair decision flows. In the rest of this chapter, we define causal perception in Section 5.2 and its two kinds (unfaithful and inconsistent) in Section 5.3. We discuss the relationship of

perception to fairness in Section 5.4. We conclude in Section 5.5 with potential implementations and limitations of the framework. Finally, Appendix B contains additional supporting material.

### 5.1.1 Motivation: What's the Problem, Linda?

Our interest in perception started with a talk by the late Daniel Kahneman, a leading psychologist, at NeurIPS 2021 [158]. In particular, it started with his discussion of the Linda Problem (Example 5.1.1), one of his and Amos Tversky's most famous experiment. The Linda Problem studies the conjunction fallacy, which occurs when the probability of a conjunction is considered higher than the probability of one of its parts. The conjunction fallacy is caused by the representativeness heuristic [279] where an event is made to be more representative of a class than what it actually is, as measured by a higher probability, due to an individual's perception of the event. Tversky and Kahneman [280, 281] tested the conjunction fallacy via a series of experiments in which participants were provided with fictitious profiles and asked to rank the statements that best described each profile. Linda's profile remains the most famous one. The fallacy occurs as participants overwhelmingly rank the conjunction, option *(b)*, over one of the conjunction's parts, option *(a)*, violating basic laws of probability (see Appendix B.1) that, in theory, describe rational decision-making.

**Example 5.1.1** (The Linda Problem)**.** Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. What is more probable today?
*(a)* Linda is a bank teller; or
*(b)* Linda is a bank teller and is in the feminist movement.

Why would participants *perceive* option *(b)* as more representative than option *(a)* of Linda? The answer, as Kahneman [158] argued, was in the description of Linda. After further research, what struck us about the Linda Problem was how little it had changed and how effective it has remained since it first appeared in the 1980s (see Appendix B.2): *all versions describe Linda as a young, single, educated female.* This pattern resonated with us given the recent ML works calling for special attention to sensitive attributes when reasoning about fairness in decision-making (e.g., Álvarez and Ruggieri [8], Hanna et al. [128], Hu and Kohler-Hausmann [149]). Perception to us appeared to be a phenomenon driven by the sensitive attribute (here, gender) and the stereotypes associated with it.

### 5.1.2 Related Work

Perception has been studied mainly by psychologists. Tversky and Kahneman [279, 280, 281] formulate the representativeness heuristic and other cognitive biases using probabilistic reasoning. Bayesian modeling–where the baseline representativeness of an event (the prior) is adjusted by the agent based on her experience (the posterior)–remains the common approach for modeling explicitly the representativeness heuristic and, thus, implicitly perception (e.g., Bordalo et al. [44], Costello [71], Tentori [275]). We are the first to formulate perception explicitly and using SCM, extending it into causal probabilistic

reasoning. We discuss further our modeling choice in Section 5.2.3. In doing so, we join recent works, like Beckers et al. [29] who analyze harmful decision-making, that use SCM to understand decision flows involving humans and ML models.

Within ML there is some interest in cognitive biases (e.g., Bengio [31], Booch et al. [43]). These works, however, focus mainly on how to create intelligent systems that improve over and potentially replace human decision-making. We discuss critically this growing line of work in Appendix B.3. Simultaneously, works studying the human-in-the-loop problem from a fairness and accuracy perspective (e.g., De et al. [79], Mozannar et al. [204]), which includes LtD and xAI, consider the human's interaction with the ML model but do not give much agency to the human within the problem formulation. These works often treat the human as an additional and costly decision-maker, ignoring any influence from the cognitive biases. We note, however, recent human-in-the-loop works (e.g., Caraban and Karapanos [56], Rastogi et al. [230], Yang et al. [311]) that design ML applications that improve human decision-making under these cognitive biases.

The role of cognitive biases in fair ML tools and applications, with some exceptions (e.g., Bertrand et al. [33], Echterhoff et al. [89]), remains unexplored. Given the problem of subjective or context-aware fairness, where we recognize that fairness can have different meanings across humans, works like Srivastava et al. [272] and Yaghini et al. [309] design experiments to test for the human perception of fairness. Different from these works, we focus more on formalizing perception itself through causal reasoning. We come back to this line of work in Section 5.4.

## 5.2 Problem Formulation

We present the problem of causal perception in its most basic form. The goal is to formulate when two individuals, or decision makers (DM), interpret differently the same information from a decision flow. Let us use Example 5.2.1.

**Example 5.2.1.** (College Admissions) An admissions officer (the DM) is tasked with choosing the incoming class based on the applicants' profiles. Assume a decision flow in which the officer admits, $Y = 1$, or rejects, $Y = 0$, applicants based on their SAT results, $X_1$, high-school GPA, $X_2$, and suitability scores $f(X_1, X_2) = G \in [0, 1]$ where $f$ is a ML model trained by the college. The officer, through the applicants' motivation letters, also has access to their address, $Z$. Consider three scenarios in which the officer relies on $f$ with varying degrees for the decision:
*(i)* $f$ abstains from classifying an applicant on which it is not confident and the officer must classify this applicant;
*(ii)* $f$ provides the same score for two applicants and the officer must choose one between these two applicants;
*(iii)* $f$ alone derives $Y$ using $G$ to rank applicants and admits the top-$k$ ones where $k$ is set by the officer.

Scenarios *(i)* and *(ii)* represent a partially automated decision flow where the ML model $f$ aids the admissions officer, while scenario *(iii)* represents a fully automated decision flow where the ML model $f$ replaces the admissions officer as the DM. These are high-levels scenarios. Further context is provided for each when as we use Example 5.2.1.

### 5.2.1 Setting and Background

We represent the *information* as a random variable $\mathbf{X}$. Let $P(\mathbf{X})$ denote the joint probability distribution of a set of $p$ random variables or *classes* $\mathbf{X} = X_1, \ldots, X_p$, with $P(\mathbf{x})$ representing the probability that $\mathbf{X}$ equals the $p$ realizations or *instances* $\mathbf{x} = x_1, \ldots, x_p$. Further, we represent individuals as *agents*. Borrowing from Spence [269], we define two types of agents: a *sender* $S \in \mathcal{S}$, with $\mathcal{S}$ denoting the set of senders, is an agent that provides information while a *receiver* $R \in \mathcal{R}$, with $\mathcal{R}$ denoting the set of receivers, is an agent that interprets the information provided.

For instance, in Example 5.2.1 the admissions officer is clearly the receiver $R$. Who or what is the sender $S$, though, depends on which information $R$ decides to use from the decision flow. Consider, e.g., scenario *(ii)*. If $R$ breaks the tie by evaluating the applicants' profiles $\mathbf{X}$, then the applicants are the senders. If $R$, instead, breaks the tie using the two explanations provided by $f$ for each applicant's score $f(\mathbf{X})$, then the xAI method behind the explanations is the sender. This classification of agents is the only aspect we borrow from the signaling games literature as we do not conceive strategic behavior between $S$ and $R$ [266].

Under this setting, perception can occur once we consider a second $R$ (e.g., two officers disagreeing on the tiebreaker). Perception can also occur if we allow for $R$ to change its interpretation over time (e.g., an officer interpreting differently two similar tiebreakers). We focus on the former case though we stress that our framework handles also the latter case.

**Definition 5.2.1.** (Perception) For receivers $R_i, R_j \in \mathcal{R}$, given the information by sender $S \in \mathcal{S}$ in the form of the instances $\mathbf{x}$ of the class $\mathbf{X}$, perception occurs when

$$d\big(P_{R_i}(\mathbf{X}), P_{R_j}(\mathbf{X})\big) > \epsilon \tag{5.1}$$

where the probability distributions $P_{R_i}(\mathbf{X})$ and $P_{R_j}(\mathbf{X})$ represent the interpretations by $R_i$ and $R_j$, respectively, of the information $\mathbf{X}$ and $d(\cdot, \cdot)$ denotes a suitable probability distance metric with $\epsilon \in \mathbb{R}^+$. Implicit to (5.1) are the corresponding probabilities $P_{R_i}(\mathbf{X} = \mathbf{x})$ and $P_{R_j}(\mathbf{X} = \mathbf{x})$ representing the interpretation of the information $\mathbf{x}$.

To illustrate Definition 5.2.1, consider Example 5.1.1. If we define Linda's description as the information, with Kahneman and Tversky as the sender $S$, then we can formulate the conjunction fallacy as a consequence of perception based on two receivers: the rational participant $R_1$ that views option *(a)* more likely than *(b)* and the average participant $R_2$ that views option *(b)* more likely than *(a)*. We will argue in Section 5.2.2 our choice to model information in terms of probabilities and in Section 5.2.3 our choice to use causal reasoning to handle these probabilities. Now we present the necessary causal background before defining causal perception.

**Structural causal models** We model causality using the SCM $\mathcal{M}$ (2.2) presented in Definition 2.1.1. The SCM $\mathcal{M}$ has a corresponding causal graph $\mathcal{G}$ that we assume to be acyclical. Hence, we refer to $\mathcal{G}$ also as a directed acyclical graph, or DAG. Refer to Section 2.1 in Chapter 2 for details.

**Causal reasoning and its implied distributions.** The SCM $\mathcal{M}$ allows to reason about $P(\mathbf{X})$ in terms of observed and hypothetical scenarios.[1] For the observed scenario, or *what is*, it is possible to disentangle the joint probability distribution $P(\mathbf{X})$ by factorizing it as a product of cause-effect pairs given the SCM $\mathcal{M}$:

$$P(\mathbf{X}) = \prod_{i=1}^{p} P\Big(X_i \mid X_{pa(i)}\Big) \tag{5.2}$$

which simplifies reasoning about $P(\mathbf{X})$, as it states that $X_i$ is conditionally independent of all other variables given its parents $X_{pa(i)}$. This property is known as the *Markovian condition* [224].

For the hypothetical scenarios, or *what if*, it is possible to generate new distributions of $P(\mathbf{X})$ by intervening the SCM $\mathcal{M}$. An intervention on a single variable $X_i$ is done via the *do-operator*, $do(X_i = X_i)$, which replaces the structural equation in $\mathbf{F}$ for the variable $X_i$ with the value $x_i$. Interventions apply similarly for multiple variables, $do(X_i = x_i, X_j = x_j)$, replacing the structural equations for each variable individually.

Let $\mathcal{I}_{\mathbf{X}}$ denote the *set of all interventions*, which is an index set with each index representing a specific intervention on the variables $\mathbf{X}$. We use $\emptyset \in \mathcal{I}_{\mathbf{X}}$ to denote the null intervention. As Rubenstein et al. [241] point out, $\mathcal{I}_{\mathbf{X}}$ has a *natural partial ordering*, in which for interventions $i, j \in \mathcal{I}_{\mathbf{X}}$, $i \leq_{\mathbf{X}} j$ if and only if $i$ intervenes on a subset of the variables that $j$ intervenes on and sets them equal to the same values as $j$.[2] Each intervention implies a well-defined single joint distribution of $\mathbf{X}$ variables $P(\mathbf{X})^{do(i)}$ for $i \in \mathcal{I}_{\mathbf{X}}$. Following Rubenstein et al. [241], we define the *poset of all distributions implied* by the SCM $\mathcal{M}$, where $\leq_{\mathbf{X}}$ is the natural partial ordering inherited from $\mathcal{I}_{\mathbf{X}}$, as:

$$\mathcal{P}_{\mathbf{X}} := \Big(\Big\{P(\mathbf{X})^{do(i)} : i \in \mathcal{I}_{\mathbf{X}}\Big\}, \leq_{\mathbf{X}}\Big) \tag{5.3}$$

We note that, by definition, $P(\mathbf{X}) \in \mathcal{P}_{\mathbf{X}}$. Further, $\mathcal{P}_{\mathbf{X}}$ is a singleton comprised of $P(\mathbf{X})$ when $\mathcal{I}_{\mathbf{X}} = \{\emptyset\}$. Intuitively, $\mathcal{P}_{\mathbf{X}}$ *represents all possible ways of reasoning about variables* $\mathbf{X}$ *as implied by a* SCM $\mathcal{M}$.

**On faithfulness and consistency.** All causal reasoning tied to a SCM $\mathcal{M}$ is subject to the structure of the model. We focus on two key SCM properties: faithfulness and consistency. Regarding *faithfulness* [224], because multiple graphs can describe the same joint probability distribution, the goal is to work with one $\mathcal{G}$ for $P(\mathbf{X})$ to derive a single factorization of $P(\mathbf{X})$ (5.2). We say that $\mathcal{G}$ is faithful to $P(\mathbf{X})$, as it is non-trivial to show that $\mathcal{G}$ offers the only factorization for $P(\mathbf{X})$ [38]. All implied distributions in $\mathcal{P}_{\mathbf{X}}$ (5.3) are assumed with respect to a faithful $\mathcal{G}$ for $P(\mathbf{X})$.

Regarding *consistency* [241], the goal is for the reasoning to be consistent across all levels of model abstraction. We focus on low (or micro) to high (or macro) modeling levels. If we picture each node in $\mathcal{G}$ as a *molecular structure*, then we can conceive different levels of structural representations that manifest when "zooming in and out" of

[1] Pearl and Mackenzie [220] present three levels of reasoning: observational (*what is*), interventional (*what if*), and counterfactual (*what would have been if*). The latter two represent the hypothetical scenario. For our purposes, we use a simpler distinction.

[2] E.g., $do(X_i = x_i) \leq_{\mathbf{X}} do(X_i = x_i, X_j = x_j)$. Intuitively, the $j$ intervention can be done after the $i$ intervention without needing to change the modifications done by $i$ on the SCM.

$\mathcal{G}$. Consistency requires that, when reasoning about $\mathbf{X}$ variables via interventions, the conclusion is the same regardless of the modeling level. Consistency has implications on $\mathcal{P}_{\mathbf{X}}$ (5.3). We come back to it in Section 5.3.3.

**Definition 5.2.2.** (Causal Perception) For receivers $R_i, R_j \in \mathcal{R}$ with SCM $\mathcal{M}_{R_i}$ and $\mathcal{M}_{R_i}$ for the information $\mathbf{X}$ provided by sender $S \in \mathcal{S}$, causal perception occurs when

$$\overline{d}\big(\mathcal{P}_{\mathbf{X}_{R_i}}, \mathcal{P}_{\mathbf{X}_{R_j}}\big) > \epsilon \tag{5.4}$$

where $\mathcal{P}_{\mathbf{X}_{R_i}}$ and $\mathcal{P}_{\mathbf{X}_{R_j}}$ represent the poset of all distributions according to $\mathcal{M}_{R_i}$ and $\mathcal{M}_{R_i}$ of $\mathbf{X}$ and $\overline{d}(\cdot,\cdot)$ denotes a suitable aggregated distance measure between two sets of probability distributions with $\epsilon \in \mathbb{R}^+$.

Definition 5.2.2 extends perception into the realm of causal reasoning. Notice that unlike Definition 5.2.1, it defines perception beyond a disagreement on the representation of $\mathbf{X}$ in terms of probabilities by also accounting for (any) disagreement on reasoning about $\mathbf{X}$ in terms of probabilities. This is why we introduce $\overline{d}$ as an aggregated distance, such as an average or a maximum (see Goldenberg and Webb [108]). The choice of $\overline{d}$ will depend on the context, meaning the kind of disagreement, we wish to capture. We illustrate Definition 5.2.2 in Section 5.3, where we describe what it means for a receiver to be equipped with SCM $\mathcal{M}_R$ and how that determines the kind of causal perception.

### 5.2.2 The Probabilistic Problem of Representation

Probabilities allow to quantify *how representative* an instance $x$ is of a class $X$. Therefore, as formulated in Definitions 5.2.1 and 5.2.2, perception coalesces into two individuals judging differently the representativeness of the same instance $x$ (or instances $\mathbf{x}$) of a class $X$ (or classes $\mathbf{X}$).

**Degree of representativeness.** Perception is driven by the representativeness heuristic, one of three *judgment heuristics* that causes biased decision-making in humans under uncertainty [279].[3] The *representativeness heuristic* is used in scenarios where we are asked to evaluate the degree to which one instance $x$ is representative of another instance $x'$, leading us to evaluate poorly the representativeness of $x$ in terms of the probability $P(x)$. In such scenarios, we dwell into the question of *resemblance between instances*: the more one instance resembles another, the more representative it is of the other instance. The question "what is the probability that $x$ belongs to the class $X$?" asks "to what degree the instance $x$ resembles other known instances in the class $X$?" If the resemblance is high, then we judge the probability $P(x)$ that $x$ belongs to $X$ or, equivalently, that $x$ is generated by $X$ to be high. Here, it helps to reason in terms of *degree of representativeness*, which translates naturally into how we understand probabilities.

**Definition 5.2.3.** (Degree of Representativeness) For the class $X$ with known instance $x'$, the probability $P(x)$ measures the degree of representativeness of $x$ as an instance of $X$ based on its resemblance to $x'$. Given a distance $d(\cdot,\cdot)$ between instances, we have $P(x) \approx P(x')$ as $d(x,x') \approx 0$. This definition extends to the joint probability of the collection of classes $\mathbf{X}$, with instances $\mathbf{x}$ and known instances $\mathbf{x}'$.

[3] The availability and anchoring heuristics being the others.

It follows that if the known instance $x'$ is viewed as *representative of* $X$, denoted, e.g., by a high-enough $P(x')$ within the relevant context, then $x$ is also representative of $X$ as captured by $P(x)$. The issue with Definition 5.2.3 is that the degree of representativeness is clearly a function of what someone (or something) considers representative of $X$ as captured by $x'$ and $P(x')$. This degree varies across individual experience [279, 280, 281]. It is what underpins the representative heuristic and leads to, e.g., the conjunction fallacy in Example 5.1.1.

**Example 5.2.2.** In Example 5.2.1, scenario *(i)*, the admissions officer resorts to the degree of representativeness when evaluating an applicant that is deferred by the abstaining ML model $f$. To evaluate such an applicant, with profile $\mathbf{x} = \langle x_1, x_2 \rangle$, the officer implicitly compares it to a past successful (or an imagined ideal) applicant, with profile $\mathbf{x}' = \langle x_1', x_2' \rangle$. The closer $\mathbf{x}$ is $\mathbf{x}'$, the higher the probability for applicant $i$ to be admitted by the officer, $P(Y = 1|\mathbf{x})$. What the officer chooses to (un)consciously use as $\mathbf{x}'$ influences the classification of all deferred applicants.

**Evocation.** We view the act of judging the representativeness of some instance $x$ based on the known instance $x'$ as an act of *interpretation* and, in turn, an act specific to receiver $R$. This act is what underpins perception, linking it to the probabilistic problem of representation. In Definition 5.2.1, the discrepancy between $P_{R_i}(\mathbf{X})$ and $P_{R_j}(\mathbf{X})$ comes from receivers $R_i$ and $R_j$ eliciting different representations for $\mathbf{X}$. The same holds in Definition 5.2.2, only that the receivers $R_i$ and $R_j$ are eliciting different modes of reasoning causally about $\mathbf{X}$. We use the term *evocation* to refer to this implicit process behind all receivers. Each receiver is equipped with its own pre-conceived notion of what a known representative instance is, representing its individual experience.

Given our focus on perception, we do not study the process of evocation, taking it for granted for all receivers. We do, however, borrow some stylised facts from Kahneman and Miller [159]'s *norm theory*, which theorizes how individuals respond to an event by recruiting and creating alternative scenarios. Under this theory, a receiver $R$ recruits a number of representations about an event. These representations are based on what $R$ views as a normal, with each scenario having a set of elements and each element having a set of features. These representations can be aggregated into a single scaled representation denoting the most common alternative, or *the norm*, according to $R$. In Definition 5.2.1, $R$ works with a single representation of $\mathbf{X}$, $P_R(\mathbf{X})$, denoting what is normal. In Definition 5.2.2, $R$ works with multiple representations of $\mathbf{X}$, $\mathcal{P}_{\mathbf{X}_R}$, as implied by the SCM $\mathcal{M}_R$, denoting what are the normal ways of reasoning.

**Example 5.2.3.** Continuing with Example 5.2.2, the admission officer clearly evokes the reference profile $\mathbf{x}'$ for classifying the deferred applicant with profile $\mathbf{x}$. Such reference profile is independent of the ML model $f$.

### 5.2.3 Assuming Causal Reasoning

Why rely on causality, in particular, SCM to model perception? Our modeling choice is based on three factors. *First*, under the premise that humans use probabilistic reasoning for decision-making under uncertainty [157], the properties of a SCM $\mathcal{M}$ (2.2) offer a structured way to describe the information $\mathbf{X}$ as represented by $P(\mathbf{X})$. A SCM $\mathcal{M}$ allows to disentangle $P(\mathbf{X})$ (5.2) and to reason under hypothetical scenarios about $P(\mathbf{X})$

(5.3), in principle, similar to how humans interpret information and accumulate knowledge [303]. *Second*, SCM are being increasingly used by ML researchers to approximate human-like reasoning [254, 255]. Given this trend, our proposed framework would be compatible with the next wave of ML applications (e.g., Dittadi et al. [86], van Steenkiste et al. [283]). *Third*, SCM, in particular through the DAG $\mathcal{G}$, can be useful tools to engage multiple stakeholders (e.g., Álvarez and Ruggieri [8], Baumann et al. [27], Kusner et al. [177]), forcing to "draw" the assumptions about $\mathbf{X}$ and its data generating model.

We do not, however, view SCM as equivalent to human reasoning and discourage such an interpretation. Based on the above factors, we view a SCM $\mathcal{M}$ as a useful tool for formalizing the reasoning about $\mathbf{X}$ by $R$. We are aware that this is not a view widely held within the fairness community (e.g., Hu and Kohler-Hausmann [149]). Whether causal reasoning is or not the best framework for formalizing perception is a valid point worth exploring in future work.

## 5.3 The Framework

We now formalize the framework by defining how receivers $R_i, R_j \in \mathcal{R}$ can have different causal interpretations of the same information $\mathbf{X}$ provided by a sender $S \in \mathcal{S}$. Each receiver is "equipped" with its own SCM $\mathcal{M}$ describing $P(\mathbf{X})$ (5.2). Hence, the receivers disagree on the poset of possible distributions $\mathcal{P}_{\mathbf{X}}$ (5.3). Under Definition 5.2.2, we define two kinds of causal perception:

- ***Unfaithful causal perception***, when the receivers disagree on the cause-effect pairs. For instance, for $\mathbf{X} = \{X_1, X_2\}$ receiver $R_i$ views $X_1 \to X_2$ while receiver $R_j$ views $X_1 \leftarrow X_2$ as the causal graph.
- ***Inconsistent causal perception***, when the receivers agree on the cause-effect pairs but disagree on the nature of the effects. For instance, for $\mathbf{X} = \{X_1, X_2\}$ both receivers agree on $X_1 \overset{w}{\to} X_2$, which reads as "$X_1$ causes $X_2$ with effect $w$," but $R_i$ views the causal effect $w > 0$ while $R_j$ views it as $w < 0$.

### 5.3.1 Equipping the Receiver

Two processes are central to causal perception: categorization and signification [186]. *Categorization* entails sorting instances (or classes) into categories. *Signification* entails representing the social meanings of the categories describing the instances (or classes) of interest. Using Example 5.1.1 to illustrate the two, describing Linda as female, single, and 31 years old implies a different process from that of imagining Linda based on the combination of these descriptors. We define each process below.

**Definition 5.3.1.** (Categorization) Let $\Theta^R(X) = \{\theta_1^X, \ldots, \theta_n^X\}$ denote a *conceptualization mapping* in the form of a set of $n$ *descriptors* (or labels) of the variable (or class) $X \in \mathbf{X}$ according to the receiver $R \in \mathcal{R}$. We define the *categorization set* as:

$$\vartheta^R = \left\{\Theta^R(X_i)\right\}_{i=1}^{p} \tag{5.5}$$

where $p = |\mathbf{X}|$. Each $R \in \mathcal{R}$ comes with its own categorization set. It is implied that $X$ is a variable in the SCM $\mathcal{M}$ used by $R$. It is possible for $\Theta^R(X_i) = \emptyset$. A descriptor $\theta^X$, as the name suggests, is a label that describes $X$.

Through categorization $R$ incorporates additional information for each variable $X$ in the form of its descriptors. We view categorization as a movement between different levels of abstraction by $R$, capturing the movement between low (or micro) and high (or macro) levels of modeling $X$. The descriptors $\theta^X$ in Definition 5.3.1 are a low-level representation of the high-level representation $X$.

**Example 5.3.1.** In Example 5.2.1, scenario *(ii)*, as discussed by Kleinberg et al. [174], assume that the admissions officer $R_1$ breaks the tie between the two applicants with the same suitability score $G$ from the ML model $f$ using the applicants' SAT scores $X_1$, high-school GPA $X_2$, and address $Z$. Further assume that both applicants have the same $X_1$ and $X_2$. We define the categorization set of applicants' $X_1$, $X_2$, and $Z$ for $R_1$ as:

$$\begin{aligned}\vartheta^{R_1} =&\{\Theta^{R_1}(X_1), \Theta^{R_1}(X_2), \Theta^{R_1}(Z)\}\\ =&\{\{\text{tutoring, expensive, performative}\},\\ &\{\text{discipline, school funding, potential}\},\\ &\{\text{family income, school district}\}\}\end{aligned}$$

where the first line states the variables $R_1$ is categorizing; and the second line states the descriptors for each variable. For later use, let us also define another admissions officer $R_2$ that has a similar categorization set to $R_1$ but with $\Theta^{R_2}(X_1) = \emptyset$, meaning $\vartheta^{R_2} = \{\Theta^{R_1}(X_2), \Theta^{R_1}(Z)\}$.

**Definition 5.3.2.** (Signification) Let $\Phi^R(X_i, X_j) = \phi\big(\Theta^R(X_i) \wedge \Theta^R(X_j)\big)$ denote an *operationalization mapping* in the form of a *causal relational statement* $\phi$ between a pair of variables (or classes) $X_i, X_j \in \mathbf{X}$ and/or their descriptors according to the receiver $R \in \mathcal{R}$. We define the *signification set* as:

$$\varphi^R = \Big\{\big\{\Phi^R(X_i, X_j)\big\}_{j\neq i, j=1}^{p}\Big\}_{i=1}^{p} \tag{5.6}$$

where $p = |\mathbf{X}|$. Each $R \in \mathcal{R}$ comes with its own signification set that is based on its own categorization set. It is implied that $X_i, X_j$ are variables in the SCM $\mathcal{M}$ used by $R$. It is possible that $\Phi^R(X_i, X_j) = \emptyset$.

We view signification as a more complex process as it aims to formalize the reasoning (read, interpretation) of $R$ based on the information provided and the potential additional information from the categorization process. Intuitively, (5.5) captures the process in which $R$ lists the elements that constitute a variable $X \in \mathbf{X}$ while (5.6) captures the process in which $R$ reasons (causally via $\phi$) about two variables $X_i, X_j \in \mathbf{X}$ based on their lists of elements. Now what do we mean by *the causal relational statement* $\phi\big(\Theta^R(X_i) \wedge \Theta^R(X_j)\big)$ (or just $\phi$) in Definition 5.3.2? Such statement and, thus, the signification process itself varies in meaning based on the kind of causal perception. We present signification under unfaithful perception in Section 5.3.2 and under inconsistent perception in Section 5.3.3.

**Definition 5.3.3.** (Receiver Profile) Both categorization and signification equip the receiver with its experience in the form of causal knowledge. When speaking of a $R \in \mathcal{R}$ in the causal perception framework we imply the object:

$$R = \big(\vartheta^R, \varphi^R\big). \tag{5.7}$$


Figure 5.1: Unfaithful causal perception based on Example 5.3.2. LHS is the causal graph for $R_1$; RHS is the causal graph for $R_2$. Both graphs describe $P(Y, X_1, X_2, Z)$.

### 5.3.2 Perception due to Unfaithfulness

Unfaithful perception occurs when receivers cannot agree on the causal graph $\mathcal{G}$ behind $P(\mathbf{X})$. Hence, $R_i$'s causal graph, $\mathcal{G}_{R_i}$, is unfaithful to $R_j$'s causal graph, $\mathcal{G}_{R_j}$, and vice versa. Different graphs imply different factorizations for $P(\mathbf{X})$ (5.2) and, in turn, different posets of implied distributions $\mathcal{P}_{\mathbf{X}}$ (5.3), leading to causal perception. Therefore, under unfaithful causal perception, the *causal relational statement* $\phi$ in Definition 5.3.2 represents a *cause-effect ordering* between the variables $X_i, X_j \in \mathbf{X}$. We denote it with an arrow $\rightarrow$, representing what causes what between $X_i$ and $X_j$. Let us consider Example 5.3.2 below.

**Example 5.3.2.** Continuing with Example 5.3.1, based on the categorization sets $\vartheta^{R_1}$ and $\vartheta^{R_2}$, we define the corresponding signification sets below. Assume (e.g., due to the college's bylaws) that an admissions officer in this situation can directly determine $Y$ using $X_1$ and $X_2$ only. Hence, $X_1 \rightarrow Y$ and $X_2 \rightarrow Y$ are assumed and $Z \rightarrow Y$ is not allowed, meaning these cause-effect pairs are provided to and shared by both $R_1$ and $R_2$. For $R_1$:

$$\begin{aligned}
\varphi^{R_1} =&\{\Phi^{R_1}(Z, X_1), \Phi^{R_1}(Z, X_2), \Phi^{R_1}(Z, Y),\\
&\quad \Phi^{R_1}(X_1, X_2), \Phi^{R_1}(X_1, Y), \Phi^{R_1}(X_2, Y)\}\\
=&\{\phi^{R_1}(\{\text{family income}\} \wedge \{\text{tutoring, expensive}\}),\\
&\quad \phi^{R_1}(\{\text{school district}\} \wedge \{\text{school funding}\}), \emptyset,\\
&\quad \phi^{R_1}(\{\text{performative}\} \wedge \{\text{discipline, knowledge}\}),\\
&\quad\quad X_1 \rightarrow Y, X_2 \rightarrow Y\}\\
=&\{Z \rightarrow X_1, Z \rightarrow X_2, X_2 \rightarrow X_1, X_1 \rightarrow Y, X_2 \rightarrow Y\}
\end{aligned}$$

where the first line states the pair of variables $R_1$ is signifying; the second line states the descriptors of each variable within each pair; and the third line states the cause-effect ordering for each pair based on the combination of these descriptors. For $R_2$ we have a similar signification set $\varphi^{R_2}$ with the exception of $\Phi^{R_2}(Z, X_1) = \emptyset$ as $\Theta^{R_2}(X_1) = \emptyset$. We present the resulting graphs $\mathcal{G}_{R_1}$ and $\mathcal{G}_{R_2}$ in Figure 5.1.

Having different $\mathcal{G}_{R_1}$ and $\mathcal{G}_{R_2}$ conditions how each officer reasons about applicants in Example 5.3.2. The factorization of $P(Y, X_1, X_2, Z)$ is different for each officer:

$$P(Y|X_1, X_2)P(X_2|Z)P(X_1|X_2, Z)P(Z)$$

under $\mathcal{G}_{R_1}$ and

$$P(Y|X_1, X_2)P(X_2|Z)P(X_1|X_2)P(Z)$$

under $\mathcal{G}_{R_2}$, leading to unfaithful perception. If they wanted to reason about SAT scores and zip code via $do(Z := z)$, e.g.: "what would be the average SAT score if all applicants where from neighborhood $z$?", $R_1$ considers $P(X_1|X_2, z)P(X_2|z)P(z)$ while $R_2$ considers $P(X_1|X_2)P(X_2|z)P(z)$. Hence, each officer might arrive at a different tie breaker for the two applicants. Further, we could include the suitability score $G$ from the ML model $f$ into the signification sets. Similar to $X_1$ and $X_2$, though, the score would be another variable of the decision flow in scenario *(ii)* to be signified by the officers.

### 5.3.3 Perception due to Inconsistency

Inconsistent perception occurs when receivers agree on the causal graph $\mathcal{G}$ behind $P(\mathbf{X})$, but disagree on the nature of the causal effects, like an effect's sign or magnitude. Hence, $R_i$ and $R_j$ are faithful to one factorization of $P(\mathbf{X})$ (5.2) under $\mathcal{G}$, yet reason differently about it (i.e., are inconsistent w.r.t. each other) because of the set of structural equations $\mathbf{F}$ in their SCM $\mathcal{M}_{R_i}$ and $\mathcal{M}_{R_j}$. It implies different posets of distributions $\mathcal{P}_{\mathbf{X}}$ (5.3), leading to causal perception.

What does the causal relational statement $\phi$ represent exactly under this kind of causal perception? Consistency requires for a receiver to reach the same causal conclusions regardless of the modeling level. Let us formalize it further. Consider the the low-level SCM $\mathcal{M}_L$ for the random variable $L$ and the high-level SCM $\mathcal{M}_H$ for the random variable $H$ with corresponding interventions sets $\mathcal{I}_L$ and $\mathcal{I}_H$. Formally, consistency requires that:

$$\tau(P(L))^{do(i)} = P(H)^{\omega(do(i))} \quad \forall i \in \mathcal{I}_L \tag{5.8}$$

where $\tau : L \to H$ is an *exact transformation* between $\mathcal{M}_L$ and $\mathcal{M}_H$, meaning it includes a corresponding *order-preserving, surjective mapping* $\omega : \mathcal{I}_L \to \mathcal{I}_H$ such that $\tau(P(L))^{do(i)}$ is the distribution of the variable $\tau(L)$ with $L \sim P(L)^{do(i)}$. The key idea behind (5.8) is that *it requires for low-level interventions, either through $\tau$ or $\omega$, to hold when moving up to the high-level model.* Consistency, thus, preserves causal reasoning between model abstractions. See Rubenstein et al. [241, Def. 3; Thm. 6] and Beckers et al. [28, Def. 3.1] for technical details.

Let us focus on the mapping from low to high level abstractions, meaning $\tau$. Inconsistency between $R_i$ and $R_j$ then implies that each receiver has its own $\tau$ (and $\omega$): both receivers are faithful to one $\mathcal{G}$, but each reasons differently about it when moving between low (i.e., the descriptors of $X$) and high modeling levels (i.e., $X$). For concrete results, we consider the simplest functional form for $\tau$ under this setting: *an average*. Intuitively, *considering the average of all low-level causal forces should be equivalent to considering the high-level causal force.* Under Rubenstein et al. [241, Thm. 11], this formulation holds if we assume a *linear, additive noise* SCM, such that $f_j := f_j(X_{pa(j)}, U_j)$ in (2.2) becomes

$$f_j := \sum_{i=1}^{|pa(j)|} \alpha_i \cdot X_{pa(j)_i} + U_j \tag{5.9}$$

where $\alpha_i$ is the causal weight or coefficient of the $i$-th parent $pa(j)_i$ of $X_j$.

Such SCM can be restrictive for capturing complex reasoning, though it serves our conceptual goal. Additionally, causal consistency theory has been formulated only for this type of SCM [28, 193, 241]. Therefore, under inconsistent causal perception the

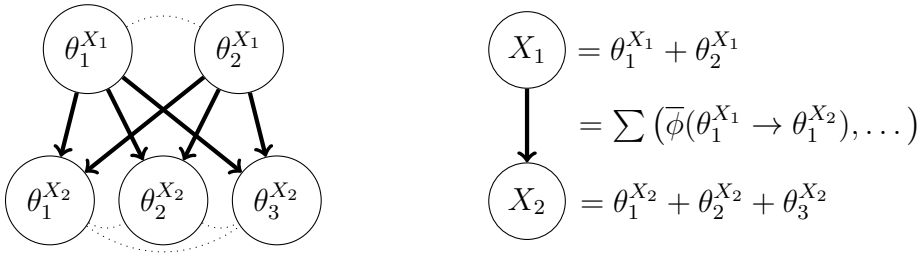



Figure 5.2: An illustration of the low-level to high-level modeling levels under $\tau_R = \sum$. LHS shows the descriptors for $X_1$ and $X_2$, illustrating the low-level representation. The dotted lines represent a simultaneous relationship between the descriptors. RHS shows the variables $X_1$ and $X_2$, illustrating the high-level representation. Both follow the cause-effect $X_1 \overset{w}{\rightarrow} X_2$.

*causal relational statement* $\phi$ in Definition 5.3.2, thus, represents *causal aggregated effect* of an agreed cause-effect pair. We denote by $\tau_R$, which represents the exact transformation used by $R$ to aggregate the causal effects of the low-level representations into a high-level representation.

Figure 5.2 illustrates the role of $\tau_R$. Let us unpack it further before we consider Example 5.3.3. Under (5.8) and (5.9), given a cause-effect pair $X_i \rightarrow X_j$ in $\mathcal{G}$, we allow for receivers to disagree on how $X_i$ causes $X_j$ in terms of $\beta_{i,j}$ where $\beta_{i,j}$ represents the aggregated causal effect of $X_i$ in $X_j$. Formally, each receiver will have its own categorization set for each causal-effect pair $X_i \rightarrow X_j$. This process takes place at the low-level of model abstraction. Each receiver, by having different categorization sets, reasons differently on how the descriptors of $X_i$ associate to those of $X_j$ and will aggregate all of these low-level associations by computing $\beta_{i,j}$ through their own $\tau_R$. This process takes place at the high-level of model abstraction. The simplest aggregation function is a summation, meaning each receiver adds up all the associations between the descriptors of each variable as captured by $\overline{\phi}(.)$ to obtain the causal weight between the variables as captured by $\beta_{i,j}$. Under this formulation, receivers will agree on $X_i \rightarrow X_j$ but disagree on $\beta_{i,j}$. We provide a first formulation of $\tau_R$ in Appendix B.4.

**Example 5.3.3.** Similar to Example 5.3.2, consider scenario *(ii)* in Example 5.2.1. Assume receivers $R_1$ and $R_2$ agree on a causal graph $\mathcal{G}$ describing $P(Y, X_1, X_2, Z)$ that leads to the factorization $P(Y|X_1, X_2)P(X_1|X_2, Z)P(X_2|Z)P(Z)$ (LHS in Figure 5.2). Here, we formalize the scenario where two admissions officers consider the potential of the same applicant differently. Assume that both officers have similar signification sets $\varphi^{R_1}$ and $\varphi^{R_2}$, except for $Z \rightarrow X_1$:

$$\begin{aligned}
\Phi^{R_1}(Z, X_1) &= \phi^{R_1}(\{\text{family income}\} \wedge \\
&\qquad \{\text{practice hrs, pri. tutoring, aptitude}\})) \\
&= \tau_{R_1}\big(\overline{\phi}(\text{fam. income} \rightarrow \text{pri. tutoring})\big) = \beta_1 \\
\Phi^{R_2}(Z, X_1) &= \phi^{R_2}(\{\text{family income}\} \wedge \\
&\qquad \{\text{practice hrs, onl. resources, aptitude}\}) \\
&= \tau_{R_2}\big(\overline{\phi}(\text{fam. income} \rightarrow \text{onl. resources})\big) = \beta_2
\end{aligned}$$

such that $\beta_1 > \beta_2$ and where (for both receivers) the first line states the categorization

of the $Z \rightarrow X_1$ cause-effect pair and the second line states the associations between descriptors with weights $\overline{\phi}(.)$ and the exact transformation $\tau_R(.)$ that aggregates them into the high-level causal effect $\beta$. $\beta_1$ and $\beta_2$ are conceived as illustrated in Figure 5.2.

Both $R_1$ and $R_2$ in Example 5.3.3 see $Z$ as a proxy for socioeconomic background, but draw different associations between it and SAT scores. Both receivers also associate the SAT score to the applicant's aptitude and time spent on practicing for the test, which are qualities specific to the applicant. There is a shared baseline interpretation between receivers. Disagreement occurs when reasoning about how $Z$ causes $X_1$. Intuitively, $R_1$ views it as: the higher the family income, the greater the access to private tutoring, the higher the SAT score. Instead, $R_2$ views it as: family income determines access to online resources, a less divisive view along income brackets as most applicants will have, e.g., a computer. Hence, $R_1$ is less impressed by a high score from a wealthy applicant relative to the same score by a poor applicant, while $R_2$ sees similarly the two applicants' scores. These two distinct causal weights for $Z \rightarrow X_1$ lead $R_1$ and $R_2$ to calculate differently, due the SCM $\mathcal{M}_{R_1}$ and $\mathcal{M}_{R_2}$, the probability $P(Y = 1|x_1, x_2, z)$ for an applicant. As noted in Example 5.3.2, here we could also include $R$ from the ML model $f$ into the signification sets. Similar to $X_1$ and $X_2$, though, it would be another variable of the decision flow in scenario *(ii)* to be signified by the officers.

## 5.4 Relationship of Perception to Fairness

The problem of perception, as motivated in Section 5.1.1 and illustrated through the examples used so far, is relevant for developing fair ML applications. In our view, designing, testing, and auditing fair ML applications requires treating perception as a parameter of interest. In this section, we argue for this relationship. We position perception as a parameter of interest, in particular, within the fair representing learning problem in Section 5.4.1; present the role of sensitive attributes as drivers of perception in Section 5.4.2; and discuss relevant ML applications in Section 5.4.3.

### 5.4.1 Perception-Induced Bias

We avoided using the term bias in Sections 5.2 and 5.3. Indeed, the disagreement in the interpretation of information $\mathbf{X}$ by two receivers $R_i, R_j \in \mathcal{R}$ can lead to biased and, potentially, unfair decision-making as we understand it within the fairness literature. When $R_i$ and $R_j$ disagree, broadly, we can speak of a *perception-induced bias* (PIB). The Linda Problem is a clear example of PIB. However, to account formally for PIB under causal perception and, in turn, link it to fairness, we need to prioritize one receiver over the other. This is because, under Definition 5.2.2, $R_i$ and $R_j$ are unfaithful or inconsistent only relative to each other.

The PIB is specific only to the receivers involved: $R_i$ is biased w.r.t. $R_j$, and vice-versa. Such PIB, we believe, is different from speaking of a PIB, whether it involves a ML application or not, that leads to unfair decision-making. For instance, if $R_i$ and $R_j$ are in equal standing (like the two admissions officers in Examples 5.3.2 and 5.3.3), they can be biased w.r.t. each other but how can we claim that one receiver is correct (read, unbiased) and the other receiver wrong (read, biased)? To link PIB to fairness, it is

important to define a reference for interpretation: i.e., a receiver profile (Definition 5.3.3) that represents a preferred decision maker within the decision flow. It is convenient to view such receiver as a *representative receiver*. This remark applies also to general perception (Definition 5.2.1).

**Example 5.4.1.** In Examples 5.3.2 and 5.3.3 we could define $R_1$ as the representative (fair) receiver given its reasoning behind $X_1$ and $Z$. $R_1$ would represent the desired decision-maker to break potential ties between applicants in terms of the suitability score $G$ of the ML model $f$.

The need to define a reference when addressing fairness under causal perception, we further argue, motivates reconsidering the standard fair ML problem formulation. Since each interpretation is itself a representation of the information $\mathbf{X}$ as captured by $P(\mathbf{X})$ and $\mathcal{P}(\mathbf{X})$, Definition 5.2.2 captures the *causal representation learning problem* from the perspective of two receivers $R_i, R_j$ that interpret the information of a sender $S \in \mathcal{S}$. Formally, $R_i, R_j$ *have learned to represent* the information $\mathbf{X}$ by constructing a representation $P(\mathbf{X})$ that summarizes essential features of $\mathbf{X}$ using probabilities.[4] $P_{R_i}(\mathbf{X})$ and $P_{R_j}(\mathbf{X})$ denote competing representations of $\mathbf{X}$, and $\mathcal{P}_{R_i}(\mathbf{X})$ and $\mathcal{P}_{R_j}(\mathbf{X})$ denote competing sets of representations of $\mathbf{X}$ based on causal reasoning. This problem becomes a fairness one once we fear that the learned representations are influenced by sensitive information, such as gender or race, [88, 317], especially when the receivers are decision makers tasked with classifying $\mathbf{X}$ using $Y = \{0, 1\}$ with the help of a ML model $f$ like in, e.g., deferring systems.

The role of perception in ML-enabled decision-making is largely understudied. Here, we stress how perception might expand the fair ML problem formulation moving forward. Intuitively, if one of the receivers embodies a desired learned representation of $\mathbf{X}$, then any deviation from that receiver by another receiver would represent a form of PIB that leads to unfair decision-making. Implicit to this formulation is that one receiver's interpretation is preferred over the other receiver's interpretation of $\mathbf{X}$. Since the fair causal learning representation literature (e.g., Louizos et al. [184], Madras et al. [187], McNamara et al. [196]) is based on obtaining a single, objective classifier $f$, we would expect for future works to aim at minimizing the risk of perception among the decision makers in a decision flow. The standard fairness goal would be to minimize $\epsilon$ under $d$ for perception (5.1) and $\overline{d}$ for causal perception (5.4), meaning we would want for all receivers to agree on one desired representation of $\mathbf{X}$.

Alternatively, we suggest, the fairness goal could be not to minimize $\epsilon$ but to embrace it, meaning we would address fairness given the disagreement between receivers on the representation of $\mathbf{X}$ [129]. This formulation remains unexplored, and captures realistic scenarios in which multiple decision makers use $f$. Implicit to this formulation is that all receivers have an equal standing. Here, *we would move from modeling a single, objective decision maker to allow for multiple, subjective decision makers.* In both formulations, perception becomes a parameter of interest.

[4] Formally, $P(\mathbf{X})$ represents a learned low-dimensional representation of high-dimensional data $\mathbf{X}$. See, e.g., Wang and Jordan [295] for an introduction to (causal) representation learning.

### 5.4.2 Loaded Attributes

We argue that sensitive attributes, such as gender and race, are more prone than other attributes to induce perception among receivers. These attributes, as discussed in previous works (e.g., Bonilla-Silva [42], Sen and Wasow [261]), are *summaries of historical processes* and, thus, are likely to influence the receiver. For instance, here we are referring to the conceptual difference between describing an individual as female versus feminine [149], both of which are based on the attribute gender. Female refers to a category of gender while feminine refers to a set of behavioral expectations attributed to females: i.e., the phenotype versus the construct. We describe these attributes as *loaded* because they almost surely, be it alone or in combination with other attributes, lead to different interpretations among the receivers as they extrapolate from the phenotype to the construct. That said, loaded attributes are context-specific and the term also applies to other attributes that carry the same meaning among receivers.

If $X$ is a loaded attribute, then it should be easier for a receiver $R$ to evoke its own pre-conceived information about $X$. *We view loaded attributes as attributes that thrive on stereotypes of social categories shared and maintained by the receivers.* A *social category* is the result of classifying people into groups over shared perceived identities [46]. We refer to a social category as a *social construct* when the classification is also used purposely to enforce exclusionary policies [192]. Sensitive attributes are clear examples of social constructs. A *stereotype* refers to the cognitive representation people develop about a particular social category, based around beliefs and expectations about probable behaviors, features and traits, which can translate into implicit or explicit attitudes that materialize into bias [37, 155].

The role of loaded attributes in perception is illustrated in Example 5.1.1, where Tversky and Kahneman [281] admittedly wrote Linda to be *representative* of an active feminist and *unrepresentative* of a bank teller. We attribute the longstanding success of the Linda Problem to the fact that Linda is described as some sort of female socially-held stereotype that resonates with participants. For instance, we believe that there is a significant difference in describing the fictitious profile as "female" versus as "female, single, 31 years, and educated."

**Example 5.4.2.** Recall that in Example 5.2.1, scenario *(ii)*, the admissions officer must choose between applicants one and two with profiles $\mathbf{x}_1$ and $\mathbf{x}_2$, with $\mathbf{x} = \langle x_1, x_2 \rangle$, where $P(Y = 1|\mathbf{x}_1) = P(Y = 1|\mathbf{x}_2)$. The ML model $f$, by constructing, scored equally both candidates. As implied in both Examples 5.3.2 and 5.3.3, suppose the officer considers the address, $Z$, of each applicant for the tiebreaker, inferring that applicant one lives in a wealthy neighborhood while applicant two lives in a poor neighborhood. Here, $Z$ is a loaded attribute as it acts as a proxy for socioeconomic background. By linking $Z$ with $X_1$ and $X_2$ and relying, e.g., on the stereotypes between wealth and SAT scores, the officer favors applicant two, meaning $P(Y = 1|\mathbf{x}_1) < P(Y = 1|\mathbf{x}_2)$.

### 5.4.3 Future Work: Relevant Applications

We consider three areas within fairness that would benefit from the causal perception framework. For a comprehensive discussion, we consider Example 5.4.3, which addresses a fully automated decision flow.

**Example 5.4.3.** Consider scenario *(iii)* in Example 5.2.1. Given the top-$k$ applicants by the ML model $f$, suppose that the admissions officer is asked by the college to account for the chosen applicants by $f$. The officer uses xAI tools, like counterfactual explanations, to understand the model's decisions and explain them to a supervisor. Assume that the supervisor also has access to $f$ and the xAI tools.

This new example, although it addresses a case of automated decision-making (ADM), still illustrates the problem of interpretation as in the previous examples for scenarios *(i)* and *(ii)*. Here, we can conceive the admissions officer as one receiver and the supervisor as another receiver. The natural question to ask then is "what happens if they disagree on th chosen $k$ applicants by $f$?" Again, this question is often overlooked as we tend to treat the human expert as one-dimensional, taking for granted its interaction with the ML application be it ADM or not.

Works like Srivastava et al. [272] and Yaghini et al. [309], through user experiments, have shown that the fairness of an outcome can be judged differently depending on who is judging. Similarly, other works have shown how human cognitive biases can be exacerbated by xAI techniques used by the expert to explain $f$ [33], or how they can affect how the expert evaluates the output (like an applicant ranking) from $f$ [89]. Our point is that, to develop fair decision flows robust to these human-specific biases, including perception, it is necessary to first formulate these mental phenomenons in a way suitable for ML applications. We believe that the proposed causal perception framework provides such a formalization for incorporating perception as a parameter of interest in the following areas.

**Fairness under competing graphs.** The causal graph $\mathcal{G}$ clearly impacts the fairness of a causal problem as it conditions how we reason about it. Binkyte et al. [38], e.g., study how multiple causal discovery algorithms, which are algorithms designed for drawing a causal graph from data, derive at different causal graphs for the same dataset. This chapter further shows how it is possible to obtain different fairness conclusions for the same problem depending on which $\mathcal{G}$ is used. We view these results as evidence for unfaithful causal perception.

Suppose in Example 5.4.3 that we want to evaluate the counterfactual fairness (CF) [177] of $f$. What happens if the officer and her supervisor disagree on $\mathcal{G}$? The standard approach is to assume or discover a single $\mathcal{G}$ to work on for the causal fairness problem, which conditions the fairness conclusions [38]. Russell et al. [246], e.g., explore the robustness of CF under multiple graphs while Kilbertus et al. [168] do so under the threat of hidden confounders. Both of these works aim on having a single $\mathcal{G}$: all robustness claims are relative to that graph. Such approach is fine if we aim for an objective view of the problem, meaning both the officer and her supervisor must agree on the $\mathcal{G}$ considered. Future work involving the causal perception framework, thus, could explore the CF of $f$ under multiple stakeholders by defining CF in terms of competing $\mathcal{P}_R(\mathbf{X})$ (that contains the factual and counterfactual distributions used for estimating CF) given a $\overline{d}$. A similar applies for testing discrimination [8] or synthetic data generation [27] under more than one $\mathcal{G}$ per stakeholder.

**Humans-in-the-loop.** As we have argued throughout the examples, decision flows involving ML applications eventually interact with a human expert. In particular to fair-

ness, designing frameworks that can account for subjective or context-aware fairness [156, 272, 309] is important to ensure that such frameworks are robust to multiple fairness views on the same problem.

Suppose in Example 5.4.3, as studied by Bertrand et al. [33], that the officer and her supervisor interpret differently the counterfactual explanations. How could this setting be avoided or, at least, highlighted as a risk by the xAI method? Similar to nudging in behavioural economics [107], which was inspired by Tversky and Kahneman's work on cognitive biases [276], future work should be able to create xAI tools that preemptively account for a set of potential interpretations and aim to provide explanations that ensure that (most) users align on a desired interpretation. One way could be to run user experiments like Srivastava et al. [272], Yaghini et al. [309] and construct the receiver profiles under the causal perception framework. Alternatively, we could define them ourselves and use the framework to calculate the probability of a user's response given an explanation. We can do this since $\mathcal{P}_R(\mathbf{X})$ contains all possible distributions given the set of interventions. Intuitively, it contains all possible ways of reasoning about a counterfactual explanation as captured by a SCM $\mathcal{M}_R$.

**Modeling social stereotypes.** As shown with inconsistent causal perception, it is possible for receivers to agree on a causal graph but disagree on its internal interpretations. We argue that this occurs by considering more granular levels of information for a given variable, which, in practice, means incorporating additional information that characterizes the variable. Such process is linked to sensitive attributes, which we classify as loaded attributes due to their role in evoking perception. In Example 5.4.3, similar to previous examples, what happens when the officer and her supervisor judge the ML model's chosen candidates by incorporating social stereotypes based on the zip code $Z$?

Hu and Kohler-Hausmann [149] are the first to formulate the causal complexity behind modeling sensitive attributes and propose treating these variables as molecules. Inconsistent causal perception extends their analysis under causal perception and formalizes it using causal consistency [28, 241]. Future work should explore further this link between inconsistent perception and sensitive attributes, in particular, cases where consistency breaks for a given receiver.

## 5.5 Conclusion

In this work we have introduced a framework that formulates how experience shapes the interpretation of information per individual under causal reasoning. Causal perception is useful in cases where multiple interpretations (as well as representations) of information are allowed, and enables to position bias in terms of who is interpreting a problem and how their interpretation differs from others involved. It is, in turn, useful for tackling fairness problems involving human experts interacting with ML applications.

**Implementation.** Moving beyond the conceptual framework, it is possible to implement causal perception under current ML methods. The simplest implementation is to assume the set of receivers, with each receiver already equipped with a categorization and signification set. This implementation is similar to what we have done throughout

the examples. In fact, this implementation in practice comes down to specifying the SCM $\mathcal{M}$ for each receiver.

The hardest implementation is to construct the set of receivers, from human-subject experiments or data, and derive each receiver's categorization and signification sets. Here, on top of the SCM $\mathcal{M}$, we would need to implement the conceptualization and operationalization mappings from, respectively, Definitions 5.3.1 and 5.3.2. For instance, we could define the conceptualization mapping using *ontologies* [115] or *knowledge graphs* [141], among other techniques for representing knowledge. Similarly, we could define the operationalization mapping using *logic argumentation* [36] or *relational learning* [249], among other techniques for argumenting knowledge.

**Limitations.** By being this work's main formalization tool, causality is also its main limitation. Besides our modeling choice, our work is conceptual and, thus, limited by its potential implementation in a modern decision flow. Some aspect of the framework, such as defining the aggregated distance function $\overline{d}$, should be easier to implement than others, such as deriving from data a receiver's categorization and signification set, under the current ML techniques. Given our conceptual objective of formulating causal perception, we leave this for future work.

# Chapter 6

# Data Science Applications under Unrepresentative Data

This chapter is based on the journal paper, M. Lazzari, J. M. Álvarez, and S. Ruggieri. Predicting and explaining employee turnover intention. *Int. J. Data Sci. Anal.*, 14(3):279–292, 2022, and the conference paper, J. M. Álvarez, K. M. Scott, B. Berendt, and S. Ruggieri. Domain adaptive decision trees: Implications for accuracy and fairness. In *FAccT*, pages 423–433. ACM, 2023.

In this chapter, we move our focus from causality for Fair ML to data science applications that benefit from using causality as auxiliary knowledge. Here, in particular, we look at the problem of working with an unrepresentative sample when learning the model $\hat{f}$.[1] This problem is often referred to as *sample selection bias*, and it can be formulated using structural causal models.

We present two popular data science applications under the threat of this problem and our proposed solutions. In Section 6.2, we study the problem of interpreting a black-box model using partial dependence plots (PDP) under the threat of unrepresentative survey data. We introduce the *weighted PDP* (WPDP) to account for this problem. Following Zhao and Hastie [323], we also use causal reasoning explicitly to interpret the plots causally. This work falls under the field of *interpretable AI*, or *xAI*. In Section 6.3, we study the problem of learning a decision tree for a classification task when the training data follows a different distribution from the intended test data. We introduce *domain adaptive decision trees* (DADT) to account for this problem. This work falls under the field of *domain adaptation*, or *transfer learning*.

Both of the proposed solutions rely on *(re)weighting the unrepresentative data*, in particular, the set of predictive attributes $\mathbf{X}$ as they are used for each application. With WPDP, we weight each data instance used for drawing the plots; with DADT, we weight the information gain provided by each data instance when deciding the next split. In the next section, Section 6.1, we formulate causally this (re)weighting step.

[1] The sample, e.g., is *i.i.d.* but it does not follow the distribution of interest.

## 6.1 Unrepresentative Data: A Causal Problem

It is a common problem that the training sample used for learning the model $\hat{f}$ (2.1) is not representative of the model's intended population. The data is often assumed (implicitly or explicitly) to be independently drawn from such population. This situation, however, is rarely the case in practice when we do not have control over the data gathering process nor the data generating model. This problem leads to learning a biased model $\hat{f}$.

Recall from Chapter 2 that, under a representative data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, meaning that the *i.i.d.* assumption holds, the learned model $\hat{f}$ follows the distribution $\mathcal{D}$ with domain $\mathbf{X} \times Y$. When this assumption fails, it helps to distinguish two domains with different distributions. Under unrepresentative data, let the training data represent the *source domain* with distribution $\mathcal{D}_S$ over $\mathbf{X} \times Y$ and let the test data represent the *target domain* with distribution $\mathcal{D}_T$ over $\mathbf{X} \times Y$, such that $\mathcal{D}_S \neq \mathcal{D}_T$. Hence, the learned model $\hat{f}$ follows the distribution $\mathcal{D}_S$ but is used over instances that follow the distribution $\mathcal{D}_T$. We are interested, since $\hat{f}$ is intended to be used on the target domain, in learning a model that follows $\mathcal{D}_T$ instead of $\mathcal{D}_S$.

Indeed, if we had access to a representative data (i.e., one that follows the target distribution $\mathcal{D}_T$), then training an unbiased $\hat{f}$ would require using such data. That is also not the case in practice. The common setting, at best, is having partial (or limited) knowledge, in the form of data, of the target domain. Therefore, solving for sample selection bias reduces to trying to derive an unbiased $\hat{f}$ under the potentially biased training data combined with what is known about the target domain.

We can generalize further this setting by simply stating that the sample used for learning the model $\hat{f}$ follows the distribution $\mathcal{D}'$ (i.e., $\mathcal{D}_S$) instead of following the distribution $\mathcal{D}$ (i.e., $\mathcal{D}_T$). Both notations are equivalent and tackle the sample selection bias problem. We will use one or the other depending on the application. This is because the problem of sample selection bias brings together different fields that have developed their own word choice for formulating the problem. The distinction between source and target domain is specific to the domain adaptation literature (e.g., [189, 318]) while the broader sample selection bias literature simply refers to a biased sample under $\mathcal{D}'$ with respect to a population under $\mathcal{D}$ (e.g., [70, 90, 134, 135, 313]).

Let us formalize this general scenario by considering the *sample selection mechanism* $S$, such that $S = 1$ when a tuple is drawn from $\mathbf{X} \times Y$ into the training sample and $S = 0$ otherwise. It follows that the sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n = \{(\mathbf{x}_i, y_i, s_i = 1)\}_{i=1}^n$. If $S$ ensures random sampling or, overall, is an independent mechanism within the context of interest, then the sample is representative of its population. Recall from Chapter 2 that the learned model $\hat{f}$ (2.1) minimizes the expected loss over the sample. Under a representative sample, the learned model does so over a sample with distribution $\mathcal{D}$ with domain $\mathbf{X} \times Y \times S$; while under a non-representative sample, the learned model does so over a sample with distribution $\mathcal{D}'$ with domain $\mathbf{X} \times Y \times S$ and $\mathcal{D}'$ representing the biased distribution with respect to $\mathcal{D}$.[2] Note that, by definition of the sample selection bias, the support of the biased distribution $\mathcal{D}'$ is included in the support of the true distribution $\mathcal{D}$. This assumption is what allows us to draw information about the population through the sample despite the biased influence of $S$. Similarly, as a specific domain adaptation

[2]Formally, given that $s_i = 1$ for all $i$ instances in the sample: $\mathbb{E}_{\mathbf{X},Y,S\sim\mathcal{D}}\big[\ell\big(f(\mathbf{X}), Y\big)|S = 1\big]$ under a representative sample, and $\mathbb{E}_{\mathbf{X},Y,S\sim\mathcal{D}'}\big[\ell\big(f(\mathbf{X}), Y\big)|S = 1\big]$ under a non-representative sample.

problem, $S$ simply represents the domain of interest.

Now, before moving forward, how is the above scenario a causal problem? We emphasize that, in principle, sample selection bias and the overall problem of unrepresentative data, does not have to be formalized as a causal problem. In fact, the majority of the literature does not do so: consider, e.g., from the earlier works by economists [134, 135] to the recent works by computer scientists [70, 313]. As these works illustrate, correcting for sample selection bias can be carried out without any auxiliary causal knowledge. However, as more recent causal works argue (e.g., [23, 318]), formulating the sample selection bias problem, in particular, using structural causal models (SCM) is helpful for understanding the source of the bias and, in turn, mitigating its effects. We will come back to this point at the end of this section.

It helps to view the sample selection mechanism (or, equivalently, domain mechanism) $S$ *as a random variable that causes what is and is not selected into our sample.* In other words, $S$ causes the change in distributions over the domain $\mathbf{X} \times Y$. Intuitively, when $S = 1$, the sample represented by $P(\mathbf{X}, Y)$ follows the distribution $\mathcal{D}'$ (or, similarly, $\mathcal{D}_S$) and when $S = 0$, the sample represented by $P(\mathbf{X}, Y)$ follows the distribution $\mathcal{D}$ (or, similarly, $\mathcal{D}_T$). As we will discuss below, unrepresentative data can be unrepresentative in different ways depending on how $S$ causes, respectively, $\mathbf{X}$ and $Y$. The Figure 6.1 shows different kinds of unrepresentative data due to different kinds of sample selection bias. Each SCM $\mathcal{M}$ allows us to formalize in a causal but also clear and intuitive way the problem of sample selection bias and, in turn, its remedy.

The training sample is the only source of bias [70]. A common bias correction in machine learning is using a *weighted training sample* $\{\mathbf{x}_i, y_i, w_i\}_{i=1}^n$ under cost-sensitive learning [90, 314]. The non-negative weights $w_i \geq 0$ (de-)emphasize the individual contribution of each observation in the training sample when learning the model $\hat{f}$, accounting for the *cost of an error on a tuple* $t_i = (\mathbf{x}_i, y_i)$.

Here, the error behind each $t_i$ is that of drawing it from the observed but biased distribution $\mathcal{D}'$ instead of the true but unobserved distribution $\mathcal{D}$. We can relate the two distributions via the sample selection mechanism by

$$P_{\mathcal{D}}(t_i|s_i = 1) = P_{\mathcal{D}'}(t_i) \tag{6.1}$$

Assuming that all points $t_i$ in the support of $\mathcal{D}$ can be sampled with a non-zero probability, we can write for all $t$ that

$$P_{\mathcal{D}}(t) = \frac{P(t|s = 1)P(s = 1)}{P(s = 1|t)} = \frac{P(s = 1)}{P(s = 1|t)} P_{\mathcal{D}'}(t) \tag{6.2}$$

where, under a weigh-sensitive algorithm, we re-weight each tuple by

$$w_i = \frac{P(s_i = 1)}{P(s_i = 1|t_i)} \tag{6.3}$$

to correct for the bias selection in the training sample. Given access to $P(s_i = 1)$ and $P(s_i = 1|t_i)$, we can re-weight the training sample such that the expected empirical error is the same as if we were learning the model $\hat{f}$ on the true distribution [70], thus, correction for the bias. Under (6.2), an observation with a higher probability of being sampled regardless of its characteristics ($P(s_i = 1) > P(s_i = 1|t_i)$), which approaches

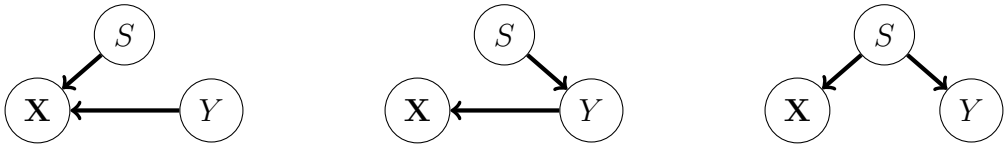



Figure 6.1: A SCM $\mathcal{M}$ for each sample selection (or domain adaptation) scenario [318], where $S$ denotes the sample selection (or domain) mechanism. Left: A causal model for Scenario S1 (or covariate shift). Center: A causal model for scenario S2 (or target shift). Right: A causal model for scenario S3 (or dataset shift).

random sampling, receives a larger weight $w_i$ when learning $\hat{f}$ than an observation with a higher probability of being sampled because of its characteristics ($P(s_i = 1) < P(s_i = 1|t_i)$), which approaches non-random sampling.

Following Zadrozny [313], we distinguish three sample selection bias scenarios for any tuple in the training sample:

- S1: $s$ is independent of $y$ given $\mathbf{x}$, or $P(s|t) = P(s|\mathbf{x})$, meaning the training sample is biased but the biasedness depends only on $\mathbf{x}$ (see, e.g., [90]);
- S2: $s$ is independent of $\mathbf{x}$ given $y$, or $P(s|t) = P(s|y)$, meaning the training sample is biased but the biasedness depends only on $y$ (see, e.g., [313]); and
- S3: there is no independence assumption between $s$, $\mathbf{x}$, and $y$, and we can only address the biasedness of the sample through additional information in the form of $\mathbf{x}_s$, which represents any feature variable that only affects selection into the training sample (see, e.g., [135]).

In this chapter, we address the first scenario, which is common for predictive modeling scenarios (under supervised learning) where we have incoming unlabeled data to be used by the learned model. In practice, this scenario requires that the input $\mathbf{x}$ to the model includes the variables that affect the sample selection. This is a strong assumption that cannot be verified. Under S1, the sample weights $w$ (6.3) simplify to

$$P_{\mathcal{D}}(t) = \frac{P(s=1)}{P(s=1|\mathbf{x})} P_{\mathcal{D}'}(t) \tag{6.4}$$

which means that *we can use unlabeled data drawn from the true distribution $D$ to correct for the selection bias.* Note that we can address the other scenarios as long as we modify the sample weights accordingly.

There are, unsurprisingly, strong similarities between bias in sample selection and in domain adaptation in terms of problem formulation. Domain adaptation [201, 228] tackles problems in which the distribution of the training (or source) and test (or target) data are not the same, leading to a biased model when deployed. Formally, we write it as $P_S(\mathbf{X}, Y) \neq P_T(\mathbf{X}, Y)$ where $P_S$ and $P_T$ denote the source and target probabilities that follow, respectively, $\mathcal{D}_S$ and $\mathcal{D}_T$. Clearly, as previously noted, the sample selection bias, where we end up learning a model on the biased distribution $\mathcal{D}'$ instead of the true distribution $\mathcal{D}$, can be formalized as a domain adaptation problem.

The sample selection mechanism $S$ represents the change in domain, with $S = 1$ denoting the source domain and $S = 0$ the target domain. The domain adaptation

literature distinguishes between three scenarios for explaining the difference between the source and target distributions:

- *covariate shift*, or $P_S(Y|\mathbf{X}) = P_T(Y|\mathbf{X})$ but $P_S(\mathbf{X}) \neq P_T(\mathbf{X})$, meaning the source of bias comes from a shift in the feature space;
- *label shift*, or $P_S(Y|\mathbf{X}) = P_T(Y|\mathbf{X})$ but $P_S(Y) \neq P_T(Y)$, meaning the source of bias comes from a shift in the label space; and
- *dataset shift*, where we do not assume any independence between the spaces and the domains.

where each type of shift aligns with the selection sample bias scenarios previously discussed. Under S1, we find ourselves in a covariate shift scenario in terms of domain adaptation; similarly, label shift coincides with S2, and dataset shift with S3.

The wording used for the three scenario in both sample selection and domain adaption, which relies on what causes the bias, translates naturally into specific SCM. Figure 6.1, which is based on Zhang et al. [318], illustrates how these scenarios can be formalized using causal reasoning. Each causal model is one possible representation of each scenario (we could, e.g., in S1 not draw the cause-effect pair $\mathbf{X} \leftarrow Y$). Hence, each causal model represents structurally the assumptions made between $S$ and its influence on $\mathbf{X}$ and $Y$ and, thus, its influence on the domain $\mathbf{X} \times Y$.

Regarding the sample weights (6.3), viewing them under the prism of domain adaptation, we argue, adds a new meaning to their implementation for bias correction. It expands the focus from just bias correction during model training to also bias correction during model deployment. It also allows to interpret such weights causally, even if just at a conceptual level, given that the weights are motivated by the causal structures motivating the data gathering or data generation process.

## 6.2 Weighted Partial Dependence Plots

This section is based on the journal paper *Predicting and explaining employee turnover intention* by Lazzari et al. [178], which studies *turnover intention*. The term turnover intention refers to an employee's reported willingness to leave her organization within a given time horizon, and it is used to study actual employee turnover. Since employee turnover can have a detrimental impact on business and the labor market at large, it is important to understand its determinants. Using a unique European-wide survey on employee turnover intention, as the title suggests, Lazzari et al. [178] aims for two objectives: first, predicting, and second, explaining employee turnover.

For the first objective, we compare the state-of-the-art of Machine Learning models for predicting employee turnover and find the logistic regression and LightGBM as the top-two performing models. We incorporate country-specific weights, using official EU census data, to account for potential sample selection bias in the survey. For the second objective, which includes the *weighted partial dependence plot* (WPDP), we investigate the importance of the predictive features for these two models by ranking their driving features using a novel cross-validation approach. We then use the WPDP, which incorporates the country-specific weights, to mimic policy interventions regarding employee turnover and, based on Zhao and Hastie [323], use structural causal models (SCM) to interpret the results causally.

Given the focus of Chapter 6, we prioritize the WPDP and its application under auxiliary causal knowledge. See Appendix C for additional material specific to Lazzari et al. [178]. The country-specific correction applied to the models for predicting employee turnover and, consequently, to the WPDP are based on scenario S1 (or the covariate shift) in Section 6.1.

### 6.2.1 Introduction

Employee turnover refers to the situation where an employee leaves an organization. It can be *voluntary*, when it is the employee who decides to terminate the working relationship, or *involuntary*, when it is the employer who decides [144]. Voluntary turnover is divided further into *functional* and *dysfunctional* [112], which refer to, respectively, the exit of low-performing and high-performing workers. Here, we focus on voluntary dysfunctional employee turnover (henceforth, employee turnover) as the departure of a high-performing employee can have a detrimental impact on the organization itself [307] and the labor market at large [144].

It is important for organizations to retain their talented workforce as this brings stability and growth [139]. It is also important for governments to monitor whether organizations are able to do so as changes in employee turnover can be symptomatic of an ailing economic sector.[3] For instance, the European Commission includes it in its annual joint employment report to the European Union (EU) [69]. Understanding why employees leave their jobs is crucial for both employers and policy makers, especially when the goal is to prevent this from happening.

[3]Consider, e.g., the recent wave of workers quitting their jobs during the pandemic due to burn-out. See "Quitting Your Job Never Looked So Fun" [link to NYT article] and "Why The 2021 'Turnover Tsunami' Is Happening And What Business Leaders Can Do To Prepare" [link to Forbes article].

Turnover intention, which is an employee's reported willingness to leave the organization within a defined period of time, is considered the best predictor of actual employee turnover [145]. Although the link between the two has been questioned [67], it is still widely used for studying employee retention as detailed quit data is often unavailable due to, e.g., privacy policies. Moreover, since one precedes the other, the correct prediction of intended turnover enables employers and policy makers alike to intervene and thus prevent actual turnover.

We model employee turnover intention using a set of traditional and state-of-the-art Machine Learning (ML) models and a unique cross-national survey collected by Effectory,[4], which contains individual-level information. The survey includes sets of questions (called *items*) organized by *themes* that link an employee's working environment to her willingness to leave her work. Our objective is to train accurate predictive models, and to extract from the best ones the most important features with a focus on such items and themes. This allows the potential employer and/or policy maker to better understand intended turnover and to identify areas of improvement within the organization to curtail actual employee turnover.

We train three interpretable (k-nearest neighbor, decision trees, and logistic regression) and four black-box (random forests, XGBoost, LightGBM, and TabNet) classifiers. We analyze the main features behind our two best performing models (logistic regression and LightGBM) across multiple folds on the training data for model robustness. We do so by ranking the features using a new procedure that aggregates their model importance across folds. Finally, we go beyond correlation-based techniques for feature importance by using a novel causal approach based on structural causal models and their link to partial dependence plots (PDP). This in turn provides an intuitive visual tool for interpreting our results.

Throughout our ML pipeline, we account for the potential of sample selection bias and weight the survey data using country-specific weights. In doing so, when interpreting the PDP we adjust for the weight of each instance accordingly, leading to the weighted PDP, or WPDP.

**Summary of our contributions.** We highlight the (causal) contributions of Lazzari et al. [178] from two perspectives. First, from a data science perspective:

- we analyze a real-life, European-wide, and detailed survey dataset to test state-of-the-art ML techniques, finding a new top-performing model (LightGBM) for predicting turnover intention;
- we carefully study the importance of predictive features, which have causal policy-making implications; and
- we present the weighted partial dependence plot (WPDP), which is an extension of the standard PDP under the threat of sample selection bias.

Second, from a method-wise perspective:

[4]Effectory is a leading European provider of employee feedback solutions. Visit `https://www.effectory.com` for more information.

- we devise a robust ranking method for aggregating feature importance across many folds during cross-validation; and
- we are the first within the employee turnover literature to use structural causal models for interventional analysis of ML model predictions.

**Related work.** We group the interdisciplinary related work by themes. Given the focus on WPDP, here we highlight the related work relevant to Lazzari et al. [178]'s second objective: that of explaining employee turnover intention. Section C.1 in Appendix C contains the additional related work.

*Turnover determinants.* The study of both actual and intended employee turnover has a long tradition within the fields of human resource management [210] and psychology [145], where research focuses mostly on what factors influence and predict employee turnover [113]. Similarly, a complementary line of research focuses on job embeddedness, or why employees stay within a firm [199, 300]. A number of determinants have been identified for losing, or conversely, retaining employees [267], including demographic ones (such as gender, age, marriage), economic ones (working time, wage, fringe benefits, firm size, carrier development expectations) and psychological ones (carrier commitment, job satisfaction, value attainment, positive mood, emotional exhaustion), among other determinants. Most of this literature has centered on the United States or on just a few European countries. See, for instance, [267] and [274], respectively. Our work is the first to cover almost all of the European countries.

*Determining feature importance.* Beyond predictive performance, we are interested in determining the main features behind employee turnover. We build on the explainable AI (xAI) research [119], in particular xAI for tabular data [247], for extracting from ML models a ranking of the features used for making predictions. ML models can either explain and present in understandable terms the logic of their predictions (white-boxes) or they can be obscure or too complex for human understanding (black-boxes).

The k-nearest neighbor, logistic regression, and decision trees models we use are white-box models. All the other models are black-box models. For the latter group, we use the built-in model-specific methods for feature importance. We, however, add to this line of work in two ways. First, we device our own ranking procedure to aggregate each feature's importance across many fold. Second, following Zhao and Hastie [323] we use structural causal models (SCM) [218] to equip the partial dependence plot (PDP) [102] with causal inference properties. PDP is a common XAI visual tool for feature importance. Under our approach, we are able to test causal claims around drivers of turnover intention. Further, our work extends Friedman [102] by considering a simple weighted version of the PDP to account for potential sample selection bias (Section 6.1).

We highlight Loftus et al. [182] that extends the Friedman [102]'s PDP and Zhao and Hastie [323]'s causal interpretation of the PDP under counterfactual reasoning. As we will discuss later, Zhao and Hastie [323] use only a causal graph to guide and, in turn, equip the PDP with causal meaning. Loftus et al. [182] instead use the structural equations corresponding to the causal graph to generate a counterfactual distribution for the analysis performed by the PDP. Under this approach, the plots also show the downstream effects of the feature of interest on all other features, which are usually kept constant, revealing non-linear relationships. Our work precedes Loftus et al. [182]. Future work could extended that paper's treatment of the PDP under our WPDP.

| Attribute | Type | Attribute | Type |
|---|---|---|---|
| Age | ordinal | Industry | nominal |
| Country | nominal | Job function | nominal |
| Continent | nominal | Time in company | ordinal |
| Education level | ordinal | Type of business | binary |
| Gender | binary | Work status | binary |

Table 6.1: Contextual information and data type in the GEEI Survey.

*Causal analysis.* This is not the first work to approach employee turnover from a causality perspective, but, to the best of our knowledge, it is the first to do so using SCM. Other papers such as Goodman et al. [110] and Price [227] use causal graphs as conceptual tools to illustrate their views on the features behind employee turnover. However, these papers do not equip their causal models with any interventional properties. Allen and Shanock [6], Firth et al. [98], Wunder et al. [306], e.g., go further than other works by testing the consistency of their conceptual models with data using path analysis techniques. Still, none of these three papers use SCM, meaning that they cannot reason about or care for causal interventions.

### 6.2.2 The GEEEI Survey

We use data derived from Effectory's 2018 Global Employee Engagement Index (GEEI) survey. The GEEI is a labor market questionnaire that covered a sample of 18,322 employees from 56 countries. It is is composed of 123 questions that inquire *contextual information*, items related to a number of HR *themes* (also called, constructs), and a target question. The target question (or *target variable*, i.e., the one to be predicted) is the intention of the respondent to leave the organization within the next three months. It takes values *leave* (positive) and *stay* (negative). Contextual information is reported in Table 6.1, together with type of data encoded – binary for two-valued domains (male/female gender, profit/non-profit type of business, full/part time work status), nominal for multi-valued domains (e.g., country name), and ordinal for ranges of numeric values (e.g., age band) or for ordered values (e.g., primary/secondary/higher education level).

The design and validation of the GEEI questionnaire followed the approach of [82]. After reviewing the social science literature, the designers defined the relevant themes, and items for each theme. Then they ran a pilot study in order to validate psychometric properties of questions to assess their internal consistency, and to test convergent and discriminant validity of questions.[5]

Items refer to questions related to a theme. Consider, e.g., the items for the *Trust* theme shown in Table 6.2. There are 112 items in total. Each item belongs to one and only one theme. Each item admits answers in Likert scale. A score from 0 to 10 is assigned to an answer by a respondent as follows: *strongly agree* equals 10; *agree* equals 7.5; *neither agree nor disagree* equals 5; *disagree* equals 2.5; and *strongly disagree* equals 0. The direction of the response scale is uniform throughout all the items [250]. Table 6.3

[5]Two items belonging to a same theme are highly correlated (convergence), whilst two items from different themes are almost uncorrelated (discrimination). For more information on construct validity, see `https://en.wikipedia.org/wiki/Construct_validity`.

| | |
|---|---|
| *Trust* | I have confidence in my colleagues |
| | I have confidence in my organisation's management |
| | I have confidence in the future of my organisation |
| | I have confidence my manager |
| | My colleagues stick to agreements |
| | My organisation trusts that I do my job in the best way possible |

Table 6.2: The items of the *Trust* theme in the GEEI Survey.

| | |
|---|---|
| Adaptability | Motivation |
| Alignment | Productivity |
| Attendance Stability | Psychological Safety |
| Autonomy | Retention factor |
| Commitment | Role Clarity |
| Customer Orientation | Satisfaction |
| Effectiveness | Social Safety |
| Efficiency | Sustainable Employability |
| Employership | Trust |
| Engagement | Vitality |
| Leadership | Work climate |
| Loyalty | |

Table 6.3: Themes in the GEEI Survey.

shows the list of all 23 themes. For a respondent, a score from 0 to 10 is also assigned to a theme as the average score of the items of the theme.

From the raw data of the GEEI survey, we construct two tabular datasets, both including the contextual information. The dataset with the scores of the themes is called the *themes dataset*. The dataset with the scores of the items is called the *items dataset*.

The datasets are restricted to respondents from 30 countries in Europe. The GEEI survey includes 303 to 323 respondents per country, with the exception of Germany which has 1342 respondents. We sampled 323 German respondents stratifying by the target variable. Therefore, the datasets have an approximately uniform distribution per country. Also, gender is uniformly distributed with 50.9% of males and 49.1% of females. See, e.g., Figure 6.2 (left) for further details. These forms of selection bias, however, do not take into account the (working) population size of countries. Caution will be mandatory when making conclusions about inferences on those datasets.

In summary, the two datasets have a total of 9,296 rows each, one row per respondent. Only 51 values are missing (out of a total of 1.1M cell values), and they have been replaced by the mode of the column they appear in. The positive rate is 22.5% on average, but it differs considerably across countries, as shown in Figure 6.2 (right). In particular, it ranges from 12% of Luxembourg up to 30.6% of Finland.

### 6.2.3 From predicting to explaining

Here, we first learn the best model $\hat{f}$ (2.1) by minimizing the empirical risk (recall Chapter 2) for predicting employee turnover, and then extrapolate from these learned models

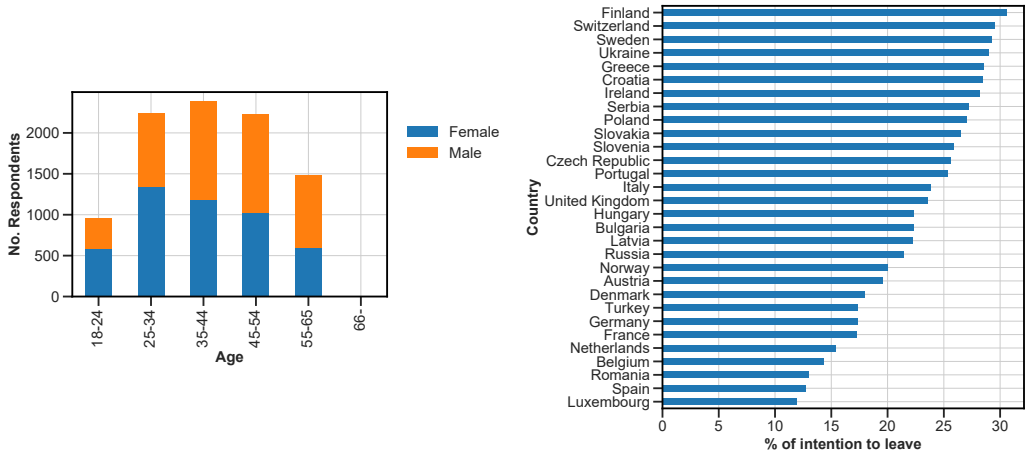



Figure 6.2: Two plots describing the composition of the GEEI survey. Left: Distribution of respondents by Age and Gender. Right: Target variable by Country.

the main drivers of employee turnover. From this section we obtain the top models, which are the logistic regression (LR) and LightGBM (LGBM), to consider for the WPDP and its causal analysis in the next section. Regarding the results, we include most tables and figures in Section C.2. This is because, from this section, we are mainly interested in finding the classifier to focus on the next section.

**Predictive modeling.** The objective is to compare the predictive performances of state-of-the-art ML classifiers on both items and themes the datasets, which, as observed, are quite imbalanced [48]. We experiment with interpretable classifiers, namely k-nearest neighbors (KNN), decision trees (DT), and ridge logistic regression (LR), as well as with black-box classifiers, namely random forests (RF), XGBoost (XGB), LightGBM (LGBM), and TabNet (TABNET). We use the *scikit-learn*[6] implementation of LR, DT, and RF, and the *xgboost*[7], *lightgbm*[8], and *pytorch-tabnet*[9] Python packages of XGB, LGBM, and TABNET. Parameters are left as default except for the ones set by the hyper-parameter search.

We adopt a repeated stratified 10-fold cross validation as testing procedure to estimate the performance of classifiers. Cross-validation is a nearly unbiased estimator of the generalization error [175], yet highly variable for small datasets. Kohavi [175] recommends to adopt a stratified version of it. Variability of the estimator is accounted for by adopting repetitions [170]. Cross-validation is repeated 10 times. At each repetition, the available dataset is split into 10 folds, using stratified random sampling. An evaluation metric is calculated on each fold for the classifier built on the remaining 9 folds used as training set. The performance of the classifier is then estimated as the average evaluation metric over the 100 classification models: 10 models times 10 repetitions.

An hyper-parameter search is performed on each training set by means of the *Optuna*[10] library [4] through a maximum of 50 trials of hyper-parameter settings. Each

[6] https://scikit-learn.org/
[7] https://xgboost.readthedocs.io/
[8] https://lightgbm.readthedocs.io/
[9] https://github.com/dreamquark-ai/tabnet
[10] https://optuna.org/

trial is a further 3-fold cross-validation of the training set to evaluate a given setting of hyper-parameters. The following hyper-parameters are searched for: (LR) the inverse of regularization strength; (DT) the maximum tree depth; (RF) the number of trees and their maximum depth; (XGBoost) the number of trees, number of leaves in trees, the stopping parameter of minimum child instances, and the re-balancing of class weights; (LightGBM) minimum child instances, L1 and L2 regularization coefficients, number of leaves in trees, feature fraction for each tree, data (bagging) fraction, and frequency of bagging; (TABNET) the number of shared Gated Linear Units.

As the evaluation metric, we consider the Area Under the Precision-Recall Curve (AUC-PR) [166], which is more informative than the Area Under the Curve of the Receiver operating characteristic (AUC-ROC) on imbalanced datasets [78, 252]. A random classifier achieves an AUC-PR of $0.225$ (positive rate), which is then the reference baseline. A point estimate of the AUC-PR is the mean average precision over the 100 folds [47]. Confidence intervals are calculated using a normal approximation over the 100 folds [83]. We refer to Boyd et al. [47] for details and for a comparison with alternative confidence interval methods.

For the predictive modeling, we first consider the *themes dataset*, and then repeat the procedure for the *items dataset*. For each datasets, we evaluate a *unweighted* (the original) and a *weighted* version using the country-specific weights based on the workforce size of each country. We run a total of four predictive modeling pipelines. For all four, the nominal contextual features from Table 6.1, namely Country, Industry, and Job Function, are one-hot encoded.

For the weighted datasets, overall, we wanted to answer *how the performance would change if the datasets were weighted to reflect the workforce of each country?* We collected the employment figures for all the countries in our training dataset for 2018. The country-specific employment data was obtained from Eurostat (for the EU member states as well as for the United Kingdom) and from the World Bank (for Russia and Ukraine).[11] The numbers correspond to the country's total employed population between the ages of 15 and 74. For Russia and Ukraine, however, the number corresponds to the total employed population at any age. *We assigned a weight to each instance in our datasets proportional to the workforce in the country of the employee.* We assigned a weight to each instance in our datasets proportional to the workforce in the country of the employee. *Weights are considered both in training of classifiers and in the evaluation metric.*

Let us first concentrate on the case of the themes dataset. As a feature selection preprocessing step, we run a logistic regression for each theme, with the theme as the only predictive feature. Figure 6.3 (left) reports the achieved AUC-PRs (mean $\pm$ standard deviation over the $10\times10$ cross-validation folds). We found that the top three themes, which are Retention Factor, Loyalty, and Commitment, include among their items a question close or exactly the same as the target question. For this reason, we removed these themes (and their items, for the item dataset) from the set of predictive features.

The performances of the classifiers are shown in the Table C.1 for the unweighted (top) and the weighted (bottom) themes dataset in Appendix C. The table includes the AUC-PR (mean $\pm$ standard deviation), the 95% confidence interval of the AUC-PR, and the elapsed time (mean $\pm$ standard deviation), including hyper-parameter search, over the $10\times10$ cross-validation folds. AUC-PRs for all classifiers are considerably better than

[11] For the Eurostat data, visit this link. For the World Bank data, visit this link.

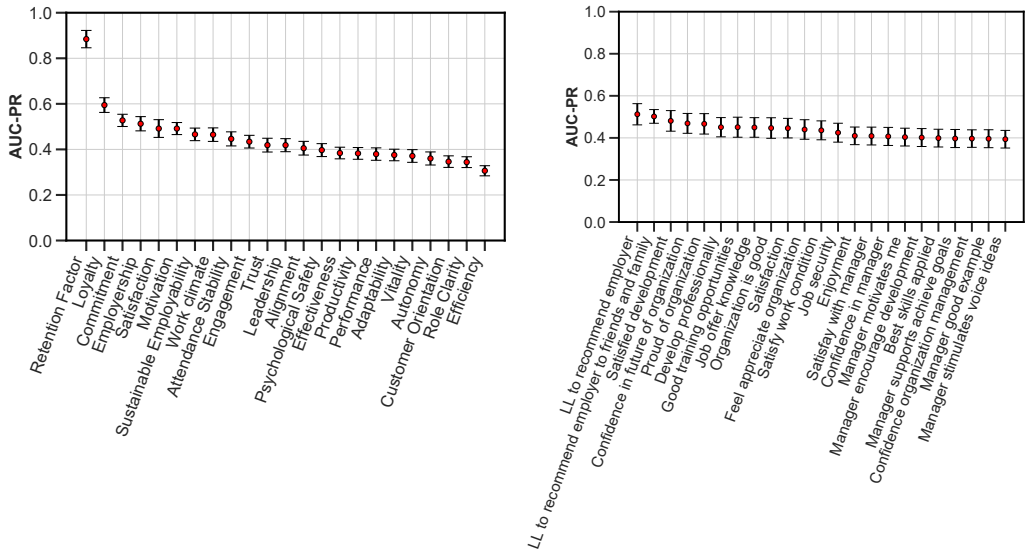



Figure 6.3: Pre-processing step for themes and items dataset. Bars show mean $\pm$ standard deviation over $10\times10$ cross-validation folds. Left: AUC-PR of logistic regression based on a single theme. Right: AUC-PR of logistic regression based on a single item.

the baseline (more than twice the baseline even for the lower limit of the confidence interval). Refer to Table C.1 for details.

For the unweighted themes dataset, Table C.1 (top), DT is the fastest classifier[12], but, together with KNN, also the one with lowest predictive performance. LGBM has the best AUC-PR values and an acceptable elapsed time. LR is the runner up, but it is almost as fast as DT. RF has a performance close to LGBM and LR but it is slower. XGB is in the middle as per AUC-PR and elapsed time. Finally, TABNET has intermediate performances, but it is two orders of magnitude slower than its competitors.

The statistical significance of the difference of mean performances between classifiers is assessed with two-way ANOVA if values are normally distributed (i.e., the Shapiro test) and homoscedastic (i.e., the Bartlett test). Otherwise, the non-parametric Friedman test is adopted [81, 143]. For the theme dataset, ANOVA was used. The test shows a statistically significant difference among the mean values (family-wise significance level $\alpha = 0.001$). The post-hoc Tukey HSD test shows a no significant difference between LGBM and LR. All other differences are significant.

For the weighted themes dataset, Table C.1 (bottom), we observe similar results to its unweighted counterpart. The mean AUC-PR is now smaller for most classifier, the same for LGBM, and slightly better for RF. Standard deviation has increased in all cases. The post-hoc Tukey HSD test shows a small significant difference between LGBM and LR.

Let us now consider the items dataset. Figure 6.3 (right) shows the predictive performances of single-feature logistic regressions. The Table C.2 in Appendix C reports the performances of the classifiers on all features for both the unweighted (top) and the weighted (bottom) data.

Overall, performances of each classifier improve over the themes dataset. Elapsed

[12]Notice that the implementations of DT and LR are single-threaded, while the ones of RF, XGB, LGBM, and TABNET are multi-threaded.

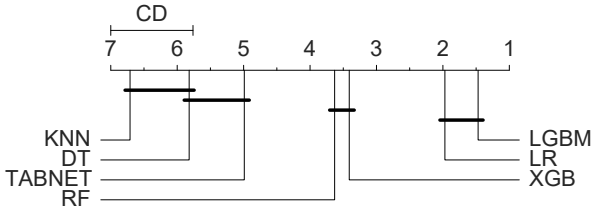



Figure 6.4: Unweighted items dataset: Critical Difference (CD) diagram for the post hoc Nemenyi test at 99.9% confidence level [81].

times also increase due to the larger dimensionality of the dataset. Differences are statistically significant. LGBM and LR are the best classifiers for both the unweighted and the weighted datasets. The Figure 6.4 shows the critical difference diagram for the post-hoc Nemenyi test for the unweighted dataset following a significant Friedman test. An horizontal line that crosses two or more classifier lines means that the mean performances of those features are not statistically different. Refer to Table C.2 for details.

To summarize, we conclude that the LR and LGBM classifiers have the highest predictive power and are, thus, the best performing models for predicting turnover intention based on all four pipelines examined: unweighted/weighted $\times$ items/themes datasets. We focus only on these two classifiers moving forward.

**Explanatory factors.** We examine the driving features behind the LGBM and LR classifiers. We use each model's specific method for determining *feature importance* and aggregate the feature importance into rankings over the 100 experimental folds. This novel approach yields more robust estimates (a.k.a., lower variance) of importance ranks than using a single hold-out set. We note that *we do so for the weighted version of both the themes and items datasets.*

For a fixed fold, the feature importance of the LR model is determined as the absolute value of a feature's coefficient in the model. The importance of a feature in the LGBM model is measured as the number of times the feature is used in a split of a tree in the model. We aggregate feature importance using their ranks, as in non-parametric statistical tests [143]. For instance, LR absolute coefficients $(|\beta_1|, |\beta_2|, |\beta_3|, |\beta_4|) = (1, 2, 3, 0.5)$ lead to the ranking $(3, 2, 1, 4)$.

The top-10 features w.r.t. the mean rank over the 100 folds are shown in Figure 6.5 to Figure 6.8 for the themes/items datasets and LR/LGBM models. For the themes dataset (respectively, the items dataset), LR and LGBM share almost the same set of top features with slight differences in the mean ranks. For instance, the *Sustainable Employability*, *Employership*, and *Attendance Stability* themes are all within the top-five features for both LR and LGBM. For the items dataset, we observe *Time in Company*, *Satisfied Development*, and *Likelihood to Recommend Employer to Friends and Family* to be among the top-five shared features. Interestingly, *Gender*, a well-recognized determinant of turnover intention, is not among the top features for both datasets. Also, no country-specific effect emerges.

The Friedman test shows significant differences among the importance measures in all four cases in Figure 6.5 to Figure 6.8. Further, the figures show the critical difference

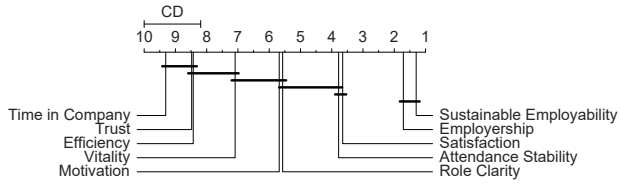



Figure 6.5: Weighted theme dataset: CD diagram for the post-hoc Nemenyi test at 99.9% confidence level for the top-10 LR feature importances.

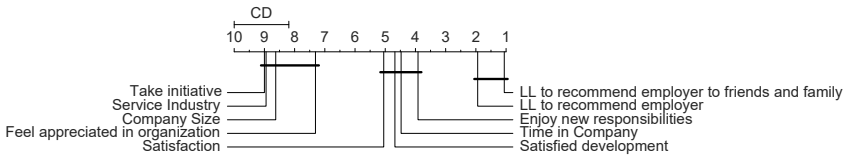



Figure 6.6: Weighted item dataset: CD diagram for the post hoc Nemenyi test at 99.9% confidence level for the top-10 LR feature importances.

diagrams for the post-hoc Nemenyi test, therefore answering the question whether there is any statistical difference among them. An horizontal line that crosses two or more feature lines means that the mean importances of those features are not statistically different. In Figure 6.7, e.g., the *Motivation*, *Vitality*, and *Attendance Stability* themes are grouped together.

Statistical significance of different feature importance is valuable information when drawing potential policy recommendations as we are able to prioritize policy interventions. For instance, given these results, a company interested in employee retention could focus on improving either motivation or vitality, as they strongly influence LGBM predictions and, *a fortiori*, turnover intention. However, the magnitude and direction of the influence is not accounted for in the feature importance plots of Figure 6.5 to Figure 6.8. This is not actually a limitation of our (non-parametric) approach. Any association measure between features and predictions (such as the coefficients in regression models) does not allow for causal conclusions.

### 6.2.4 Causal Analysis through the WPDP

Now we want to assess whether a specific theme $T$ has a causal effect on the target variable, written $T \rightarrow Y$, given the trained model $\hat{f}$ (our options being either the LR or the LGBM) and the contextual attributes in Table 6.1. We use $T^*$ to denote the set of remaining themes and $\tau$ to denote the set of all themes, such that $\tau = \{T\} \cup T^*$. Establishing evidence for a direct causal link between $T$ and $Y$ would allow our model $\hat{f}$ to answer intervention-policy questions related to the theme scores. Given our focus on $T$, here we work only with the theme dataset.

**Auxiliary causal knowledge.** We divide all contextual attributes into three distinct groups based on their level of specificity: individual-specific attributes, $I$, where we

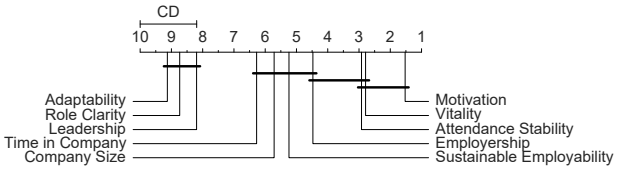



Figure 6.7: Weighted theme dataset: CD diagram for the post-hoc Nemenyi test at 99.9% confidence level for the top-10 LGBM feature importances.

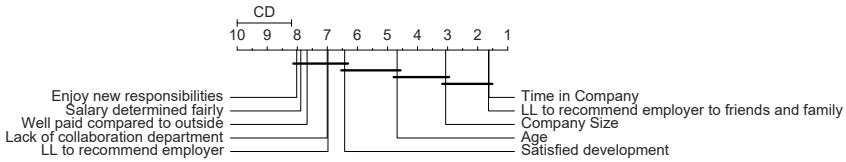



Figure 6.8: Weighted item dataset: CD diagram for the post-hoc Nemenyi test at 99.9% confidence level for the top-10 LGBM feature importances.

include attributes such as *Age* and *Gender*; work-specific attributes, $W$, where we include attributes such as *Work Status* and *Industry*; and geography-specific attributes, $G$, where we include the attribute *Country*.[13] We summarize the causal relationships across the contextual attributes, a given theme's score $T$, the remaining themes $T^*$, and the target variable $Y$ using the causal graph $\mathcal{G}$ in Figure 6.9 (left). The nodes on the graph represent groupings of random variables, while the edges represent causal claims across the variable groupings.

Within each of these contextual nodes, we picture the corresponding variables as their own individual nodes independent from each other but with the same causal effects with respect to the other groupings. For instance, under the causal graph $\mathcal{G}$, $I \to W$ implies the causal relationships $Age \to Industry$, $Gender \to Industry$, $Age \to Work\ Status$, $Gender \to Work\ Status$, but not $Age \to Gender$ nor $Gender \to Age$.

Note that in Figure 6.9 (left) two edges go from $\tau$ to $Y$. This is because we have defined $\tau = \{T\} \cup T^*$, and are interested in identifying the edge between $T$ and $Y$ (in red), while controlling for the edges from $T^*$ to $Y$ (in black as the rest). Our objective becomes clearer in Figure 6.9 (right) where we detail the internal structure of $\tau$.

In Figure 6.9 (right) we assume independence between whatever theme is chosen as $T$ and the remaining themes in $T^*$. This is a strong assumption, but the alternative would be to drop all themes except $T$ and fit $\hat{f}$ on that subset of the data, which would have considerable risks of overestimating the effect of $T$ on $Y$. Further, $T^*$ represents the grouping of all themes in $\tau$ but $T$ where each theme is its own node and independent of each other while have the same inward and outward causal effects. To use the proper causal terminology, all themes have the same parents (the incoming edges from the variables in $I$, $G$, and $W$) and the same child ($Y$). No given theme is the parent or child of any other theme in $\tau$.

[13]Given that we focus only on European countries, the attribute *Continent* is fixed and thus controlled for. We can exclude it from $G$.

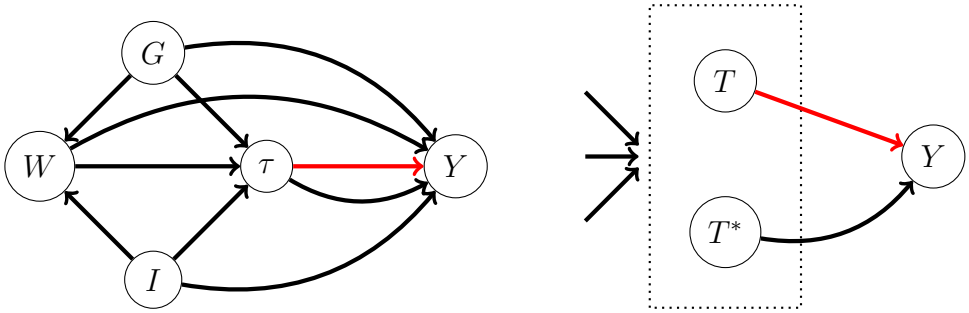



Figure 6.9: The auxiliary causal knowledge (in the form of a SCM) for the themes dataset. Left: Causal graph $\mathcal{G}$ showing three groups of contextual attributes (individual $I$, geographic $G$, and working $W$), the collection of themes ($\tau$), and the target variable $Y$. We are interested in the edge going from $\tau$ to $Y$, highlighted in red. Right: A more detailed look into $\tau$ (dashed-black rectangle) where we can see the distinct edges going from $T$ and $T^*$ into $Y$. For illustrative purposes, we ignore the nodes $W$, $G$, and $I$ and their edges going into $Y$; we do show their edges going into $\tau$.

Under $\mathcal{G}$, all three contextual attribute groups act as confounders between $T$ and $Y$ and, thus, need to be controlled for (along with $T^*$) to identify the causal effect of $T$ on $Y$. Otherwise, e.g., observing a change in $Y$ cannot be attributed to changes in $T$ as $G$ (or, similarly, $I$ or $W$) could have influenced both simultaneously, resulting in an observed association that is not rooted on a causal relationship. Therefore, controlling for $G$, as for the rest of the contextual attributes insures the identification of $T \rightarrow Y$. This is formalized by the back-door adjustment formula [218], where $X_C = I \cup W \cup G \cup T^*$ is the set for all contextual attributes:

$$P(Y|do(T := t)) = \sum_{x_C} P(Y|T = t, X_C = x_C)P(X_C = x_C) \tag{6.5}$$

where the term $P(X_C = x_C)$ is shorthand for $P(I = i, W = w, G = g, T^* = t^*)$. The set $X_C$ satisfies the *back-door criterion* as none of its nodes are descendants of $T$ and it blocks all back-door paths between $T$ and $Y$ [218].

Given $X_C$, under the back-door criterion, the direct causal effect $T \rightarrow Y$ is identifiable. Further, (6.5) represents the joint distribution of the nodes in Figure 6.9 (left) after a $t$ intervention on $T$, which is illustrated by the $do$-operator. If $T$ has a causal effect on $Y$, then the original distribution $P(Y)$ and the new distribution $P(Y|do(T := t))$ should differ over different values of $t$. The goal of such interventions is to mimic what would happen to the system if we were to intervene it in practice. Consider, e.g., a European-wide initiative to improve confidence among colleagues, such as providing subsidies to team-building courses at companies. Then the objective of this action would be to improve the *Trust* theme's score to a level $t$ with the hopes of affecting $Y$.

The causal structure of Figure 6.9 (left) is motivated both from the data and expert knowledge. We argue that $I$, $W$, and $G$ are potential confounders of $T$ and $Y$. Let us consider, e.g., the *Country* attribute, which belongs to $G$. It is sensible to picture that *Country* affects $T$ as employees from different cultures can have different views on the same theme. Similarly, *Country* can affect $Y$ as different countries have different labor laws that could make some labor markets more dynamic (reflected in the form of higher turnover rates) than others. We also observe this in the data. The *Country* attribute is

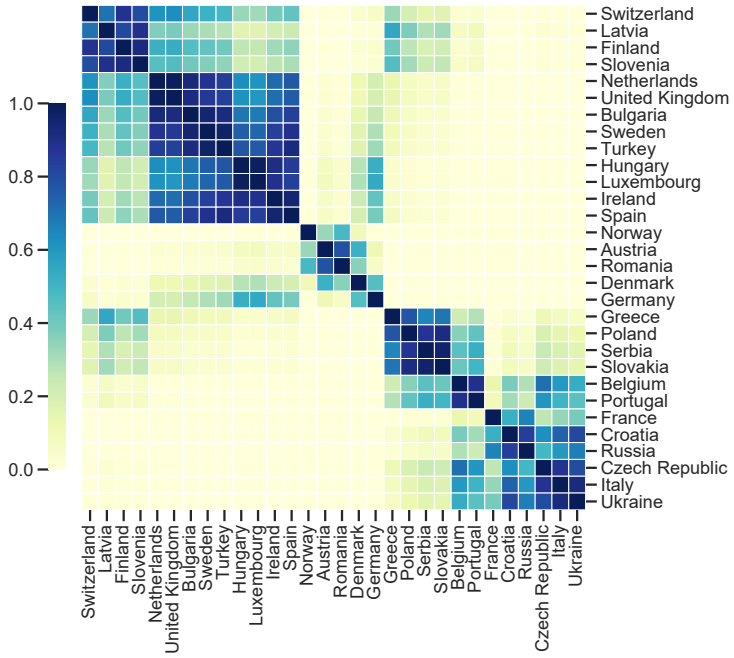



Figure 6.10: The pairwise Conover–Iman post-hoc test p-value for *Trust* theme versus *Country* in a clustered map. The map clusters together countries whose score distributions are similar.

correlated to each of the themes: the non-parametric Kruskal-Wallis H test [143] shows a p-value close to 0 for all themes, which means that we reject the null hypothesis that the scores of a theme in all countries originate from the same distribution.

Let us consider, e.g., the *Trust* theme. To understand which pair of countries have similar/different *Trust* score distributions, we run the Conover–Iman post hoc test pairwise. The p-values are shown in the clustered map of Figure 6.10. The groups of countries highlighted by dark colors (e.g., Switzerland, Latvia, Finland, Slovenia) are similar among them in the distribution of *Trust* scores, and dissimilar from the countries not in the group.[14] Such clustering shows that the societal environment of a specific country has some effect on the respondents' scores of the *Trust* theme. Similar conclusions hold for all other themes.

Further, both $G$ and $I$ have a direct effect also on $W$. We argue that country-specific traits, from location to internal politics, affect the type of industries that developed nationally. Countries with limited natural resources, e.g., will prioritize non-commodity-intensive industries. Similarly, individual-specific attributes will determine the type of work that an individual performs. Individuals with higher education, e.g., where education is among the attributes in $I$, can apply to a wider range of industries than an individual with lower levels of educational attainment.

**A causal (weighted) PDP.** Given the causal knowledge behind the themes dataset in Figure 6.9 (left), we are missing a procedure for estimating (6.5) over our sample and

[14]The clustered map adopts a hierarchical clustering. Therefore, groups can be identified at different levels of granularity.

given our model $\hat{f}$ to test our causal claim. We follow the procedure by Zhao and Hastie [323] and use the partial dependence plot (PDP) tools by Friedman [102] to test visually the causal claim.

The PDP is a model-agnostic xAI method that shows the marginal effect one feature has on the predicted outcomes generated by the model, often used for interpreting black-box models. If changing the former leads to changes in the latter, then we have evidence of a partial dependency between the feature of interest and the outcome variable that is manifested through the model output.[15] We define the partial dependence of feature $T$ on the outcome variable $Y$ given the model $\hat{f}$ and the complementary set $X_C$ as:

$$\begin{aligned} b_T(t) &= E[\hat{f}(T = t, X_C)] \\ &= \sum_{x_C} \hat{f}(T = t | X_C = x_C) P(X_C = x_C) \end{aligned} \tag{6.6}$$

If there exist a partial dependence between $T$ and $Y$, then $b_T(t)$ should vary over different values of $T$, which could be visually inspected by plotting the values via the PDP. If $X_C$ satisfies the back-door criterion, Zhao and Hastie [323] argue that then (6.6) is equivalent to (6.5), [16] and we can use the PDP to check visually our causal claim. Under this setting, the PDP would have a stronger claim than partial dependence between $T$ and $Y$, as it would also allow for causal claims of the sort $T \rightarrow Y$. Therefore, we could assess the claim $T \rightarrow Y$ by estimating (6.6) over our sample of $n$ respondents using:

$$\hat{b}_T(t) = \frac{1}{n} \sum_{j=1}^{n} \hat{f}(T = t, X_C = x_C^{(j)}) \tag{6.7}$$

where we can visually assess the causal effect of $T$ on $Y$ by plotting $\hat{b}_T$ against values of $T$. As Zhao and Hastie [323] argue, if $\hat{b}_T$ varies across the values of $t$, meaning $\hat{b}_T$ is indeed a function of $t$, then we have evidence for causal claim $T \rightarrow Y$.

Under (6.7), however, we assume representative data. We already address the potential sample selection bias when training the classifiers by weighting the datasets using country-specific weights representing their respective 2018 labor force. The standard estimation of the PDP assumes that that any $j$ element in $X_C^{(j)}$ is equiprobable.[17] This is often assumed because we expect random sampling when creating our dataset. The probability, e.g., of sampling a German worker and a Belgian worker would be same. This is a very strong assumption (and one that is hard to prove or disprove in practice), which can become an issue if we were to deploy the trained model $\hat{f}$ as it may suffer from selection bias and could hinder the policy maker's decisions.

To account for this potential issue, one approach is to estimate $P(X_C = x_c)$ from other data sources such as official statistics. Ideally, we would estimate it across the

[15] This under the assumption that the model that is generating the predicted outcomes approximates the "true" relationship between the feature of interest and the outcome variable. This is why Zhao and Hastie [323] emphasize the importance of having a good performing model for applying this approach.

[16] To be more precise, (6.6) is equivalent to the expectation over (6.5), which would allow us to rewrite (6.5) in terms of expectations rather than in terms of probabilities and thus formally derive the equivalence between the two.

[17] Under this assumption, we can apply a simple average as done in (6.7).

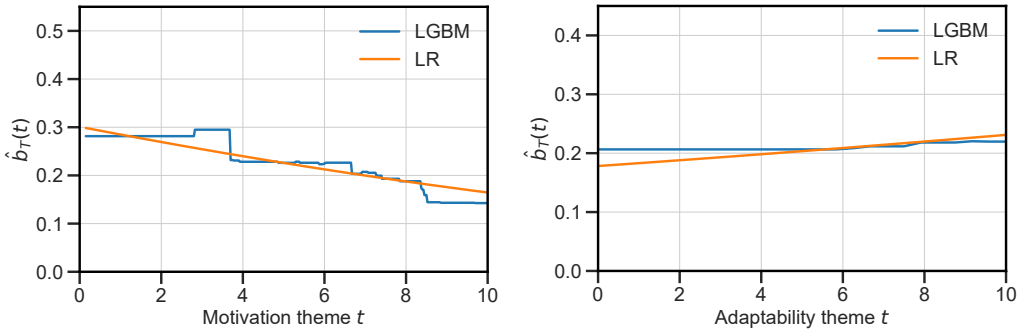



Figure 6.11: Weighted PDP for LR and LGBM classifiers, under the weighted themes datasets, for the themes *Motivation* (left) and *Adaptability* (right).

entirety of the complementary set.[18] However, such estimation was not possible. The main complication we found for estimating the weight of the complementary set was that there is no one-to-one match between the categories used in the survey and the EU official statistics. Therefore, it is important to keep this limitation in mind when interpreting the results beyond the context of our research.

In turn, this potential issue leads us to define a weighted version of partial dependence and, thus, the weighted PDP (or WPDP). By using the country-weighted theme dataset, we can rewrite (6.7) as a country-specific weighted average:

$$\hat{b}_T(t) = \frac{1}{\alpha} \sum_{j=1}^{n} \alpha^{(j)} \hat{f}(T = t, i^{(j)}, w^{(j)}, g^{(j)}, t^{*(j)}) \tag{6.8}$$

where $\alpha_j$ is the weight assigned to $j$'s country, and $\alpha = \sum_{j=1}^{n} \alpha^{(j)}$. Under this approach, we are still using the causal graph $\mathcal{G}$ in Figure 6.9. Note that the weighted partial dependence 6.8 is independent of this work and can be used beyond it as long as we are able to provide the relevant weights.

We estimate the weighted PDP using (6.8). We define as $T$ our top feature from the LGBM model in the weighted theme dataset, which was the *Motivation* (see Figure 6.7). We then use the corresponding top LGBM hyper-parameters and retrain the classifier on the entire dataset. We note that it is common to use the PDP on the training dataset [323] and since we are not interested here in testing performance, we use the entire dataset for fitting the model. Finally, we compute the PDP for *Motivation* theme as shown in Figure 6.11 (left). We do the same for the LR model for comparison.

From Figure 6.11 (left), under the causal graph $\mathcal{G}$, we conclude that there is evidence for the causal claim $T \to Y$ for the *Motivation* theme. For the LGBM model, the theme score (*x-axis*), which ranges from 0 to 10, as it increases the corresponding predicted probabilities of employee turnover decrease, meaning that a higher motivation score leads to a lower employee turnover intention. We see a similar, though smoother, behaviour with the LR model. This is expected as the LGBM can capture non-linear relationships between the variables better than the LR.

We repeat the procedure on a non-top-ranked theme for both models, the *Adaptability* theme (the capability to adapt to changes), to see how the weighted PDPs compare. The

[18] For instance, by estimating the (joint) probability of being a German worker who is also female and has also a college degree.

| Theme | $\Delta\hat{b}_T$ LR | $\Delta\hat{b}_T$ LGBM |
|---|---|---|
| Sustainable Emp. | 0.349 | 0.103 |
| Employership | 0.340 | 0.208 |
| Satisfaction | 0.260 | 0.116 |
| Attendance Stability | 0.205 | 0.119 |
| Motivation | 0.151 | 0.163 |
| Trust | 0.111 | 0.014 |
| Leadership | 0.063 | 0.024 |
| Alignment | 0.038 | 0.006 |
| Work climate | 0.025 | 0.005 |
| Effectiveness | 0.022 | -0.014 |
| Psychol. Safety | 0.017 | 0.004 |
| Productivity | 0.006 | -0.007 |
| Engagement | -0.009 | -0.007 |
| Performance | -0.017 | -0.001 |
| Autonomy | -0.046 | -0.009 |
| Adaptability | -0.067 | -0.005 |
| Customer Focus | -0.078 | -0.016 |
| Efficiency | -0.095 | -0.044 |
| Vitality | -0.111 | -0.024 |
| Role Clarity | -0.127 | -0.092 |

Table 6.4: $\Delta\hat{b}_T$ per theme for LR and LGBM.

results are shown in Figure 6.11 (right). In the case of the LGBM, the PDP is essentially flat and implies a potential non-causal relationship between this theme and employee turnover intention. For the LR, however, we see a non-flat yet narrower PDP, which also seems to support a potential non-causal link. This might be due again to the non-linearity in the data, where the more flexible model (LGBM) can better capture the effects in the changes of $T$ than the less flexible one (LR) that can tend to overestimate them.

We summarize our approach for all themes by calculating *the change in (weighted PDP)*, which we define under (6.8) as:

$$\Delta\hat{b}_T = \hat{b}_T(0) - \hat{b}_T(10) \tag{6.9}$$

and perform this calculation for all themes across the LGBM and LR models. The results are shown in the Table 6.4. The themes are ordered based on the LGBM's deltas. We note that the deltas across models tend to agree: the signs (and for some themes like *Motivation* even the magnitudes) coincide. This is inline with previous results in other sections where the LR's behaviour is comparable to the LGBM's.

Further, comparing the ordering of the themes in Table 6.4 with the feature rankings in Figures 6.5 and 6.7, we note that some of the themes with the largest deltas (such as *Sustainable Emp.* and *Employership*) are also among the top-ranked features. Although there is no clear one-to-one relationship between the two approaches, it is comforting to see the top-ranked themes also having the higher causal impact on employee turnover as it implies some potential shared underlying mechanism.

The Table 6.4 also provides a view on how each theme causally affects employee turnover, where themes with a positive delta cause a decrease in employee turnover. As

the theme's score increases, the probability of turnover decreases. The reverse holds for negative deltas. We recognize that some of these results are not fully aligned with findings by other papers, mainly from the managerial and human resources fields. For instance, we find *Role Clarity* to cause employee turnover to increase, which is the opposite effect found in other studies like Hassan [132]. These other claims, though, we note, are not causal. Moreover, such discrepancies are possible already by taking into account that those findings are based on US data while ours are based on European data. As we argued when motivating Figure 6.9 (left), we believe that the interaction between geographical and work variables affect employee turnover. Hence, the transportability of these previous results into a European context was not expected.

Overall, Table 6.4 along with both Figure 6.11 (left and right) are useful to inform a policy maker. These results can, e.g., serve as evidence for justifying a specific policy intervention. Here, based on our results, we would advise prioritizing policies that foster employee motivation over employee and organization adaptability.

### 6.2.5 Conclusion

We had the opportunity to analyze a unique cross-national survey of employee turnover intention, covering 30 European countries. The analytical methodologies adopted in Lazzari et al. [178] followed three perspectives. The first perspective is from the human resource predictive analytics, and it consisted of the comparison of state-of-the art machine learning predictive models. Logistic Regression (LR) and LightGBM (LGBM) resulted the top performing models. The second perspective is from the explainable AI literature, consisting in the ranking of the determinants (themes and items) of turnover intention by resorting to feature importance of the predictive models. Moreover, a novel composition of feature importance rankings from repeated cross-validation was devised, consisting of critical difference diagrams. The output of the analysis showed that the themes *Sustainable Employability*, *Employership*, and *Attendance Stability* are within the top-five determinants for both LR and LGBM. From the XAI strand of research, we also adopted partial dependency plots, but with a stronger conclusion than correlation/importance using auxiliary causal knowledge as pioneered by Zhao and Hastie [323]. Finally, the third perspective, in fact, is a novel causal approach in support of policy interventions which is rooted in causal structural models. The output confirms those from the second perspective, where highly ranked themes showed PDPs with higher variability than lower ranked themes. The value added from the third perspective here is that we quantify the magnitude and direction for the causal claim $T \rightarrow Y$.

Implicit to this work, motivated by the risk of sample selection bias in the survey data, was our usage of country-specific weights, leading to weighted versions of our two main datasets of themes and items. This effort lead to the weighted PDP, which can also benefit from causal knowledge as shown in three previous section. That said, the weighted PDP, as the reader might infer from the discussion of Lazzari et al. [178], was not the central contribution of this work. Further research is needed.

Three overall limitations of the conclusions of our analysis should be highlighted. The first one is concerned with comparison with related work. Due to the specific set of questions and the target respondents of the GEEI survey, it is difficult to compare our results with related works that use other survey data, which cover a different set of ques-

tions and/or respondents. The second limitation of our results consists of a weighting of datasets, to overcome selection bias, which is limited to country-specific workforce. Either the dataset under analysis should be representative of the workforce, or a more granular weighting should be used to account for country, gender, industry, and any other contextual feature. The final and third limitation of our results concern the causal claims. Our analysis is based on a specific and by far non-unique causal view of the problem of turnover intention where, for example, variables such as *Gender* and *Education level* that belong to the same group node $I$ are considered independent. The interventions carried out to test the causal claim are reliant on the specified causal graph, which limits our results.

Further interdisciplinary research like this work can be beneficial for tackling employee turnover. One possible extension would be to collect country's national statistics to avoid selection bias in survey data or, alternatively, to align the weights of the data to a finer granularity level. Another extension would be to carry out the causal claim tests using a causal graph derived entirely from the data using causal discovery algorithms. In fact, an interesting combination of these two extensions would be to use methods for causal discovery that can account for shifts in the distribution of the data. All of these we consider for future work.

## 6.3 Domain Adaptive Decision Trees

This section is based on the conference paper *Domain Adaptive Decision Trees: Implications for Accuracy and Fairness* by Álvarez et al. [12]. The paper studies the problem of learning a decision tree under unrepresentative data, in particular, under covariate shift (or scenario S1) as discussed in Section 6.1. The paper further formalizes the learning problem and provides a learning algorithm to solve it.

We focus on decision trees given their growing popularity due to their interpretability and performance relative to other more complex models. With *domain adaptive decision trees* (DADT) we aim to improve the accuracy of models trained in a source domain (or training data) that differs from the target domain (or test data). We propose an in-processing step that adjusts the information gain split criterion with outside information corresponding to the distribution of the target population. We demonstrate DADT on real data and find that it improves accuracy over a standard decision tree when testing in a shifted target population. We also study the change in fairness under demographic parity and equal opportunity. Results show an improvement in fairness with the use of DADT. See Appendix D for additional material specific to Álvarez et al. [12].

### 6.3.1 Introduction

In uses of pre-trained machine learning models, it is a known issue that the target population in which the model is being deployed may not have been reflected in the data with which the model was trained. There are many reasons why a training set would not match the target population, including sampling bias [273], concept drift [108], and domain shift [108]. This situation can lead to a reduction in model performance in the target domain. One risk is that, as the demographic distribution of the population changes, certain groups will be under-served by model performance, even as they become more represented in the target population: a type of representation bias [273]. Lack of representation of this kind can be unfair, and adequate visibility can be a prerequisite for fairness [84, Chapter 4]. A classic example is that of female and darker-skinned people being underrepresented in computer-vision datasets, hence scarcely visible to the learning algorithm, with consequences like high error rates in facial recognition and consequent denials of benefits (such as authentication) or imposition of harms (such as arrests) [53]. One, often advisable, approach for dealing with this is to train a new model with updated or improved training data. However, in the case of supervised learning, this may not be possible, as label information for these additional members of the target population may not yet exist. Additionally, while collection of representative data is important, it does come at a cost, including a time cost, so that some shift in the target is likely to occur before updated data is collected or a shift is even identified. The field of domain adaptation proposes techniques for addressing these situations [232].

In this work we contribute to the domain adaptation literature by introducing *domain adaptive decision trees* (DADT). With DADT we aim to improve accuracy of decision tree models trained in a source domain (or training data) that differs from the target domain (or test data), as it may occur when we do not have labeled instances for the target domain. We do this by proposing an in-processing step that adjusts the information gain (IG) split criterion with outside information in the form of unlabeled data, corresponding

to the distribution of the target population we aim for. The approach works by adapting probability estimation to the target domain and, thus, making parts of the feature space more visible to the learning algorithm. We investigate the conditions in which this strategy can lead to increases in performance, fairness, or both.

As an illustrative example, consider the case of a sports retail store looking to target new clients ($D_T$) using what it knows about its current clients ($D_S$) based on geographical regions. The store only has information on the purchasing habits ($Y$) of $D_S$ . Imagine that the store wants to use a classifier to inform its inventory on women's football shoes. If the two client populations differ by region, which is likely, the classifier trained on $D_S$ and intended to predict purchasing patterns ($\hat{Y}$) on $D_T$ could lead to biased predictions when used. For instance, if there is less demand for women's football shoes in the source region relative to the target region, the classifier could underestimate the stocks of women's football shoes needed, under-serving the potential new clients. This could lead to lower service or higher prices for some social groups, and the lost opportunity by the store to gain or even retain customers in the target region. To break such feedback loops, the store could improve the classifier by amplifying some of the knowledge about football shoes purchases in the source region. It could, for instance, use knowledge about the demographics in the target region to better approximate the demand for football shoes by women.

We focus on decision trees for domain shift because decision trees are accessible, interpretable, and well-performing classification models that are commonly used. In particular, we study decision trees rather than more complex classifiers when using tabular data for three reasons. First, these models are widely available across programming languages and are standard in industry and academic communities [242]. Second, these models are inherently transparent [243], which may facilitate the inclusion of stakeholders in understanding and assessing model behaviour. Third, ensembles of these models still outperform the deep learning models on tabular data [114]. For these reasons, and as proposed AI regulations include calls for explainable model behaviour [92, 298], decision trees are a relevant choice when training a classifier and it is therefore important to address issues specific to them.

There are different types of domain shift [228] (recall Section 6.1) and they have different implications for suitable interventions. We focus on the *covariate shift* case of domain shift. This is the case where only the distribution of the attributes change between the source and target, not the relationship between the attributes and the label.

In Section 6.3.2, we introduce the problem setting as a domain adaptation problem, focusing on the covariate shift type of this problem. We also present the necessary background before presenting our proposed intervention to the information gain and introduce the domain adaptive decision trees in Section 6.3.3.

Then in Section 6.3.4, we present the results of our experiments. In the experiments reported we utilize the *ACSPublicCoverage* dataset—an excerpt of US Census data [85], with the prediction task of whether or not a low income individual is covered by public health coverage. The dataset provides the same feature sets for each of the US states. This design allows us to set up an experimental scenario that mirrors our retail example of having no labeled data for the target domain, but some knowledge of the distribution of the attributes in the target domain.

With these experiments we aim not only to improve overall accuracy, but also to pro-

duce sufficient accuracy for different demographic groups. This is important because the distribution of these groups may be different in the target population and even shift over time in that population. For example, Ding et al. [85] found that naïvely implementing a model trained on data from one US state and using it in each of the other states resulted in unpredictable performance in both overall accuracy and in *demographic parity* (a statistically defined metric of model fairness performance based on treatment of members of a selected demographic group compared to members of another demographic group). We therefore also test the impact of our intervention on the results of a post-processing fairness intervention [131], which we measure using two common fairness metrics: *demographic parity* and *equal opportunity* [26, Chapter 3].

We examine those results in relation to the covariate shift assumption between source and target populations. We see that our intervention leads to an increase in accuracy when the covariate shift assumption holds. Section 6.3.5 closes this work and gives an outlook on future work.

**Summary of our contributions.** Our main contributions are the following:

- we are the first to formulate the decision tree learning problem as a domain adaptation problem, focusing on covariate shift;
- we propose a simple and intuitive solution: an in-processing step based on reweighting the information gain split criteria using target domain information; and
- we introduce a new line of work for domain adaption in the form of domain adaptive decision trees.

**Related work.** The related work situates our approach in the literature on domain adaptation in decision trees and adjusting the information gain of decision trees.

Domain adaptation (DA) studies how to achieve a robust model when the training (source domain) and test (target domain) data do not follow the same distribution [232]. Here, we focus on the covariate shift type, which occurs when the attribute space $\mathbf{X}$ is distributed differently across domains [201, 228, 318]. DADT is the first framework to address domain DA as an in-processing problem specific to decision tree classifiers.

Previous work on adjusting entropy estimation has been conducted largely outside of machine learning, as well as in the context of information gain (IG) in decision trees. Here, too, DADT is the first work to look at entropy estimation under DA. Guiasu [117] proposes a general form for a weighted entropy equation for adjusting the likelihood of the information being estimated. Other works study the estimation properties behind using frequency counts for estimating the entropy [19, 209, 213, 258]. In relation to decision trees, [265] proposes a weighted IG based on the the risk of the portfolio of financial products that the decision tree is trying to predict. Similarly, [320, 321] reweight IG with the fairness metric of statistical parity. Vieira and Antunes [285] adjust the IG calculation with a gain ratio calculation for the purpose of correcting a bias of against attributes that represent higher levels of abstraction in an ontology.

Recent work has started to examine the relationship between DA and fairness. Mukherjee et al. [205] show that domain adaptation techniques can enforce individual fairness notions. Maity et al. [190] show that enforcing risk-based fairness minimization notions

can have an ambiguous effect under covariate shift for the target population, arguing that practitioners should check on a per-context basis whether fairness is improved or harmed. This is line with the findings of [85] who test both standard and fairness adjusted gradient boosting machines across numerous shifted domains and find that both accuracy and fairness metric measures are highly variable across target domains. These works call for further work to understand the impact of domain drifts and shifts.

Similarly, as discussed in Section 6.1, causal works have started to formulate domain adaptation as a causal problem [189, 318]. Zhang et al. [318] use structural causal models to formulate the different kinds of domain adaptation. An important contribution of their work is the connection between structure (or invariance) and changes in the distribution: the presence of a shift in distribution implies the existence of a causal relation that should be recoverable once the shift is identified. [189] use the previous causal formulation to predict invariant conditional distributions to measure the treatment effect of an intervention in a changing population. In the future, we hope to explore our work more explicitly using causality.

### 6.3.2 Problem Setting

Let $\mathbf{X}$ denote the set of discrete/continuous *predictive attributes*, $Y$ the *class attribute*, and $f$ the *decision tree classifier* such that $\hat{Y} = f(\mathbf{X})$ with $\hat{Y}$ denoting the *predicted class attribute*. We assume a scenario where the population used for training $f$ (the *source domain* $D_S$) is not representative of the population intended for $f$ (the *target domain* $D_T$). Formally, we write it as $P_S(\mathbf{X}, Y) \neq P_T(\mathbf{X}, Y)$, where $P_S(\mathbf{X}, Y)$ and $P_T(\mathbf{X}, Y)$, respectively, denote the source and target domain joint probability distributions. We tackle this scenario as a *domain adaptation* (DA) problem [232] as it allows us to formalize the difference between distributions in terms of distribution shifts.

There are three types of distribution shifts in DA: covariate, prior probability, and dataset shift. Here, we focus on *covariate shift* [201, 228, 318] in which the conditional distribution of the class, $P(Y|\mathbf{X})$, remains constant but the marginal distribution of the attributes, $P(\mathbf{X})$, changes across the two domains:

$$P_S(Y|\mathbf{X}) = P_T(Y|\mathbf{X}) \text{ but } P_S(\mathbf{X}) \neq P_T(\mathbf{X}) \tag{6.10}$$

We focus on covariate shift because we assume, realistically, to have some access only to the predictive attributes $\mathbf{X}$ of the target domain.[19] Under this *unsupervised setting*, we picture a scenario where a practitioner needs to train $f$ on $D_S$ to be deployed on $D_T$. Aware of the potential covariate shift, the practitioner wants to avoid training a biased model relative to the target domain that could result in poor performance on $\hat{Y}$.

What can be done here to address the DA problem depends on what is known about $P_T(\mathbf{X})$. In the ideal case in which we know the whole covariate distribution $P_T(\mathbf{X})$, being under (6.10) allows for computing the full joint distribution due to the multiplication rule of probabilities:

$$P_T(Y, \mathbf{X}) = P_T(Y|\mathbf{X}) \cdot P_T(\mathbf{X}) = P_S(Y|\mathbf{X}) \cdot P_T(\mathbf{X}) \tag{6.11}$$

[19] The other two settings require information on $Y$ being available in $D_T$, with *prior probability shift* referring to cases where the marginal distribution of the class attribute changes, $P_S(\mathbf{X}|Y) = P_T(\mathbf{X}|Y)$ but $P_S(Y) \neq P_T(Y)$, and *dataset shift* referring to cases where neither covariate nor prior probability shifts apply but the joint distributions still differ, $P_S(\mathbf{X}, Y) \neq P_T(\mathbf{X}, Y)$.

where we can exchange $P_T(Y|\mathbf{X})$ for $P_S(Y|\mathbf{X})$, which is convenient as we know both $Y$ and $\mathbf{X}$ in $D_S$. In reality, however, the right-hand-side of (6.11) can be known to some extent due to three issues:

(P1) $P_T(\mathbf{X})$ is not fully available, meaning the marginal distributions of some of the attributes $X \in \mathbf{X}$ are known;

(P2) $P_T(Y|\mathbf{X}) \approx P_S(Y|\mathbf{X})$ but not equal, meaning the covariate shift holds in a relaxed form; and

(P3) $P_T(\mathbf{X})$ and $P_S(Y|\mathbf{X})$ are estimated given sample data from the respective populations, and, as such, the estimation can have some variability.

Issue P3 is pervasive in statistical inference and machine learning. We do not explicitly[20] consider it in our problem statement. Therefore, the main research question that we intend to address in this paper is:

*RQ1. With reference to the decision tree classifier, which type and amount of target domain knowledge (issue P1) help reduce the loss in accuracy at the variation of relaxations of covariate shift (issue P2)?*

As domain shift can have a detrimental impact on performance of the model for some demographic groups over others, a subsequent question to address in this paper is:

*RQ2. How does the loss in accuracy by the decision tree classifier, based on the issues P1 and P2, affect a fairness metric used for protected groups of interest in the target domain?*

The knowledge relevant for (6.11) and RQ1 and RQ2 is bounded by two border cases. ***No target domain knowledge:*** it consists of training $f$ on the source data and using it on the target data without any change or correction. Formally, we estimate $P_T(\mathbf{X})$ as $P_S(\mathbf{X})$ and $P_T(Y|\mathbf{X})$ as $P_S(Y|\mathbf{X})$. ***Full target domain knowledge:*** it consists of training $f$ on the source data and using it on the target data, but exploiting full knowledge of $P_T(\mathbf{X})$ in the learning algorithm to replace $P_S(\mathbf{X})$. ***Partial target domain knowledge:*** consequently, the in-between case consists of training a decision tree on the source data and using it on the target data, but exploiting partial knowledge of $P_T(\mathbf{X})$ in the learning algorithm and complementing it with knowledge of $P_S(\mathbf{X})$.

The form of partial knowledge depends on the information available on $\mathbf{X}$, or subsets of it. Here, we consider a scenario where for $\mathbf{X}' \subseteq \mathbf{X}$, an estimate of $P(\mathbf{X}')$ is known only for $|\mathbf{X}'| \leq 2$ (or $|\mathbf{X}'| \leq 3$), namely we assume to know bi-variate (resp., tri-variate) distributions only, but not the full joint distribution. This scenario occurs, for example, when using cross-tabulation data from official statistics. We specify how to exploit the knowledge of $P_T(\mathbf{X})$ for a decision tree classifier in Section 6.3.3, introducing what we refer to as a *domain-adaptive decision tree* (DADT). We introduce the required technical background in the remainder of this section.

**Decision tree learning.** Top-down induction algorithms grow a decision tree classifier [133] from the root to the leaves. At each node, either the growth stops producing a leaf, or a split condition determines child nodes that are recursively grown. Common

[20] We tackle it implicitly through the Law of Large Numbers by restricting to estimation of probabilities in contexts with a minimum number of instances. This is managed in decision tree learning by a parameter that stops splitting a node if the number of instances at a node is below a minimum threshold.

stopping criteria include node purity (all instances have the same class value), data size (the number of instances is lower than a threshold), and tree depth (below a maximum depth allowed). Split conditions are evaluated based on a split criterion, which selects one of them or possibly none (in this case the node becomes a leaf).

We assume binary splits of the form:[21] $X = t$ for the left child and $X \neq t$ for the right child, when $X$ is a discrete attribute; or $X \leq t$ for the left child and $X > t$ for the right child, when $X$ is a continuous attribute. We call $X$ the *splitting attribute*, and $t \in X$ the *threshold value*. Together they form the *split condition*. Instances of the training set are passed from a node to its children by partitioning them based on the split condition. The conjunction of split conditions from the root to the current node being grown is called the *current path* $\varphi$. It determines the instances of the training dataset being considered at the current node. The predicted probability of class $y$ at a leaf node is an estimation of $P(Y = y|\varphi)$ obtained by the relative frequency of $y$ among the instances of the training set reaching the leaf or, equivalently, satisfying $\varphi$.

**The information gain split criterion.** We focus on the information gain split criterion. It is, along with Gini, one of the standard split criteria used. It is also based on information theory via entropy [73], which links the distribution of a random variable to its information content. The *entropy* ($H$) measures the information contained within a random variable based on the uncertainty of its events. The standard is *Shannon's entropy* [194] where we define $H$ for the class random variable $Y$ at $\varphi$ as:

$$H(Y|\varphi) = \sum_{y \in Y} -P(Y = y|\varphi) \log_2(P(Y = y|\varphi)) \tag{6.12}$$

where $-\log_2(P(Y = y|\varphi)) = I(y|\varphi)$ represents the *information* ($I$) of $Y = y$ at current path $\varphi$. Therefore, entropy is the expected information of the class distribution at the current path. Intuitively, the information of class value $y$ is inversely proportional to its probability $P(Y = y|\varphi)$. The more certain $y$ is, reflected by a higher $P(Y = y|\varphi)$, the lower its information as $I(y|\varphi)$ (along with its contribution to $H(Y|\varphi)$). The general idea is that there is little new information to be learned from an event that is certain to occur. We picture it as *low entropy implies no surprises*.

The *information gain* ($IG$) for a split condition is the difference between the entropy at a node and the weighted entropy at the child nodes determined by the split condition $X$ and $t$ under consideration. $IG$ uses (6.12) to measure how much information is contained under the current path. For a discrete splitting attribute $X$ and threshold $t$, we have:

$$\begin{aligned} IG(X, t|\varphi) = H(Y|\varphi) - P(X = t|\varphi)H(Y|\varphi, X = t) \\ - P(X \neq t|\varphi)H(Y|\varphi, X \neq t) \end{aligned} \tag{6.13}$$

and for a continuous splitting attribute $X$ and threshold $t$:

$$\begin{aligned} IG(X, t|\varphi) = H(Y|\varphi) - P(X \leq t|\varphi)H(Y|\varphi, X \leq t) \\ - P(X > t|\varphi)H(Y|\varphi, X > t) \end{aligned} \tag{6.14}$$

[21] There are other forms of binary splits, as well as multi-way and multi-attribute split conditions [216].

where the last two terms in each (6.13) and (6.14) represent the total entropy obtained from adding the split condition on $X$ and $t$ to $\varphi$.[22] The selected split attribute and threshold are those with maximum $IG$, namely $arg\,max_{X,t}\, IG(X, t|\varphi)$.

Intuitively, we aim at maximizing the reduction of weighted average entropy from the parent to the child nodes. This is because we aim at constructing leafs that are homogeneous when learning a decision tree. The more homogeneous a node is, the less information it contains, and the lower its entropy. This allows us to sort (or group) instances similar on the attribute space given the class as we move top down.

**On estimating probabilities.** Probabilities and, thus, $H$ (6.12) and $IG$ (6.13)–(6.14) are defined for random variables. As the decision tree grows, the probabilities are estimated on the subset of the training set $D$ reaching the current node satisfying $\varphi$ by frequency counting:

$$\begin{aligned} \hat{P}(X = t \,|\, \varphi) &= \frac{|\{w \in D \mid \varphi(w) \wedge w[X] = t\}|}{|\{w \in D \mid \varphi(w)\}|}\,, \\ \hat{P}(Y = y \,|\, \varphi) &= \frac{|\{w \in D \mid \varphi(w) \wedge w[Y] = y\}|}{|\{w \in D \mid \varphi(w)\}|} \end{aligned} \tag{6.15}$$

The denominator represents the number of instances $w$ in $D$ that satisfy the condition $\varphi$ (written $\varphi(w)$) and the numerator the number of those instances that further satisfy $X = t$ (respectively, $Y = y$). We use the estimated probabilities (6.15) to estimate $H$ (6.12) and $IG$ (6.13)–(6.14).

The hat in (6.15) differentiates the estimated probability, $\hat{P}$, from the population probability, $P$. Frequency counting is supported by the Law of Large Numbers. Assuming that the training set $D$ is an *i.i.d.* sample from the $P$ probability distribution, we expect for $\hat{P}(X = t|\varphi) \approx P(X = t \,|\, \varphi)$ and $\hat{P}(Y = y|\varphi) \approx P(Y = y \,|\, \varphi)$ as long as we have enough training observations in $D$, which is often the case when training $f$. A key issue is whether $D$ is representative of the population of interest. This is important as $\hat{P}$ will approximate the $P$ behind $D$.

When training any classifier, the key assumption is that the $P$ probability distribution is the same for the data used for growing the decision tree (the training dataset) and for the data on which the decision tree makes predictions (the test dataset). Under covariate shift (6.10) this assumption does not hold. Instead, the training dataset belongs to the source domain $D_S$ with probability distribution $P_S$ and the test dataset belongs to the target domain $D_T$ with probability distribution $P_T$, such that $P_T \neq P_S$. To stress this point, we use the name *source data* for training data sampled from $D_S$ and *target data* for training data sampled from $D_T$. The estimated probabilities (6.15), and the subsequent estimations for $H$ (6.12) and $IG$ (6.13)–(6.14) based on source data alone can be biased, in statistical terms, relative to the intended target domain. This can result, among other issues, in poor model performance from the classifier. This is why we propose extending (6.15) by embedding target-domain knowledge into these estimated probabilities.

[22] Formally, together these last two terms represent the conditional entropy $H(Y, X|\varphi)$ written for the binary split case we are considering such that:

$$H(Y, X|\varphi) = \sum_{x \in X} -P(X = x|\varphi) \sum_{y \in Y} P(Y = y|\varphi, X = x) \log_2(P(Y = y|\varphi, X = x))$$

Measuring the distance between the probability distributions $P_S$ and $P_T$ is relevant for detecting distribution shifts. We resort to the *Wasserstein distance* $W$ between two probability distributions to quantify the amount of covariate shift and the robustness of target domain knowledge. See Appendix D.1 for details.

Under covariate shift (6.10), it is assumed that $P_S(Y|\mathbf{X}) = P_T(Y|\mathbf{X})$, which allows one to focus on the issue of $P_S(\mathbf{X}) \neq P_T(\mathbf{X})$. This equality is often not verified in practice. We plan to use $W$, along with an approximation of $P_T(Y|\mathbf{X})$ (since $Y$ is unavailable in $D_T$), to measure the distance between these two conditional probabilities to ensure that our proposed embedding with target domain knowledge is impactful . Measuring this distance will allow us to evaluate how relaxations of $P_S(Y|\mathbf{X}) = P_T(Y|\mathbf{X})$ affect the impact of our proposed target domain embedding.

### 6.3.3 Domain Adaptive Decision Trees

We present our approach for addressing covariate shift by embedding target domain knowledge when learning the decision tree classifier. We propose an *in-processing step* under the information gain split criterion, motivating what we refer to as *domain adaptive decision trees* (DADT) learning.

As discussed in the previous section, when growing the decision tree, the estimated probabilities (6.15) used for calculating $H$ (6.12) and thus $IG$ (6.13)–(6.14) at the current path $\varphi$ are derived over a training dataset, which is normally a dataset over the source domain $D_S$. For the split condition $X = t$, it follows that $\hat{P}(X = t|\varphi) \approx P_S(X = t|\varphi)$, which is an issue under covariate shift. We instead want that $\hat{P}(X = t|\varphi) \approx P_T(X = t|\varphi)$. We propose to embed in the learning process knowledge from the target domain $D_T$, reducing the potential bias in the estimation of the probabilities and, in turn, reducing the bias of the trained decision classifier.

**Embedding target domain knowledge.** There are two probability forms that are to be considered when growing a decision tree for the current path $\varphi$: $P(X = t|\varphi)$ in (6.13) (and, respectively, $P(X \leq t|\varphi)$ in (6.14)) and $P(Y = y|\varphi)$ in (6.12)–(22). In fact, the formulas of entropy and information gain only rely on those two probability forms, and on the trivial relation $P(X \neq t|\varphi) = 1 - P(X = t|\varphi)$ for discrete attributes (and, respectively, $P(X > t|\varphi) = 1 - P(X \leq t|\varphi)$ for continuous attributes). It follows that we can easily estimate $\hat{P}_S(X = t|\varphi)$ and $\hat{P}_S(Y = y|\varphi)$ using the available source domain knowledge.

*Estimating* $P(X = t|\varphi)$. We assume that some target domain knowledge is available, from which we can estimate $\hat{P}_T(X|\varphi)$, in the following cases:[23]

$$\hat{P}_T(X = x|\varphi) \approx P_T(X = x|\varphi) \text{ for } X \text{ discrete,}$$
$$\hat{P}_T(X \leq x|\varphi) \approx P_T(X \leq x|\varphi) \text{ for } X \text{ continuous} \quad (6.16)$$

[23]Actually, since we consider $x$ in the (finite) domain of $X$ the two forms are equivalent, due to basic identities $P_T(X \leq x|\varphi) = \sum_{X \leq x} P_T(X = x|\varphi)$ and $P_T(X = x|\varphi) = P_T(X \leq x|\varphi) - P_T(X \leq x'|\varphi)$, where $x'$ is the element preceding $x$ in the domain of $X$. Moreover, by definition of conditional probability, we have $P(X = t|\varphi) = P(X = t, \varphi)/P(\varphi)$ and then target domain knowledge boils down to estimates of probabilities of conjunction of equality conditions. Such form of knowledge is, for example, provided by cross-tables in official statistics data.

In case $\hat{P}_T(X = x|\varphi)$ is not directly available in the target domain knowledge $D_T$, we adopt an affine combination for discrete and continuous attributes using the source domain knowledge $D_S$:

$$\hat{P}(X = x|\varphi) = \alpha \cdot \hat{P}_S(X = x|\varphi) + (1 - \alpha) \cdot \hat{P}_T(X = x|\varphi') \tag{6.17}$$

$$\hat{P}(X \leq x|\varphi) = \alpha \cdot \hat{P}_S(X \leq x|\varphi) + (1 - \alpha) \cdot \hat{P}_T(X \leq x|\varphi') \tag{6.18}$$

where $\varphi'$ is a maximal subset of split conditions in $\varphi$ for which $\hat{P}_T(X = x|\varphi')$ is in the target domain knowledge, and $\alpha \in [0, 1]$ is a tuning parameter to be set. In particular, setting $\alpha = 1$ boils down to estimating probabilities based on the source data only. With such assumptions, $P(X = t|\varphi)$ in (6.13) (respectively $P(X \leq t|\varphi)$ in (6.14)) can be estimated as $\hat{P}(X = t|\varphi)$ (resp. $\hat{P}(X \leq t|\varphi)$) to derive $IG$.

*Estimating* $P(Y = y|\varphi)$. Let us consider the estimation of $P(Y = y|\varphi)$ in (6.12) over the target domain. Since $Y$ is unavailable in $D_T$, it is legitimate to ask whether $P(Y = y|\varphi)$ is the same probability in the target domain as in the source domain when growing the decision tree classifier?

If yes, then we would simply estimate $P_T(Y = y|\varphi) \approx \hat{P}_S(Y = y|\varphi)$.

Unfortunately, however, the answer is no. To illustrate this last point, recall that the covariate shift assumption (6.10) states that $P_S(Y|\mathbf{X}) = P_T(Y|\mathbf{X})$, namely that the probability of $Y$ conditional on fixing *all of the* variables in $\mathbf{X}$ is the same in the source and target domains:

$$\forall \mathbf{x} \in \mathbf{X}, \forall y \in Y,\ P_S(Y = y|\mathbf{X} = \mathbf{x}) = P_T(Y = y|\mathbf{X} = \mathbf{x}) \tag{6.19}$$

The above equality, however, may not hold when growing the tree because the current path $\varphi$ does not necessarily fix all of the $\mathbf{X}$'s. In other words, (6.19) does not necessarily imply $\forall \varphi\ P_T(Y = y|\varphi) = P_S(Y = y|\varphi)$. This situation is an instance of Simpson's paradox [264]. Example D.2.1 in Appendix D.2.

Given the potential violation of the equality (6.19), we rewrite $P_T(Y = y|\varphi)$ using the law of total probability as follows:

$$\begin{aligned} P_T(Y = y|\varphi) &= \sum_{\mathbf{x} \in \mathbf{X}} P_T(Y = y|\mathbf{X} = \mathbf{x}, \varphi) \cdot P_T(\mathbf{X} = \mathbf{x}|\varphi) \\ &= \sum_{\mathbf{x} \in \mathbf{X}} P_S(Y = y|\mathbf{X} = \mathbf{x}, \varphi) \cdot P_T(\mathbf{X} = \mathbf{x}|\varphi) \end{aligned} \tag{6.20}$$

where the final equation exploits the covariate shift assumption (6.19) when it holds for a current path $\varphi$. Instead of taking $P_S(Y|\varphi) = P_T(Y|\varphi)$ for granted, which we should not do under DADT learning, we rewrite $P_T(Y = y|\varphi)$ in terms of probabilities over source domain, $P_S(Y = y|\mathbf{X} = \mathbf{x}, \varphi)$, and target domain, $P_T(\mathbf{X} = \mathbf{x}|\varphi)$, knowledge.

Varying $\mathbf{x} \in \mathbf{X}$ over all possible combination as stipulated in (6.20), however, is not feasible in practice as it would require extensive target domain knowledge to estimate $P_T(\mathbf{X} = \mathbf{x}|\varphi)\ \forall \mathbf{x} \in \mathbf{X}$. This would still be a practical issue in the ideal case in which we have a full sample of the target domain, as it would require the sample to be large enough for observing each value $\mathbf{x} \in \mathbf{X}$ in $D_T$.

We approximate (6.20) by varying values with respect to a *single attribute* $X_w \in \mathbf{X}$ and relying on (6.17) for an estimate of $P_T(X_w = x|\varphi)$. Let us then define the estimate

of $P(Y = y|\varphi)$ as:

$$\hat{P}(Y = y|\varphi) = \sum_{x \in X_w} \hat{P}_S(Y = y|X_w = x, \varphi) \cdot \hat{P}(X_w = x|\varphi) \tag{6.21}$$

where we now use target domain knowledge only about $X_w$ instead of spanning the entire attribute space $\mathbf{X}$.

To account for the above instance of Simpson's paradox, the attribute $X_w$ should be chosen such that $\hat{P}_S(Y = y|X_w = x, \varphi) \approx P_T(Y = y|X_w = x, \varphi)$. Such an attribute $X_w$, however, may be specific to the current path $\varphi$. Hence, we only consider the empty $\varphi$, and choose $X_w$ such that the average distance between $\hat{P}_S(Y|X_w = x)$ and an estimate $\hat{P}_T(Y|X_w = x)$ of $P_T(Y|X_w = x)$ is minimal:

$$\begin{aligned} X_w &= \arg\min_X \mathcal{W}(X) \text{ where} \\ \mathcal{W}(X) &= \sum_{x \in X} W(\hat{P}_S(Y|X = x), \hat{P}_T(Y|X = x)) \cdot \hat{P}_T(X = x) \end{aligned} \tag{6.22}$$

$\mathcal{W}(X)$ is the average Wasserstein distance between $\hat{P}_S(Y|X)$ and $\hat{P}_T(Y|X)$. In terms of target domain knowledge, computing (6.22) requires knowledge of $\hat{P}_T(Y|X)$, an estimate of the conditional distribution of the class in the target domain.

In other words, in (6.22), *we depart slightly from our assumption that no knowledge is available of $Y$ in $D_T$*. If calculation of (6.22) is not feasible, we assume some expert input on an attribute $X$ such that $\hat{P}_S(Y = y|X = x) \approx \hat{P}_T(Y = y|X = x)$, as a way to minimize the first term of the summation (6.22). In such a case, we do not actually compute $\mathcal{W}(X)$.

To summarize, we use $X_w$ as from (6.22) (or as provided by a domain expert) to derive (6.21) as an empirical approximation to (6.20). This approach is how we estimate $P(Y|\varphi)$ over the target domain.

**How much target domain knowledge?** We can now formalize the range of cases based on the availability of $P_T(\mathbf{X})$ described in Section 6.3.2.

Under ***no target domain knowledge***, we have no information available on $D_T$, which means that $\hat{P}(X = x|\varphi) = \hat{P}_S(X = x|\varphi)$. This amounts to setting $\alpha = 1$ in (6.17) and (6.18), and, whatever $X_w$ is, (6.21) boils down to $\hat{P}(Y = y|\varphi) = \hat{P}_S(Y = y|\varphi)$. Both probability estimations boil down to growing the DADT classifier using the source data $D_S$ without any modification or, simply, growing a standard decision tree classifier.

Similarly, under ***full target domain knowledge*** we have target domain knowledge for all attributes in $D_T$ along with enough instances to estimate both probabilities. This scenario amounts to setting $\alpha = 0$ in (6.17)–(6.18), and to know which attribute $X_w$ minimizes (6.22).

The full target domain knowledge is the strongest possible assumption within our DADT approach, but not in general. For validation purposes (Section 6.3.4), we move away from our unsupervised setting and assume $Y \in D_T$ to set up an *additional baseline* under the full knowledge of $D_T$: ***target-to-target baseline.*** In this scenario, the decision tree is grown *and* tested exclusively on the target data. Such a scenario does not require covariate shift, since probabilities $P(X = t|\varphi)$ and $P(Y = y|\varphi)$ are estimated directly

over the target domain. This scenario is *the ideal case* as we train the classifier on the intended population.

Finally, under ***partial target domain knowledge*** we consider cases where we have access to estimates of $P(\mathbf{X}')$ only for some subsets $\mathbf{X}' \subseteq \mathbf{X}$. This allows us to estimate $P(X = x|\varphi)$ only if $X$ and the variables in $\varphi$ are in one of those subsets $\mathbf{X}'$. When the target domain information is insufficient, DADT resorts to the source domain information in (6.17)–(6.18) by an affine combination of both. The weight $\alpha$ in such an affine combination should be set proportional to the contribution of the source domain information. We refer to Section 6.3.4 for our experimental setting of $\alpha$.

### 6.3.4 Experiments

We consider the *ACSPublicCoverage* dataset—an excerpt from the 2017 US Census data [85]—that provides the same feature sets for different geographical regions based on the US states, which may have different distributions. This allows us to examine the impact of our method given a wide range of distribution shifts. We utilize the prediction task, constructed by the dataset creators, of whether or not a low-income individual is covered by public health coverage.

Inspired by this experimental setting, we imagine a task where a public administrator wants to identify individuals who do not receive the public benefits they are entitled to. Information about who does and does not receive these benefits, however, is only available for a population different from the target population: for example, the population from another state. This administrator is likely to have some information about the target population distribution; information that they realistically may have on population breakdown by demographics such as age, race and gender.

Recall our two research questions from Section 6.3.1. To address RQ1 we now test whether, with DADT, we can utilize that information to train an improved model in the new state, compared to blindly applying a model trained in the other state. Additionally, we address RQ2 by testing the impact of using DADT instead of standard decision trees on two fairness metrics: demographic parity and equal opportunity.

**Experimental setup.** The design of the *ACSPublicCoverage* dataset allows us to set up a scenario that mirrors our example of the retail store in Section 6.3.1: we have unlabeled data for the target domain, but some knowledge of the (unconditional) distribution of the target domain. Here, however, we extend the scenario by having access to the labeled data for each state. We note that the implementation of the DADT does not require this information, but we utilize our access to the target domain labeled data in this section to test our assumption that DADTs are suitable for addressing covariate shift.

Given the dataset design, we are able to utilize the distribution of the predictive attributes in the target domain as our source of outside knowledge, to adjust the information gain calculation. Unless otherwise stated, we consider the attributes: SCHL (educational attainment), MAR (marital status), AGEP (age), SEX (male or female), CIT (citizenship status), RAC1P (race), with AGEP being continuous and all others discrete. Data was accessed through the Python package Folktables.[24]

[24] `https://github.com/zykls/folktables`

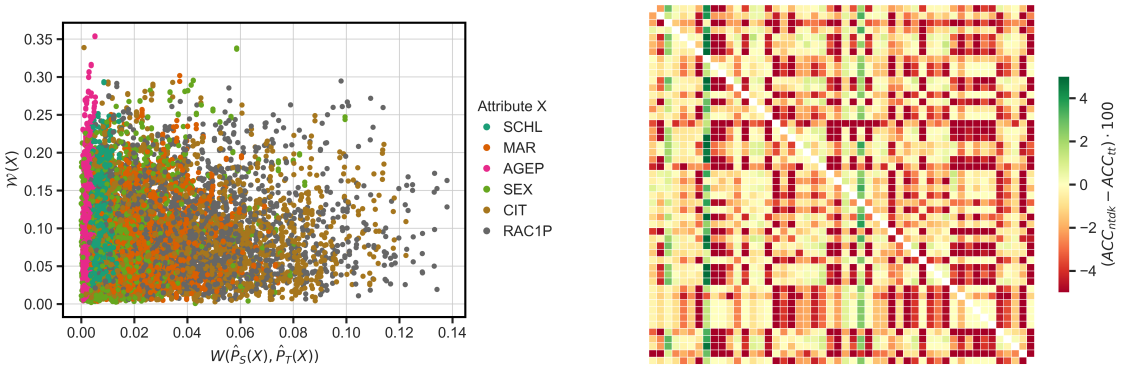



Figure 6.12: The scatter-plot (left) relates the Wasserstein distances each attribute and source-target US state pair. The x-axis shows the distance of each attribute's marginal distributions between source $\hat{P}_S(X)$ and target domains $\hat{P}_T(X)$, while the y-axis shows the average distance between conditional $\hat{P}_S(Y|X)$ and $\hat{P}_T(Y|X)$, as in (6.22). The heat-map (right) shows the difference in accuracy between the cases no target domain knowledge $ACC_{ntdk}$ and target-to-target baseline $ACC_{tt}$ for each source-target US state pair. Both figures show a lack of an overall pattern across all states in *ACSPublicCoverage*. The dataset does not in general satisfy the covariate shift assumption (left).

We consider pairs of source and target datasets consisting of data from different US states, with a model trained in each of the fifty states being tested on every state, for a total of 2500 train / test pairs. The decision trees are all trained on 75% of source data $D_S$, and tested on 25% of the target data $D_T$. Stopping criteria include the following: a node must have at least 5% of the training data, and not all instances have the same class value (purity level set to 100%), the maximum tree depth is 8.[25]

To address RQ2, in particular, we undertake a post-processing approach to fairness based on the known link between a model's performance and its fairness [87, 179, 282]. The public administrator wants to evaluate the performance of the trained classifier on certain demographic groups in the target population. The administrator thus resorts to applying a post-processing method around the classifier that adjusts the predictions under the chosen fairness metric. In practice, this comes down to using a wrapper function based on [131].

We focus on this model agnostic post-processing fairness intervention to measure the impact of DADT on a fairness intervention. Post-processing methods rely on the non-DA setting, meaning that $P_S(\mathbf{X}, Y) = P_T(\mathbf{X}, Y)$. Classifiers are a statement on the joint probability distribution of the training data. Under DA, post-processing methods are essentially only modifying $P_S(\mathbf{X}, Y)$. Granted the user trains an oracle-like standard decision tree, the issue remains that the post-processing fairness intervention would only be addressing issues on the source and not the target population. Therefore, DADT is expected to positively affect the fairness measure.

**Results on accuracy.** We now address RQ1 (Section 6.3.2). The scatter-plot Figure 6.12 (left) relates the Wasserstein distances for each attribute and source-target pair. On the x-axis, there is the distance between the marginal attribute distributions, in other

[25]The code, data, and run are available at this repository.

words: $W(\hat{P}_S(X), \hat{P}_T(X))$. On the y-axis, there is the average distance between conditional $\hat{P}_S(Y|X)$ and $\hat{P}_T(Y|X)$, i.e., $\mathcal{W}(X)$ from (6.22). Notice that the distances between the marginal attribute distributions are rather small, with the exception of CIT and RAC1P. The distances between class conditional distributions are instead much larger, for all attributes.

The plot shows that the *ACSPublicCoverage* dataset does not in general satisfy the covariate shift assumption (at least when conditioning on a single attribute), but rather the opposite: close attribute distributions and distant conditional class distributions. This fact will help us in exploring how much our approach relies on the covariate shift assumption. Below we report accuracy at varying levels of target domain knowledge (issue P1 Section 6.3.2), as defined in Section 6.3.3.

***Case 1: no target domain knowledge (ntdk) vs target-to-target baseline (tt).*** Let us consider the scenario of no target domain knowledge, meaning that training a decision tree on the source training data and testing it on the target test data. We compare the decision tree accuracy in this scenario (let us call $ACC_{ntdk}$) to the accuracy of training a decision tree on the target training data and testing on the target test data ($ACC_{tt}$), i.e., the target-to-target baseline.

Recall that accuracy estimates on a test set (of the target domain) the probability that the classifier prediction $\hat{Y}$ is correct with respect to the ground truth $Y$. Let us define *accuracy* formally as:

$$ACC = P_T(\hat{Y} = Y) \tag{6.23}$$

which is standard definition of accuracy.

The heat-map plot Figure 6.12 (right) shows for each source-target pair of states the difference in accuracy $(ACC_{ntdk} - ACC_{tt}) \cdot 100$ between the no target domain knowledge scenario and the target-to-target baseline. In most of the cases the difference is negative, meaning that there is an accuracy loss in the no target domain knowledge scenario.

***Case 2: full target domain knowledge (ftdk) vs no target domain knowledge (ntdk).*** The decision tree in this scenario is grown on the source (training) data but probabilities are estimated by full target domain knowledge using (6.16), and (6.21) with $X_w$ minimizing (6.22). In the experiments, $\hat{P}_T(X = t|\varphi)$ and $\hat{P}_T(X \leq t|\varphi)$ are calculated from the target training data, for each $X$, $t$, and $\varphi$.

Let us compare the accuracy of the decision tree grown using full target domain knowledge (let us call it $ACC_{ftdk}$) to the decision tree grown using with no target domain knowledge ($ACC_{ntdk}$). In 48% of the source-target pairs, the accuracy of the full target domain knowledge scenario is better than the one of the no target domain knowledge scenario ($ACC_{ftdk} > ACC_{ntdk}$), and in 26% of the pairs they are equal ($ACC_{ftdk} = ACC_{ntdk}$). This gross comparison needs to be investigated further.

Let us define the *relative gain in accuracy* as:

$$rACC = \frac{ACC_{ftdk} - min(ACC_{ntdk}, ACC_{tt})}{|ACC_{tt} - ACC_{ntdk}|} \cdot 100 \tag{6.24}$$

where $ACC_{tt}$ is the accuracy in the target-to-target baseline. The relative gain quantifies how much of the loss in accuracy in the no target domain knowledge scenario has been recovered in the full target domain knowledge scenario.

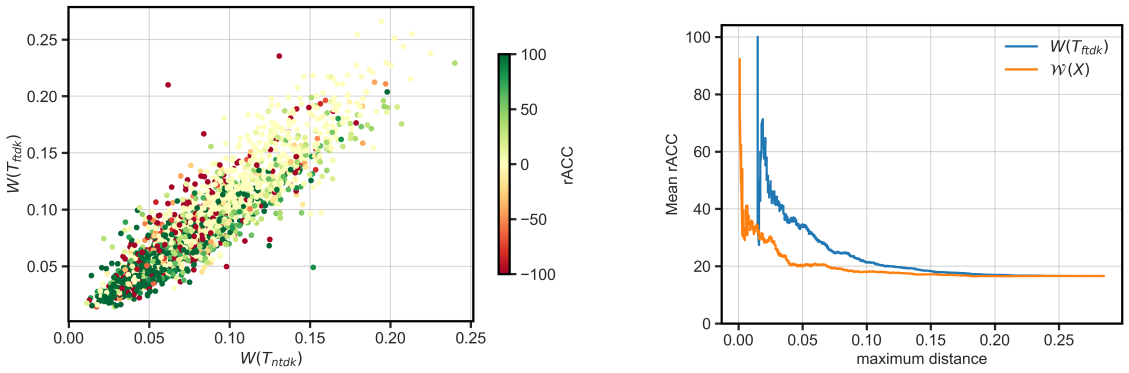



Figure 6.13: The scatter-plot (left) shows the relative gain in accuracy $rACC$, with a greener dot indicating a greater gain derived from the full target domain knowledge (*ftdk*) relative to the no target domain knowledge (*ntdk*). The x- and y-axis, respectively, shows the covariate shift measured by the Wasserstein distance between the source-target domain pairs used for a decision tree grown in the *ntdk*, $W(T_{ntdk})$, and in the *ftdk*, $W(T_{ftdk})$, scenarios. It shows that a greater gain in accuracy from access to the full target domain knowledge is achieved when the covariate shift assumption is (strictly) met. Similarly the plot (right) shows how model performance (mean $rACC$) deteriorates as the covariate shift assumption is relaxed (shown by a larger Wasserstein distance).

The above definition quantifies the recovered loss in accuracy also in the case that $ACC_{ntdk} > ACC_{tt}$, which may occur by chance. Moreover, to prevent outliers due to very small denominators, we cap $rACC$ to the $-100$ and $+100$ boundaries. The mean value of $rACC$ over all source-target pairs is $16.6$, i.e., on average our approach recovers $16.6\%$ of the loss in accuracy. However, there is a large variability, which we examine further in the next section.

***Case 3: partial target domain knowledge.*** We reason on partial target domain knowledge under the assumption that we only know an estimate of the distribution of some subsets of $\mathbf{X}$'s but not of the full joint probability distribution $P_T(\mathbf{X})$. We experiment assuming to know $\hat{P}_T(\mathbf{X}')$ for $\mathbf{X}' \subseteq \mathbf{X}$, only if $|\mathbf{X}'| \leq 2$ (resp., $|\mathbf{X}'| \leq 3$). Equivalently, we assume to know $\hat{P}(X = x|\varphi')$ only if $\varphi'$ contains at most one (respectively, two) variables.

Formulas (6.17)–(6.18) mix such a form of target domain knowledge with the estimates on the source: for $\hat{P}(X = x|\varphi)$, we compute $\varphi'$ as the subset of split conditions in $\varphi$ regarding at most the first (resp., the first two) attributes in $\varphi$ – namely, the attributes used in the split condition at the root (resp., at the first two levels) of the decision tree, which are the most critical ones.

The weight $\alpha$ in (6.17)–(6.18) is set dynamically as the proportion of attributes in $\varphi$ which are not in $\varphi'$. This value is $0$ when $\varphi$ tests on at most one variable (resp., two variables), and greater than $0$ otherwise. We consider the proportion of attributes and not of the number of split conditions, since continuous attributes may be used in more than one split along a decision tree path.

**Covariate shift and accuracy.** We test whether the difference in model performance is due to the fact that different pairs match or do not match the covariate shift assump-

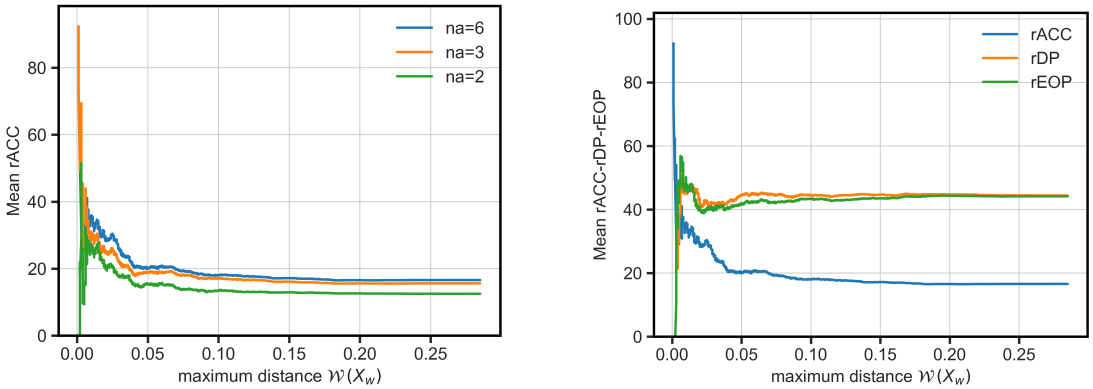



Figure 6.14: The plot on the left shows results of DADT, across all state pairs, with partial target domain knowledge; we show the mean $rACC$ for pairs with bounded $\mathcal{W}(X_w)$ for the cases of having knowledge of $na = 6$ attributes (i.e., *ftdk*), $na = 3$, and $na = 2$. The plot on the right shows the change in relative demographic parity, relative equalized odds and accuracy over bounded $\mathcal{W}(X_w)$ in the full target domain knowledge scenario.

tion. To quantify the covariate shift (issue P2), we define for a decision tree $T$:

$$W(T) = \sum_{\varphi \text{ path of a leaf of } T} W(\hat{P}(Y|\varphi), \hat{P}_T(Y|\varphi)) \cdot \hat{P}_T(\varphi) \tag{6.25}$$

as the average Wasserstein distance between the estimated (through (6.21)) and target domain class distributions at leaves of the decision tree, weighted by the leaf probability in the target domain.

Notice that, as $P_T$ is unknown, we estimate the probabilities in the above formula on the test set of the target domain. We write $W(T_{ntdk})$ and $W(T_{ftdk})$, respectively, for denoting the amount of covariate shift for the decision tree grown in the no target domain knowledge and with full target domain knowledge scenarios.

The scatter plot Figure 6.13 (left) shows the relative accuracy (in color) at the variation of $W(T_{ntdk})$ and $W(T_{ftdk})$[26]. We make the following qualitative observations:

- when $W(T_{ftdk})$ is small, say smaller than $0.05$, i.e., when the covariate shift assumption holds, the relative accuracy is high, i.e., using target domain knowledge allows for recovering the accuracy loss;
- when $W(T_{ftdk})$ is large, in particular, larger than $W(T_{ntdk})$, then the gain is modest or even negative.

Let us consider how to determine quantitatively on which pairs there is a large relative accuracy. Figure 6.13 (right) reports the mean $rACC$ for source-target pairs sorted by two different distances. Ordering by $W(T_{ftdk})$ allows to identify more source-target pairs for which our approach works best than ordering by the average class conditional distance $\mathcal{W}(X_w)$, where $X_w$ is from (6.22). However:

[26] $W(T_{ntdk})$ and $W(T_{ftdk})$ appear to be correlated. While they are specific of their respective decision trees, they both depend on the distribution shift between the source and target domain.

- $W(T_{ftdk})$ requires target domain knowledge on $P_T(Y|\varphi)$ for each leaf in $T_{ftdk}$, which is impractical to obtain.

- $\mathcal{W}(X_w)$ is easier to calculate/estimate, as it regards only the conditional distribution $P_T(Y|X)$. The exact knowledge of which attribute is $X_w$ is not required, as, by definition of $X_w$, using any other attribute instead of $X_w$ provides an upper bound to $\mathcal{W}(X_w)$.

In summary, Figure 6.13 (right) shows that DADT is able to recover a good proportion of loss in accuracy, and it provides a general guidance for selecting under how much the covariate shift assumption can be relaxed. Finally, Figure 6.14 (left) contrasts the $rACC$ metric of the full target domain knowledge scenario to the two cases of the partial target domain knowledge scenario when we have knowledge of only pairs or triples of variables. There is, naturally, a degradation in the recovery of accuracy loss in latter scenarios, e.g., for a distance of up to $0.03$, we have the mean $rACC$ equal to $25.3\%$ for full target domain knowledge, to $21.6\%$ when using triples, and to $17.5\%$ when using pairs of variables[27]. Even with partial target domain knowledge in the form of cross-tables, we can achieve a moderate recovery of the loss in accuracy.

**Results on fairness.** We now address experiments on RQ2 (Section 6.3.2). Other quality metrics beyond accuracy can degrade in presence of covariate shift. There is also a risk that certain demographic groups are more impacted by drops in accuracy than others. This issue can occur even if an overall minimal accuracy drop is seen.

To test the impact of DADT on specific groups, and answer RQ2, we utilize two *fairness metrics* commonly used in fair machine learning literature, *demographic parity* and *equal opportunity*. We consider here the fairness metrics in reference to the protected attribute SEX. We study how DADT compares to the standard decision tree under the same post-processing fairness step. We hypothesize that under a DA scenario, said step is more impactful under a DADT classifier as it can account for the target domain information during training.

From Section 2.3, Chapter 2, recall that *demographic parity* (DP) quantifies the disparity between predicted positive rate for men and women:

$$DP = |P(\hat{Y} = 1|\text{SEX=women}) - P(\hat{Y} = 1|\text{SEX=men})|$$

The lower the $DP$ the better is the fairness performance. We consider this metric in the context of our women's football shoes example, where one measure of whether a model is addressing our identified feedback loop is whether the positive rate for the question of "will buy football shoes" moves towards parity for men and women. Similarly, recall that *equal opportunity* (EOP) quantifies the disparity between true positive rate for men and women:

$$EOP = |P(\hat{Y} = y|\text{SEX=women}, Y = 1) - P(\hat{Y} = 1|\text{SEX=men}, Y = 1)|$$

[27]Notice that the extension of $rACC$ to partial target domain knowledge is immediate by replacing $ACC_{ftdk}$ in its definition with the accuracy $ACC_{ptdk}$ of the decision tree grown by using partial target domain knowledge.

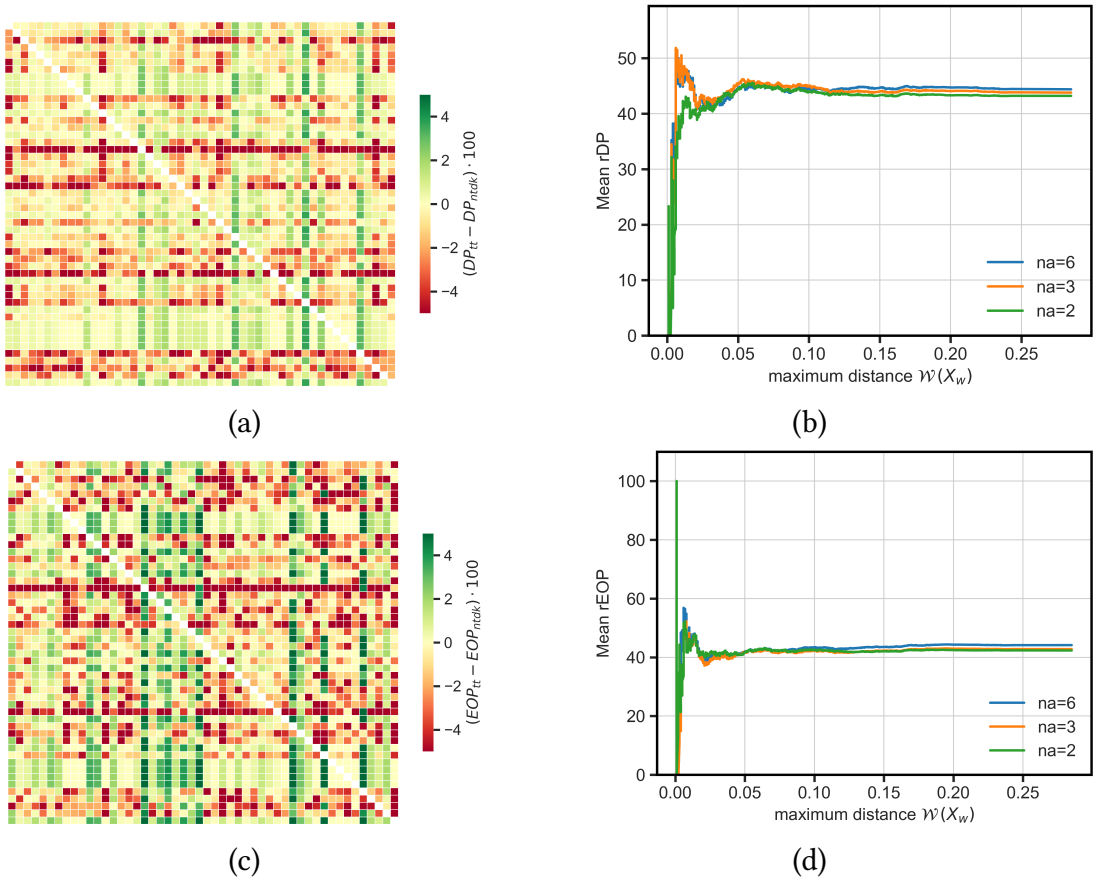



Figure 6.15: The heatmap (a) shows the difference in DP between the cases target-to-target baseline $DP_{tt}$ and no target domain knowledge $DP_{ntdk}$ for each source-target US state pair. Similarly, (c) for EOP. The plots (b) and (d) show the mean $rDP$ and $rEOP$ respectively, for the cases of having knowledge of $na = 6$ attributes (i.e., *ftdk*), $na = 3$, and $na = 2$, over bounded $\mathcal{W}(X_w)$.

We consider this metric in the context of the example of the public administrator who is identifying people who do not receive benefits to which they are entitled to. Here, our concern is that the model is equally performant for all groups as prescribed by SEX.

Fairness-aware classifiers control for these metrics. We use here a classifier-agnostic post-processing method by Hardt et al. [131] that specializes the decision threshold for each protected group. The correction is applied after the decision tree is trained. Figure 6.15 (a) confirms a degradation of the DP metric from the target-to-target scenario to the no target domain knowledge scenario. Figure 6.15 (c) shows a less marked degradation for the EOP metric.

We mimic the reasoning done for the accuracy metric and introduce the *relative gain in demographic parity* ($rDP$) and *the relative gain in equal opportunity* ($rEOP$):

$$rDP = \frac{max(DP_{ntdk}, DP_{tt}) - DP_{ftdk}}{|DP_{tt} - DP_{ntdk}|} \cdot 100 \tag{6.26}$$

$$rEOP = \frac{max(EOP_{ntdk}, EOP_{tt}) - EOP_{ftdk}}{|EOP_{tt} - EOP_{ntdk}|} \cdot 100 \tag{6.27}$$

where note that, since DP and EOP improve when they become smaller, the definitions of relative gain are symmetric if compared to the one of $rACC$.

Figure 6.14 (right) substantiates also for $rDP$ and $rEOP$ the conclusions for $rACC$ mentioned in for the accuracy results. The distance $\mathcal{W}(X_w)$ provides a guidance on when DADT works the best. For DP and EOP, however, for large values of such a distance, we do not observe a degradation as in the case of ACC. In other words, when the assumption of covariate shift is strictly met, DADT works the best (relative to the post-processing step), but when it is not, the recovery of the DP and EOP does not degrade.

Finally, Figure 6.15 (b) confirms the degradation of the DADT performances in the case of partial target domain knowledge. E.g., for a distance of $0.03$ we have that the mean $rDP$ equal to $41.8\%$ for full target domain knowledge, $42.2\%$ when using triples, and $40.8\%$ when using pairs of variables. This result is much less marked for $rEOP$, for which DADT performs very well also with knowledge of pairs of variables, as shown in Figure 6.15 (d).

### 6.3.5 Conclusion

In answer to RQ1 and RQ2 in Section 6.3.2, we see that domain-adaptive decision trees (DADT) result in both increased accuracy and better performance on fairness metrics over our baseline standard decision tree trained in $D_S$ and tested in $D_T$.

Looking more closely at our experimental results, we see that the improvements are best when the covariate shift assumption holds in, at least, a relaxed form (P2). We also see this increase when we only have partial domain knowledge (P1), though a greater amount of domain knowledge, as we define it, results in greater improvements in those metrics. Interestingly, our post-processing fairness intervention does not have a worse performance over a standard decision tree even when the covariate shift assumption does not hold.

Back to the example inspired by the experimental setting in Section 6.3.4, we have demonstrated that DADT are an effective method for using existing information about a target state. We can also think back to our retail example in Section 6.3.1, wherein we identified a potential feedback loop leading to a lack of stock in women's football shoes. We propose that DADTs are a method for intervening on this feedback loop; if the store identified a pool of potential customers (such as the population living near the store), which had a higher rate of women than their existing customer base, DADT provides an accessible, interpretable, and performative classification model which can incorporate this additional information. In future work, different definitions of outside information should be explored as the outside information may not have the same structure as the source and target datasets.

While we see that the benefits are clear, we want to be explicit about the limitations of our method. Firstly, we show that DADT is most effective when the covariate shift assumption holds. We consider a strength of our work that we specify and test this assumption and encourage future work on domain adaptation methods to similarly specify the conditions under which a method is suitable to be used. Secondly, we emphatically acknowledge that DADT are not intended as a replacement for collecting updated and improved datasets. However, our solution is a low cost improvement that can be made over blindly applying to a new or changing context. Additionally, there are cases in

which labelled data simply does not exist yet. Finally, DADT are not a complete solution for achieving or ensuring fair algorithmic decision making; rather they are an easy to use method for improving accuracy, and fairness metric performance in the commonly occurring case of distribution shift between source and target data.

# Chapter 7

# Final Discussion

> Unlike correlation and most of the other tools of mainstream statistics, causal analysis requires the user to make a subjective commitment. She must draw a causal diagram that reflects her qualitative beliefs—or, better yet, the consensus belief of researchers in her field of expertise—about topology for the causal processes at work. She must abandon the centuries-old dogma of objectivity for objectivity's sake. Where causation is concerned, a grain of wise subjectivity tells us more about the real world that any amount of objectivity.
>
> *The Book of Why* by Pearl and Mackenzie [220, p. 89]

We finish with this final chapter. First, we discuss the contributions of this thesis. Second, we discuss the general limitations of this thesis while provide a position/warning on taking causality too seriously. Third and finally, we present potential future work based on the topics covered in this thesis.

## 7.1 Contributions

In the previous chapters, I introduced causality for Fair ML (Chapter 2); discussed and experimented with its implications for testing (algorithmic) discrimination (Chapters 3 and 4); explored its role for formalizing the problem of perception (Chapter 5); and studied its use for understanding unrepresentative data through two common data science applications (Chapter 6). In doing so, I addressed the three research questions posed in Chapter 1, which I discuss below.

***Q1: How can we use causal reasoning to test for discrimination so that we capture the role of protected attributes, such as race and gender, on the other seemingly neutral attributes that are used for the decision-making process?*** From the start of this work, I viewed causal reasoning as a necessary tool for understanding discrimination and wanted to expand Kohler-Hausmann [176]'s critiques on the counterfactual model of discrimination from a casual analysis perspective. My contributions to this question are Chapters 3 and 4.

With Chapter 3, by distinguishing for the first time two kinds of discrimination com-

parators (the cp and mm comparators), I offer two ways for testing discrimination under casual reasoning based on how the protected attribute influences other attributes. The mm-comparator represents a formal move away from the standard cp-comparator that Kohler-Hausmann [176] criticizes. I make the case for why the counterfactual model of discrimination is used and why we do not have a better option. I also present a causal desiderata for future discrimination tools. The comparators I defined illustrate the normative choice behind testing for discrimination. This is a choice that modelers are not exempt from as it is represented by how we decide to model similarity. Moving forward, the mm-comparator (and the overall notion of *mutatis mutandis*) is a new and viable option given the available Fair ML methods when testing for discrimination.

With Chapter 4, I present counterfactual situation testing (CST) as a new framework for testing (algorithmic) discrimination and study it under a k-NN implementation. CST implements the "fairness given the difference" notion, which is a statement on how the protected attribute's downstream effects affect the non-protected attributes used by the decision-maker. CST, thus, carries out the counterfactual model of discrimination but in a way that it is aware that when the complainant's protected attribute changes, other attributes should change as well, which conditions how the comparator is derived. CST merges the earlier attempts of discrimination discovery [244] with recent Fair ML efforts based on SCM like Kusner et al. [177]. The goal was to develop a tool that would be appealing to both Fair ML researchers and lawyers/regulators. CST explicitly implements the mm-comparator, showing that it detects more cases of individual discrimination than the standard cp-comparator used by Thanh et al. [277]'s k-NN situation testing implementation. These results are not surprising (the mm-comparator is much more flexible than the cp-comparator) and show what it would mean to implement substantive equality [288] goals for discrimination testing. Further, I show that a counterfactually fair (algorithmic) decision-maker can be discriminatory, which is a setting that we have not yet considered within Fair ML. Furthermore, following Xenidis [308], I provide evidence for how multiple discrimination fails to account for intersectional discrimination, where, recall, only the former is considered under EU non-discrimination law.

***Q2: How can we use causal reasoning to formalize scenarios where fairness is, essentially, subjective as in dependent on who is making the decision?*** My contribution to this question is Chapter 5 in the form of the causal perception (CP) framework. Perception occurs when two individuals (or agents) interpret the same information differently. It has been largely studied by psychologists [279, 280, 281], but overlooked by the Fair ML community. Perception is, by definition, a form of subjective knowledge as objective knowledge can only be interpreted one way. With this chapter, starting from the premise that causal reasoning best describes human reasoning [220], I propose a SCM formalization of perception. I define two kinds of CP, unfaithful and inconsistent, and, in doing so, rephrase the violations of the faithfulness [224] and consistency [241] causal properties as a function of who (or what) is interpreting the knowledge. With CP, I implement, literally, the view that causality can be subjective. I then revisit fairness as a subjective notion due to the occurrence of perception.

Fairness problems are a natural implementation for CP since fairness can mean different things to different stakeholders. CP provides a framework able to operationalize the subjectivity of fairness and to account for it using causal reasoning. Overall, CP has considerable Fair ML implications once we consider human-centered AI pipelines

in which the a set of human decision-makers interpret the information provided by a learned model. As I argue, even under a fair model, perception can lead to disagreement on the fairness of the setting or even to distinct decisions with equal weight that are at odds with each other in terms of fairness. CP offers a move from the single, objective view dominating Fair ML problems to a partial, subjective view.

***Q3: How can we use causal reasoning to mitigate the potential bias in a learned model from using an unrepresentative sample as training data?*** My contribution to this question is Chapter 6 in the form of two common data science applications modified, respectively, to account for unrepresentative training data.

In Section 6.1, I use SCM to formulate the problem of unrepresentative data and present the problem under the lens of sample selection bias and domain adaptation. Using causal analysis, I am able to focus on covariate shift and justify the proposed modifications to, respectively, partial dependence plots [102] and decision tree learning [133] in the form of representative weights to the training data's and incoming test data's covariate space. Under covariate shift, the source of bias comes only from the predictive variables, which allows us to tackle the supervised learning setting where we have potentially learned a biased model and plan to use it on incoming, unlabeled data that follows a different distribution from the training data.

With Section 6.2, I introduce the weighted partial dependence plot (WPDP), which is a modified version of the standard PDP. In WPDP, the instances used for drawing the plots are weighted to account for a possible sample selection bias. Additionally, based on Zhao and Hastie [323], I use SCM to interpret the WPDP causally. The WPDP is the first weighted version of the PDP. With Section 6.3, I introduce domain adaptive decision trees (DADT). I first define the decision tree learning problem under domain adaptation, focusing on the information gain split criterion. I then propose a modification to such criterion, resulting in the domain adaptation version of the decision tree learning problem. The modification proposes weighting the entropy gain for each instance being considered at the time of learning a split. The DADT is the first treatment of decision trees under domain adaptation. Both modified data science applications are implemented, tested, and evaluated using real world datasets.

## 7.2 Challenges and Limitations

In terms of challenges, the main one was drawing a common narrative for causality given its range across multiple fields. Causality, as illustrated through this work, covers fields such as Philosophy [303], Law [176], Economics [16], and Computer Science [255], all of which contribute to Fair ML. Each field has its own treatment of and use for causality. For instance, although both philosophers and computer scientists use SCM, the former rely on logic-based cases (see, e.g., [29]) while the latter on ML-based cases (see, e.g., [286]). When speaking of causality in general, even within Fair ML, it is difficult to reduce the discussion to a common narrative. Coming myself from an Economics background, my views on causality were based on the potential outcomes framework [16] and embracing Pearl [218]'s SCM was a slow process. I also struggled with moving between the two modeling cultures [49], often falling in between the inferential and prediction driven problem formulations.

In this work, I have tried to create a common narrative, at least, useful enough for a discussion on causal Fair ML problems like discrimination. My approach from the beginning has been to focus on the problem first, like discrimination, and then move backwards among the multiple fields using causality for that problem. Hence, why this thesis relies on works that do not agree with each other, such as Hu and Kohler-Hausmann [149], Kohler-Hausmann [176] versus Schölkopf [254], Woodward [303].

Another challenge faced, mainly toward the ML literature, was coming to terms with the usage of causality within the current deep learning hype. I am fully aware of the technical limitations or, more concretely, of the technical simplicity behind most of the work presented in this thesis. Besides the range of ML models used in Section 6.2, which includes neural networks, I mainly used traditional ML models. This is in part due to my educational and professional background, which preceded the deep learning boom. This is also due, though, to the problems tackled in this thesis, which often do not require such complex models.[1] Most ADM problems involving tabular data have few variables. Further, causal analysis for tabular data mainly focuses on issues around agreeing or not with the SCM given. Of course, I do see the advantage of deep learning models (mainly in terms of relaxing model-specification assumptions and for generating counterfactual distributions under causal insufficiency; see, e.g., Javaloy et al. [154], Zhao et al. [324]), and I plan to work more with these models in the near future.

Overall, what I found and still find challenging with mainstream causal ML literature are its "objective" claims. I struggle with works that take causality too seriously. To me, causality for ADM involving humans represents a useful tool for expressing domain knowledge. Its main purpose is to declare information useful to the ML problem. In that sense, causality is inherently subjective: the causal models used describe a point of view for or an informed take on a ML problem. And I am fine with that. The feeling I get from causal ML these days, mainly from works relying on deep learning methods, is that we have somehow arrived at the missing piece for ultimate AI [254, 255, 295]. This might be true for causal ML for Physics or Chemistry applications, but I am skeptical with works that claim similar results for human behavior. As a field, causal ML needs to be more humble and ware of its limitations to remain useful, especially for Fair ML purposes.

Similarly, within Fair ML, it is challenging to discuss causality beyond the modeling camp. I would argue that this challenge is mostly due to Hu and Kohler-Hausmann [149], which is a popular work within Fair ML. Although I too view it as an important work for causal Fair ML, I disagree with many of its claims against using causal knowledge to discuss fairness. Mainly, Hu and Kohler-Hausmann [149] reduce counterfactual fairness [177] to a *ceteris paribus* causal intervention (recall, $A$ changes but $\mathbf{X}$ remains the same) when it often is a *mutatis mutandis* causal intervention (under the right causal structure, when $A$ changes $\mathbf{X}$ changes too). Later works, like Kasirzadeh and Smart [164], take this statement for granted and further criticize counterfactual fairness for performing idealized comparisons between the factual and counterfactual worlds. This statement is not true as shown by the results in Chapter 4, though still it has managed to make counterfactual fairness and SCM into a niche topic within Fair ML. I would argue that Hu and Kohler-Hausmann [149] and subsequent fairness works fail to grasp how counterfactual generation works under a SCM.

[1] Here, I have a similar take to Grinsztajn et al. [114] that show that decision trees still outperform deep learning models for tabular data.

There is this overall prevalent narrative against causality within Fair ML because it flattens the meaning of variables into nodes and arrows as argued by Hu and Kohler-Hausmann [149]. I am not saying this is not a valid point, but this point is also true for all quantitative and qualitative approaches to science.[2] Such prevalent narrative should bring us back to Cox's "all models are wrong, but some are useful" saying, instead of forcing us to discard completely causal modeling for Fair ML. Although not discussed in Hu and Kohler-Hausmann [149], Pearl himself (see, e.g., his quote at the start of this chapter) points out at these same limitations of SCM and stresses the subjective role of SCM. In short, it has been challenging to work on fairness problems that clearly require auxiliary causal knowledge under certain environments within the Fair ML community. As a field, we need to consider that Fair ML is still a ML problem; we cannot keep targeting the modeling camp just for the sake of it.

Now, beyond the limitations discussed in each chapter, this work is clearly limited to an interventionist account of causality [303] and, in turn, to a SCM implementation of this view on causality [218]. I did not consider other accounts of causality given that the interventionist account is the one used within ML. Similarly, for the same reason, I did not consider other implementations of the interventionist account beyond SCM. I recognize, though, that more ML works are relying on potential outcomes (e.g., [72]) and that said implementation could be useful for the topics covered in this work.

This work is limited to the European context, in particular, to the EU context. This choice was deliberate given my location and funding through the NoBIAS ITN. Objectively, though, the EU is at an advantage relative to other countries/regions in terms of regulating AI and ADM. The AI Act [92] is a clear example of this trend. Focusing on the EU makes it more likely that the methods proposed here go beyond academic circles. A consequence of this focus, however, is that most of the discussions in this work are suitable to EU non-discrimination law, which aims for substantive equality [288], and do not necessarily translate to other regions.

Finally, in line with the challenges previously discussed, I have limited myself to exploring mainly standard ML models for implementing this work. I acknowledge the growing focus of deep learning models on causality (e.g., causal normalizing flows [154]) and the role they will play in modeling Fair ML problems. This limitation is mainly due to the timing of this work. Under more time, I would have liked to explore further the causal works presented here under deep learning methods.

To conclude this section, now a few word on causality for Fair ML. I believe that causality is and will continue to be important for Fair ML. I also believe that causality is a loaded term. After all, it is a strong claim to say that $X$ is a cause of $Y$ and, similarly, that $Y$ is an effect of $X$, especially when referring to humans. Even historians, who enjoy the benefit of hindsight, tend to disagree on the causes of the effects (or, similarly, the effects of the causes) that took places decades (or centuries) ago.

Causality denotes, depending on the audience, the presence of some sort of universal structure, representing the culmination, if you will, of objective thought. This sort of objectivism, although ideal and desirable for Fair ML and overall ML research, rarely holds when modeling humans. This might be the only universal (i.e., transportable) characteristic of any causal analysis: that it is context-specific.

In that sense, works such as those by Kohler-Hausmann [176] and Hu and Kohler-

[2] Jorge Luis Borges's *On Exactitude in Science* captures this beautifully.

Hausmann [149] have been correct in going after the use of causality for answering questions on discrimination, overall fairness, and similar themes. With some exceptions [8, 177, 246], most works using causal reasoning for these topics and applications show a lack of awareness, going straight to defining (or discovering) the cause-effect pairs of a Fair ML problem without acknowledging the subjectivity of the task at hand. Kilbertus et al. [167] and Plecko and Bareinboim [226] are clear examples of this practice.

Pearl himself, as the quote at the start of this chapter shows, is a strong proponent of causality as a form of subjective knowledge representation. Moving forward, as a field, we need to keep in mind that causal models represent both our greatest strengths and weakness when addressing Fair ML problems. Here, *causality is useful not because it allows to communicate universal truths to ML models, but because it allows to communicate our "truths" to ML models*, serving as a counterpart to the purely data-driven methods used by researchers and practitioners of ML.

## 7.3 Prospects

We covered several topics and applications in which causality is useful for Fair ML. Here, we propose and argue for future research directions involving causality, ML, and fairness with a focus on ADM systems. Naturally, it is not an exhaustive list. Each direction summarizes work of interest to the Fair ML community.

**Causal graphs as participatory objects.** As discussed in Chapters 3 and 4, establishing discrimination requires inputs from multiple stakeholders. In practice, this setting implies the risk of excluding certain stakeholders if the way in which discrimination is tested is too abstract or unfamiliar, which is a likely risk when using statistical learning models. Further, this setting implies the risk of having more than one valid view on the discrimination context, leading to perception as discussed in Chapter 5, which can lead to different fairness results.

Pipelines for establishing discrimination must be aware of and account for this participatory setting. Given the prevalence of causal reasoning in testing for discrimination, we believe that causal graphs and the structural causal models they imply can help us meet this participatory goal. Causal graphs are intuitive and, as argued by, e.g., Mulligan [206], can empower stakeholders by allowing them to discuss through these graphical objects the assumptions being made about the discrimination case. The fairness community, which has taken a predominantly strong stance against causality and causal graphs [149, 164, 176], should view them instead as useful participatory tools and study their impact via *participatory design* frameworks [96, 259, 294].

The causal interpretation of the partial dependence plot (PDP) (Section 6.2 in Chapter 6) [178, 182, 323] is another example of how causal graphs enable stakeholder participation within a ML problem. By taking a participatory view on the PDP, it would be interesting to study how different stakeholders interpret the same black-box model causally, especially when the stakeholders themselves can each stipulate how the input variables to the black-box model relate causally to each other: i.e., by letting each stakeholder draw his or her own causal graph.

**Substantive equality goals: algorithms as societal shapers.** Wachter et al. [288]

argues that ML models can be classified into "bias preserving" and "bias transforming" in which the former refers to models that assume and preserve the status quo (i.e., formal equality) while the latter refers to models that view the status quo as biased and aim to change it (i.e., substantive equality). Wachter et al. [288] also argues, at least for the EU, that non-discrimination law should implement substantive equality goals, meaning that the law should test for decisions that fail to transform the non-neutral status quo. Similar arguments, in particular Hardt et al. [131]'s critique of statistical parity through the introduction of equal opportunity, have hinted at the ML model's inherent capacity to perpetuate patterns from the past, in turn, sustaining the status quo. See also, e.g., Kim and Hardt [171].

We can all probably agree on substantive equality as a desired goal, but to make such a goal useful (read, realistic) we also need to agree on ways in which we can formalize and operationalize it. Similar to the discussions on the long-term effects of fairness interventions and how what might seem fair today may lead to unfair outcomes in the near future (see, e.g., [148, 257, 322]), substantive equality is a subjective gamble on what the future is or should be. Hence, to speak of substantive equality meaningfully we need to be able to imagine it in a tangible way: i.e., we need to model it. We believe that counterfactual reasoning is well suited for that purpose.

Recent work on treating decisions by ML models as societal interventions, in particular, the concepts of *performative predictions* [130] and *counterfactual risk assessment* [72] show how causal reasoning can be used for tuning the present to reach a certain outcome in the future. The mm-comparator, defined in Chapter 3 and used for the counterfactual situation testing in Chapter 4, for instance, could reflect some sort of long-term fairness/substantive equality goal: the difference in "fairness given the difference" would reflect the difference between today's status quo and where we, as a group of stakeholders, would want that status quo to be in the near future.

Similar to the first prospect, the objective of defining a better status quo and agreeing on how substantive equality looks like is also a participatory process. We should be open to modeling a set of viable future outcomes with their own status quo, which would position causal perception from Chapter 5 as a useful framework for modeling substantive equality goals across a set of heterogeneous stakeholders.

**Testing for algorithmic discrimination in a changing society.** What if, when testing for algorithmic discrimination, we were to have access to the ML model's training data? Given the nature of ML models, the average algorithmic decision-maker classifies new instances based on patterns from past instances. Hence, if the incoming data is representative of the training data, then we should be able to use the training data to compare incoming (unlabeled) instances to "similar" training (labeled) instances. Under this setting, tools like counterfactual situation testing from Chapter 4 could test for algorithmic discrimination based on both $Y$ and $\hat{Y}$.

Now, under incoming data that is not representative of the training data, which is likely the case given how, e.g., a model can be trained under one setting and used under another completely different setting (for instance, train the model using Italian customer data to then use the model on Portuguese customers), the causal problem of unrepresentative data from Chapter 6 enters the wider problem of discrimination testing. Therefore, if we have access to the training data, then through *domain adaptation* and/or *sample se-*

*lection bias techniques* we could modify the training data and use it to pair the incoming unlabeled data with labels from similar past instances in the training labeled data.

Overall, we need to consider to what extend is the discrimination claim applicable to the ML model when the incoming data is not representative of the ML model's training data. So far the legal works on algorithmic discrimination [122, 288, 296, 308] have considered, broadly, the setting in which a model is biased but have yet to address the implications of the bias being due to deploying the model on the wrong population. Would the ML model still be liable for discrimination if that were the case? Similarly, Fair ML works [8, 41, 222, 277] have yet to consider the case of incoming data that is unrepresentative of the training data used for learning the algorithmic decision-maker being audited.

**On individualized justice, or removing the comparator.** Binns [40] argues that, at the conceptual level at least, there is no conflict between individual and group fairness. According to Binns [40], both families of definitions are rooted in the notions of *consistency* and *egalitarianism* and tackle the same problem. Binns [40] argues that such debate would end if, instead of asking what kind of fairness definition to use, we would ask "what kind of injustice do we believe may be in operation in this context that may reflect in and perpetuated by the model being used?" As we already argued in Chapter 2, this distinction of individual versus group fairness is sometimes self-defeating and at odds with, e.g., the idea of discrimination in which an individual can make a claim but his or her claim will be judged by his or her membership to a group.

In the spirit of moving beyond this group-vs-individual-fairness debate, Binns [40] discusses the notion of *individualized justice*, which is another of Aristotle's maxims that is at odds with the one motivating individual fairness, i.e., the "like cases" maxim. Under individualized justice, essentially, the complainant is all the evidence we need; we are not required to provide for a comparator. Here, discrimination becomes non-comparative. Hence, we judge the merits of the complainant's discrimination case based only on his or her profile.

Individualized justice is neither a group nor an individual fairness concern and, in practice, removes the role of the comparator when testing for the unfairness of a decision. It may also, as argued by Binns [40], potentially remove the appeal of developing a ML model for ADM to begin with: if we are willing to carry out a pure individual-by-individual case, we should then be less inclined to use a ML model that aims to generalize individual instances from group-wise information. Individualized justice presents an interesting prospect for algorithmic discrimination, one that would force us to reconsider what was discussed in Chapters 3 and 4 as well as what we often discuss within Fair ML.

Further, individualized justice would position perception at the center of discrimination testing. If we were to remove the comparator and, in turn, the requirement for gathering evidence of *prima facie* discrimination, then all judgment of the individual case of the complainant would fall on whoever (or whatever) is judging the specific individual case of the complainant. Causal perception from Chapter 5 could play an important role within this non-comparative approach for discrimination testing.

**Protected attributes: equally protected, but unequally created.** Finally, within Fair ML, at least from the modeling camp, we have a tendency of treating all protected

attributes alike. This thesis, e.g., uses $A$ to refer to a protected attribute and presents solutions applicable, in principle, to any $A$.

What is interesting to Fair ML and often overlooked, though, is how protected attributes can have different behavioral effects on their own members based on the social expectations surrounding and/or associated to each protected attribute. It becomes even more interesting when we consider the intersection of multiple protected attributes and how that intersection in-itself creates another groups with its own characteristics, social expectations, and behavioral norms [74]. Concretely, what we are hinting at here is at the possibility that $A$'s causal effects (and the Fair ML problem's general causal structure) on the non-protected attributes $\mathbf{X}$ is conditional on the type of $A$. Hence, defining $A$ as gender or race would change, modeling-wise, the problem and the proposed solution. At the moment, this is not the case within Fair ML since we always define the problem in general terms as though the solution applied for any kind of $A$. Maybe that is the case and we do not need to worry further in terms of Fair ML problem formulations, but it would be interesting to study the case in which we are wrong.

An example of how a specific $A$ hinders what we see through $\mathbf{X}$ is formalized by Austen-Smith and Fryer [22] and tested by Fryer Jr and Torelli [104]. Based on ethnographic studies, these works hypothesize on the impact of "acting white" in the case of black students and their school performance, highlighting the role of culture and social expectations centered around the protected attribute race. Briefly, in the context considered by these two papers, black students that performed well in class were deemed as "white" or "acting as white" by other black students. The average black student had to choose between school performance at the expense of being ostracized by their own racial group or validate group membership at the expense of a better future through education. As a result, depending of course on the school's racial composition and the aptitude of the student (see Austen-Smith and Fryer [22], Fryer Jr and Torelli [104] for details), some black students unperformed on purpose in class. Here, the protected attribute race not only represented systematic discrimination (under which we can treat all protected attributes the same), but it also represented some sort of identity badge that conditioned the personal choices of its members (which would be harder to define as discriminatory given that it is up to the individual).

This ambivalence between what is imposed over the individual and what the individual imposes over his or herself because of the protected attribute $A$ offers an interesting prospect for Fair ML. The first part could be seen as the goal of non-discrimination law: it is clearly unfair that an individual has a negative outcome because of being a member of $A = a$. This setting is the one used for most, if not all, Fair ML formulations. The second part, however, is less clear: how is it unfair for an individual to have reached a negative outcome because of his or her own choices based on $A = a$?

If we believe in the agency of the individual, then we must separate what falls under the responsibility of the individual from what falls under the responsibility of the society. As individuals, we all should strive for a fairer, more just society and aim for ML models that correct for systematic issues that affect certain groups of individuals. But also as individuals, we all should have some degree of individual responsibility for our own choices and their consequences, granted that the opportunities offered are fairly distributed among all groups of individuals. It is difficult to disentangle where society's responsibility ends and the individual's responsibility beings, granted that it is even pos-

sible to do so as a ML modeling problem. However, ignoring the role of individual choice on individual outcome is dangerous and, to an extent, patronizing.

Heidari et al. [137], e.g., make a similar distinction between an individual's circumstances and an individual's choices and argue that fairness efforts needs to be aimed at the former. We believe that this is an interesting debate that has been largely avoided within Fair ML. Can we speak about unfairness when the problem is due to the individual's choices (i.e., things that he or she had controlled over) or is this discussion restricted only to the individual's circumstances (i.e., things that he or she could not have controlled for)? This tension is intrinsically linked to how protected attributes were formed and are currently used as identity markers, and it opens the possibility of (or, more precisely, the need for) treating each protected attribute, as a modeling problem, by its own effects on individual circumstances and individual choice.

And with these future research directions for causal Fair ML, we conclude this section and, thus, this thesis. This thesis concludes four years of research. It is one attempt among many to show the usefulness of causality for Fair ML, in particular, for addressing the pressing problem of algorithmic discrimination under ADM systems. I have enjoyed this deep dive into causality, what it means, and why it is useful for the Fair ML problems that will inevitably appear in the next decade or so. It is my hope that this thesis adds positively to the relevant literature, even if just in a minor way.

# Appendix A

# Supplementary Material for Chapter 4

## A.1 Additional Experiments

We re-run Section 4.4.1 and 4.4.2 for $\tau = 0.05$, keeping all other parameters in Section 4.4 equal. We do so to check for the robustness of CST. The results align with the ones we present in the main body. Here, we still define individual discrimination as prescribed in Definition 4.3.4.

Table A.1: Number (and %) of detected individual discrimination cases for the illustrative example based on gender.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 288 (16.8%) | 307 (17.9%) | 331 (19.3%) | 360 (21.0%) |
| ST [277] | 0 | 55 (3.2%) | 60 (3.5%) | 75 (4.4%) | 79 (4.6%) |
| CST | 0 | 420 (24.5%) | 309 (18.1%) | 334 (19.5%) | 363 (21.2%) |
| CF [177] | 376 (22%) | 376 (22%) | 376 (22%) | 376 (22%) | 376 (22%) |

Table A.1 shows the same pattern between CST versions relative to ST and CF as in Table 4.1 illustrating the robustness of our framework. Two points we want to raise from Table A.1. First, CF, as expected, detects the same number of cases as it always looks for the strict equality between the factual and counterfactual quantities. In that sense, CST and ST too are more flexible due to a larger parameter space. These methods, e.g., could accommodate for a situation where $\Delta p$ must be larger than some non-zero threshold. Second, under $\tau = 0.05$, CST and CST (w/o) align in the number of cases for smaller $k$ sizes. This shows how influential $\tau$ can be for detecting discrimination, but also shows that either CST version can tackle the discrimination problem.

Table A.2: Number (and %) of individual discrimination cases for the law school admissions scenario based on race.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 256 (7.30%) | 301 (8.59%) | 323 (9.21%) | 376 (10.72%) |
| ST [277] | 0 | 33 (0.94%) | 48 (1.37%) | 57 (1.63%) | 46 (1.31%) |
| CST | 0 | 286 (8.16%) | 301 (8.59%) | 323 (9.21%) | 376 (10.72%) |
| CF [177] | 231 (6.59%) | 231 (6.59%) | 231 (6.59%) | 231 (6.59%) | 231 (6.59%) |

Table A.3: Number (and %) of individual discrimination cases in for the law school admissions scenario based on gender.

| Method | $k = 0$ | $k = 15$ | $k = 30$ | $k = 50$ | $k = 100$ |
|---|---|---|---|---|---|
| CST (w/o) | 0 | 78 (0.82%) | 105 (1.10%) | 224 (2.35%) | 231 (2.42%) |
| ST [277] | 0 | 77 (0.81%) | 92 (0.96%) | 181 (1.90%) | 185 (1.94%) |
| CST | 0 | 99 (1.04%) | 105 (1.10%) | 224 (2.35%) | 231 (2.42%) |
| CF [177] | 56 (0.59%) | 56 (0.59%) | 56 (0.59%) | 56 (0.59%) | 56 (0.59%) |

We observe similar results in Tables A.2 and A.3, which are the $\tau = 0.05$ counterparts of Tables 4.4 and 4.5. Overall, for both experiments, unsurprisingly, the number of cases drops under $\tau = 0.05$ as we have increased the difficulty of proving the individual discrimination claims. Similar to passing from $\tau = 0.0$ to $\tau = 0.05$, the results under the Wald confidence intervals (Definition 4.3.5), would lead to a drop in the the number of discrimination cases.

# Appendix B

# Supplementary Material for Chapter 5

## B.1 The Conjunction Fallacy

For any two random variables $X_1$ and $X_2$, it holds that their intersection cannot be more probable than any of its parts:

$$P(X_1 \cap X_2) \leq P(X_1) \text{ (or } P(X_2)).$$

Here, the intersection denotes the conjunction. We could have also used the $\wedge$ or the comma over $\cap$. At a high-level, conceptually, they all represent *and*. This rule is known as *the conjunction rule* and it comes from the basic laws of probability: *the probability of what is contained cannot be more than the probability of what contains it.* It follows, in fact, from the *extension rule*:

$$\text{If } X_1 \subseteq X_2, \text{then } P(X_1) \leq P(X_2).$$

We state the *conjunction fallacy* as the contradiction to the conjunction rule:

$$P(X_1 \cap X_2) \ > \ P(X_1) \text{ (or } > \ P(X_2)) \tag{B.1}$$

which holds regardless of whether $X_1$ and $X_2$ are independent. As with the conjunction rule, the conjunction fallacy follows for two or more random variables.

## B.2 The Original Linda Problem

Below we present the full version of the Linda Problem as presented by Tversky and Kahneman [280, 281]. The order of the response options is meaningless as these were randomized during the experiments. In the first experiments, 80-90% of participants ranked the conjunction "Linda is a bank teller and is an active feminist" as more probable than its less representative constituent "Linda is a bank teller".

**Example B.2.1** (The Linda Problem)**.** Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

What is more probable today?
*(a)* Linda is a teacher in elementary school.
*(b)* Linda works in a bookstore and takes Yoga classes.
*(c)* Linda is active in the feminist movement.
*(d)* Linda is a psychiatric social worker.
*(e)* Linda is a member of the League of Women Voters.
*(f)* Linda is a bank teller.
*(g)* Linda is an insurance salesperson.
*(h)* Linda is a bank teller and is active in the feminist movement.

The version in Example 5.1.1 is commonly known as the *reduced form* of the Linda Problem. In the latest version of the Linda Problem, Linda is described as an accountant and works at an NGO. See, e.g., Kahneman [158].

Here, we also want to note Tversky and Kahneman [281] attempt at using causal reasoning to formulate the conjunction fallacy. Their formulation, which we present in Figure B.1, was conceptual and mainly intuitive, without using any causal framework. Still, we find it interesting as it clearly draws parallels with the proposed causal perception framework, at least, in terms the role of causality for human reasoning. Our treatment of perception, causality-wise, is much deeper than theirs. Structural causal models (SCM) [218] had yet to become an accepted framework in ML. Figure B.1 reinforces our modeling choice to use causality as discussed in Section 5.2.3.

## B.3 Additional Related Work

With the ML works using Kahneman and Tversky work, the focus mostly has been on developing systems that approximate human-like reasoning to improve over it. We believe that there is a genuine push from AI researchers in using their work to create systems that improve over biased human decision-making. But we also fear that within that same push there is little interest in understating the contextual forces behind these biases, as though the intelligent systems were to be used without humans. We fear in the long term a sort of *Physics envy* [208] by the AI field, as (once) experienced by economics, that struggles to admit that mathematical formulas alone may explain how particles move in a system but are not enough to describe complex human behavior. This paper aims to counteract this narrative by formulating perception in its full complexity and presenting a framework that embraces it as such.

Here, in particular, we cannot ignore the parallels between ML and economics. Kahneman and Tversky's work challenged the rational homo-economicus agent present in economic models. Kahneman received the Nobel Prize in 2002; Tversky had passed away in 1996. This is no minor achievement. Economics, arguably the most influential field in governance since the 1920s [18], is an insular social science [127, 225]. [158] speaking at NeurIPS is a clear sign of the field's interest on topics that seemed out-of-scope for ML decades before.

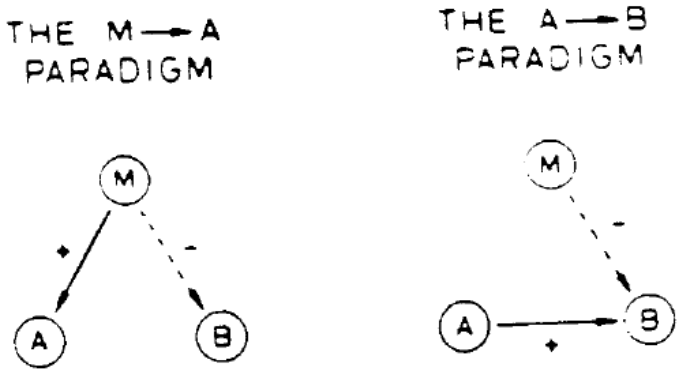



*Figure 1* Schematic representation of two experimental paradigms used to test the conjunction rule. (Solid and broken arrows denote strong positive and negative association, respectively, between the model M, the basic target B, and the added target A.)

Figure B.1: Figure from Tversky and Kahneman [281] regarding *causal conjunctions*. Neither Tversky or Kahneman, to the best of our knowledge developed this causal interpretation of the conjunction fallacy beyond this figure.

## B.4 Additional Discussion for Section 5.3.3

Recall the implementation of the causal relational statement $\phi$ in Definition 5.3.2 discussed for the inconsistent causal perception. We present $\tau_R$ (5.8) under the functional form (5.9) as an aggregation of the associations between the descriptors of a causal-effect pair. Given $X_i \to X_j$ in $\mathcal{G}$, we define the operationalization mapping $\phi$ between $X_i, X_j$ as a function that transforms the *associations* between the descriptors of each of these variables into real-valued causal weights:

$$\begin{aligned}
\phi\big(\Theta^R(X_i) \wedge \Theta^R(X_j)\big) = \\
&= \phi\big(\{\theta_1^{X_i}, \ldots, \theta_n^{X_i}\}, \{\theta_1^{X_j}, \ldots, \theta_m^{X_j}\}\big) \\
&= \tau_R\Big(\overline{\phi}(\theta_1^{X_i} \to \theta_1^{X_j}), \ldots, \overline{\phi}(\theta_n^{X_i} \to \theta_m^{X_j})\Big) \\
&= \overline{\phi}(\theta_1^{X_i} \to \theta_1^{X_j}) + \ldots + \overline{\phi}(\theta_n^{X_i} \to \theta_m^{X_j}) \\
&= \beta_{X_i \to X_j} \quad \big(\text{or } = X_i \xrightarrow{\beta} X_j\big)
\end{aligned} \tag{B.2}$$

where $\overline{\phi}(\cdot) \in \mathbb{R}^{n \times m}$ represents the weights of the associations between descriptors, which can be zero; $\tau_R$ represents that $\tau$-exact transformation of $R$, which we have made into a summation; and $\beta_{X_i \to X_j}$ represents the causal aggregated effect of $X_i$ on $X_j$. To illustrate (B.2), see Figure 5.2.

# Appendix C

# Supplementary Material for Chapter 6: Section 6.2

## C.1 Additional Related Work

Here, we present the other related work themes discussed in Lazzari et al. [178] that relate to the first objective of predicting employee turnover.

*Modeling approaches.* Traditional approaches for testing the determinants of employee turnover have focused largely on statistical significance tests via regression and ANOVA analysis, which are tools commonly used in applied econometrics. See, e.g., [113, 267]. This line of work has embraced causal inference techniques as it works often with panel data, resorting to other econometric tools such as instrumental variables and random/fixed effects models. For a recent example see [140]. For an overview on these approaches see [16].

There has been a recent push for more advanced modeling approaches with the raise of human resource (HR) predictive analytics, where ML and data mining techniques are used to support HR teams [211]. This work falls within this line of work. Most ML approaches use classification models to study the predictors of turnover. See, e.g., [5, 93, 105, 153]. The common approach among papers in this line of work is to test many ML models and to find the best one for predicting employee turnover. However, despite the fact that some of these papers use the same datasets, there is no consensus around the best models. Using the same synthetic dataset, e.g., [5] finds the support vector machine (SVM) to be the best-performing model while [93] finds it to be the naive Bayes classifier. We note, however, that similar to [105] we find the logistic regression to be one of our top-performing models. This work adds to the literature by introducing a new top-performing model to the list, the LightGBM.

Similarly, this line of work does not agree on the top data-driving factors behind employee turnover either. For instance, [5] identifies overtime as the main driver while [105] identifies it to be the salary level. This work adds to this aspect in two ways. First, rather than reporting feature importance on a final model, we do so across many folds for the same model, which gives a more robust view on each feature's importance within a specific model. Second, we go beyond the limited correlation-based analysis [7] by incorporating causality into our feature importance analysis.

Among the classification models used in the literature and from the recent state-of-

the-art in ML, we will experiment with the following models: logistic regression [146], k-nearest neighbor [260], decision trees [51], random forests [50], XGBoost [61], and the more recent LightGBM [165], which is a gradient boosting method [103]. Ensemble of decision trees achieve very good performances in general, with few configuration parameters [80], and especially when the distribution of classes is imbalanced [48], which is typically the case for turnover data. Recent trends in (deep) neural networks are showing increasing performances of sub-symbolic models for tabular data (see the survey [45]). We will experiment with TabNet [20], which is one of the top recent approaches. Implementations of all of the approaches are available in Python with uniform APIs.

*Modeling intent.* A parallel and growing line of research focuses on predicting individual desire or want (i.e., intent or intention) over time using graphical and deep learning models. These approaches require sequential data detailed per individual. The adopted models allow to account for temporal dependencies within and across individuals for identifying patterns of intent. Intention models have been used, for example, to predict driving routes for drivers [263], online consumer habits [292, 293], and even for suggesting email [262] and chat bot responses [256]. Our survey data has a static nature, and therefore we cannot directly compare with those models, which would be appropriate for longitudinal survey data.

*Turnover data.* Predictive models are built from survey data (questionnaires) and/or from data about workers' history and performances (roles covered, working times, productivity). Given its sensitive information, detailed data on actual and intended turnover is difficult to obtain. For instance, all of the advanced modeling approaches previously mentioned either use the IBM Watson synthetic data set[1] or the the Kaggle *HR Analytics* dataset[2].This work contributes to the existing literature by applying and testing the latest in ML techniques to a unique, relevant survey data for turnover intention. Hence, through the analysis of this survey we provide useful information to both employers and policy makers, which allows this work to have a potential policy impact.

## C.2 Predictive Modeling: Results

We recall the models considered. We experiment with interpretable classifiers, namely k-nearest neighbors (KNN), decision trees (DT), and ridge logistic regression (LR), as well as with black-box classifiers, namely random forests (RF), XGBoost (XGB), LightGBM (LGBM), and TabNet (TABNET).

The tables include the AUC-PR (mean $\pm$ standard deviation), the 95% confidence interval of the AUC-PR, and the elapsed time (mean $\pm$ standard deviation), including hyper-parameter search, over the $10\times10$ cross-validation folds. The tests were performed on a PC with Intel 8 cores-16 threads i7-6900K at 3.7 GHz, 128 Gb RAM, and Windows Server 2016 OS. Python version 3.8.5.

[1] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
[2] https://www.kaggle.com/c/sm/overview

**Themes dataset**

| Classifier | AUC-PR | 99.9% CI | Magn. | Elapsed (s) |
|---|---|---|---|---|
| DT | 0.511 ± 0.026 | [0.505, 0.516] | large | 12.5 ± 2.9 |
| KNN | 0.498 ± 0.027 | [0.492, 0.504] | large | 51.6 ± 0.6 |
| LGBM | **0.588 ± 0.029** | [0.583, 0.594] | negl. | 26.4 ± 7.1 |
| LR | **0.583 ± 0.031** | [0.578, 0.589] | negl. | 13.2 ± 1.8 |
| RF | 0.577 ± 0.027 | [0.571, 0.583] | small | 61.4 ± 12.9 |
| TABNET | 0.529 ± 0.034 | [0.520, 0.538] | large | 5603 ± 554 |
| XGB | 0.556 ± 0.032 | [0.550, 0.562] | large | 32.6 ± 12.6 |

**Weighted themes dataset**

| Classifier | AUC-PR | 99.9% CI | Magn. | Elapsed (s) |
|---|---|---|---|---|
| DT | 0.483 ± 0.048 | [0.472, 0.493] | large | 13.6 ± 0.2 |
| KNN | 0.410 ± 0.049 | [0.400, 0.421] | large | 53.1 ± 0.5 |
| LGBM | **0.588 ± 0.054** | [0.577, 0.599] | negl. | 26.3 ± 3.2 |
| LR | 0.577 ± 0.054 | [0.566, 0.587] | small | 10.2 ± 0.4 |
| RF | **0.588 ± 0.053** | [0.578, 0.599] | negl. | 55.8 ± 3.9 |
| TABNET* | 0.436 ± 0.059 | [0.419, 0.452] | large | 2005 ± 23.8 |
| XGB | 0.539 ± 0.055 | [0.528, 0.549] | large | 28.5 ± 8.4 |

Table C.1: Predictive performances over the theme dataset: unweighted (top) and weighted data (bottom). Best and runner-up in bold. (*) no hyper-parameter search due to very large running times.

**Items dataset**

| Classifier | AUC-PR | 95% CI | Magn. | Elapsed (s) |
|---|---|---|---|---|
| DT | 0.538 ± 0.035 | [0.531, 0.545] | large | 23.3 ± 0.7 |
| KNN | 0.513 ± 0.028 | [0.508, 0.519] | large | 55.± 0.5 |
| LGBM | **0.641 ± 0.028** | [0.636, 0.647] | negl. | 35.1 ± 3.0 |
| LR | **0.635 ± 0.029** | [0.630, 0.641] | small | 13.6 ± 0.4 |
| RF | 0.613 ± 0.028 | [0.607, 0.618] | large | 64.9 ± 3.0 |
| TABNET | 0.561 ±0.038 | [0.553, 0.568] | large | 7489 ± 576 |
| XGB | 0.614 ± 0.032 | [0.608, 0.621] | large | 49.6 ± 10.3 |

**Weighted items dataset**

| Classifier | AUC-PR | 95% CI | Magn. | Elapsed (s) |
|---|---|---|---|---|
| DT | 0.502 ± 0.055 | [0.491, 0.513] | large | 28.8 ± 1.9 |
| KNN | 0.492 ± 0.056 | [0.481, 0.502] | large | 58.8 ± 1.8 |
| LGBM | **0.624 ± 0.051** | [0.613, 0.635] | negl. | 46.5 ± 15.7 |
| LR | **0.627± 0.052** | [0.616, 0.637] | negl. | 12.7 ± 1.0 |
| RF | 0.610 ± 0.053 | [0.599, 0.621] | small | 63.3 ± 5.0 |
| TABNET* | 0.471 ± 0.050 | [0.455, 0.488] | large | 2854 ± 124 |
| XGB | 0.585 ± 0.052 | [0.574, 0.595] | large | 81.9 ± 31.7 |

Table C.2: Predictive performances over the items dataset: unweighted (top) and weighted data (bottom). Best and runner-up in bold. (*) no hyper-parameter search due to very large running times.

# Appendix D

# Supplementary Material for Chapter 6: Section 6.3

## D.1 Distance between Probability Distributions

We resort to the *Wasserstein distance* $W$ between two probability distributions to quantify the amount of covariate shift and the robustness of target domain knowledge. In the former case, we quantify the distance between $P_S(Y|\mathbf{X})$ and $P_T(Y|\mathbf{X})$. In the latter case, the distance between $P_S(X|\varphi)$ and $P_T(X|\varphi)$. We define $W$ between $P_S$ and $P_T$ as:

$$W(P_S, P_T) = \int_{-\infty}^{+\infty} |\mathcal{P}_S - \mathcal{P}_T|$$

where $\mathcal{P}_S$ and $\mathcal{P}_T$ are the cumulative distribution functions (CDFs) of $P_S$ and $P_T$.[1] We can estimate $\mathcal{P}_S$ and $\mathcal{P}_T$ from the data using (6.15). The smaller $W$ is, the closer are the two distributions, indicating similar informational content.

## D.2 Additional Theoretical Discussion

Recall the equality (6.19), which is central to covariate shift. Under a decision tree learning setting, it does not necessarily imply $P_T(Y = y|\varphi) = P_S(Y = y|\varphi)$ for a current path $\varphi$. Consider the example below.

**Example D.2.1.** Let $\mathbf{X} = X_1, X_2$ and $Y$ be binary variables, and $\varphi$ be $X_1 = 0$. Since $P(X_1, X_2, Y) = P(Y|X_1, X_2) \cdot P(X_1, X_2)$, the full distribution can be specified by stating $P(Y|X_1, X_2)$ and $P(X_1, X_2)$. Let us consider any distribution such that:

$$P_S(X_1, X_2) = P_S(X_1) \cdot P_S(X_2) \quad P_T(X_1 = X_2) = 1 \quad Y = I_{X_1 = X_2}$$

i.e., $X_1$ and $X_2$ are independent in the source domain, while they are almost surely equal in the target domain. Notice that $Y = I_{X_1=X_2}$ readily implies that $P_S(Y|X_1, X_2) = P_T(Y|X_1, X_2)$, i.e., the covariate shift condition (6.19) holds. Using the multiplication

[1] See Scipy's Wasserstein distance for implementation details.

rule of probabilities, we calculate:

$$\begin{aligned}
&P_S(Y|\varphi) = P_S(Y|X_1 = 0) = \\
&\quad P_S(Y|X_1 = 0, X_2 = 0) \cdot P_S(X_2 = 0|X_1 = 0) + \\
&\qquad\qquad P_S(Y|X_1 = 0,X_2 = 1) \cdot P_S(X_2 = 1|X_1 = 0) = \\
&\quad P_S(Y|X_1 = 0, X_2 = 0) \cdot P_S(X_2 = 0) + \\
&\qquad\qquad P_S(Y|X_1 = 0,X_2 = 1) \cdot P_S(X_2 = 1)
\end{aligned}$$

where we exploited the independence of $X_1$ and $X_2$ in the source domain, and

$$\begin{aligned}
&P_T(Y|\varphi) = P_T(Y|X_1 = 0) = \\
&\quad P_T(Y|X_1 = 0, X_2 = 0) \cdot P_T(X_2 = 0|X_1 = 0) + \\
&\qquad\qquad P_T(Y|X_1 = 0,X_2 = 1) \cdot P_T(X_2 = 1|X_1 = 0) = \\
&\quad P_T(Y|X_1 = 0, X_2 = 0)
\end{aligned}$$

were we exploited the equality of $X_1$ and $X_2$ in the target domain. $P_S(Y|\varphi)$ and $P_T(Y|\varphi)$ are readily different when setting $X_1, X_2 \sim Ber(0.5)$ because $P_S(Y = 1|\varphi) = 1 \cdot 0.5 + 0 \cdot 0.5 \neq 1 = P_T(Y = 1|\varphi)$.

# Bibliography


[1] J. Adams-Prassl, R. Binns, and A. Kelly-Lyth. Directly discriminatory algorithms. *The Modern Law Review*, 2022.

[2] J. S. Adler. *Murder in New Orleans: the creation of Jim Crow policing*. University of Chicago Press, 2019.

[3] A. Aggarwal, P. Lohia, S. Nagar, K. Dey, and D. Saha. Automated test generation to detect individual discrimination in AI models. *CoRR*, abs/1809.03260, 2018.

[4] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *KDD*, pages 2623–2631. ACM, 2019.

[5] S. S. Alduayj and K. Rajpoot. Predicting employee attrition using machine learning. In *IIT*, pages 93–98. IEEE, 2018.

[6] D. G. Allen and L. R. Shanock. Perceived organizational support and embeddedness as key mechanisms connecting socialization tactics to commitment and turnover among new employees. *J. of Organizational Behavior*, 34(3):350–369, 2013.

[7] D. G. Allen, J. I. Hancock, J. M. Vardaman, and D. N. Mckee. Analytical mindsets in turnover research. *J. of Organizational Behavior*, 35(S1):S61–S86, 2014.

[8] J. M. Álvarez and S. Ruggieri. Counterfactual situation testing: Uncovering discrimination under fairness given the difference. In *EAAMO*, pages 2:1–2:11. ACM, 2023.

[9] J. M. Álvarez and S. Ruggieri. Causal perception. *CoRR*, abs/2401.13408, 2024.

[10] J. M. Álvarez and S. Ruggieri. Uncovering algorithmic discrimination: An opportunity to revisit the comparator. *CoRR*, abs/2405.13693, 2024.

[11] J. M. Alvarez, A. Fabris, C. Heitz, C. Hertweck, M. Loi, and M. Zehlike, editors. *Proceedings of the 2nd European Workshop on Algorithmic Fairness, Winterthur, Switzerland, June 7th to 9th, 2023*, volume 3442 of *CEUR Workshop Proceedings*, 2023. CEUR-WS.org.

[12] J. M. Álvarez, K. M. Scott, B. Berendt, and S. Ruggieri. Domain adaptive decision trees: Implications for accuracy and fairness. In *FAccT*, pages 423–433. ACM, 2023.

[13] J. M. Álvarez, A. Bringas-Colmenarejo, A. Elobaid, S. Fabbrizzi, M. Fahimi, A. Ferrara, S. Ghodsi, C. Mougan, I. Papageorgiou, P. Reyero, et al. Policy advice and best practices on bias and fairness in ai. *Ethics and Information Technology*, 26(2): 31, 2024.

[14] J. M. Álvarez, A. Mastropietro, and S. Ruggieri. The initial screening order problem. *CoRR*, abs/2307.15398, 2024.

[15] T. Anbinder. *Five Points: The 19th-century New York City neighborhood that invented tap dance, stole elections, and became the world's most notorious slum*. Simon and Schuster, 2001.

[16] J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 2008.

[17] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica*, 2016.

[18] B. Appelbaum. *The Economists' Hour: False Prophets, Free Markets, and the Fracture of Society*. Little Brown, 2019.

[19] E. Archer, I. M. Park, and J. W. Pillow. Bayesian entropy estimation for countable discrete distributions. *J. Mach. Learn. Res.*, 15(1):2833–2868, 2014. doi: 10.5555/2627435.2697056.

[20] S. Ö. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning. In *AAAI*, pages 6679–6687. AAAI Press, 2021.

[21] S. Athey and G. W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

[22] D. Austen-Smith and R. G. Fryer. An economic analysis of "acting white". *The Quarterly Journal of Economics*, 120(2):551–583, 2005.

[23] E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *AISTATS*, volume 22 of *JMLR Proceedings*, pages 100–108. JMLR.org, 2012.

[24] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference*, volume 36 of *ACM Books*, pages 507–556. ACM, 2022.

[25] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

[26] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

[27] J. Baumann, A. Castelnovo, R. Crupi, N. Inverardi, and D. Regoli. Bias on demand: A modelling framework that generates synthetic data with bias. In *FAccT*, pages 1002–1013. ACM, 2023.

[28] S. Beckers, F. Eberhardt, and J. Y. Halpern. Approximate causal abstractions. In *UAI*, volume 115 of *Proceedings of Machine Learning Research*, pages 606–615. AUAI Press, 2019.

[29] S. Beckers, H. Chockler, and J. Y. Halpern. A causal analysis of harm. In *NeurIPS*, 2022.

[30] M. Bendick. Situation testing for employment discrimination in the United States of America. *Horizons stratégiques*, 3(5):17–39, 2007.

[31] Y. Bengio. NeurIPS 2019 Posner Lecture: From System 1 Deep Learning to System 2 Deep Learning. `https://nips.cc/Conferences/2019/ScheduleMultitrack?event=15488`, 2019.

[32] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *NIPS*, pages 899–907, 2013.

[33] A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell. How cognitive biases affect xai-assisted decision-making: A systematic review. In *AIES*, pages 78–91. ACM, 2022.

[34] M. Bertrand and E. Duflo. Field experiments on discrimination. *Handbook of Economic Field Experiments*, 1:309–393, 2017.

[35] M. Bertrand and S. Mullainathan. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review*, 94(4):991–1013, 2004.

[36] P. Besnard, M. Cordier, and Y. Moinard. Arguments using ontological and causal knowledge. In *FoIKS*, volume 8367 of *Lecture Notes in Computer Science*, pages 79–96. Springer, 2014.

[37] C. Beukeboom and C. Burgers. How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (scsc) framework. *Review of Communication Research*, 7:1–37, 2019. ISSN 2255-4165.

[38] R. Binkyte, K. Makhlouf, C. Pinzón, S. Zhioua, and C. Palamidessi. Causal discovery for fairness. In *AFCP*, volume 214 of *Proceedings of Machine Learning Research*, pages 7–22. PMLR, 2022.

[39] R. Binns. Fairness in machine learning: Lessons from political philosophy. In *FAT*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR, 2018.

[40] R. Binns. On the apparent conflict between individual and group fairness. In *FAT**, pages 514–524. ACM, 2020.

[41] E. Black, S. Yeom, and M. Fredrikson. Fliptest: fairness testing via optimal transport. In *FAT**, pages 111–121. ACM, 2020.

[42] E. Bonilla-Silva. Rethinking racism: Toward a structural interpretation. *American Sociological Review*, pages 465–480, 1997.

[43] G. Booch, F. Fabiano, L. Horesh, K. Kate, J. Lenchner, N. Linck, A. Loreggia, K. Murugesan, N. Mattei, F. Rossi, and B. Srivastava. Thinking fast and slow in AI. In *AAAI*, pages 15042–15046. AAAI Press, 2021.

[44] P. Bordalo, K. Coffman, N. Gennaioli, and A. Shleifer. Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794, 2016.

[45] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci. Deep neural networks and tabular data: A survey. *CoRR*, abs/2110.01889, 2021.

[46] G. Bowker and S. Star. *Sorting Things Out: Classification and Its Consequences*. MIT Press, 1999.

[47] K. Boyd, K. H. Eng, and C. D. P. Jr. Area under the precision-recall curve: Point estimates and confidence intervals. In *ECML/PKDD (3)*, volume 8190 of *LNCS*, pages 451–466. Springer, 2013.

[48] P. Branco, L. Torgo, and R. P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2):31:1–31:50, 2016.

[49] L. Breiman. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001.

[50] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[51] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[52] R. Briggs. Interventionist counterfactuals. *Philosophical studies*, 160:139–166, 2012.

[53] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 2018.

[54] E. Calvano, G. Calzolari, V. Denicolo, and S. Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–97, 2020.

[55] A. Calvi and D. Kotzinos. Enhancing AI fairness through impact assessment in the european union: a legal and computer science perspective. In *FAccT*, pages 1229–1245. ACM, 2023.

[56] A. K. Caraban and E. Karapanos. The '23 ways to nudge' framework: designing technologies that influence behavior subtly. *Interactions*, 27(5):54–58, 2020.

[57] A. N. Carey and X. Wu. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers Big Data*, 5:892837, 2022.

[58] N. Cartwright. *The dappled world: A study of the boundaries of science*. Cambridge University Press, 1999.

[59] N. Cartwright. Modularity: It can-and generally does-fail. *Stochastic Causality*, 2001.

[60] N. Cartwright. Against modularity, the causal markov condition, and any link between the two: Comments on hausman and woodward. *British Journal for the Philosophy of Science*, 53(3), 2002.

[61] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794. ACM, 2016.

[62] R. Chetty, N. Hendren, M. R. Jones, and S. R. Porter. Race and economic opportunity in the united states: An intergenerational perspective. *The Quarterly Journal of Economics*, 135(2):711–783, 2020.

[63] S. Chiappa. Path-specific counterfactual fairness. In *AAAI*, pages 7801–7808. AAAI Press, 2019.

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

[65] E. Chzhen and N. Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, 2022.

[66] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with wasserstein barycenters. In *NeurIPS*, 2020.

[67] G. Cohen, R. S. Blake, and D. Goodman. Does turnover intention matter? Evaluating the usefulness of turnover intention rate as a predictor of actual turnover rate. *Review of Public Personnel Administration*, 36(3):240–263, 2016.

[68] J. Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[69] E. Commission. Joint employment report 2021. `https://ec.europa.eu/social/BlobServlet?docId=23156&langId=en`, 2021.

[70] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *ALT*, volume 5254 of *Lecture Notes in Computer Science*, pages 38–53. Springer, 2008.

[71] F. J. Costello. How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, 22:213–234, 2009.

[72] A. Coston, A. Mishler, E. H. Kennedy, and A. Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *FAT**, pages 582–593. ACM, 2020.

[73] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2001.

[74] K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989:139–167, 1989.

[75] C. Criado-Perez. *Invisible Women*. Vintage, 2019.

[76] A. D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *CoRR*, abs/1902.10286, 2019.

[77] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 2018.

[78] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 233–240. ACM, 2006.

[79] A. De, P. Koley, N. Ganguly, and M. Gomez-Rodriguez. Regression under human assistance. In *AAAI*, pages 2611–2620. AAAI Press, 2020.

[80] M. F. Delgado, E. Cernadas, S. Barro, and D. G. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, 15(1): 3133–3181, 2014.

[81] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

[82] R. F. DeVellis. *Scale development: Theory and applications*. Sage, 2016.

[83] T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:31895—-1923, 1998.

[84] C. D'Ignazio and L. F. Klein. *Data Feminism*. The MIT Press, 2020.

[85] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. In *NeurIPS*, pages 6478–6490, 2021.

[86] A. Dittadi, F. Träuble, F. Locatello, M. Wuthrich, V. Agrawal, O. Winther, S. Bauer, and B. Schölkopf. On the transfer of disentangled representations in realistic settings. In *ICLR*. OpenReview.net, 2021.

[87] S. Dutta, D. Wei, H. Yueksel, P. Chen, S. Liu, and K. R. Varshney. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 2803–2813. PMLR, 2020.

[88] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *ITCS*, pages 214–226. ACM, 2012.

[89] J. M. Echterhoff, M. Yarmand, and J. J. McAuley. Ai-moderated decision-making: Capturing and balancing anchoring bias in sequential decision tasks. In *CHI*, pages 161:1–161:9. ACM, 2022.

[90] C. Elkan. The foundations of cost-sensitive learning. In *IJCAI*, pages 973–978. Morgan Kaufmann, 2001.

[91] European Parliament and Council of the European Union. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.

[92] European Parliament and Council of the European Union. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts, 2021.

[93] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. D. Luca. Predicting employee attrition using machine learning techniques. *Comput.*, 9(4):86, 2020.

[94] S. Federici. *Caliban and the Witch.* Autonomedia, 2004.

[95] S. Federici. *Revolution at point zero: Housework, reproduction, and feminist struggle.* PM press, 2020.

[96] M. Feffer, M. Skirpan, Z. C. Lipton, and H. Heidari. From preference elicitation to participatory ML: A critical survey & guidelines for future research. In *AIES*, pages 38–48. ACM, 2023.

[97] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268. ACM, 2015.

[98] L. Firth, D. J. Mellor, K. A. Moore, and C. Loquet. How can managers reduce employee intention to quit? *J. of Managerial Psychology*, pages 170–187, 2004.

[99] M. Fix and R. J. Struyk. *Clear and Convincing Evidence: Measurement of Discrimination in America.* Urban Institute Press, 1993.

[100] E. U. A. for Fundamental Rights and C. of Europe. Handbook on European non-discrimination law. `https://fra.europa.eu`, 2018. Downloaded in 2023.

[101] S. R. Foster. Causation in antidiscrimination law: Beyond intent versus impact. *Houston Law Review*, 41(5):1469–1548, 2004.

[102] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

[103] J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4): 367–378, 2002.

[104] R. G. Fryer Jr and P. Torelli. An empirical analysis of 'acting white'. *Journal of Public Economics*, 94(5-6):380–396, 2010.

[105] G. Gabrani and A. Kwatra. Machine learning based predictive model for risk assessment of employee attrition. In *ICCSA (4)*, volume 10963 of *Lecture Notes in Computer Science*, pages 189–201. Springer, 2018.

[106] S. Galhotra, Y. Brun, and A. Meliou. Fairness testing: testing software for discrimination. In *ESEC/SIGSOFT FSE*, pages 498–510. ACM, 2017.

[107] G. Gigerenzer. The bias bias in behavioral economics. *Review of Behavioral Economics*, 5(3-4):303–336, 2018.

[108] I. Goldenberg and G. I. Webb. Survey of distance measures for quantifying concept drift and shift in numeric data. *Knowl. Inf. Syst.*, 60(2):591–615, 2019. doi: 10.1007/s10115-018-1257-z.

[109] C. Goldin and C. Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, September 2000.

[110] A. Goodman, J. M. Mensch, M. Jay, K. E. French, M. F. Mitchell, and S. L. Fritz. Retention and attrition factors for female certified athletic trainers in the national collegiate athletic association division I football bowl subdivision setting. *J. of Athletic Training*, 45(3):287 – 298, 2010.

[111] W. H. Greene. *Econometric Analysis*. Prentice Hall, 5ht edition, 2002.

[112] R. Griffeth and P. Hom. *Retaining Valued Employees*. Sage, 2001.

[113] R. W. Griffeth, P. W. Hom, and S. Gaertner. A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *J. of Management*, 26(3):463–488, 2000.

[114] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS*, 2022.

[115] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum. Comput. Stud.*, 43:907–928, 1995.

[116] E. Grunberg and F. Modigliani. The predictability of social events. *Journal of Political Economy*, 62(6):465–478, 1954.

[117] S. Guiasu. Weighted Entropy. *Reports on Mathematical Physic*, 2(3), 1971.

[118] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intell. Syst.*, 34(6):14–23, 2019.

[119] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.

[120] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.*, 53(4):75:1–75:37, 2021.

[121] T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1), 1941.

[122] P. Hacker. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, 55 (4), 2018.

[123] J. Y. Halpern. A modification of the halpern-pearl definition of causality. In *IJCAI*, pages 3022–3033. AAAI Press, 2015.

[124] J. Y. Halpern. *Actual Causality*. MIT Press, 2016.

[125] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach - Part I: Causes. In *UAI*, pages 194–202. Morgan Kaufmann, 2001.

[126] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach - Part II: Explanations. In *IJCAI*, pages 27–34. Morgan Kaufmann, 2001.

[127] D. S. Hamermesh. Citations in economics: Measurement, uses, and impacts. *Journal of Economic Literature*, 56(1):115–156, 2018.

[128] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *FAT**, pages 501–512. ACM, 2020.

[129] D. Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599, 1988.

[130] M. Hardt and C. Mendler-Dünner. Performative prediction: Past and future. *CoRR*, abs/2310.16608, 2023.

[131] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[132] S. Hassan. The importance of role clarification in workgroups: Effects on perceived role clarity, work satisfaction, and turnover rates. *Public administration review*, 73 (5):716–725, 2013.

[133] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[134] J. J. Heckman. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for such Models. *Annals of Economic and Social Measurement*, 5(4):475 – 492, 1976.

[135] J. J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1): 153 – 161, 1979.

[136] J. J. Heckman. Detecting discrimination. *Journal of Economic Perspectives*, 12(2): 101–116, 1998.

[137] H. Heidari, M. Loi, K. P. Gummadi, and A. Krause. A moral framework for understanding fair ML through economic models of equality of opportunity. In *FAT*, pages 181–190. ACM, 2019.

[138] M. Heikkila. Dutch scandal serves as a warning for europe over risks of using algorithms. *POLITICO*, 2022.

[139] H. G. Heneman, T. A. Judge, and J. Kammeyer-Mueller. *Staffing organizations*. McGraw-Hill Higher Education, 9 edition, 2018.

[140] M. Hoffman and S. Tadelis. People management skills, employee attrition, and manager rewards: An empirical analysis. *Journal of Political Economy*, 129(1): 243–285, 2021.

[141] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, and A. Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4):71:1–71:37, 2022.

[142] P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[143] M. Hollander, D. A. Wolfe, and E. Chicken. *Nonparametric Statistical Methods*. Wiley, 3 edition, 2014.

[144] B. C. Holtom, T. R. Mitchell, T. W. Lee, and M. B. Eberly. Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future. *The Academy of Management Annals*, 2(1):231–274, 2008.

[145] P. Hom, T. Lee, J. Shaw, and J. Hausknecht. One hundred years of employee turnover theory and research. *J. of Applied Psychology*, 102, 01 2017.

[146] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, 2 edition, 2000.

[147] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, pages 689–696. Curran Associates, Inc., 2008.

[148] L. Hu and Y. Chen. Fair classification and social welfare. In *FAT**, pages 535–545. ACM, 2020.

[149] L. Hu and I. Kohler-Hausmann. What's sex got to do with machine learning? In *FAT**, page 513. ACM, 2020.

[150] B. Hutchinson and M. Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *FAT*, pages 49–58. ACM, 2019.

[151] G. W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, 2020.

[152] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[153] N. Jain, A. Tomar, and P. K. Jana. A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. *J. Intell. Inf. Syst.*, 56(2):279–302, 2021.

[154] A. Javaloy, P. Sánchez-Martín, and I. Valera. Causal normalizing flows: from theory to practice. *CoRR*, abs/2306.05415, 2023.

[155] G. M. Johnson. The structure of bias. *Mind*, 129(516):1193–1236, 2020.

[156] C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu. An algorithmic framework for fairness elicitation. In *FORC*, volume 192 of *LIPIcs*, pages 2:1–2:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[157] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[158] D. Kahneman. Neurips 2021: A conversation on human and machine intelligence. `https://nips.cc/virtual/2021/invited-talk/22284`, 2021.

[159] D. Kahneman and D. T. Miller. Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93(2):136–153, 1986.

[160] D. Kahneman, A. M. Rosenfield, L. Gandhi, and T. Blaser. Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*, October 2016.

[161] F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

[162] F. Kamiran, I. Zliobaite, and T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Inf. Syst.*, 35(3):613–644, 2013.

[163] A. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *FAccT*, pages 353–362. ACM, 2021.

[164] A. Kasirzadeh and A. Smart. The use and misuse of counterfactuals in ethical machine learning. In *FAccT*, pages 228–236. ACM, 2021.

[165] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *NIPS*, pages 3146–3154, 2017.

[166] J. Keilwagen, I. Grosse, and J. Grau. Area under precision-recall curves for weighted and unweighted data. *PLOS ONE*, 9(3):1–13, 2014.

[167] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *NIPS*, pages 656–666, 2017.

[168] N. Kilbertus, P. J. Ball, M. J. Kusner, A. Weller, and R. Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *UAI*, volume 115 of *Proceedings of Machine Learning Research*, pages 616–626. AUAI Press, 2019.

[169] N. Kilbertus, M. G. Rodriguez, B. Schölkopf, K. Muandet, and I. Valera. Fair decisions despite imperfect predictions. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 277–287. PMLR, 2020.

[170] J.-H. Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745, 2009.

[171] M. P. Kim and M. Hardt. Is your model predicting the past? In *EAAMO*, pages 5:1–5:8. ACM, 2023.

[172] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[173] J. M. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[174] J. M. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein. Discrimination in the age of algorithms. *CoRR*, abs/1902.03731, 2019.

[175] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pages 1137–1145. Morgan Kaufmann, 1995.

[176] I. Kohler-Hausmann. Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination. *Northwestern University Law Review*, 113(5):1163, 2019.

[177] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *NIPS*, pages 4066–4076, 2017.

[178] M. Lazzari, J. M. Álvarez, and S. Ruggieri. Predicting and explaining employee turnover intention. *Int. J. Data Sci. Anal.*, 14(3):279–292, 2022.

[179] A. Liang, J. Lu, and X. Mu. Algorithmic design: Fairness versus accuracy. In *EC*, pages 58–59. ACM, 2022.

[180] K. Lippert-Rasmussen. The badness of discrimination. *Ethical Theory and Moral Practice*, 9:167–185, 2006.

[181] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva. Causal reasoning for algorithmic fairness. *CoRR*, abs/1805.05859, 2018.

[182] J. R. Loftus, L. E. J. Bynum, and S. Hansen. Causal dependence plots for interpretable machine learning. *CoRR*, abs/2303.04209, 2023.

[183] M. Loi, F. Nappo, and E. Viganò. How i would have been differently treated. discrimination through the lens of counterfactual fairness. *Res Publica*, 29(2):185–211, 2023.

[184] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. S. Zemel. The variational fair autoencoder. In *ICLR*, 2016.

[185] C. Louizos, U. Shalit, J. M. Mooij, D. A. Sontag, R. S. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *NIPS*, pages 6446–6456, 2017.

[186] G. Loury. Why does racial inequality persist? Culture, causation, and responsibility. *The Manhattan Institute*, 2019.

[187] D. Madras, E. Creager, T. Pitassi, and R. S. Zemel. Learning adversarially fair and transferable representations. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 3381–3390. PMLR, 2018.

[188] D. Madras, T. Pitassi, and R. S. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *NeurIPS*, pages 6150–6160, 2018.

[189] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Causal transfer learning. *CoRR*, abs/1707.06422, 2017.

[190] S. Maity, D. Mukherjee, M. Yurochkin, and Y. Sun. Does enforcing fairness mitigate biases caused by subpopulation shift? In *NeurIPS*, pages 25773–25784, 2021.

[191] K. Makhlouf, S. Zhioua, and C. Palamidessi. Survey on causal-based machine learning fairness notions. *CoRR*, abs/2010.09553, 2020.

[192] R. Mallon. A field guide to social construction. *Philosophy Compass*, 2(1):93–108, 2007.

[193] R. Massidda, A. Geiger, T. Icard, and D. Bacciu. Causal abstraction with soft interventions. In *CLeaR*, volume 213 of *Proceedings of Machine Learning Research*, pages 68–87. PMLR, 2023.

[194] T. Maszczyk and W. Duch. Comparison of Shannon, Renyi and Tsallis entropy used in decision trees. In *ICAISC*, volume 5097 of *Lecture Notes in Computer Science*, pages 643–651. Springer, 2008.

[195] L. C. McCandless, P. Gustafson, and A. Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, 26(11): 2331–2347, 2007.

[196] D. McNamara, C. S. Ong, and R. C. Williamson. Costs and benefits of fair representation learning. In *AIES*, pages 263–270. ACM, 2019.

[197] A. P. Miller. Want less-biased decisions? Use algorithms. *Harvard Business Review*, July 2018.

[198] A. Mishler, E. H. Kennedy, and A. Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *FAccT*, pages 386–400. ACM, 2021.

[199] T. R. Mitchell, B. C. Holtom, T. W. Lee, C. J. Sablynski, and M. Erez. Why people stay: Using job embeddedness to predict voluntary turnover. *The Academy of Management Journal*, 44(6):1102–1121, 2001.

[200] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[201] J. G. Moreno-Torres, T. Raeder, R. Alaíz-Rodríguez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognit.*, 45(1):521–530, 2012.

[202] A. C. Morgan, S. F. Way, M. J. Hoefer, D. B. Larremore, M. Galesic, and A. Clauset. The unequal impact of parenthood in academia. *Science Advances*, 7(9), 2021.

[203] C. Mougan, J. M. Álvarez, S. Ruggieri, and S. Staab. Fairness implications of encoding protected categorical attributes. In *AIES*, pages 454–465. ACM, 2023.

[204] H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. A. Sontag. Who should predict? exact algorithms for learning to defer to humans. In *AISTATS*, volume 206, pages 10520–10545. PMLR, 2023.

[205] D. Mukherjee, F. Petersen, M. Yurochkin, and Y. Sun. Domain adaptation meets individual fairness. and they get along. In *NeurIPS*, 2022.

[206] D. Mulligan. Invited talk: Fairness and privacy. https://www.afciworkshop.org/afcp2022, 2022. At the NeurIPS 2022 Workshop on Algorithmic Fairness through the Lens of Causality and Privacy.

[207] T. B. Nachbar. Algorithmic fairness, algorithmic discrimination. *Florida State University Law Review*, 48:50, 2021.

[208] R. R. Nelson. Physics envy: Get over it. *Issues in Science and Technology*, 31(3): 71–78, 2015.

[209] I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. In *NIPS*, pages 471–478. MIT Press, 2001.

[210] P. E. Ngo-Henha. A review of existing turnover intention theories. *Int. J. of Economics and Management Engineering*, 11(11):2760 – 2767, 2017.

[211] S. Nijjer and S. Raj. *Predictive Analytics in Human Resource Management: A Hands-on Approach*. Routledge India, 2020.

[212] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, and J. Gama. Methods and tools for causal discovery and causal inference. *WIREs Data Mining Knowl. Discov.*, 12(2), 2022.

[213] S. Nowozin. Improved information gain estimates for decision tree induction. In *ICML*. icml.cc / Omnipress, 2012.

[214] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernández, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab. Bias in data-driven artificial intelligence systems - an introductory survey. *WIREs Data Mining Knowl. Discov.*, 10(3), 2020.

[215] F. Palomba, A. Pugnana, J. M. Alvarez, and S. Ruggieri. A causal framework for evaluating deferring systems. *CoRR*, abs/2405.18902, 2024.

[216] V. K. Pang-Ning Tan, Michael Steinbach. *Introduction to Data Mining*. Addison Wesley, 2006.

[217] C. Panigutti, R. Hamon, I. Hupont, D. F. Llorca, D. F. Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. S. Garrido, and E. Gómez. The role of explainable AI in the context of the AI act. In *FAccT*, pages 1139–1150. ACM, 2023.

[218] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[219] J. Pearl. Haavelmo and the emergence of causal calculus. *Econometric Theory*, 2014.

[220] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[221] J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

[222] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *KDD*, pages 560–568. ACM, 2008.

[223] J. C. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 2020.

[224] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[225] R. Pieters and H. Baumgartner. Who talks to whom? Intra- and interdisciplinary communication of economics journals. *Journal of Economic Literature*, 40(2):483–509, 2002.

[226] D. Plecko and E. Bareinboim. Causal fairness analysis. *CoRR*, abs/2207.11385, 2022.

[227] J. L. Price. Reflections on the determinants of voluntary turnover. *Int. J. of Manpower*, 22(7):600 – 624, 2001.

[228] J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer. *Dataset shift in machine learning*. MIT Press, 2009.

[229] B. Qureshi, F. Kamiran, A. Karim, S. Ruggieri, and D. Pedreschi. Causal inference for social discrimination reasoning. *J. Intell. Inf. Syst.*, 54(2):425–437, 2020.

[230] C. Rastogi, Y. Zhang, D. Wei, K. R. Varshney, A. Dhurandhar, and R. Tomsett. Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW1):83:1–83:22, 2022.

[231] J. Rawls. Justice as fairness. *The philosophical review*, 67(2):164–194, 1958.

[232] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani. A survey on domain adaptation theory: Learning bounds and theoretical guarantees. *arXiv:2004.11829 [cs, stat]*, Aug. 2020.

[233] K. T. Rodolfa, H. Lamba, and R. Ghani. Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. *Nat. Mach. Intell.*, 3(10): 896–904, 2021.

[234] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.*, 29(5):582–638, 2014.

[235] D.-O. Rooth. Correspondence testing studies. *IZA World of Labor*, 58, 2021.

[236] I. Rorive. Proving discrimination cases: The role of situation testing. *Centre for Equal Rights and MPG*, 2009.

[237] E. K. Rose. A Constructivist Perspective on Empirical Discrimination Research. *Working Manuscript*, 2022.

[238] R. Rothstein. *The Color of Law: A Forgotten History of How our Government Segregated America*. Liveright Publishing, 2017.

[239] A. Roy, J. Horstmann, and E. Ntoutsi. Multi-dimensional discrimination in law and machine learning - A comparative overview. In *FAccT*, pages 89–100. ACM, 2023.

[240] D.-H. Ruben. *Explaining Explanation*. Routledge, 1990.

[241] P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *UAI*. AUAI Press, 2017.

[242] C. Rudin. A renaissance for decision tree learning. `https://www.youtube.com/watch?v=bY7WEr6lcuY`, 2016. Keynote at PAPIs 2016.

[243] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

[244] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Trans. Knowl. Discov. Data*, 4(2):9:1–9:40, 2010.

[245] S. Ruggieri, J. M. Álvarez, A. Pugnana, L. State, and F. Turini. Can we trust fair-AI? In *AAAI*, pages 15421–15430. AAAI Press, 2023.

[246] C. Russell, M. J. Kusner, J. R. Loftus, and R. Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *NIPS*, pages 6414–6423, 2017.

[247] M. Sahakyan, Z. Aung, and T. Rahwan. Explainable artificial intelligence for tabular data: A survey. *IEEE Access*, 9:135392–135422, 2021.

[248] T. Salimans, D. P. Kingma, and M. Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1218–1226. JMLR.org, 2015.

[249] B. Salimi, H. Parikh, M. Kayali, L. Getoor, S. Roy, and D. Suciu. Causal relational learning. In *SIGMOD Conference*, pages 241–256. ACM, 2020.

[250] T. Salzberger and M. Koller. The direction of the response scale matters – accounting for the unit of measurement. *European Journal of Marketing*, 53(5):871–891, 2019.

[251] P. Sánchez-Martín, M. Rateike, and I. Valera. VACA: designing variational graph autoencoders for causal queries. In *AAAI*, pages 8159–8168. AAAI Press, 2022.

[252] T. Sato and M. Rehmsmeier. Precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015.

[253] E. C. Schneider. *Smack: Heroin and the American city*. University of Pennsylvania Press, 2008.

[254] B. Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference*, volume 36 of *ACM Books*, pages 765–804. ACM, 2022.

[255] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Towards causal representation learning. *CoRR*, abs/2102.11107, 2021.

[256] J. Schuurmans, F. Frasincar, and E. Cambria. Intent classification for dialogue utterances. *IEEE Intell. Syst.*, 35(1):82–88, 2020.

[257] P. Schwöbel and P. Remmers. The long arc of fairness: Formalisations and ethical discourse. In *FAccT*, pages 2179–2188. ACM, 2022.

[258] T. Schürmann. Bias analysis in entropy estimation. *Journal of Physics A: Mathematical and General*, 37(27):L295–L301, 2004.

[259] K. M. Scott, S. M. Wang, M. Miceli, P. Delobelle, K. Sztandar-Sztanderska, and B. Berendt. Algorithmic tools in public employment services: Towards a jobseeker-centric perspective. In *FAccT*, pages 2138–2148. ACM, 2022.

[260] T. Seidl. Nearest neighbor classification. In *Encyclopedia of Database Systems*, pages 1885–1890. Springer, 2009.

[261] M. Sen and O. Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19(1): 499–522, 2016.

[262] K. Shu, S. Mukherjee, G. Zheng, A. H. Awadallah, M. Shokouhi, and S. T. Dumais. Learning with weak supervision for email intent detection. In *SIGIR*, pages 1051–1060. ACM, 2020.

[263] R. G. Simmons, B. Browning, Y. Zhang, and V. Sadekar. Learning to predict driver route and destination intent. In *ITSC*, pages 127–132. IEEE, 2006.

[264] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[265] G. Singer, R. Anuar, and I. Ben-Gal. A weighted information-gain measure for ordinal classification trees. *Expert Syst. Appl.*, 152:113375, 2020. doi: 10.1016/j.eswa.2020.113375.

[266] J. Sobel. Signaling games. *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, pages 251–268, 2020.

[267] A. Sousa-Poza and F. Henneberger. Analyzing job mobility with job turnover intentions: An international comparative study. *J. of Economic Issues*, 38(1):113–137, 2004.

[268] J. S. Speagle. A conceptual introduction to markov chain monte carlo methods. *CoRR*, abs/1909.12313, 2019.

[269] M. Spence. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973.

[270] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press, 2000.

[271] A. Srinivasan. *The Right to Sex*. BLOOMSBURY, 2021.

[272] M. Srivastava, H. Heidari, and A. Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *KDD*, pages 2459–2468. ACM, 2019.

[273] H. Suresh and J. V. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *EAAMO*, pages 17:1–17:9. ACM, 2021.

[274] C. Tanova and B. C. Holtom. Using job embeddedness factors to explain voluntary turnover in four European countries. *The Int. J. of Human Resource Management*, 19(9):1553–1568, 2008.

[275] K. Tentori. What can the conjunction fallacy tell us about human reasoning? In *Human-Like Machine Intelligence*, pages 449–464. Oxford University Press, 2022.

[276] R. Thaler and C. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008.

[277] B. L. Thanh, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *KDD*, pages 502–510. ACM, 2011.

[278] M. C. Tschantz. What is proxy discrimination? In *FAccT*, pages 1993–2003. ACM, 2022.

[279] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[280] A. Tversky and D. Kahneman. Judgments of and by Representativeness. Technical Report 3, Defense Technical Information Center, 1981.

[281] A. Tversky and D. Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293, 1983.

[282] A. Valdivia, J. Sánchez-Monedero, and J. Casillas. How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *Int. J. Intell. Syst.*, 36 (4):1619–1643, 2021.

[283] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem. Are disentangled representations helpful for abstract visual reasoning? In *NeurIPS*, pages 14222–14235, 2019.

[284] S. Verma and J. Rubin. Fairness definitions explained. In *FairWare@ICSE*, pages 1–7. ACM, 2018.

[285] J. Vieira and C. Antunes. Decision tree learner in the presence of domain knowledge. In *CSWS*, volume 480 of *Communications in Computer and Information Science*, pages 42–55. Springer, 2014.

[286] J. von Kügelgen, A. Mohamed, and S. Beckers. Backtracking counterfactuals. In *CLeaR*, volume 213 of *Proceedings of Machine Learning Research*, pages 177–196. PMLR, 2023.

[287] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841, 2017.

[288] S. Wachter, B. Mittelstadt, and C. Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

[289] S. Wachter, B. Mittelstadt, and C. Russell. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 41, 2021.

[290] D. F. Wallace. This Is Water. `https://www.youtube.com/watch?v=8CrOL-ydFMI&ab_channel=LynnSkittle`, 2005. Commencement speech to the 2005 graduating class at Kenyon College, Ohio, USA.

[291] A. Wang, V. V. Ramaswamy, and O. Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *FAccT*, pages 336–349. ACM, 2022.

[292] S. Wang, L. Hu, Y. Wang, Q. Z. Sheng, M. A. Orgun, and L. Cao. Intention nets: Psychology-inspired user choice behavior modeling for next-basket prediction. In *AAAI*, pages 6259–6266. AAAI Press, 2020.

[293] S. Wang, L. Hu, Y. Wang, Q. Z. Sheng, M. A. Orgun, and L. Cao. Intention2basket: A neural intention-driven approach for dynamic next-basket planning. In *IJCAI*, pages 2333–2339. ijcai.org, 2020.

[294] S. M. Wang, K. M. Scott, M. Artemenko, M. Miceli, and B. Berendt. "we try to empower them" - exploring future technologies to support migrant jobseekers. In *FAccT*, pages 972–983. ACM, 2023.

[295] Y. Wang and M. I. Jordan. Desiderata for representation learning: A causal perspective. *CoRR*, abs/2109.03795, 2021.

[296] H. J. P. Weerts, R. Xenidis, F. Tarissan, H. P. Olsen, and M. Pechenizkiy. Algorithmic unfairness through the lens of EU non-discrimination law: Or why the law is not a decision tree. In *FAccT*, pages 805–816. ACM, 2023.

[297] P. Westen. The empty idea of equality. *Harvard Law Review*, pages 537–596, 1982.

[298] White House. Blueprint for an AI Bill of Rights. URL https://www.whitehouse.gov/ostp/ai-bill-of-rights/, 2022. Accessed on January 2nd, 2023.

[299] L. F. Wightman. *LSAC national longitudinal bar passage study*. Law School Admission Council, 1998.

[300] T. William Lee, T. C. Burch, and T. R. Mitchell. The story of why we stay: A review of job embeddedness. *Annual Review of Organizational Psychology and Organizational Behavior*, 1(1):199–216, 2014.

[301] D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. *J. Artif. Intell. Res.*, 6:1–34, 1997.

[302] L. Wittgenstein. *Philosophical Investigations*. Wiley-Blackwell, New York, NY, USA, 1953.

[303] J. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford university press, 2005.

[304] J. M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 2015.

[305] S. Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.

[306] R. S. Wunder, T. W. Dougherty, and M. A. Welsh. A casual model of role stress and employee turnover. In *Academy of Management Proceedings*, volume 1982, pages 297–301, 1982.

[307] J. Wynen, W. V. Dooren, J. Mattijs, and C. Deschamps. Linking turnover to organizational performance: the role of process conformance. *Public Management Review*, 21(5):669–685, 2019.

[308] R. Xenidis. Tuning eu equality law to algorithmic discrimination: Three pathways to resilience. *Maastricht Journal of European and Comparative Law*, 27(6):736–758, 2020.

[309] M. Yaghini, A. Krause, and H. Heidari. A human-in-the-loop framework to construct context-aware mathematical notions of outcome fairness. In *AIES*, pages 1023–1033. ACM, 2021.

[310] K. Yang, J. R. Loftus, and J. Stoyanovich. Causal intersectionality and fair ranking. In *FORC*, volume 192 of *LIPIcs*, pages 7:1–7:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[311] S. C.-H. Yang, T. Folke, and P. Shafto. The inner loop of collective human–machine intelligence. *Topics in Cognitive Science*, 2023.

[312] G. U. Yule. Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134, 1903.

[313] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.

[314] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *ICDM*, page 435. IEEE Computer Society, 2003.

[315] M. Zehlike, K. Yang, and J. Stoyanovich. Fairness in ranking, part I: score-based ranking. *ACM Comput. Surv.*, 55(6):118:1–118:36, 2023.

[316] M. Zehlike, K. Yang, and J. Stoyanovich. Fairness in ranking, part II: learning-to-rank and recommender systems. *ACM Comput. Surv.*, 55(6):117:1–117:41, 2023.

[317] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 325–333. JMLR.org, 2013.

[318] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 819–827. JMLR.org, 2013.

[319] L. Zhang, Y. Wu, and X. Wu. Situation testing-based discrimination discovery: A causal inference approach. In *IJCAI*, pages 2718–2724. IJCAI/AAAI Press, 2016.

[320] W. Zhang and A. Bifet. FEAT: A fairness-enhancing and concept-adapting decision tree classifier. In *DS*, volume 12323 of *Lecture Notes in Computer Science*, pages 175–189. Springer, 2020.

[321] W. Zhang and E. Ntoutsi. FAHT: an adaptive fairness-aware decision tree classifier. In *IJCAI*, pages 1480–1486. ijcai.org, 2019.

[322] X. Zhang, R. Tu, Y. Liu, M. Liu, H. Kjellström, K. Zhang, and C. Zhang. How do fair decisions fare in long-term qualification? In *NeurIPS*, 2020.

[323] Q. Zhao and T. Hastie. Causal interpretations of black-box models. *J. of Business & Economic Statistics*, 39(1):272–281, 2021.

[324] X. Zhao, K. Broelemann, S. Ruggieri, and G. Kasneci. Causal fairness-guided dataset reweighting using neural networks. In *IEEE Big Data*, pages 1386–1394. IEEE, 2023.