Advancing Ethical and Responsible AI : Exploring Fairness, Privacy, and Explainability through Causal Perspectives

Karima Makhlouf

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAX057

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

IP PARIS

Advancing Ethical and Responsible AI: Exploring Fairness, Privacy, and Explainability through Causal Perspectives

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Mathématiques et informatique

Thèse présentée et soutenue à Palaiseau, le 7 octobre 2024, par

KARIMA MAKHLOUF

Composition du Jury :

Daniel AUGOT Professeir, INRIA, CNRS and École polytechnique	Président
Salvatore RUGGIERI Professeur, University of Pisa	Rapporteur
Jean-Michel LOUBES Professeur, Institut de Mathématiques de Toulouse	Rapporteur
Miguel COUCEIRO Professeur, Université de Lorraine	Examinateur
Sébastien GAMBS Professeur, Université du Québec à Montréal	Examinateur
Michael PERROT Chercheur, Centre INRIA Lille	Examinateur
Catuscia Palamidessi Directrice de recherche, Inria Paris-Saclay (Comète)	Directrice de thèse
Héber H. Arcolezi Chercheur, Inria Grenoble (Privatics)	Co-directeur de thèse

Abstract

This dissertation explores the intersection of privacy, fairness, and causality within the realm of machine learning (ML) and data-driven decision-making. The dissertation's main contributions can be summarized as follows: (1) we investigate the applicability of statistical and causality-based fairness notions in diverse application domains, evaluating their alignment with stakeholder preferences and societal norms in algorithmic decision-making systems; (2) we conduct a systematic and formal study on the impact of local differential privacy (LDP) on fairness. We quantitatively assess how ML model decisions change under varying levels of privacy and data distributions. Additionally, we empirically examine the fairness implications of collecting multiple sensitive attributes under LDP; (3) we study causal discovery through the lens of algorithmic fairness, analyzing how the causal discovery process influences the structure of causal graphs and, consequently, fairness conclusions. Furthermore, we propose a novel data generation mechanism to produce biased synthetic datasets based on causal graphs and specified bias levels, exploring the influence of different causal discovery algorithms on various causal structures and the degree of introduced bias. Overall, this thesis contributes to the growing body of literature on ethical and responsible artificial intelligence by offering theoretical insights complemented by practical considerations for policymakers, practitioners, and researchers seeking to develop fairer algorithmic systems that adhere to privacy and explainability principles.

Résumé

Cette thèse explore l'intersection complexe et multidimensionnelle de la confidentialité, de l'équité et de la causalité dans le cadre de l'apprentissage automatique et des systèmes de prise de décision algorithmiques. L'objectif est de comprendre comment ces trois notions fondamentales interagissent et de proposer des méthodes pour améliorer la conception de systèmes plus éthiques et responsables. Dans ce contexte, nous apportons plusieurs contributions majeures: (1) nous examinons l'applicabilité des notions d'équité statistiques et basées sur la causalité dans divers domaines d'application, en particulier dans des systèmes où les décisions sont partiellement ou entièrement automatisées. Nous analysons leur alignement avec les préférences des parties prenantes, ainsi qu'avec les normes et attentes sociétales. Cette analyse s'appuie sur des exemples concrets issus de différents secteurs tels que les ressources humaines, la finance et la santé. L'objectif est d'évaluer dans quelle mesure les méthodes existantes d'évaluation de l'équité peuvent être ajustées ou améliorées pour garantir des décisions algorithmiques plus transparentes et justes; (2) nous menons une étude systématique et formelle sur l'impact de la confidentialité différentielle locale sur l'équité des décisions issues de modèles d'apprentissage automatique. Dans ce cadre, nous analysons la relation entre les niveaux de confidentialité imposés par les mécanismes de confidentialité différentielle locale et la performance équitable des modèles. En quantifiant les changements dans les décisions algorithmiques en fonction de différents niveaux de protection de la confidentialité et en examinant les variations induites par des distributions de données hétérogènes, nous mettons en lumière des compromis importants entre confidentialité et équité. Par ailleurs, nous abordons une question particulièrement peu explorée : l'impact de la collecte de plusieurs attributs sensibles, tels que la race et le genre, sous les contraintes de la confidentialité différentielle locale. Nous montrons empiriquement que la collecte simultanée de ces attributs, dans un contexte de protection de la confidentialité, peut accentuer ou atténuer les disparités observées dans les décisions algorithmiques; (3) nous explorons la découverte causale à travers le prisme de l'équité algorithmique. Plus précisément, nous analysons comment les méthodes de découverte causale influencent la structure des graphes causals et, par extension, les conclusions sur l'équité des décisions algorithmiques. Nous proposons un cadre théorique qui permet de comprendre comment l'incertitude ou les erreurs dans la découverte de relations causales peuvent impacter l'équité de manière significative. En examinant les implications de la découverte causale sur les relations de dépendance entre variables sensibles et résultats, nous mettons en évidence les défis liés à l'utilisation de graphes causals pour

garantir l'équité dans les systèmes de prise de décision. Pour approfondir cette analyse, nous développons un mécanisme innovant de génération de données synthétiques. Ce mécanisme est conçu pour produire des ensembles de données biaisées, basées sur des graphes causals spécifiques et des niveaux de biais contrôlés. Ces ensembles de données synthétiques permettent d'étudier l'influence de différents algorithmes de découverte causale sur des structures causales complexes et le degré de biais introduit dans les décisions algorithmiques. En fournissant un cadre d'expérimentation reproductible, nous contribuons à une meilleure compréhension des méthodes de correction du biais dans les graphes causals. Dans l'ensemble, cette thèse contribue de manière significative au corpus croissant de la littérature sur l'intelligence artificielle éthique et responsable. Elle propose des perspectives théoriques et pratiques visant à aider non seulement les chercheurs, mais aussi les décideurs politiques et les praticiens dans le développement de systèmes algorithmiques plus justes et plus transparents. Nos travaux soulignent l'importance de concevoir des modèles d'apprentissage automatique qui respectent à la fois les principes de confidentialité et d'équité, tout en maintenant un haut degré d'explicabilité. En intégrant ces trois dimensions, nous offrons des solutions qui permettent de répondre aux préoccupations sociétales croissantes concernant l'utilisation des algorithmes dans des domaines sensibles. Nos résultats peuvent ainsi servir de guide pour la conception de politiques publiques et de meilleures pratiques dans le développement et le déploiement d'algorithmes, avec pour objectif ultime de promouvoir une intelligence artificielle au service de tous.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Catuscia Palamidessi, and co-supervisor, Héber Arcolezi, for their invaluable guidance, unwavering support, and insightful feedback throughout this journey. Their expertise and encouragement have been instrumental in completing this work. I was privileged to have Catuscia as my supervisor. She is, without question, one of the most exceptional and committed individuals I have ever encountered. Catuscia has a unique blend of intellect, compassion, and attentiveness that truly distinguishes her. I am deeply thankful that she allowed me to pursue the research direction I am truly passionate about. She has consistently had faith in my capabilities and respected my independence. Beyond research, Catuscia has taught me valuable lessons on balancing being a successful woman with being a loving wife and mother—values I deeply cherish. I have been consistently impressed by her passion for her work and her openness to learning new things, even with the remarkable achievements she has already made in her career. Her kindness and generosity have left a lasting impact on me. I want to extend my heartfelt thanks to my co-supervisor, Héber. His expertise has contributed significantly to the direction and depth of this research. His encouragement and constructive feedback have been a source of inspiration and growth throughout this journey. Beyond his professional guidance, I appreciate him as a genuinely good person—kind, approachable, and always supportive. I truly appreciate the time and energy he spent in guiding and helping me achieve this goal.

To my parents, words cannot express the depth of my gratitude. Your unwavering love, encouragement, and support have been the foundation upon which I've built this journey. You have been my biggest cheerleader, offering strength during the toughest times and celebrating with me in moments of achievement. I am deeply thankful for all your sacrifices and for always believing in me. This accomplishment is as much yours as it is mine.

To my wonderful husband, Sami, I am deeply grateful for your endless support, patience, and understanding throughout this long journey. Your unwavering belief in

me, even when I doubted myself, has been my greatest source of strength. Thank you for being my partner in every sense of the word—always there to listen, encourage, and lift me up. This achievement would not have been possible without your love and constant reassurance.

To my amazing children, Mohamed, Hajer, and Amine, you have been my greatest source of joy and motivation. Your love kept me grounded and reminded me of what truly matters. Even when I was busy with work, and so sorry for that, your understanding and patience were beyond your years. Through this journey, I hope I've shown you the value of perseverance and hard work. Thank you for being my biggest inspiration.

To my dear sister, Marwa. Thank you for being my constant source of support and encouragement. Your love and endless words of motivation have helped me a lot. You've always been there for me when I needed it most. I'm truly grateful to have you by my side, and thank you for believing in me.

I want to thank my colleagues at Team Comète for their collaboration, support, and shared experiences during this journey. Whether through academic discussions, shared challenges, or simply offering words of encouragement, you have made this process both enriching and rewarding. Your insights and friendship have been invaluable, and I am truly grateful for the sense of community we have built together.

To my wonderful colleague and friend, Ruta, thank you for being my constant source of joy, laughter, and unwavering support. Your encouragement and belief in me, even from afar, have helped me stay grounded throughout this process.

I would like to extend my sincere thanks to all the Laboratoire d'Informatique de l'Ecole Polytechnique (LIX) staff at Inria Paris-Saclay for the quality of their service and for their continuous support and help.

Financial support This work was supported by the European Research Council (ERC) project HYPATIA under the European Union's Horizon 2020 research and innovation program. Grant agreement n. 835294.

Contents

List of figures

List of tables

Chapter 1

Introduction

1.1 The Ethics of Artificial Intelligence

August 2016.

Allegheny County, Pennsylvania, USA.

An automated system called the Allegheny Family Screening Tool (AFST) assigned a risk score of child maltreatment for a 14-year-old living in a bad-conditioned house three times higher than for a 6-year-old potentially facing abuse and homelessness. Why does this discrepancy occur?

AFST was found to be unfair as it uses referral calls made by neighbors to report child abuse or maltreatment to make its predictions. The problem is that the community calls the child abuse hotline to report non-white families at a much higher rate than it does to report white families. In other words, AFST uses a hypothetical proxy for child harm, namely referral calls, not actual child abuse [82]. As a result, these discriminated families are subject to greater scrutiny and more requirements to satisfy. Eventually, they will likely fall short of these requirements and confirm the system's predictions.

As we all know, artificial intelligence (AI) is already significantly impacting society, and this impact will only increase in the future. With the rise of AI, key questions span a variety of ethical, technical, and societal aspects. What ethical guidelines should govern the development and deployment of AI? How can we ensure fairness and avoid bias in AI algorithms? What are the implications of AI on privacy, and how can we protect individuals' data? How can we make AI systems more understandable and interpretable? What regulatory frameworks are needed to ensure

responsible AI development and use? What are the potential long-term societal and cultural impacts of widespread AI adoption? How can we ensure that AI benefits humanity rather than exacerbating existing inequalities?

Addressing these questions involves an interdisciplinary joint effort among computer scientists, socialists, ethicists, policymakers, and the broader public to shape the future development and integration of AI responsibly and beneficially.

Many organizations have launched initiatives to establish ethical principles that focus on maximizing the societal benefits of AI while minimizing potential harm. Examples of such organizations include the AI Ethics Guidelines by the European Commission [83], the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [109], the NoBIAS project funded by European Union's Horizon 2020 research and innovation program [7], the China Global AI Governance [49], and Partnership on AI [105].

Some of the central principles of AI ethics include, among others, fairness, privacy, explainability, and safety. In the following sections, we present the three specific subdomains of AI ethics that are the focus of this thesis, namely fairness, privacy, and causality[1].

1.1.1 Fairness

Automated systems are increasingly making decisions across various domains. These systems strive to make optimal decisions by analyzing relevant historical data and employing Machine Learning (ML) algorithms. However, to maximize efficiency, ML algorithms can systemize discrimination against a specific population, typically minorities. In one notable case, researchers at MIT found that commercially available facial recognition systems showed higher error rates when identifying darker-skinned individuals and women compared to lighter-skinned individuals and men [43]. This bias may lead to unfair outcomes, such as misidentification by law enforcement or denial of access to services based on flawed facial recognition technology. Another example is the use of automated resume screening systems by companies during the hiring process. Studies have shown that these systems can inadvertently discriminate against candidates based on gender, race, or socioeconomic background [67]. For instance, if historical hiring data is biased toward certain demographics, the algorithm may learn and propagate these biases, leading to unfair outcomes for underrepresented groups.

[1]Causality is closely linked to explainability because it can help identify the root causes behind AI decisions or predictions.

These biases in data can pose more significant risks when applied to other critical or sensitive contexts. For instance, numerous cases exist in medical fields where the data used are biased toward particular demographics, posing potential hazards for marginalized populations. One example is the case of a widely used commercial algorithm used to predict healthcare needs and costs, which was found to allocate healthcare resources away from Black patients disproportionately [176].

Types of Bias in AI. Various types of bias may manifest, potentially resulting in unfairness across diverse downstream learning tasks. The data used for the model training introduces the most known type of bias. Common examples of bias in data include *historical* bias, *measurement* bias, *representation* bias, etc. *Historical* bias may reflect systemic inequalities and disparities that have existed in society over time and is influenced by stereotypes, prejudices, or subjective judgments that reflect societal attitudes and biases prevalent at the time of data collection. *Measurement* bias refers to a type of bias that arises from inaccuracies or inconsistencies in the measurement process used to collect data. It can occur because proxies are generated differently across sub-populations. An example of this type of bias was observed in the recidivism risk prediction tool COMPAS [56], where prior arrests and friend/family arrests were used as proxy variables to measure an individual's likelihood of reoffending. This stems in part from the increased scrutiny and policing of minority sub-populations, resulting in higher arrest rates [227]. *Representation* bias arises when the data used to train or test a model does not accurately represent the real-world population or distribution of interest. In other words, the data is not sufficiently representative, yielding incomplete representations of certain groups [133, 53]. For instance, in 2016, Beauty.AI, an automatic face analysis system supported by Microsoft, was used to identify the most attractive contestants based on facial symmetry, wrinkles, and other features. The competition garnered approximately 6000 entrants from over 100 countries. Of the 44 winners, the overwhelming majority were white, with a small number of Asian winners, and only one winner had dark skin. While Beauty.AI did not intend to favor light skin as a beauty standard, the input data biases effectively steered the algorithmic judges to this conclusion [125].

ML algorithms can also introduce bias called *algorithmic* bias, mainly if the algorithms are not designed to account for fairness considerations. That is, *algorithmic* bias occurs when the bias is introduced solely by the algorithm, independent of bias in the input data [21].

Bias can also occur during the deployment and use of AI systems, mainly if applied in contexts where they were not appropriately trained or validated. [164] includes a more comprehensive list of the types of bias.

Algorithmic Discrimination. Like bias, discrimination is also a source of unfairness. While bias can stem from factors such as data collection, sampling, and measurement, discrimination refers to the unfair or prejudicial treatment of individuals or groups based on specific sensitive attributes such as race, gender, age, disability, or religion. Discrimination can take on various forms, namely, direct and indirect. Another type of discrimination worth noting is explainable discrimination. These three forms will be briefly introduced in what follows.

- **Direct Discrimination.** Direct discrimination refers to explicit disparate treatment of individuals or groups based on specific attributes. Typically, these attributes are recognized by law as traits against which discrimination is prohibited. Computer science literature refers to them as protected or sensitive attributes. An example of direct discrimination is when an employer refuses to hire a qualified candidate solely because of their race. In this scenario, the employer's decision to discriminate against the candidate is explicit and based entirely on their race, without considering their qualifications or abilities.

- **Indirect Discrimination.** Indirect discrimination occurs when seemingly neutral rules, practices, or policies disproportionately disadvantage individuals or groups with particular sensitive attributes, even if those attributes are not explicitly used to make decisions, leading to unequal outcomes. An example of indirect discrimination is using residential zip codes as a proxy for determining creditworthiness or eligibility for financial services, such as loans or mortgages. This practice, known as *redlining* [187], involves denying or limiting access to financial products and services based on the neighborhood or area in which individuals reside. Although zip code appears to be a non-sensitive attribute, it may correlate with race because of the population of residential areas [197].

- **Explainable Discrimination.** Explainable discrimination refers to instances of discrimination with an apparent and identifiable reason or explanation for the disparate treatment or unequal outcomes experienced by individuals or groups. Unlike direct and indirect discrimination, where discrimination is evident,

explainable discrimination may appear justified based on seemingly legitimate attributes [119]. For instance, hiring practices may unduly benefit candidates from privileged professional backgrounds or specific educational qualifications. Consequently, these attributes might be cited to justify and elucidate the discrepancies between different groups.

Fairness Notions. Given the subjectivity inherent in understanding the concept of fairness, the literature has introduced several diverse and nuanced notions to grasp its multifaceted dimensions better. Fairness notions are formally defined to evaluate and measure discrimination adeptly within data or algorithmic decisions, facilitating the identification of biased outcomes. These notions can be broadly categorized into three main categories: group fairness, individual fairness, and causality-based fairness, each addressing distinct aspects of fairness in decision-making processes.

- **Group Fairness Notions.** This type of fairness notion aims to ensure fair treatment for entire groups, particularly those distinguished by specific sensitive attributes such as race, gender, or age. Group fairness notions can be characterized by the properties of the joint distribution of the sensitive attribute, the true decision, and the prediction. This means we can write them as some statement involving properties of these three attributes resulting in the following fairness criteria: independence, separation, and sufficiency [24]. Independence means that the sensitive attribute is statistically independent of the prediction. Separation refers to a category of fairness notions that, to varying extents, ensure conditional independence between the prediction and the sensitive attribute given the true decision. Finally, sufficiency represents a category of fairness notions that, to different degrees, ensure conditional independence between the true decision and the sensitive attribute given the prediction.

  Overall, while group fairness notions play a crucial role in promoting fairness and transparency in algorithmic decision-making and are relatively easy to apply, they may oversimplify the complexities of real-world contexts, overlooking individual differences and nuances within demographic groups.

- **Individual Fairness Notions.** This type of fairness notion emphasizes fairness at the individual level, focusing on treating similar individuals similarly [75]. In other words, individual fairness seeks to ensure that similar individuals receive similar outcomes or decisions, regardless of their membership in any particular group or demographic category. One issue in the practical application of individual

fairness is determining what constitutes fair treatment at the individual level, which can be subjective and context-dependent. In particular, it is difficult to establish a similarity measure and identify the relevant attributes for which individuals should present a resemblance. Another problem is that individual fairness can lead to overlooking systemic or structural factors contributing to broader inequalities between groups.

- **Causality-based Fairness Notions.** This third type of fairness notion differs from the aforementioned statistical fairness (group and individual) approaches in that they are not totally based on data but consider additional knowledge about the structure of the world in the form of a causal model. This broader understanding enables us to grasp how data is generated and how variable changes propagate within a system. Many of these fairness notions are framed in terms of non-observable factors, such as interventions (to emulate random experiments) and counterfactuals (which contemplate hypothetical scenarios beyond the actual world). The main challenge in applying causality-based fairness notions is that they may require extensive, high-quality data to construct accurate causal models, which are not always readily available, particularly in applications of the real world. Another problem regarding the applicability of these notions is the calculation of unobservable quantities (interventions and counterfactuals), which may prove impossible in some cases. This problem is called (un)identifiability [181].

Challenges and Complexities of Achieving Fairness in AI. Achieving fairness in AI presents several challenges and complexities that could make fairness difficult to achieve in practice. One of the problems relates to legal and ethical considerations. Specifically, achieving fairness in AI involves navigating the legal and ethical frameworks that govern discrimination, privacy, and civil rights. Critical challenges include complying with regulations such as the General Data Protection Regulation (GDPR)[2] and ensuring alignment with ethical principles such as privacy, transparency, and accountability. Another question concerns the choice of evaluation measure used to assess fairness. Defining and measuring fairness in AI systems is complex and context-dependent. Developing robust assessment measures that capture different aspects of fairness, such as group fairness and individual fairness, is a significant challenge, primarily because some notions of fairness are incompatible and cannot be carried out simultaneously [24, 166, 8]. Another challenge related to the opaque aspect of group bias mitigation arises from the reliance on complex and non-transparent models to

[2]https://gdpr-info.eu/

reduce the influence of sensitive attributes in ML. This includes deep learning, ensemble models, or sophisticated data projections [132]. Moreover, there are trade-offs between fairness and other desirable properties of AI systems, such as accuracy, efficiency, and interpretability [4, 135]. Balancing these trade-offs requires careful consideration of AI decisions' societal impact and ethical implications. Thus, addressing these challenges requires a multidisciplinary approach involving researchers, policymakers, ethicists, and technologists to develop fair and reliable AI systems that benefit society.

1.1.2 Privacy

In the data context, privacy protects sensitive information, such as personal identifiers, medical records, financial transactions, and online activities, from unauthorized access, use, or disclosure. In 2014, Anthem Inc., one of the largest health insurers in the U.S., experienced a data breach affecting nearly 80 million individuals, compromising their personal and medical information[3]. In 2018, it was revealed that Cambridge Analytica, a political consulting firm, harvested the personal data of millions of Facebook users without their consent. This data was used to create psychological profiles for targeted political advertising during the 2016 U.S. presidential election. The scandal raised concerns about using personal data for political manipulation and led to investigations into Facebook's data privacy practices[4].

Data anonymization is a process used to protect the privacy of individuals by changing or removing identifiable information from data sets. This technique ensures that data cannot be used to identify specific individuals while preserving its usefulness for analysis and research purposes. Anonymization techniques such as k-anonymity [228], l-diversity [149], and t-closeness [142] ensure that individual records in a dataset cannot be distinguished from each other w.r.t identifying attributes specifically, thus reducing the risk of re-identification. While these techniques provide a degree of privacy by obscuring individual identities, they may not always offer strong guarantees against re-identification attacks. In 2006, Netflix released a large dataset containing anonymous movie ratings of thousands of users and their demographic information for contest entrants. Despite efforts to anonymize the dataset by removing personally identifiable information, such as names and addresses, the researchers demonstrated that it was possible to re-identify individuals by combining the Netflix dataset with publicly available information from another movie platform, the Internet Movie Database (IMDb) [171].

[3] https://coverlink.com/case-study/anthem-data-breach/
[4] https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

In response to these challenges, researchers and practitioners have turned to a more robust framework for privacy called differential privacy (DP) [76].

Differential Privacy (DP). DP is a rigorous framework that guarantees strong privacy by adding carefully calibrated noise to query responses or data. The key concept in DP is the notion of *plausible deniability*. In other words, under the DP framework, any query or analysis performed on a dataset should not reveal sensitive information about any individual, regardless of what other information an adversary may have. Essentially, DP assures that including or excluding an individual's data from a dataset will not substantially alter the likelihood of any given outcome [74]. This assurance is quantified by a parameter ε, indicating the privacy protection degree. A smaller ε corresponds to a narrower gap between datasets with and without a specific record, thereby indicating a higher level of privacy. Other beneficial properties of DP include sequential composition and robustness to post-processing [77]. Sequential composition allows for quantification of the privacy level ε such that the total privacy level when crossing information from several sources is always bounded by the sum of the privacy levels of each individual source. The robustness to post-processing implies that regardless of any subsequent data processing or function applied to differentially private data, the output remains differentially private. Central DP (CDP) and Local DP (LDP) are two variants of DP that differ in their approach to protecting privacy.

- **Central Differential Privacy (CDP).** CDP is the initial design of DP, which operated assuming the central server responsible for data collection and processing could be trusted. In CDP, the central server applies privacy-preserving mechanisms to the aggregated data before releasing the results or performing further analysis. The primary goal of CDP is to provide strong privacy guarantees while enabling centralized data collection and analysis. However, the assumption of trust in central servers became increasingly problematic as concerns about data breaches, unauthorized access, and misuse of personal information grew. This led to the need for a privacy setting where the server is assumed to be untrusted.

- **Local Differential Privacy (LDP) [121].** LDP emerged as a solution that decentralizes privacy protection, allowing users to perturb their data locally before sharing it with a central authority. This approach reduces reliance on centralized trust and minimizes the risk of data breaches. In other words, in the context of LDP, the server is assumed to be untrusted. This means that data is obfuscated

before it is sent to the server, ensuring that it cannot infer any individual's private information directly from the data it receives. Users have greater visibility and control over data handling, promoting trust and cooperation in data-sharing ecosystems. Several high-tech companies such us Google [79], Apple [229], and Microsoft [69] have shown interest in and adopted LDP techniques to enhance user privacy while maintaining data utility.

1.1.3 Causality

The study of causality has evolved over centuries, drawing on insights from philosophy, science, statistics, and other disciplines. In recent decades, advances in statistical methods and computational techniques have led to significant progress in causality. Researchers have developed sophisticated models, such as structural equation modeling (SEM) [37], instrumental variable analysis [223], and Bayesian networks [104], to address complex causal questions in epidemiology, public health, and ML.

Several approaches have been developed to elucidate and quantify causal relationships between variables. The main two approaches are experimental studies and observational studies. Experimental studies, including randomized controlled trials (RCTs) [224], are the gold standard for establishing causality. In an experimental study, researchers manipulate an independent variable (the treatment) and measure its effect on a dependent variable (the outcome) while controlling for potential confounding variables[5]. However, in some cases, conducting RCTs may be unethical or impractical. For example, it may be unethical to withhold a potentially life-saving treatment from participants, especially if there is strong evidence supporting its efficacy. In such cases, observational studies can be a good alternative. Observational studies involve observing naturally occurring phenomena without intervention by the researcher. While observational studies cannot establish causality as definitively as experimental studies, they can still provide valuable insights into causal relationships.

Causal inference and causal discovery are two related but distinct areas within the broader study of causality. While they share the common goal of uncovering causal relationships between variables, their approaches and methodologies differ. We present each of them in the following.

Causal Inference. Causal inference is the process of making inferences about causal relationships between variables based on observed data. This involves determining

[5]Confounding variables are variables that are common causes of both the treatment and the outcome.

whether one variable causes changes in another variable and estimating the magnitude and direction of the causal effects. Causal inference often relies on statistical methods and frameworks, such as Pearl's structural causal models (SCMs) [182] and Rubin's potential outcomes framework [110]. Formally, the two frameworks are equivalent [183, 167]. However, each is more equipped to address different problems in particular situations. For example, accounting for the many causal pathways in real applications can be more straightforward using SCM. On the other hand, potential outcome framework is preferred when estimating individual-level causal effects.

- **Structural Causal Model (SCM) Framework.** The SCM framework is a mathematical framework used to represent causal relationships among variables in a system. It provides a formal way to describe how variables causally influence one another and how interventions or changes to the system affect its behavior. The SCM framework is widely used in economics, epidemiology, social sciences, and ML for causal inference, prediction, and decision-making. Causal assumptions between variables are captured by a directed acyclic graph (DAG) $\mathcal{G} = (\mathbf{V}, \mathcal{E})$, where vertices $\mathbf{V}$ represent variables and directed edges $\mathcal{E}$ represent functional relationships between the variables. The DAG relates causal structure and joint distribution in the data through the Markov condition, where every variable is conditionally independent of its nondescendants given its parents. Directed edges can have two interpretations. A probabilistic interpretation where the edge represents a dependency among the variables such that the direction of the edge is irrelevant. A causal interpretation where the edge represents a causal influence between the corresponding variables such that the direction of the edge matters. In the presence of a cause-effect relation between two variables A and Y, a confounder is a third variable C, which affects both the cause A and the effect Y.

  The SCM framework allows researchers to predict the effects of interventions or changes to the system by simulating counterfactual outcomes under different intervention scenarios. Counterfactuals represent what would have happened if certain variables had taken different values, allowing researchers to assess the causal impact of interventions on the system. Interventional and counterfactual quantities are causal quantities that can be computed from observational data under some identifiability conditions [181]. These conditions outline the criteria dictating when and how causal quantities can be computed from observational data.

- **Potential Outcomes Framework.** Unlike the SCM framework, expressing causal relations in the potential outcome framework starts at the unit level. A unit i is the atomic research object. For example, every patient corresponds to a unit i in a clinical trial investigating the effectiveness of a new drug (treatment) for reducing blood pressure in patients with hypertension. The potential outcome aims to compare observed "factual" outcomes with hypothetical "counterfactual" outcomes that would have occurred under different treatment conditions. It defines causal effects as the difference between the observed and counterfactual outcomes. For the clinical trial example, the potential outcome framework allows us to compare the observed outcomes (blood pressure measurements) with the counterfactual outcomes (what would have happened if the patient received a different or no treatment at all?). In observational studies (in contrast to experimental studies), only one potential outcome can be observed: the factual outcome. The counterfactual potential outcome is usually impossible to observe. The potential outcome framework is widely applied to various fields, including epidemiology, social sciences, and economics.

SCM vs. Potential Outcomes. Although both causal frameworks are considered equivalent [183], interesting differences exist between them. Depending on the task at hand, one framework might be more appropriate to use than the other. For example, reasoning about causal effects at the individual (unit) level is more straightforward with the potential outcome framework [167]. On the other hand, considering the different paths of causal effects (direct, indirect, and spurious) is much easier to achieve using SCMs and causal graphs. More generally, the potential outcome framework is more suitable for causal inference problems where the goal is to narrowly estimate the causal (treatment) effect of a cause variable A on an outcome variable Y. There are two justifications for this point. First, developing estimators of causal effects and counterfactuals can be more straightforward using the potential outcome framework [260]. Second, the potential outcome framework can decompose the sources of inconsistency and bias into unaccounted-for baseline differences between individuals and treatment effect bias [167]. SCMs and causal graphs, however, are more suitable in causal discovery problems where the goal is to learn the causal relations among a set of variables [98]. The potential outcome framework is not well equipped for such issues because the causal effect of variables other than the treatment (sensitive attribute) is not defined.

Causal Discovery. The main impediment to causal inference is the unavailability of the true causal graph, which can be set manually by experts in the field but is very often generated using experiments (also called interventions). Identifying the causal graph is called causal discovery or structural learning. Causal discovery methods aim to uncover causal relationships by analyzing patterns in data, identifying statistical dependencies, and inferring causal structures from observed correlations.

A large number of causal discovery algorithms exist in the literature. Most of these algorithms fall into three categories: constraint-based, score-based, and procedures that exploit semi-parametric assumptions. In the constraint-based category, algorithms such as PC [218], FCI [219], and σ-CG [88] rely mainly on the (conditional) independencies in the data to discover causal relations between variables. Therefore, their efficiency depends on the reliability of the conditional independence test procedure. Score-based algorithms, such as GES [52], FGES [193], and HCR [41] rely instead on goodness-of-fit tests. They learn causal graphs by maximizing a scoring criterion such as the Bayesian Information Criterion (BIC) [203], which trades off accuracy (fitness of graph to the data) with complexity (the number of parameters in the model). Algorithms in the third category, such as LiNGAM [207], PNL [269], and DCDI [39] use additional assumptions to learn causal relations more efficiently and in more detail. The most common assumptions relevant to the third category are the linearity of the model and the non-gaussianity of the regression residuals. Algorithms in the first two categories do not make strong assumptions about the parametric form or functions of causal connections. Therefore, they can theoretically be applied to more scenarios than the third category. However, most available implementations of constraint-based and score-based causal discovery algorithms model variables as multivariate Gaussian mixture, which implies linearity and Gaussianity of all continuous variables. Causal graphs returned by algorithms in the third category are more accurate than those of the two first categories, which are simply Markov equivalence classes. Two graphs belong to the same Markov equivalence class when they imply the same independence constraints. In such cases, researchers can leverage domain-specific background knowledge to eliminate specific causal relations, thereby reducing the set of valid causal graphs.

1.2 Summary of Contributions

The contributions of this dissertation are outlined as follows.

1. We examine prevalent fairness notions, exploring their inherent tensions and incompatibilities. Using several toy examples, we also provide how the various fairness notions can be computed in practice. [**Chapter 3/Section 3.2**]

2. We propose a decision diagram integrating a set of fairness-related features of real-world scenarios that can help researchers, practitioners, and policymakers answer the question of "which notion of fairness is most appropriate to a given real-world scenario and why?". [**Chapter 3/Section 3.3**]

3. We provide a guideline to help select a suitable causality-based fairness notion given a specific real-world scenario and a ranking of these fairness notions according to Pearl's causation ladder, indicating how difficult it is to deploy each notion in practice. [**Chapter 3/Section 3.4**]

4. We summarize the main identifiability results concerning the specific problem of discrimination discovery, emphasizing graphical criteria. In particular, we compile key findings on the identifiability of causal and counterfactual effects of particular relevance to ML fairness, including identifiability (Pearl's SCM framework) and estimation (potential outcome framework). [**Chapter 3/Section 3.4**]

5. We empirically study the impact of collecting multiple sensitive attributes under LDP on fairness by applying seven state-of-the-art LDP protocols. More specifically, we compare the impact of these LDP protocols under a homogeneous encoding when training ML binary classifiers, and we show that fairness and LDP can go hand in hand. [**Chapter 4/Section 4.2**]

6. We propose a novel privacy budget allocation scheme for LDP that considers the varying domain size of sensitive attributes. Our approach generally led to a better privacy-utility-fairness trade-off in our experiments than the state-of-the-art solution. [**Chapter 4/Section 4.2**]

7. We investigate the impact of training a model with multiple sensitive attributes, obfuscated under LDP guarantees, using two variants (independent and combined) of the widely recognized k-ary randomized response mechanism [115]. Our findings reveal that multidimensional LDP approaches (independent and combined) show differences in their impact on fairness, particularly under weak

privacy guarantees. Moreover, LDP obfuscation disproportionately affects a specific protected group, which depends heavily on the distribution of the true decision. [**Chapter 4/Section 4.3**]

8. We conduct a systematic and formal study of the effect of LDP on fairness. Specifically, we perform a quantitative study of how the fairness of the decisions made by the ML model changes under LDP for different levels of privacy and data distributions. In particular, we provide bounds in terms of the joint distributions and the privacy level, delimiting the extent to which LDP can impact the fairness of the model. We characterize the cases where privacy reduces discrimination and those with the opposite effect. We validate our theoretical findings on synthetic and real-world datasets. [**Chapter 4/Section 4.4**]

9. We conduct experimental analysis to demonstrate how the causal discovery procedure affects the structure of the causal graph, thereby influencing fairness conclusions. Specifically, we show how different causal discovery approaches can lead to diverse causal models, with even minor variations between them substantially impacting fairness conclusions. [**Chapter 5/Section 5.2**]

10. We propose a mechanism that accepts a causal graph and a specified discrimination level as inputs, producing a biased synthetic dataset that adheres to the causal graph's structure while maintaining the desired level of discrimination. Employing this mechanism, we investigate the influence of different causal discovery algorithms on various causal structures and the degree of introduced bias. [**Chapter 5/Section 5.3**]

1.3 List of Publications

The material presented in this dissertation has appeared in the following publications.

1. **Survey on Fairness Notions and Related Tensions** [8]

   This paper has been accepted and published in the EURO Journal on Decision Processes 2023 proceedings. Chapter 3/Section 3.2 presents the main contribution of this paper.

   Abstract. Automated decision systems are increasingly used to make consequential decisions in problems such as job hiring and loan granting, hoping to replace subjective human decisions with objective machine learning (ML) algorithms.

However, ML-based decision systems are prone to bias, resulting in unfair decisions. Several notions of fairness have been defined in the literature to capture the different subtleties of this ethical and social concept (e.g., statistical parity, equal opportunity, etc.). Fairness requirements must be satisfied while learning models create tensions among the notions of fairness and other desirable properties such as privacy and classification accuracy. This paper surveys the commonly used notions of fairness and discusses the tensions between them and between fairness and accuracy. Different methods to address the fairness-accuracy trade-off (classified into four approaches: pre-processing, in-processing, post-processing, and hybrid) are reviewed. The survey is consolidated with experimental analysis on fairness benchmark datasets to illustrate the relationship between fairness measures and accuracy in real-world scenarios.

2. **Machine learning fairness notions: Bridging the gap with real-world applications** [156]

   This journal paper is accepted and published in IPM (Information Processing and Management) 2021. Chapter 3/Section 3.3 presents the main findings of this paper. A preliminary version of this paper was presented at the BIAS 2020 workshop and published in the SIGKDD ACM Explorations Newsletter 2021.

   Abstract. Machine Learning (ML) based predictive systems are increasingly used to support decisions that critically impact individuals' lives, such as college admission, job hiring, child custody, criminal risk assessment, etc. As a result, fairness emerged as an important requirement to guarantee that ML predictive systems do not discriminate against specific individuals or entire sub-populations, particularly minorities. Given the inherent subjectivity of viewing the concept of fairness, several notions of fairness have been introduced in the literature. This paper is a survey of fairness notions that, unlike other surveys in the literature, addresses the question, "Which notion of fairness is most suited to a given real-world scenario and why?". Our attempt to answer this question consists in (1) identifying the set of fairness-related characteristics of the real-world scenario at hand, (2) analyzing the behavior of each fairness notion, and then (3) fitting these two elements to recommend the most suitable fairness notion in every specific setup. The results are summarized in a decision diagram that practitioners and policymakers can use to navigate the relatively large catalog of ML fairness notions.

3. **Identifiability of Causal-based ML Fairness Notions [157]**

   This paper was accepted for presentation and published in the proceedings of the 14th International Conference on Computational Intelligence and Communication Networks (CICN) 2022. This study is presented in Chapter 3/Section 3.4.3.

   Abstract. Machine learning algorithms can produce biased outcomes/predictions, typically against minorities and under-represented sub-populations. Therefore, fairness is emerging as an important requirement for the safe application of machine learning-based technologies. The most commonly used fairness notions (e.g., statistical parity, equalized odds, predictive parity, etc.) are observational and rely on mere correlation between variables. These notions fail to identify bias in the case of statistical anomalies such as Simpson's or Berkson's paradoxes. Causality-based fairness notions (e.g., counterfactual fairness, no-proxy discrimination, etc.) are immune to such anomalies and hence more reliable for assessing fairness. However, causality-based fairness notions are defined in terms of quantities (e.g., causal, counterfactual, and path-specific effects) that are not always measurable. This is known as the identifiability problem and is the topic of a large body of work in the causal inference literature. The first contribution of this paper is a compilation of the major identifiability results that are particularly relevant to machine learning fairness. To the best of our knowledge, no previous work in ML fairness or causal inference provides such systemization of knowledge. The second contribution is more general and addresses the main problem of using causality in machine learning: extracting causal knowledge from observational data in real scenarios. This paper shows how this can be achieved using identifiability.

4. **When Causality Meets Fairness: A Survey [158]**

   This journal paper is accepted and published in JLAMP (Journal of Logical and Algebraic Methods in Programming) 2024. This study is presented in Chapter 3/Section 3.4.

   Abstract. Addressing the problem of fairness is crucial to safely use machine learning algorithms to support decisions with a critical impact on people's lives, such as job hiring, child maltreatment, disease diagnosis, loan granting, etc. Several notions of fairness have been defined and examined in the past decade, such as statistical parity and equalized odds. However, the most recent notions of fairness are causal-based and reflect the now widely accepted idea that using causality is necessary to address the problem of fairness appropriately. This

paper examines an exhaustive list of causal-based fairness notions and studies their applicability in real-world scenarios. As most causal-based fairness notions are defined in non-observable quantities (e.g., interventions and counterfactuals), their deployment in practice requires computing or estimating those quantities using observational data. This paper offers a comprehensive report of the different approaches to infer causal quantities from observational data, including identifiability (Pearl's SCM framework) and estimation (potential outcome framework). The main contributions of this survey paper are (1) a guideline to help select a suitable fairness notion given a specific real-world scenario and (2) a ranking of the fairness notions according to Pearl's causation ladder, indicating how difficult it is to deploy each notion in practice.

5. **(Local) Differential Privacy has NO Disparate Impact on Fairness** [16]

   This paper was accepted for presentation and published in the proceedings of the Conference on Data and Applications Security and Privacy (DBSec) 2023. It won the best paper award at the conference. The results of this paper are presented in Chapter 4/Section 4.2.

   Abstract. In recent years, Local Differential Privacy (LDP), a robust privacy-preserving methodology, has gained widespread adoption in real-world applications. With LDP, users can perturb the data on their devices before sending it out for analysis. However, as collecting multiple sensitive information becomes more prevalent across various industries, collecting a single sensitive attribute under LDP may not be sufficient. Correlated attributes in the data may still lead to inferences about the sensitive attribute. This paper empirically studies the impact of collecting multiple sensitive attributes under LDP on fairness.

6. **On the Impact of Multi-dimensional Local Differential Privacy on Fairness** [154]

   This paper was accepted for presentation and published in the proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) 2024 – Journal track of Data Mining and Knowledge Discovery. The results of this paper are presented in Chapter 4/Section 4.3.

   Abstract. Automated decision systems are increasingly used to make consequential decisions in people's lives. Due to the sensitivity of the manipulated data and the resulting decisions, several ethical concerns need to be addressed for the

appropriate use of such technologies, particularly fairness and privacy. Unlike previous work, which focused on centralized differential privacy (DP) or on local DP (LDP) for a single sensitive attribute, in this paper, we examine the impact of LDP in the presence of several sensitive attributes (*i.e., multi-dimensional data*) on fairness. Detailed empirical analysis on synthetic and benchmark datasets revealed very relevant observations. In particular, (1) multi-dimensional LDP is an efficient approach to reduce disparity, (2) the variant of the multi-dimensional approach of LDP (we employ two variants) matters only at low privacy guarantees (high ε), and (3) the true decision distribution has an important effect on which group is more sensitive to the obfuscation. Last, we summarize our findings as recommendations to guide practitioners in adopting effective privacy-preserving practices while maintaining fairness and utility in machine learning applications.

7. **A Systematic and Formal Study of the Impact of Local Differential Privacy on Fairness: Preliminary Results** [155]

   This paper was accepted for presentation and published in the proceedings of the Computer Security Foundations Symposium (CSF) 2024. This study is presented in Chapter 4/Section 4.4.

   Abstract. Machine learning (ML) algorithms rely primarily on the availability of training data, and, depending on the domain, these data may include sensitive information about the data providers, thus leading to significant privacy issues. Differential privacy (DP) is the predominant solution for privacy-preserving ML, and the local model of DP is the preferred choice when the server or the data collector is not trusted. Recent experimental studies have shown that local DP can impact ML prediction for different subgroups of individuals, thus affecting fair decision-making. However, the results are conflicting in the sense that some studies show a positive impact of privacy on fairness while others show a negative one. In this work, we conduct a systematic and formal study of the effect of local DP on fairness. Specifically, we perform a quantitative study of how the fairness of the decisions made by the ML model changes under local DP for different levels of privacy and data distributions. In particular, we provide bounds in terms of the joint distributions and the privacy level, delimiting the extent to which local DP can impact the fairness of the model. We characterize the cases where privacy reduces discrimination and those with the opposite effect. We validate our theoretical findings on synthetic and real-world datasets. Our results are preliminary in the sense that, for now, we study only the case of one sensitive

attribute and only statistical disparity, conditional statistical disparity, and equal opportunity difference.

8. **Causal Discovery for Fairness** [34]

   This paper was accepted at the NeurIPS workshop on algorithmic fairness through the lens of causality and privacy and published in the conference proceedings 2022. The results of this paper are presented in Chapter 5/Section 5.2.

   Abstract. Fairness guarantees that ML decisions do not discriminate against individuals or minority groups. Identifying and measuring fairness/discrimination reliably is better achieved using causality, which considers the causal relation, beyond mere association, between the sensitive attribute (e.g., gender, race, religion, etc.) and the decision (e.g., job hiring, loan granting, etc.). However, the big impediment to using causality to address fairness is the unavailability of the causal model (typically represented as a causal graph). Existing causal approaches to fairness in the literature do not address this problem and assume that the causal model is available. In this paper, we do not make such an assumption, and we review the major algorithms used to discover causal relations from observable data. This study focuses on causal discovery and its impact on fairness. In particular, we show how different causal discovery approaches may result in different causal models and, most importantly, how slight differences between causal models can significantly impact fairness/discrimination conclusions.

9. **Causal Discovery on Biased Data**

   This paper will be submitted to the 39th Annual AAAI Conference on Artificial Intelligence. The results of this paper are presented in Chapter 4/Section 4.4.

   Abstract. Leveraging cause-effect relations between variables is essential to appropriately address various problems in several scientific fields. However, a crucial drawback in using causality is the lack of ground truth of the causal model that underlies the generating process for real-world data. Currently, few datasets are documented with a proper causal structure, most of which are bivariate. Moreover, these datasets often exhibit inherent biases that can impact the fairness of models trained on them. A promising way to overcome these issues is the generation of data synthetically. This paper aims to study the behavior of causal discovery algorithms in the presence of biased data. To this end, we introduce a mechanism that takes a causal graph and a discrimination level as input and generates a biased synthetic dataset satisfying the causal

structure of the graph with a desired discrimination level. Using this mechanism, we could observe how various causal discovery algorithms are impacted by the type of causal structures and the amount of injected bias. The mechanism also allowed us to study the behavior of causal discovery algorithms when outcome distribution is modified (by shifting the binarization threshold). The most notable observation is that in the presence of fair/unbiased data, causal discovery algorithms fail to correctly identify crucial parts of the causal structure (e.g., the direct edge between the sensitive and outcome nodes, confounding paths, etc.), which produces misleading fairness conclusions when causal based fairness notions are used.

1.4 List of Awards

During my Ph.D. studies, I obtained the following awards.

- IP Paris PhD best poster award in computing, data, and AI. The poster summarises the work related to our Paper: *Causal Discovery for Fairness* [34] (December 2022).
- Best paper award in the Conference on Data and Applications Security and Privacy. Paper: *(Local) Differential Privacy has NO Disparate Impact on Fairness* [16] (July 2023).
- Best poster award and three-minute pitch winner at the BigTech Tunisia Digital Summit/Tunisian AI Society Poster Session 2024. Paper: *A Systematic and Formal Study of the Impact of Local Differential Privacy on Fairness: Preliminary Results* [155] (June 2024).

1.5 Thesis Roadmap

The rest of the thesis is organized as follows.

In chapter 2, we provide background and necessary preliminaries on the three topics included in the thesis: fairness, privacy, and causality.

Chapter 3 presents the various studies we conducted on the applicability of fairness notions in real-world applications. Section 3.2 introduces our survey paper, presenting several fairness notions and discussing their tensions. Section 3.3 considers the applicability of statistical fairness notions. Finally, Section 3.4 considers the applicability and suitability of causality-based fairness notions.

Chapter 4 presents our studies on the relationship between privacy and fairness. This chapter investigates how privacy-preserving techniques can influence fairness measures. Sections 4.2 and 4.3 present two of our empirical studies on the impact of multidimensional LDP on fairness. Finally, in Section 4.4, we present our systematic and formal study on the impact of LDP on fairness.

Chapter 5 presents our studies on how causal discovery algorithms impact fairness conclusions, examining the nuanced interplay between causal models and fairness assessments. Section 5.2 presents our work on causal discovery for fairness, highlighting how slight differences between causal models can significantly impact conclusions regarding fairness and discrimination. Section 5.3 presents our study on how different causal structures and varying data bias levels impact the performance of causal discovery algorithms and, consequently, fairness conclusions.

Chapter 6 summarizes the key findings of our research and discusses their implications for the field of ML. We reflect on the contributions of our work, highlight its limitations, and propose directions for future research. This chapter aims to clearly understand how our findings can inform ongoing and future efforts to ensure fairness and privacy in ML.

Chapter 2

Preliminaries and Notations

2.1 Introduction

This chapter aims to provide readers with essential context, theoretical frameworks, and foundational knowledge crucial for understanding this dissertation's research findings and arguments. Essential terminology and concepts used throughout the thesis are introduced. The first section of the chapter is dedicated to necessary notation and preliminaries for fairness. The following section focuses on privacy, notably the local privacy setting, which is the thesis's focus. The last section provides essential terminology and concepts of causality.

Notation. Table 2.1 summarizes the notation used throughout this thesis. Note that we always consider a single sensitive attribute A and assess fairness w.r.t. that attribute. However, the LDP mechanism can be applied to a set of sensitive attributes that we denote as $\mathbf{A}$.

Symbol	*Description*
$\mathbf{A}$	Set of sensitive attributes (**privacy viewpoint**)
A	Sensitive attribute (**fairness viewpoint**), $A \in \mathbf{A}$
X	Set of non-sensitive attributes
Y	True decision, $Y \in \{0, 1\}$
$\hat{Y}$	Prediction, $\hat{Y} \in \{0, 1\}$
$\mathbf{x}_i$	i-th coordinate of vector $\mathbf{x}$
$z = \mathcal{L}(v)$	Protocol $\mathcal{L}$ perturbs v into z under ϵ-LDP
A'	Locally differentially private sensitive attribute, $A' = \mathcal{L}(A)$
$\mathbf{A'}$	Set of locally differentially private sensitive attributes, $\mathbf{A'} = \mathcal{L}(\mathbf{A})$
k_j	Domain size of the j-th attribute
d_a	Number of sensitive attributes, $d_a = \lvert\mathbf{A}\rvert$
S	Original dataset, $S = (\mathbf{A}, X, Y)$
S'	Dataset with obfuscated sensitive attributes, $S' = (\mathbf{A'}, X, Y)$

Table 2.1 Notations.

2.2 Fairness

Variables are denoted by uppercase letters, while lowercase letters denote specific values of variables (e.g., $A = a$, $Y = y$). Let V, A, and X be three random variables representing, respectively, the total set of attributes, the sensitive attributes, and the remaining attributes describing an individual such that $V = (X, A)$ and $\mathbb{P}[V = v_i]$ represents the probability of drawing an individual with a vector of values v_i from the population. For simplicity, we focus on the case where A is a binary random variable where $A = 0$ designates the protected group, while $A = 1$ designates the non-protected group. Let Y and $\hat{Y}$ be binary random variables representing, respectively, the true decision (e.g., health-care intervention, hiring, admission, releasing on parole) and the prediction of the classifier where $Y = 1$ designates a positive instance (e.g., accepting a loan), while $Y = 0$ is a negative one (e.g., denying a loan). Typically, the prediction $\hat{Y}$ is derived from a score represented by a random variable R where $\mathbb{P}[R = r]$ is the probability that the score value equals r.

2.2.1 Classification of Fairness Notions

Fairness notions are defined as a mathematical condition involving either $\hat{Y}$ or R along with the other random variables. As such, we are not concerned with the inner workings of ML systems and their fairness implications. What matters is only the score/prediction value and how fair/biased it is.

Most of the proposed fairness notions are properties of the joint distribution of the above random variables (X, A, Y, $\hat{Y}$, and R). They can also be interpreted using the confusion matrix and the related metrics (Table 2.2).

Table 2.2 Metrics based on confusion matrix.

	Actual Positive $Y = 1$	Actual Negative $Y = 0$		
Predicted Positive $\hat{Y} = 1$	**TP** (True Positive)	**FP** (False Positive) *Type I error*	**PPV** = $\frac{TP}{TP+FP}$ *Positive Predictive Value* *Precision* *PV+* *Target Population Error*	**FDR** = $\frac{FP}{TP+FP}$ *False Discovery Rate* *Target Population Error*
Predicted Negative $\hat{Y} = 0$	**FN** (False Negative) *Type II error*	**TN** (True Negative)	**FOR** = $\frac{FN}{FN+TN}$ *False Omission Rate* *Success Predictive Error*	**NPV** = $\frac{TN}{FN+TN}$ *Negative Predictive Value* *PV-*
	TPR = $\frac{TP}{TP+FN}$ *True Positive Rate* *Sensitivity* *Recall*	**FPR** = $\frac{FP}{FP+TN}$ *False Positive Rate* *Model Error*	**OA** = $\frac{TP+TN}{TP+FP+TN+FN}$ *Overall Accuracy*	**BR** = $\frac{TP+FN}{TP+FP+TN+FN}$ *Base Rate* *Prevalence (p)*
	FNR = $\frac{FN}{TP+FN}$ *False Negative Rate* *Model Error*	**TNR** = $\frac{TN}{FP+TN}$ *True Negative Rate* *Specificity*		

Group fairness notions fall into three classes defined in the properties of joint distributions: independence, separation, and sufficiency [24]. These properties are used in the literature to prove the existence of tensions between fairness notions. That is, it is impossible to satisfy all fairness notions simultaneously except in extreme, degenerate, and dump scenarios.

Independence. Independence implies that the sensitive feature A is statistically independent of the classifier $\hat{Y}$ (or the score R).

$$\hat{Y} \perp A \quad (or\ R \perp A) \tag{2.1}$$

In the case of binary classification, independence is equivalent to statistical parity as defined in Eq. (3.1). This category also includes conditional statistical parity Eq. (3.2).

Separation. Separation means that the prediction $\hat{Y}$ is conditionally independent of the sensitive feature A given the true decision Y.

$$\hat{Y} \perp A \mid Y \quad (or\ R \perp A \mid Y) \tag{2.2}$$

In the case where $\hat{Y}$ is a binary classifier, the formulation of separation is equivalent to that of the equalized odds (Eq. (3.3)). Equal opportunity (Eq. (3.4)), predictive equality (Eq. (3.5)), balance for positive class (Eq. (3.10)), and balance for negative class (Eq. (3.11)) are all relaxations of separation.

Sufficiency. Sufficiency implies that the sensitive attribute A and the target variable Y are conditionally independent given the prediction $\hat{Y}$.

$$Y \perp A \mid \hat{Y} \quad (or\ Y \perp A \mid R) \tag{2.3}$$

In the case of binary classification, sufficiency is equivalent to conditional use accuracy equality (Eq. (3.6)). Using the score R, Calibration (Eq. (3.13)), and well-calibration (Eq. (3.14)) can be considered as sufficiency [54]). Relaxation of sufficiency yields predictive parity (Eq. (3.7)), which also does not satisfy exactly the same incompatibility result as sufficiency.

Table 3.1 (Chapter 3/Section 3.2.2) lists some of the most known fairness notions and their classification.

2.3 Privacy

2.3.1 Local Differential Privacy

The focus of this thesis is solely on the local setting of DP (i.e., LDP) [121]. In other words, we assume that the centralized server in charge of aggregating data from individual users is not guaranteed to be trustworthy. Specifically, each user applies an ε-LDP mechanism to their data before submission to a central server. The server subsequently aggregates the ε-LDP data for statistical purposes (e.g., mean or frequency estimation). Formally, LDP is defined as follows:

Definition 1 **ε-LDP.** *A randomized algorithm $\mathcal{L}$ satisfies ε-LDP, where ε is a positive real number representing the privacy parameter, if for any pair of input values $v_1, v_2 \in dom(\mathcal{L})$ and any possible output z of $\mathcal{L}$:*

$$\mathbb{P}[\mathcal{L}(v_1) = z] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{L}(v_2) = z].$$

In essence, LDP guarantees that it is unlikely for the data aggregator to reconstruct the data source regardless of the prior knowledge. The privacy level ε controls the privacy-utility trade-off for which lower values of ε result in tighter privacy protection. Like central DP, LDP also has several important properties, such as immunity to post-processing and composability [77], defined as follows.

Proposition 1 Post-Processing [77]. *If $\mathcal{L}$ is ε-LDP, then for any function f, the composition of $\mathcal{L}$ and f, i.e., $f(\mathcal{L})$ satisfies ε-LDP.*

Proposition 2 Sequential Composition [77]. *Let $\mathcal{L}_1$ be an ε_1-LDP protocol and $\mathcal{L}_2$ be an ε_2-LDP protocol, then, the protocol $\mathcal{L}_{1,2}(v) = (\mathcal{L}_1(v), \mathcal{L}_2(v))$ is $(\varepsilon_1 + \varepsilon_2)$-LDP.*

2.3.2 LDP Protocols

LDP has been proposed for many tasks, including statistical analysis, data mining, ML, location privacy, etc [259, 256]. This thesis will use state-of-the-art LDP frequency estimation protocols [15, 245]. Frequency (or histogram) estimation is a primary objective of LDP as it is a building block for more complex tasks such as heavy hitter estimation [26], joint distribution estimation [126, 195], ML [150], and frequency monitoring [17, 79, 69, 14].

In this subsection, we briefly review several state-of-the-art LDP frequency estimation protocols. Note that this thesis will focus exclusively on the user-side randomization process, specifically on learning over locally differentially private data. This means we will apply ML models to users' randomized data rather than concentrating on server-side frequency estimation. Let $A = \{a_1, \ldots, a_k\}$ be a sensitive attribute with a discrete domain of size $k = |A|$.

Randomized Response (RR). RR was proposed by Warner [246] to provide "plausible deniability" to individuals responding to embarrassing (binary) questions in a survey. RR is formally defined as:

$$\mathrm{RR}(a) = \begin{cases} a & \text{with probability} \quad p = \frac{e^{\varepsilon}}{e^{\varepsilon}+1}, \\ \overline{a} & \text{with probability} \quad q = \frac{1}{e^{\varepsilon}+1}, \end{cases} \tag{2.4}$$

where $\overline{a} = 1$ if $a = 0$ and, viceversa, $\overline{a} = 0$ if $a = 1$. p denotes the probability that the reported value is the true value, and q is the probability that the value is reported at random. It is easy to prove that RR satisfies ε-LDP as $p/q = e^{\varepsilon}$.

k-Ary Randomized Response. Kairouz *et al.* [115] generalized RR to domains of arbitrary size k (with $k \geq 2$), and proposed the well-known k-RR mechanism, which is one classical technique for achieving LDP on categorical/discrete data. k-RR uses no particular encoding. Given a value $a \in dom(A)$, k-RR(a) outputs the true value a with probability p, and any other value $a' \in dom(A) \setminus \{a\}$, otherwise. More formally:

$$\forall z \in dom(A): \quad \mathbb{P}[k\text{-RR}(a) = z] = \begin{cases} p = \frac{e^{\varepsilon}}{e^{\varepsilon}+k-1} \text{ if } z = a, \\ q = \frac{1}{e^{\varepsilon}+k-1} \text{ if } z \neq a. \end{cases} \tag{2.5}$$

where z is the perturbed value sent to the server.

Binary Local Hashing (BLH). Local Hashing (LH) protocols [26, 245] can handle a large domain size k by first using hash functions to map an input value to a smaller domain size g (typically $2 \leq g \ll k$), and then applying k-RR to the hashed value. Let $\mathscr{H}$ be a universal hash function family such that each hash function $\mathrm{H} \in \mathscr{H}$ hashes a value in A into $[g]$, i.e., $\mathrm{H}: A \to [g]$. With BLH, $[g] = \{0, 1\}$, each user selects at random one hash function H, calculates $b = \mathrm{H}(v)$, and perturbs b to z as:

$$\mathbb{P}[z = 1] = \begin{cases} p = \frac{e^{\varepsilon}}{e^{\varepsilon}+1} \text{ if } b = 1, \\ q = \frac{1}{e^{\varepsilon}+1} \text{ if } b = 0. \end{cases}$$

The user sends the tuple $\langle \mathrm{H}, z \rangle$, i.e., the hash function and the perturbed value. Thus, for each user, the server can calculate the subset of all values $v \in A$ that hash to z, i.e., $S(\langle \mathrm{H}, z \rangle) = \{v | \mathrm{H}(v) = z\}$.

Optimal LH (OLH). To improve the utility of LH protocols, Wang et al. [245] proposed OLH in which the output space of the hash functions in family $\mathscr{H}$ is no longer binary as in BLH. Thus, with OLH, $g = \lfloor e^{\varepsilon} + 1 \rceil$, each user selects at random one hash function H, calculates $b = \mathrm{H}(v)$, and perturbs b to z as:

$$\forall i \in [g]: \quad \mathbb{P}[z = i] = \begin{cases} p = \frac{e^{\varepsilon}}{e^{\varepsilon}+g-1} \text{ if } b = i. \\ q = \frac{1}{e^{\varepsilon}+g-1} \text{ if } b \neq i. \end{cases}$$

Similar to BLH, the user sends the tuple $\langle \mathrm{H}, z \rangle$ and, for each user, the server can calculate the subset of all values $v \in A$ that hash to z, i.e., $S\left(\langle \mathrm{H}, z \rangle\right) = \{v | \mathrm{H}(v) = z\}$.

RAPPOR. The RAPPOR [79] protocol uses One-Hot Encoding (OHE) to interpret the user's input $v \in A$ as a one-hot k-dimensional vector. More precisely, $\mathbf{v} = \mathrm{OHE}(v)$ is a binary vector with only the bit at position v set to 1 and the other bits set to 0. Then, RAPPOR randomizes the bits from $\mathbf{v}$ independently to generate $\mathbf{z}$ as follows:

$$\forall i \in [k]: \quad \mathbb{P}[\mathbf{z}_i = 1] = \begin{cases} p = \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1} \text{ if } \mathbf{v}_i = 1, \\ q = \frac{1}{e^{\varepsilon/2}+1} \text{ if } \mathbf{v}_i = 0, \end{cases}$$

where $p + q = 1$ (i.e., symmetric). Afterward, the user sends $\mathbf{z}$ to the server.

Optimal Unary Encoding (OUE). To minimize the variance of RAPPOR, Wang et al. [245] proposed OUE, which perturbs the 0 and 1 bits asymmetrically, i.e., $p + q \neq 1$. Thus, OUE generates $\mathbf{z}$ by perturbing $\mathbf{v}$ as follows:

$$\forall i \in [k]: \quad \mathbb{P}[\mathbf{z}_i = 1] = \begin{cases} p = \frac{1}{2} \quad \text{if } \mathbf{v}_i = 1, \\ q = \frac{1}{e^{\varepsilon}+1} \text{ if } \mathbf{v}_i = 0. \end{cases}$$

Afterward, the user sends $\mathbf{z}$ to the server.

Subset Selection (SS). The SS [244, 261] protocol randomly selects $1 \leq \omega < k$ items within the input domain to report a subset of values $\Omega \subseteq A$. The user's true value v has a higher probability of being included in the subset Ω, compared to the other values in $A \setminus \{v\}$. The optimal subset size that minimizes the variance is $\omega = \max\left(1, \left\lfloor \frac{k}{e^{\varepsilon}+1} \right\rceil\right)$. Given a value $v \in A$, $\mathrm{SS}(v)$ starts by initializing an empty subset Ω. Afterwards, the true value v is added to Ω with probability $p = \frac{\omega e^{\varepsilon}}{\omega e^{\varepsilon}+k-\omega}$. Finally, it adds values to Ω as follows:

- If $v \in \Omega$, then $\omega - 1$ values are sampled from $A \setminus \{v\}$ uniformly at random (without replacement) and are added to Ω;
- If $v \notin \Omega$, then ω values are sampled from $A \setminus \{v\}$ uniformly at random (without replacement) and are added to Ω.

Afterward, the user sends the subset Ω to the server.

Thresholding with Histogram Encoding (THE). Histogram Encoding (HE) [245] encodes the user value as a one-hot k-dimensional histogram, i.e., $\mathbf{v} = [0.0, 0.0, \ldots, 1.0, 0.0, \ldots, 0.0]$ in which only the v-th component is 1.0. HE($\mathbf{v}$) perturbs each bit of $\mathbf{v}$ independently using the Laplace mechanism [76]. Two different input values $v_1, v_2 \in A$ will result in two vectors with L1 distance of $\Delta = 2$. Thus, HE will output $\mathbf{z}$ such that $\mathbf{z}_i = \mathbf{v}_i + \text{Lap}\,(2/\varepsilon)$. To improve the utility of HE, Wang et al. [245] proposed THE such that the user reports (or the server computes): $S(\mathbf{z}) = \{v \mid \mathbf{z}_v > \theta\}$, in which θ is the threshold with optimal value in $(0.5, 1)$.

2.4 Causality

We recall that a directed acyclic graph (DAG) $\mathcal{G} = (V, \mathcal{E})$ is composed of a set of variables/vertices V and a set of (directed) edges $\mathcal{E}$ between them such that no cycle is formed. Let $\mathcal{P}$ be the probability distribution over the same set of variables V. $\mathcal{G}$ and $\mathcal{P}$ are related through the Markov condition if every variable is conditionally independent of its non-descendants given its parents. Assuming the Markov condition, the joint distribution of variables $V_1, V_2, \ldots \in V$ can be factorized as:

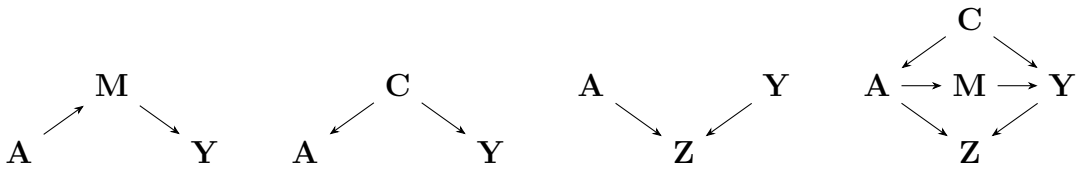
$$\mathbb{P}[V_1, V_2, \ldots] = \prod_i \mathbb{P}[V_i | Pa(V_i)] \tag{2.6}$$

where $Pa(V_i)$ denotes the set of parents of V_i.

Pairing a DAG $\mathcal{G}$ and a probability distribution $\mathcal{P}$ such that they are related with the Markov condition forms a Bayesian network (BN) $\mathcal{B} = \langle \mathcal{G}, \mathcal{P} \rangle$ [179]. A causal Bayesian network (CBN) [181] is a BN where edges have causal interpretations. That is, an edge between two variables V_i and V_j ($V_i \to V_j$) means that if all other variables are fixed to some values and we change the value of V_i, then V_j will possibly change, but never the other way around. A directed edge $V_i \to V_j$ indicates a causal relation from the cause variable V_i to the effect variable V_j. A partially directed acyclic graph (PDAG) is a particular DAG type containing directed and undirected edges. A causal graph has three basic causal structures: a mediator (Fig. 2.1a), a confounder (Fig. 3.8b), and a collider (Fig. 2.1c).

Conditional independence between variables can be graphically identified using the d-separation criterion [181]. A path[1] p is $d-$separated (or blocked) by a set of vertices W if and only if (1) if p contains a mediator structure ($X \to M \to Y$) or a confounder

[1] A path is a sequence of directed edges between two variables not necessarily pointing to the same direction. For instance, $A \leftarrow C \to Y$ in Fig. 3.8b is a path, although the edges are not pointing in the same direction.



(a) Mediator (b) Confounder (c) Collider (d) Combined

Fig. 2.1 Basic causal structures.

structure ($X \leftarrow M \rightarrow Y$), then M is in W and (2) if p contains a collider structure[2] ($X \rightarrow Z \leftarrow Y$), then the collider Z and all its descendants are not in W. If a set W d-separates (blocks) every path from X to Y, then X and Y happen to be conditionally independent given W. For instance, in Fig. 2.1d, the set $W = \{C, M\}$ $d-$separates A and Y, hence A and Y are conditionally independent given $\{C, M\}$ ($A \perp Y|C, M$). Similarliy, $A \perp Y|M$ and $A \perp Y|C$ in Fig. 2.1a and Fig. 2.1b, respectively. Note that, in the presence of a mediator or a confounder between two variables, these two variables become dependent ($A \not\perp Y$ in Figs 2.1a and 2.1b). However, in the presence of a collider Z, A and Y become independent ($A \perp Y$ in Fig. 2.1c), but when conditioning on the collider Z, they become dependent ($A \not\perp Y|Z$).

DAGs with the same d-separation properties are called Markov equivalent and imply the same conditional independence relations. Any maximal collection of DAGs, which are Markov equivalent, is called a Markov Equivalence Class (MEC). A completed partially directed acyclic graph (CPDAG) is a particular type of PDAG that serves as representative for Markov equivalence classes of DAGs.

As was stated in Section 1.1.3, there are two fundamental frameworks to mathematically represent and characterize causal relations between variables: structural causal model [181] and potential outcome [110]. The terminology and the notation for both frameworks are introduced in what follows.

Structural Causal Model (SCM) Framework. A structural causal model [181] is a tuple $M = \langle U, V, F, \mathbb{P}[U] \rangle$ where:

- U is a set of exogenous variables that cannot be observed or experimented on but constitute the background knowledge behind the model.
- V is a set of observable variables that can be experimented on.

[2] Called also v-structure.

- F is a set of structural functions where each f_i is mapping $U \cup V \rightarrow V \backslash \{V_i\}$ which represents the process by which the variable V_i changes in response to other variables in $U \cup V$.
- $\mathbb{P}[u]$ is a probability distribution over the unobservable (latent) variables U.

Unobserved variables U, typically not represented in the causal diagram, can be mutually independent (Markovian model) or dependent on each other (semi-Markovian model). In semi-Markovian models, each $U_i \in U$ is used in at most two functions in F. In causal diagrams of semi-Markovian models, dependent unobservable variables (unobserved confounders) are represented by a dotted bi-directed edge between observable variables. Fig. 2.2 shows causal graphs of Markovian model (Fig. 2.2a), semi-Markovian model (Fig.s 2.2b) and semi-Markovian model after intervening on Z (Fig. 2.2c).

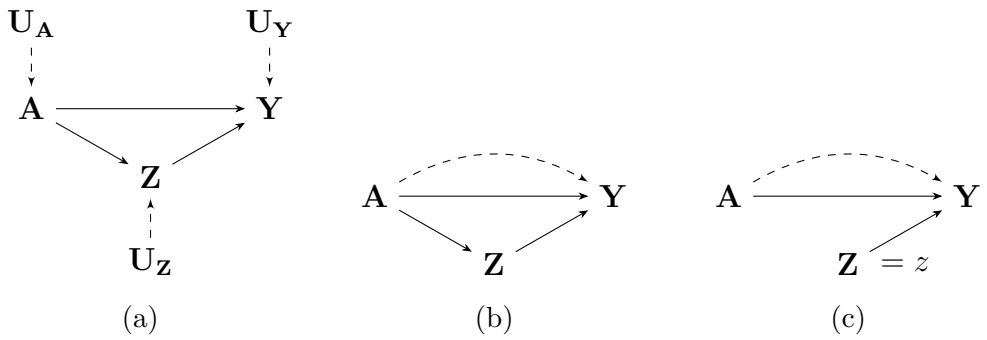


Fig. 2.2 Markovian and semi-Markovian causal models.

Potential Outcome Framework. As was mentioned in Section 1.1.3, expressing causal relations in the potential outcome framework starts at the unit level. A unit i is the atomic research object. In fairness problems, the unit typically refers to an individual. For example, every candidate corresponds to a unit i in a job hiring scenario. More specifically, the sensitive attribute of the candidate (e.g., gender) corresponds to the treatment in the potential outcome terminology. Given an outcome random variable Y, applying a treatment $A = a$ on a unit i yields a different random variable called the potential outcome $Y_i(A = a) = Y_i^a$. For example, if $A = 1$ refers to male, $A = 0$ refers to female, and Y is the hiring decision, Y_i^0 is the potential hiring decision of unit i when the gender (treatment) is female. Consequently, if the treatment variable A is binary, there are two potential outcomes Y_i^0 and Y_i^1. In observational studies (in contrast to experimental studies), only one potential outcome can be observed: the factual outcome. The other potential outcome is usually impossible to observe and is

called the counterfactual outcome. For example, if a job candidate i is female ($A = 0$) and is not hired, the potential outcome Y_i^0 is observed and is equal to 0. However, the potential outcome of that candidate i had she be male Y_i^1 is impossible to observe because this requires going back in time (impossible) and changing the sex of that individual to male (not ethical in the cases where it is possible).

2.4.1 Intervention and *do*-operator

An intervention noted $do(V = v)$ is a manipulation of the model that consists of fixing the value of a variable (or a set of variables) to a specific value regardless of the corresponding function f_v. Graphically, it consists of discarding all edges incident to the vertex corresponding to variable V. Fig. 2.2c shows the causal diagram of the manipulated model after intervention $do(Z = z)$ denoted $M_{Z=z}$ or M_z for short. The intervention $do(V = v)$ induces a different distribution on the other variables. For example, in Fig. 2.2c, $do(Z = z)$ results in a different distribution on Y, namely, $\mathbb{P}[Y|do(Z = z)]$. Intuitively, while $\mathbb{P}[Y|Z = z]$ reflects the population distribution of Y among individuals whose Z value is z, $\mathbb{P}[Y|do(Z = z)]$ reflects the population distribution of Y if *everyone in the population* had their Z value fixed at z. The obtained distribution $\mathbb{P}[Y|do(Z = z)]$ can be considered as a *counterfactual* distribution since the intervention forces Z to take a value different from the one it would take in the actual world. Such counterfactual variable is noted $Y_{Z=z}$ or Y_z for short[3]. The term counterfactual quantity is used for expressions that involve explicitly multiple worlds. In Fig. 2.2b, consider the expression $\mathbb{P}[y_{a'}|Y = y, A = a] = \mathbb{P}[y_{a'}|y, a]$. Such expression involves two worlds: an observed world where $A = a$ and $Y = y$ and a counterfactual world where $Y = y$ and $A = a'$ and it reads "the probability of $Y = y$ had A been a' given that we observed $Y = y$ and $A = a$. In the common example of job hiring, if A denotes race (a :white, a':non-white) and Y denotes the hiring decision (y:hired, y':not hired), $\mathbb{P}[y_{a'}|y, a]$ reads "given that a white applicant has been hired, what is the probability that the same applicant is still being hired had he been non-white".

Nesting counterfactuals can produce complex expressions. For example, in the relatively simple model of Fig. 2.2b, $\mathbb{P}[y_{a,z_{a'}}|y'_{a'}]$ reads the probability of $Y = y$ had (1) A been a' and (2) Z been z when A is a', given that an intervention $A = a'$ produced y'. This expression involves three worlds: a world where $A = a$, a world where $Z = z_{a'}$, and a world where $A = a'$. Such complex expressions characterize direct, indirect, and path-specific effects.

[3]The notations $Y_{Z\leftarrow z}$ and $Y(z)$ are used in the literature as well. $\mathbb{P}[Y = y|do(Z = z)] = \mathbb{P}[Y_{Z=z} = y] = \mathbb{P}[Y_z = y] = \mathbb{Y}[y_z]$ is used to define the causal effect of z on Y.

Causal inference aims to determine if the outcome of automated decision-making is fair or discriminative. Several causality-based fairness notions are defined in the literature (Section 3.4.2) and expressed in terms of joint, conditional, interventional, and counterfactual probabilities. Applying a causality-based fairness notion requires inputting a dataset D and a causal graph G. While joint probabilities (e.g., $\mathbb{P}[X = x, Y = y, Z = z]$) and conditional probabilities (e.g., $\mathbb{P}[Y = y|X = x]$) can be trivially estimated from the dataset D, probabilities involving interventions or counterfactuals cannot always be estimated from D and G. When a probability can be estimated from observable data (D), it is said to be *identifiable*. Otherwise, it is *unidentifiable* (More details on identifiability will be presented in Section 3.4.3).

2.4.2 Causal Assumptions

Causal inference relies on causal assumptions inherent in the potential outcome and Pearl's SCM frameworks.

Causal Markov Condition [181] (SCM). There is a general consensus that it is fundamental to causal inference and, hence, typically required. This assumption is already explained above (Eq. (2.6)).

Causal Faithfulness [181] (SCM). A causal graph $\mathcal{G}$ and a probability distribution $\mathcal{P}$ over the same variables V are faithful to each other if all and only the conditional independence relations that hold in $\mathcal{P}$ are entailed by the Markov condition and d-separation in $\mathcal{G}$. An example of faithfulness violation is when two variables are dependent on the causal graph but independent in the data. Consider the graph in Fig. 2.1a. If, in the data, A on Y are exactly balanced out by the indirect causal effect $A \to M \to Y$, A and Y will appear independent in the data while they are dependent in the graph. In that case, a causal discovery procedure assuming faithfulness will return a collider structure on M ($A \to M \leftarrow Y$) as a causal graph.

Causal Sufficiency [181] (SCM). Causal sufficiency implies that there are no latent (hidden) confounders between variables in $\mathbf{V}$. It is a very strong assumption, as its absence or presence may lead to very different causal graphs. Violation of causal sufficiency may sometimes be detected from data. An example of a causal sufficiency violation could be in a study examining the relationship between smoking and lung cancer. If the study only considers smoking as the sole factor influencing lung cancer and ignores other potential confounding variables such as air pollution, genetic

predisposition, or occupational exposure to carcinogens, it would violate the principle of causal sufficiency.

SUTVA [110] (Potential Outcome). SUTVA (Stable Unit Treatment Value Assumption) has two requirements. First is the absence of interference among units. In the job hiring example, it means that the hiring decision for a candidate is independent of the hiring decisions of all other candidates. Second, there is only one version of the treatment. This is more relevant in medical scenarios when a treatment (medication) has different versions (e.g., different dosages). This requirement is typically satisfied for fairness scenarios as the treatment generally corresponds to an individual's intrinsic attribute (e.g., gender, race, etc.).

Ignorability [110] (Potential Outcome). Ignorability is satisfied when the sensitive attribute A and the potential outcome variables Y^0 and Y^1 are independent given observable variables X. That is, $A \perp Y^0, Y^1 | X$[4]. This corresponds to the absence of hidden (unobservable) confounders. In the SCM framework, it is equivalent to the causal sufficiency assumption.

Positivity [110] (Potential Outcome). Positivity states that there should be a non-zero probability of receiving each treatment level for any combination of covariates. That is, $\mathbb{P}[A = a | X = x] > 0$, for all a, and x. In other words, every individual in the population should have a chance of being exposed to each treatment or intervention being studied, regardless of their covariate values. This assumption is crucial for ensuring that all relevant population subgroups are represented in the data and that causal effects can be estimated for all individuals. Violations of the positivity assumption can lead to biased estimates and unreliable causal inferences.

2.4.3 Causal Discovery Algorithms

The three main categories of causal discovery algorithms are constraint-based, score-based, and procedures that exploit semi-parametric assumptions. This section describes one representative algorithm of each category[5]. Check [98] for a more comprehensive list of causal discovery algorithms.

[4]Strong ignorability is a stronger assumption requiring independence between the potential outcomes and any covariate X ($X \perp Y^0, Y^1$).

[5]Except for the first category, as PC and FCI are both constraint-based algorithms and FCI is considered a variant of PC.

PC [218]. The PC algorithm is a constraint-based causal discovery algorithm used to identify causal relationships between variables in observational data. The algorithm is based on conditional independence tests and uses a series of statistical tests to infer the presence or absence of causal relationships between variables.

The PC algorithm consists of two main steps. The first step is the skeleton identification, where the algorithm begins by constructing an undirected graph, called the skeleton, that represents the conditional independence relationships between variables in the data. This step involves performing conditional independence tests to determine which pairs of variables are conditionally independent given other variables. The second step consists of the orientation of the edges, where the algorithm attempts to orient the edges of the graph to establish causal directions between variables. These rules are based on patterns observed in conditional independence relationships and help to infer causal directions between variables. The output of PC is a CPDAG. The conditional independence tests used to discover the skeleton of the graph for PC have an α value for rejecting the null hypothesis, which is always a hypothesis of independence or conditional independence.

FCI [221]. The FCI algorithm [219] is also a constraint-based algorithm and is considered a generalization of the PC algorithm. The main difference between PC and FCI is that the latter considers the presence of common hidden confounders between observed variables. Consequently, instead of producing a DAG, the output of FCI is a partial ancestral graph (PAG) with possibly four types of edges: $\longrightarrow, \longleftrightarrow, \circ\!\!-\!\!, \circ\!\!-\!\!\circ, \circ\!\!\longrightarrow$. The "$\circ$" mark represents an undetermined edge mark. In other words, "$\circ$" can be either a tail "$-$" or a head "$\rightarrow$". $\longleftrightarrow$ shows that there are hidden confounders between the two variables on either side of the arrow. $X \circ\!\!\longrightarrow Y$ implies that either X causes Y or there are hidden confounders between both variables. $X \circ\!\!-\!\!\circ Y$ might be: X causes Y, Y causes X, there are common hidden confounders between both variables, X causes Y and there are hidden confounders between both variables, or Y causes X and there are hidden confounders between both variables. As in the first step of the PC algorithm, FCI relies on statistical independence tests to infer the skeleton of the graph. In the second step, FCI deviates from the PC algorithm.

After orienting all the edges in the graph as $\circ\!\!-\!\!\circ$, the algorithm starts with an orientation rule to detect the v-structures in the graph.

Another rule specific to FCI is the detection of Y-structures. Four variables define a Y-structure when: $C1 \rightarrow X \leftarrow C2$ and $X \rightarrow Y$. Within the Y-structure, $C1$ and $C2$ are independent of Y conditional on X. This conditional independence helps

exclude the possibility of a latent confounder between X and Y. When FCI detects a Y-structure in the graph, no latent confounders exist between X and Y; otherwise, FCI assumes that possibly latent confounders exist [162].

Afterward, FCI applies four additional rules to direct the remaining edges. Note that FCI is limited to several thousand variables. Because FCI is a variant of PC, the same assumptions hold for FCI, except causal sufficiency, which allows FCI to work in the presence of hidden confounders.

GES [52]. Greedy Equivalence Search (GES) is a score-based algorithm that, unlike PC and FCI, starts with a completely disconnected graph and then adds, deletes, and modifies edges in a particular order until reaching the causal model that maximizes a regularized performance score, called BIC score, that stands for Bayes Information Criterion [203] (BIC) which is a likelihood-based model selection criterion. A first remark about GES is that its output is not necessarily a directed acyclic graph (DAG) but a CPDAG, representing a Markov equivalence class of causal DAGs. GES consists of searching over an abstract search space (graph) of states and transitions. Each state is an equivalence class of DAGs, all with the same BIC score and are represented as a CPDAG. The search objective is the state that maximizes BIC score, hence, the abstract output of GES is an equivalence class of DAGs.

The greedy strategy of GES consists of repeatedly following the best forward transition at each state that it encounters until a local maximum is reached, i.e., until the next state reduces the BIC score, and then, analogously, repeatedly following the best backward transition until a local maximum is reached. These two consecutive algorithms that form GES are called the forward and backward phases.

The computation of the neighboring states of a given state (for both phases) is carried out by finding edges $X \to Y$ that can be added (or removed) in such a way that the resulting PDAG can be *extended*, i.e., transformed into a DAG by smartly deciding the direction of the undirected edges. Once the DAG is computed, it is *completed* to obtain the CPDAG that represents the equivalence class containing it.

LiNGAM [208]. LiNGAM is an algorithm based on causal asymmetries that, unlike the previously discussed algorithms, yields a unique directed graph (DAG) and corresponding parameters. However, the stronger causal discovery power comes at the expense of more assumptions that must be satisfied. LiNGAM requires linearity and non-gaussianity of the variables to recover causal directions and learn functional relationships [207]. The approach is closely related to the Independent Component

Analysis (ICA) algorithm as they both base their premises on the Darmois-Skitovic theorem [177]. The theorem implies that fitting a backward model (trying to regress the cause on the effect) to the data would result in dependence between cause X and the residuals of the effect Y, allowing to correct the causal direction.

DirectLiNGAM [208] is a variant of LiNGAM which, in contrast to the ICA version, is not based on iterative search and, therefore, does not require initial guess or similar parameters and is guaranteed to converge to the right solution. The DirectLiNGAM algorithm implementation learns the causal graph in two steps. First, it finds the causal order of the variables: an ordered list, where the first is the exogenous variable (has no parents in the graph), the second is the child of the exogenous variable with the most descendants, etc. Next, the causal order is used to compute the adjacency matrix that specifies the strength of the connections. Specifically, starting from the end of the list, each variable is regressed on all the others that come before it in the causal order (potential parents).

2.4.4 Conclusion

In conclusion, this chapter has provided a comprehensive overview of the preliminaries and background information essential for understanding the context of the thesis. It began by discussing the fundamental concepts of fairness, including classifying fairness notions. Then, an overview of LDP, exploring some state-of-the-art LDP protocols, is provided. The last section of this chapter provided the necessary background for causality. In the next chapter, we will delve into our contribution to the applicability of fairness notions in real-world applications.

Chapter 3

Fairness Notions and their Applicability

3.1 Introduction

One of the key contributions of this thesis is its endeavor to narrow the gap between fairness metrics and their practical implementation in real-world contexts. Fairness, particularly in decision-making scenarios, inherently involves subjective considerations. Therefore, a critical aspect revolves around selecting appropriate evaluation metrics to assess fairness. Defining and quantifying fairness within AI systems is multifaceted and contingent on specific contexts, rendering it complex, primarily because some notions of fairness are incompatible and cannot be carried out simultaneously [24, 166]. The purpose of this chapter is threefold: first, to provide a comprehensive overview of fairness notions and their related tensions to contextualize our research within the broader academic discourse (Section 3.2); second, to introduce our work on the applicability of statistical fairness notions in real-world contexts (Section 3.3); and third, to present our study on the applicability and suitability of causality-based fairness notions in practice (Section 3.4).

3.2 Fairness Notions and Related Tensions

With the recent interest in fairness, many notions have been defined to capture different aspects of fairness. As fairness is a social construct [111] and an ethical concept [237], defining it is still prone to subjectivity. Hence, the aim of replacing subjective human decisions with objective ML-based decision systems resulted in notions and algorithms

still exhibiting unfairness. Therefore, although the different notions of algorithmic fairness appear internally consistent, several of them cannot hold simultaneously and hence are mutually incompatible [24, 166, 101, 32, 130]. As a consequence, practitioners assessing and/or implementing fairness need to choose among them.

Contributions. In this study, we examined prevalent notions of fairness, exploring the inherent conflicts within these notions. We also explored tensions between fairness and privacy. In this chapter, we solely focus on presenting fairness notions and tensions among them, and we leave the tension between fairness and privacy for Chapter 4.

Outline. Section 3.2.1 discusses related work. Section 3.2.2 briefly presents commonly used fairness notions (group and individual) and their formal definitions. Section 3.2.3 describes the tensions and incompatibilities among the various fairness notions. Section 3.2.6 draws the conclusion.

3.2.1 Related Work

With the increasing need for ethical concerns in decision-making systems that have severe implications for individuals and society, several survey papers have been proposed in the literature in recent years. In this section, we revisit these survey papers and highlight how our survey deviates from them. Mehrabi et al. [164] proposed a broader scope for their overview: in addition to concisely listing 10 definitions of fairness metrics, they discussed different sources of bias and different types of discrimination, they listed methods to mitigate discrimination categorized into pre-processing, in-processing, and post-processing, and they discussed potential directions for contributions in the field. However, they did not discuss any tensions between fairness notions, which we discuss in depth in this survey. The survey of Mitchell et al. [166] includes an exhaustive list of group and individual fairness notions and outlines most of the impossibility results among them. They also discussed in detail a "catalogue" of choices and assumptions in the context of fairness to address the question of how social goals are formulated into a prediction (ML) problem. Their survey does not tackle the problem of tensions between fairness and other ethical considerations in decision-making systems as is studied in this work.

Tsamados et al. [237] compiled an overview of the ethical problems in AI algorithms and the solutions proposed in the literature. In particular, they provided a conceptual map of six ethical concerns raised by AI algorithms, namely inconclusive, inscrutable, misguided evidence, unfair outcomes, transformative effects, and traceability. The first three concerns refer to epistemic factors, the fourth and the fifth are normative factors, and the fifth is relevant to both epistemic and normative factors. The epistemic factors

are related to the relevance of the accuracy of the data, while the informative factors refer to the ethical impact of AI systems. Although the survey explores a broad scope related to ethical concerns in AI, it remains at a conceptual level. It does not address how these ethical concerns are implemented in practice and how they conflict in detail, which we explore in depth in this study.

Other works discussing the trade-off between fairness notions include the work by Kleinberg et al.[130], which discussed the suitability of specific fairness notions in a specific setup. In particular, they discussed the applicability of calibration and balance notions. The survey of Berk et al. [32] studied the trade-offs between different group fairness notions and between fairness and accuracy in a specific context, namely: criminal justice risk assessments. They used simple examples based on the confusion matrix to highlight relationships between the notions of fairness.

In another research direction, Friedler et al. [89] discussed tensions between group and individual fairness. In particular, they defined two worldviews, WYSIWYG and WAE. The WYSIWYG (What you see is what you get) worldview assumes that the unobserved (construct) space and observed space are essentially the same. In contrast, the WAE (we're all equal) worldview implies no inherent differences between groups of individuals based on potentially sensitive attributes.

3.2.2 Fairness Notions

We recall some notations used in this chapter. Let V, A, and X be three random variables representing, respectively, the total set of attributes, the sensitive attributes, and the remaining attributes describing an individual such that $V = (X, A)$ and $\mathbb{P}[V = v_i]$ represents the probability of drawing an individual with a vector of values v_i from the population. For simplicity, we focus on the case where A is a binary random variable where $A = 0$ designates the protected (unprivileged) group, while $A = 1$ designates the non-protected (privileged) group. Let Y and $\hat{Y}$ be binary random variables representing, respectively, the true decision and the predicted outcome where $Y = 1$ designates a positive instance, while $Y = 0$ is a negative one. Typically, the predicted outcome $\hat{Y}$ is derived from a score represented by a random variable R where $\mathbb{P}[R = r]$ is the probability that the score value equals r.

All fairness notions presented in this section address the question: "Is the outcome/prediction of the ML system fair towards individuals?". As stated in Chapter 2, most of the proposed fairness notions are properties of the joint distribution of the above random variables (X, A, Y, $\hat{Y}$, and R). They can also be interpreted using the confusion matrix and the related metrics (Table 2.2).

A straightforward approach to address the fairness problem is to ignore any sensitive attribute while training the ML system. This is called *fairness through unawareness*[1]. We don't treat this approach as a notion of fairness since, given a model prediction, it does not allow us to tell whether the model is fair. Besides, it suffers from the basic problem of proxies. Many attributes (e.g., home address, neighborhood, attended college) might be highly correlated to the sensitive attributes (e.g., race) and act as proxies of these attributes. Consequently, in almost all situations, removing the sensitive attribute during the training process does not address the problem of fairness [99].

Table 3.1 depicts the fairness notions presented in this chapter along with their classification.

[1]Known also as: blindness, unawareness [166], anti-classification [57], and treatment parity [144].

Table 3.1 Classification of statistical fairness notions. ($*$ notion newly defined in this study)

Fairness Notion	Ref.	Formulation	Classification	Type
Statistical parity	[75]	$\mathbb{P}[\hat{Y}=1 \mid A=0]=\mathbb{P}[\hat{Y}=1 \mid A=1]$	Independence (equivalent or relaxed$^\bigstar$)	Group
Conditional statistical parity	[58]	$\mathbb{P}[\hat{Y}=1 \mid E=e, A=0]=\mathbb{P}[\hat{Y}=1 \mid E=e, A=1]^\bigstar$		
Equalized odds	[101]	$\mathbb{P}[\hat{Y}=1 \mid Y=y,\ A=0]=\mathbb{P}[\hat{Y}=1 \mid Y=y,\ A=1] \quad \forall y \in \{0,1\}$	Separation (equivalent or relaxed$^\bigstar$)	
Equal opportunity		$\mathbb{P}[\hat{Y}=1 \mid Y=1, A=0]=\mathbb{P}[\hat{Y}=1 \mid Y=1, A=1]^\bigstar$		
Predictice equality	[58]	$\mathbb{P}[\hat{Y}=1 \mid Y=0, A=0]=\mathbb{P}[\hat{Y}=1 \mid Y=0, A=1]^\bigstar$		
Balance for positive class	[130]	$\mathbb{E}[R \mid Y=1, A=0)]=\mathbb{E}[R \mid Y=1, A=1]^\bigstar$		
Balance for negative class		$\mathbb{E}[R \mid Y=0, A=0]=\mathbb{E}[R \mid Y=0, A=1]^\bigstar$		
Overall balance	*	$\mathbb{E}[R \mid Y=y, A=0]=\mathbb{E}[R \mid Y=y, A=1] \quad \forall y \in \{0,1\}$		
Conditional use acc. equality	[32]	$\mathbb{P}[Y=y \mid \hat{Y}=y, A=0]=\mathbb{P}[Y=y \mid \hat{Y}=y, A=1] \quad \forall y \in \{0,1\}$	Sufficiency (equivalent or relaxed$^\bigstar$)	
Predictive parity	[54]	$\mathbb{P}[Y=1 \mid \hat{Y}=1, A=0]=\mathbb{P}[Y=1 \mid \hat{Y}=1, A=1]^\bigstar$		
Negative predictive parity	*	$\mathbb{P}[Y=1 \mid \hat{Y}=0, A=0]=\mathbb{P}[Y=1 \mid \hat{Y}=0, A=1]^\bigstar$		
Calibration	[54]	$\mathbb{P}[Y=1 \mid R=r, A=0]=\mathbb{P}[Y=1 \mid R=r, A=1] \quad \forall\, r \in [0,1]$		
Well-calibration	[130]	$\mathbb{P}[Y=1 \mid R=r, A=0]=\mathbb{P}[Y=1 \mid R=r, A=1]=r \quad \forall\, r \in [0,1]$		
Overall accuracy equality	[32]	$\mathbb{P}[\hat{Y}=Y \mid A=0]=\mathbb{P}[\hat{Y}=Y \mid A=1]$	Other metrics from confusion matrix	
Treatment equality		$\frac{FN}{FP}(\mathrm{A}=0)=\frac{FN}{FP}(\mathrm{A}=1)$		
Total fairness		–	Independence, Separation and Sufficiency	
Causal discrimination	[92]	$X_{(\mathrm{A}=0)}=X_{(\mathrm{A}=1)} \wedge A_{(\mathrm{A}=0)} \neq A_{(\mathrm{A}=1)} \Rightarrow \hat{y}_{(\mathrm{A}=0)}=\hat{y}_{(\mathrm{A}=1)}$	Similarity Metric	Individual
Fairness through awareness	[75]	$D(M(v_i), M(v_j)) \leq d(v_i, v_j)$		

In Appendix A.1.1, we use a simple job hiring scenario to explain how the fairness metrics presented in what follows are computed in practice.

Statistical Parity [75]. Statistical parity (a.k.a demographic parity [136], independence [23], equal acceptance rate [280], benchmarking [214], group fairness [75]) is one of the most commonly accepted notions of fairness. It requires the prediction to be statistically independent of the sensitive attribute ($\hat{Y} \perp A$). Thus, a classifier $\hat{Y}$ satisfies statistical parity if:

$$\mathbb{P}[\hat{Y} = 1 \mid A = 0] = \mathbb{P}[\hat{Y} = 1 \mid A = 1] \tag{3.1}$$

In other words, the predicted acceptance rates for both protected and unprotected groups should be equal. Using the confusion matrix (Table 2.2), statistical parity implies that $\frac{\text{TP+FP}}{\text{TP+FP+FN+TN}}$ should be equal for both groups.

Statistical parity is appealing in scenarios where a preferred decision is over another—for example, being accepted to a job, not being arrested, being admitted to a college, etc.[2]. What really matters is a balance in the prediction rate among all groups.

Statistical parity is suitable when the true decision Y is untrustworthy due to some flawed or biased measurement[3]. An example of this type of problem was observed in the recidivism risk prediction tool COMPAS [10]. Because minority groups are more controlled, and more officers are dispatched in their regions, the number of arrests (used to assess the level of crime [227]) of those minority groups is significantly higher than that of the rest of the population. Hence, for fairness purposes, in the absence of information to precisely quantify the differences in recidivism by race, the most suitable approach is to treat all sub-populations equally w.r.t recidivism [114].

Statistical parity is also well adapted to contexts where some regulations or standards are imposed. For example, a law might impose an equal hiring or admission of applicants from different sub-populations.

The main problem of statistical parity is that it does not consider a potential correlation between the true decision Y and the sensitive attribute A. In other words, statistical parity will be misleading if the underlying base rates of the protected and unprotected groups differ. In the ideal case ($\hat{y} = y$), this will lead to loss of utility [101]. For example, Fig. 3.1 illustrates a scenario for hiring computer engineers where equal proportions of male/female applicants have been predicted hired (60%), thus satisfying

[2]This might not be the case in other scenarios such as disease prediction, child maltreatment, where imposing a parity of positive predictions is meaningless.

[3]This is also known as differential measurement error [241].

statistical parity. However, when considering the true decision and, more precisely, the base rates that differ in both groups (0.3 for men versus 0.4 for women), the classifier becomes discriminative against female applicants (50% of qualified female applicants are not predicted to be hired). More generally, statistical parity is not recommended when the ground truth is available and used during the training phase as one might justify the disparity against the unprivileged group by use of this ground truth [264].

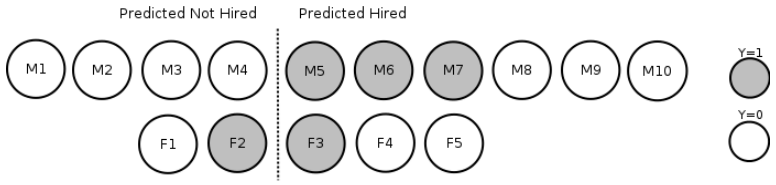


Fig. 3.1 A scenario for a hiring system where statistical parity is not recommended. F_i and M_i ($i \in [1-10]$) designate female and male applicants, respectively. The grey shaded circles indicate applicants who belong to the positive class, while white circles indicate applicants belonging to the negative class. The dotted vertical line is the prediction boundary. Thus, applicants at the right of this line are predicted to be hired, while applicants at the left are predicted not to be hired.

Another issue with this notion is its "laziness"; if we hire carefully selected applicants from the male group and random applicants from the female group, we can still achieve statistical parity, yet lead to negative results for the female group as its performance will tend to be worse than that of male group. This practice is an example of *self-fulfilling prophecy* [75] where a decision maker may simply select random members of a protected group rather than qualified ones and, hence, intentionally build a bad track record for that group. Barocas and Selbst refer to this problem as masking [25]. Masking is possible to game several fairness notions, but it is particularly easy to carry out in the case of statistical parity.

Conditional Statistical Parity [58]. Conditional statistical parity (a.k.a also conditional discrimination-aware classification in [120]) is a variant of statistical parity obtained by controlling on a set of legitimate attributes[4]. The legitimate attributes (we refer to them as E) among X are correlated with the sensitive attribute A and give some factual information about the true decision Y while leading to a *legitimate* discrimination. In other words, this notion removes the illegal discrimination, allowing the disparity in decisions to be present as long as they are explainable [58]. In a hiring system, possible explanatory factors that might affect the hiring decision for an applicant could be the education level and/or the job experience. Suppose the data comprises many highly educated and experienced male applicants and only a few

[4]Called explanatory attributes in [120].

highly educated and experienced women. In that case, one might justify the disparity between predicted acceptance rates between both groups and, consequently, does not necessarily reflect gender discrimination. Statistical parity holds if:

$$\mathbb{P}[\hat{Y} = 1 \mid E = e, A = 0] = P[\hat{Y} = 1 \mid E = e, A = 1] \tag{3.2}$$

In practice, conditional statistical parity is suitable when one or several attributes justify a possible disparate treatment between different groups in the population. Hence, choosing the legitimate attribute(s) is a very sensitive issue as it directly impacts the fairness of the decision-making process. More seriously, conditional statistical parity gives a decision-maker a tool to game the system and realize a self-fulfilling prophecy. Therefore, it is recommended to resort to domain experts or law officers to decide what is unfair and tolerable to use as legitimate discrimination attribute [120].

Equalized Odds [101]. Unlike the two previous notions, equalized odds (separation in [23], conditional procedure accuracy equality in [32], disparate mistreatment in [264], error rate balance in [54]) considers both the predicted and the actual outcomes. Thus, the prediction is conditionally independent of the sensitive attribute, given the true decision ($\hat{Y} \perp A \mid Y$). In other words, equalized odds require that both sub-populations have the same TPR and FPR (Table 2.2). In a hiring example, this means that the probability of an applicant who is actually hired to be predicted hired and the probability of an applicant who is actually not hired to be incorrectly predicted hired should be the same for men and women:

$$\mathbb{P}[\hat{Y} = 1 \mid Y = y,\ A = 0] = \mathbb{P}[\hat{Y} = 1 \mid Y = y,\ A = 1] \quad \forall y \in \{0, 1\} \tag{3.3}$$

Unlike statistical parity, equalized odds is well-suited for scenarios where the ground truth exists, such as disease prediction or stop-and-frisk [28]. It is also suitable when the emphasis is on recall (the fraction of the total number of positive instances that are correctly predicted positive) rather than precision (making sure that a predicted positive instance is actually a positive instance).

A potential problem of equalized odds is that it may not help close the gap between the protected and unprotected groups. For example, consider a group of 20 male applicants, of which 16 are qualified, and another equal size group of 20 females, of which only 2 are qualified. If the employer decides to hire 9 applicants while satisfying equalized odds, 8 offers will be granted to the male group, and only 1 offers will be granted to the female group. While this decision scheme looks fair in the short term,

in the long term, however, it will contribute to confirming this "unfair" status-quo and perpetuate this vicious cycle[5].

Because the equalized odds requirement is rarely satisfied in practice, two variants can be obtained by relaxing Eq. (3.3). The first one is called **equal opportunity** [101] (false negative error rate balance in [54]) and is obtained by requiring only TPR equality among groups:

$$\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 0] = \mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 1] \tag{3.4}$$

In a job hiring system, this is to say that we should hire an equal proportion of individuals from the qualified fraction of each group.

As $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ (Table 2.2) does not take into consideration false positives, equal opportunity is entirely insensitive to the number of false positives. This is an important criterion when considering this notion of fairness in practice. More precisely, equal opportunity should not be considered when a disproportionate number of false positives among groups has fairness implications. The scenario in Table A.3 in Appendix A.1.1 shows an extreme case of a job hiring dataset where the male group has a large number of false positives (Male $7 - 100$) while equal opportunity is satisfied.

To decide about the suitability of equal opportunity in a job hiring system, the question that should be answered by stakeholders and decision-makers is "If all other things are equal, is it fair to hire disproportionally more unqualified male candidates?". The employer shouldn't have several false positives (regardless of gender), as the company will end up with unqualified employees. For a stakeholder aiming to guarantee fairness between males and females, having more false positives in one group is not critical, provided these two groups have the same proportion of false negatives (a qualified candidate who is not hired).

In the scenario of predicting which employees to fire, however, a false positive (firing a well-performing employee) is critical for fairness. Hence, equal opportunity should not be used as a measure of fairness.

The second relaxed variant of equalized odds is called **predictive equality** [58] (false positive error rate balance in [54]), which requires only the FPR to be equal in both groups.

[5]If the job is a well-paid, male group tends to have a better living condition and affords better education for their kids, and thus enable them to be qualified for such well-paid jobs when they grow up. The gap between the groups will tend to increase over time.

In other words, predictive equality checks whether the accuracy of decisions is equal across protected and unprotected groups:

$$\mathbb{P}[\hat{Y} = 1 \mid Y = 0, A = 0] = \mathbb{P}[\hat{Y} = 1 \mid Y = 0, A = 1] \tag{3.5}$$

In a job hiring example, predictive equality holds when the probability of an applicant with an actual weak profile for the job being incorrectly predicted to be hired is the same for both men and women.

Since FPR = $\frac{\text{FP}}{\text{FP+TN}}$ (Table 2.2) is independent of false negatives, predictive equality is entirely insensitive to false negatives. Hence, predictive equality should not be used in scenarios where fairness between groups is sensitive to false negatives. Such scenarios include hiring and admission, where a false negative means qualified candidates are rejected disproportionally among groups. Predictive equality is acceptable in criminal risk assessment scenarios as false negatives (releasing a guilty person) are less critical than False positives (incarcerating an innocent person).

Predictive equality is particularly suitable for measuring the fairness of face recognition systems in crime investigations where security camera footage is analyzed. Fairness between ethnic groups with distinctive facial features is very sensitive to the FPR. A false positive means an innocent person is flagged as participating in a crime. Suppose this false identification happens at a much higher rate for a specific sub-population (e.g., dark-skinned ethnic group) than the rest of the population. In that case, it is clearly unfair for individuals belonging to that sub-population.

Looking at the problem from another perspective, choosing between equal opportunity and predictive equality depends on how the outcome/label is defined. In scenarios where the positive outcome is desirable (e.g., hiring, admission), fairness is typically more sensitive to false negatives rather than false positives; hence, equal opportunity is more suitable. In scenarios where the positive outcome is undesirable for the subjects (e.g., firing, risk assessment), fairness is typically more sensitive to false positives than false negatives, and hence predictive equality is more suitable.

Conditional Use Accuracy Equality [32]. Conditional use accuracy equality (also called sufficiency in [23]) is achieved when all population groups have equal PPV = $\frac{\text{TP}}{\text{TP+FP}}$ and NPV = $\frac{\text{TN}}{\text{FN+TN}}$ (Table 2.2). In other words, the probability of subjects with PPV truly belonging to the positive class and the probability of subjects with NPV belonging to the negative class truly should be the same:

$$\mathbb{P}[Y = y \mid \hat{Y} = y, A = 0] = \mathbb{P}[Y = y \mid \hat{Y} = y, A = 1] \quad \forall y \in \{0, 1\} \tag{3.6}$$

Intuitively, this definition implies equivalent accuracy for male and female applicants from both positive and negative predicted classes [242]. By contrast to equalized odds (Eq. (3.3)), one is conditioning on the algorithm's predicted outcome, not the actual outcome. In other words, this notion emphasizes the precision of the ML system rather than its sensitivity (a tradeoff discussed later in Section 3.3.3). A relaxation of conditional use accuracy equality requiring only equal PPV among groups is called **predictive parity** [54] (called outcome test in [214]) and is formally defined as follows:

$$\mathbb{P}[Y = 1 \mid \hat{Y} = 1, A = 0] = \mathbb{P}[Y = 1 \mid \hat{Y} = 1, A = 1] \tag{3.7}$$

In a hiring system, this is to say that the prediction used to determine the candidate's eligibility for a particular job should reflect the candidate's actual capability of doing this job, which is harmonious with the employer's benefit.

Like predictive equality (Eq. (3.5)), predictive parity is insensitive to false negatives. Hence, predictive parity should not be used in any scenario where fairness is sensitive to false negatives.

Choosing between predictive parity and equal opportunity depends on whether the scenario at hand is more sensitive to precision or recall. Typically, predictive parity is more suitable for precision-sensitive scenarios, while equal opportunity is more suitable for recall-sensitive scenarios. Precision-sensitive scenarios include disease prediction, child maltreatment risk assessment, and firing from jobs. Recall-sensitive scenarios include loan granting, recommendation systems, and hiring. Very often, precision-sensitive scenarios coincide with situations where the positive prediction ($\hat{Y} = 1$) entails a higher cost [264]. For example, a predicted child maltreatment case will result in placing the child in a foster house, which will generally entail a higher cost compared to a negative prediction (low risk of child maltreatment), in which case the child stays with the family and typically no action is taken.

Overall Accuracy Equality [32]. Overall accuracy equality is achieved when overall accuracy for both groups is the same. Thus, true negatives and true positives are equally considered and desired. Using the confusion matrix (Table 2.2), this implies that $\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}}$ is equal for both groups. In our example, it is to say that the probability of well-qualified applicants being correctly accepted for the job and non-qualified applicants being correctly rejected is the same for both male and female applicants:

$$\mathbb{P}[\hat{Y} = Y|A = 0] = \mathbb{P}[\hat{Y} = Y|A = 1] \tag{3.8}$$

Overall accuracy is closely related to conditional use accuracy equality (Eq. (3.6)). The main difference is that overall accuracy aggregates positive and negative class misclassifications (false negative and false positive). Aggregating together false positives and false negatives (and hence true positives and true negatives) without distinction is often misleading for fairness purposes. In real-world applications, it is uncommon for true positives (or false negatives) and true negatives (or false positives) to be desired simultaneously and without distinction.

Treatment Equality [32]. Treatment equality is achieved when the ratio of false positives and false negatives is the same for both protected and unprotected groups:

$$\frac{\text{FN}}{\text{FP}}^{(a=0)} = \frac{\text{FN}}{\text{FP}}^{(a=1)} \tag{3.9}$$

Treatment equality is insensitive to the numbers of true positives and true negatives, which are important to identify bias between sub-populations in most real-world scenarios. Berk et al. [32] note that treatment equality can serve as an indicator to achieve other kinds of fairness. Table A.5 in Appendix A.1.1 shows a dataset that fails to satisfy all previous notions, yet treatment equality is satisfied. Treatment equality can be used in real-world scenarios where only the type of misclassification rate matters for fairness.

Total Fairness [32]. Total fairness is another notion which holds when all aforementioned fairness notions are satisfied simultaneously, that is, statistical parity (Eq.(3.1)), equalized odds (Eq.(3.3)), conditional use accuracy equality (Eq.(3.6)), overall accuracy equality (Eq.(3.8)), and treatment equality (Eq.(3.9)). Total fairness is a very strong notion that is difficult to hold in practice. Table A.6 in Appendix A.1.1 shows a toy example where total fairness holds.

Balance. The predicted outcome $\hat{Y}$ is typically derived from a score R which is returned by the ML algorithm. All aforementioned fairness notions do not use the score to assess fairness. **Balance for positive class** [130] focuses on the applicants who constitute positive instances and is satisfied if the average score R received by those applicants is the same for both groups. The intuition behind this notion is that a balance for the positive class should be assured. Thus, a violation of this balance means that applicants belonging to the positive class in one group might receive steadily lower

predicted scores than applicants belonging to the positive class in the other group:

$$\mathbb{E}[R \mid Y = 1, A = 0] = \mathbb{E}[R \mid Y = 1, A = 1] \tag{3.10}$$

Balance of negative class [130] is an analogous fairness notion where the focus is on the negative class:

$$\mathbb{E}[R \mid Y = 0, A = 0] = \mathbb{E}[R \mid Y = 0, A = 1] \tag{3.11}$$

Both variants of balance can be required simultaneously (Eq. (3.10) and Eq. (3.11)), which leads to a stronger notion of balance. Since no previous work reported such a fairness notion, for completeness, we define it and call it **overall balance**.

Definition 2 . *Overall balance is satisfied iff:*

$$\mathbb{E}[R \mid Y = y, A = 0] = \mathbb{E}[R \mid Y = y, A = 1] \quad \forall y \in \{0, 1\} \tag{3.12}$$

Balance fairness notions are relevant in the criminal risk assessment scenario because a divergence in the score values of individuals from different races may indicate a difference in the type of crime that can be committed (a high-risk score typically means a serious crime). Balance fairness notions are also suitable in the teacher firing scenario since any discrepancy between the average evaluation scores of fired teachers in different groups is a clear indicator of bias. On the other hand, balance fairness notions can be misleading in the presence of clusters of samples sharing very similar attribute values and having score values near the positive/negative outcome threshold. In such a case, the average score of the positive/negative class can change significantly due to a slight increase/decrease in the threshold value.

Calibration [54]. Calibration (a.k.a. test-fairness [54], matching conditional frequencies [101]) relies on the score variable as follows. To satisfy calibration, for each predicted probability score $R = r$, individuals in all groups should have the same probability of actually belonging to the positive class:

$$\mathbb{P}[Y = 1 \mid R = r, A = 0] = \mathbb{P}[Y = 1 \mid R = r, A = 1] \quad \forall r \in [0, 1]^{6} \tag{3.13}$$

Eq. (3.13) is very similar to Eq. (3.7), corresponding to predictive parity. Calibration is a stronger notion of fairness than predictive parity as it does not depend on a threshold

[6] Normalizing the score value to be in the interval $[0, 1]$ makes it possible to interpret the score as the probability to predict the sample as positive.

value. If calibration is satisfied, it will remain as such, no matter which threshold value is chosen. Therefore, it is suitable for scenarios where the threshold is not fixed and will likely be tuned to accommodate a changing context. A first example is the acceptance score in loan granting applications, which may change abruptly due to economic instability. A second example is the child maltreatment risk assessment, where the threshold for intervention (withdrawing a child from his family) depends on the available seats in foster houses.

Well-Calibration [130]. Well-calibration is a stronger variant of calibration. It requires that (1) calibration is satisfied, (2) the score is interpreted as the probability of truly belonging to the positive class, and (3) for each score $R = r$, the probability of truly belonging to the positive class is equal to that particular score:

$$\mathbb{P}[Y = 1 \mid R = r, A = 0] = \mathbb{P}[Y = 1 \mid R = r, A = 1] = r \quad \forall\, r \in [0, 1] \tag{3.14}$$

Intuitively, for a set of applicants who have a certain probability r of being hired, approximately r percent of these applicants should truly be hired.

All the notions discussed above are considered group fairness, where the common objective is to ensure that groups that differ in their sensitive attributes are treated equally. These notions, mainly based on statistical measures, generally ignore all attributes of the individuals except the sensitive attribute A. Such treatment might hide unfairness. Dwork et al. [75] stated that group fairness, despite its suitability for policies among demographic sub-populations, does not guarantee that individuals are treated fairly. The fairness notions that follow attempt to address such issues by not marginalizing over non-sensitive attributes X of an individual; therefore, they are called individual fairness notions [7].

Causal Discrimination [92]. Causal discrimination implies that a classifier should produce exactly the same prediction for individuals who differ only in the sensitive attribute while possessing identical attributes X. In a hiring example, this is to say that male and female applicants with the same attributes X should have the same predictions:

[7] The term individual fairness is used in some papers to refer to fairness through awareness (Eq.(3.16)). In this thesis, individual fairness refers to notions that cannot be considered group fairness notions.

$$X_{(a=0)} = X_{(a=1)} \ \wedge \ A_{(a=0)} \ \neq A_{(a=1)} \ \Rightarrow \hat{y}_{(a=0)} = \hat{y}_{(a=1)} \tag{3.15}$$

In a hiring system, this implies that male and female applicants who otherwise have the same attributes X will either be assigned a positive prediction or both assigned a negative prediction. At a first glance, causal discrimination can be seen as an extreme case of conditional statistical parity when conditioning on all non-sensitive attributes ($E = X$). However, conditional statistical parity is a group fairness notion that is satisfied if the proportion of individuals having the same non-sensitive attribute values and predicted accepted in both groups (e.g., male and female) is the same. This is why Eq. (3.2) is expressed in terms of conditional probabilities. Causal discrimination, however, considers every individual separately regardless of her contribution to sub-population proportions.

Causal discrimination is suitable for use in decision-making scenarios where it is very common to find individuals sharing exactly the same attribute values; for example, admission decision-making based mainly on test scores and categorical attributes. To apply this fairness notion to a loan granting scenario where there are only a few individuals with exactly the same attribute values, Verma and Rubin [242] generated, for every applicant in the dataset, an identical individual of the opposite gender. The result of applying causal discrimination is the percentage of violations in the entire population (i.e., how many individuals are unfairly treated).

Fairness Through Awareness [75]. Fairness through awareness (a.k.a individual fairness [91, 136]) is a generalization of causal discrimination, which implies that similar individuals should have similar predictions. Let i and j be two individuals represented by their attribute values vectors v_i and v_j. Let $d(v_i, v_j)$ represent the similarity distance between individuals i and j. Let $M(v_i)$ represent the probability distribution over the prediction outcomes. For example, if the outcome is binary (0 or 1), $M(v_i)$ might be $[0.2, 0.8]$ which means that for individual i, $\mathbb{P}[\hat{Y} = 0] = 0.2$ and $\mathbb{P}[\hat{Y} = 1] = 0.8$. Let D be a distance metric between probability distributions. Fairness through awareness is achieved iff, for any pair of individuals i and j:

$$D(M(v_i), M(v_j)) \leq d(v_i, v_j) \tag{3.16}$$

For a hiring example, this implies that the distance between the distribution of outcomes of two applicants should be, at most, the distance between those applicants. A relevant feature to measure the similarity between two applicants might be the

education level and the job experience. Thus, the distance metric d between two applicants could be defined as the average of the normalized difference (the difference divided by the maximum difference in a dataset) of their education level and their job experience: Let N_E be the normalized difference of the education level of two applicants and N_J be the normalized difference of the job experience of two applicants. Let E_{v_i} and E_{v_j} be the education levels of individuals i and j, respectively. Let J_{v_i} and J_{v_j} be the job experiences of individuals i and j, respectively. Let m_E and m_J be the maximum differences between the education level and the job experience in the dataset. Therefore, the distance metric is defined as:

$$d(v_i, v_j) = \frac{N_E + N_J}{2},$$

where $N_E = \frac{|E_{v_i} - E_{v_j}|}{m_E}$ and $N_J = \frac{|J_{v_i} - J_{v_j}|}{m_J}$.

The distance between outcomes could be the *Hellinger distance* [173], which can be used to quantify the similarity between two probability distributions.

Fairness through awareness is more fine-grained than any group fairness notion presented earlier in this section. It is important to mention that, in practice, fairness through awareness introduces some challenges. For instance, it assumes that the similarity metric is known for each pair of individuals [129]. A challenging aspect of this approach is the difficulty of determining an appropriate metric function to measure the similarity between two individuals. Typically, this requires careful human intervention from professionals with domain expertise [136]. For instance, suppose a company intends to hire only two employees while three applicants i_1, i_2, and i_3 are eligible for the offered job. Assume i_1 has a bachelor's degree and 1 year related work experience, i_2 has a master's degree and 1 year related work experience, and i_3 has a master's degree but no related work experience (Fig. 3.2). Is i_1 closer to i_2 than i_3? If so, by how much? This is difficult to answer, especially if the company overlooked such specific cases and did not carefully define and set a suitable and fair similarity metric to rank applicants for job selection. Thus, fairness through awareness cannot be considered suitable for domains where trustworthy and fair distance metrics are unavailable.

3.2.3 Tensions Between Fairness Notions

It has been proved that there are incompatibilities between fairness notions. For instance, it is not always possible for a predictor to satisfy specific fairness notions simultaneously [24, 54, 264, 166, 66]. In the presence of such incompatibilities, the

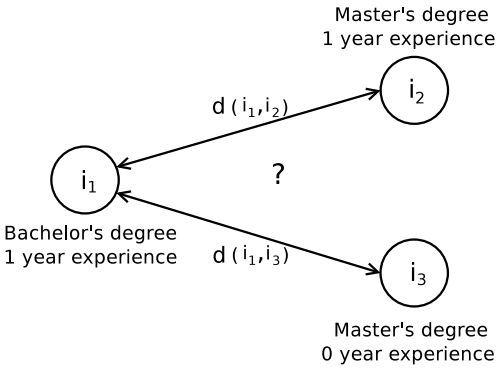


Fig. 3.2 An example showing the difficulty of selecting a distance metric in fairness through awareness.

predictor should relax some fairness notions by partially satisfying all of them. Incompatibility[8] results are well summarized by Mitchell et al. [166] as follows. Before listing the tensions, it is important to summarize the relationships between fairness notions. In addition, we define a new fairness notion for completeness, namely, **negative predictive parity** (Definition 3).

The following proposition formally states the relationship between equalized odds, equal opportunity, and predictive equality.

Proposition 3 . *Satisfying equal opportunity and predictive equality is equivalent to satisfying equalized odds:*

$$\text{Eq. (3.3)} \Leftrightarrow \text{Eq. (3.4)} \wedge \text{Eq. (3.5)}$$

Conditional use accuracy equality (Eq. (3.6)) is "symmetric" to equalized odds (Eq. (3.3)) with the only difference of switching Y and $\hat{Y}$. The same holds for equal opportunity (Eq. (3.4)) and predictive parity (Eq. (3.7)). However, there is no "symmetric" notion to predictive equality (Eq. (3.5)). We define such a notion for completeness and give it the name **negative predictive parity**.

Definition 3 . *Negative predictive parity holds iff all sub-groups have the same* $\text{NPV} = \frac{\text{TN}}{\text{FN}+\text{TN}}$*:*

$$\mathbb{P}[Y = 1 \mid \hat{Y} = 0, A = 0] = \mathbb{P}[Y = 1 \mid \hat{Y} = 0, A = 1] \tag{3.17}$$

[8]The term impossibility is commonly used as well.

Proposition 4 . *Satisfying equalized odds or conditional use accuracy equality always leads to satisfying overall accuracy.*

$$\text{Eq. (3.3)} \vee \text{Eq. (3.6)} \Rightarrow \text{Eq. (3.8)}$$

The reverse, however, is not valid. An ML system that satisfies overall accuracy does not necessarily satisfy equalized odds or conditional use accuracy equality. Check Appendix A.1.1 for an example (Table A.4) that satisfies overall accuracy but does not satisfy equalized odds or conditional use accuracy equality.

Statistical Parity (Independence) *versus* Conditional Use Accuracy Equality (Sufficiency). Independence and sufficiency are incompatible, except when both groups (protected and non-protected) have equal base rates or $\hat{Y}$ and Y are independent. Note, however, that $\hat{Y}$ and Y should not be independent since the predictor is completely useless otherwise. More formally,

$$\underbrace{\hat{Y} \perp A}_{\text{(independence)}} \quad \wedge \quad \underbrace{Y \perp A \mid \hat{Y}}_{\text{(strict sufficiency)}} \quad \Longrightarrow \quad \underbrace{Y \perp A}_{\text{(equal base rates)}} \quad \vee \quad \underbrace{\hat{Y} \perp Y}_{\text{(useless predictor)}}$$

It is important to mention that this result does not hold for the relaxation of sufficiency, particularly predictive parity. Hence, the output of a predictor can satisfy statistical parity and predictive parity between two groups having different base rates.

Statistical Parity (Independence) *versus* Equalized Odds (Separation). Similar to the previous result, independence and separation are mutually exclusive unless base rates are equal or the predictor $\hat{Y}$ is independent of the true decision Y [24]. As mentioned earlier, dependence between $\hat{Y}$ and Y is a weak assumption, as any useful predictor should satisfy it. More formally,

$$\underbrace{\hat{Y} \perp A}_{\text{(independence)}} \quad \wedge \quad \underbrace{\hat{Y} \perp A \mid Y}_{\text{(strict separation)}} \quad \Longrightarrow \quad \underbrace{Y \perp A}_{\text{(equal base rates)}} \quad \vee \quad \underbrace{\hat{Y} \perp Y}_{\text{(useless predictor)}}$$

Considering a relaxation of equalized odds, equal opportunity, or predictive equality breaks the incompatibility between independence and separation.

Equalized Odds (Separation) *versus* Conditional Use accuracy Equality (Sufficiency). Separation and sufficiency are mutually exclusive, except where groups have equal base rates. More formally:

$$\underset{\text{(strict separation)}}{\hat{Y} \perp A \mid Y} \quad \wedge \quad \underset{\text{(strict sufficiency)}}{Y \perp A \mid \hat{Y}} \quad \Rightarrow \quad \underset{\text{(equal base rates)}}{Y \perp A}$$

Both separation and sufficiency have relaxations. Considering only one relaxation will only drop the incompatibility for extreme and degenerate cases. For example, predictive parity (relaxed version of sufficiency) is still incompatible with separation (equalized odds), except in the following three extreme cases [54]:

- both groups have equal base rates.
- both groups have FPR= 0 and PPV= 1.
- both groups have FPR= 0 and FNR= 1.

The incompatibility disappears completely when considering relaxed versions of both separation and sufficiency.

3.2.4 Relaxation

Almost all fairness notions presented so far involve strict equality between quantities, mainly probabilities. In real scenarios, however, opting for an approximate or relaxed form of fairness constraint is more suitable. The need for relaxation might be due to the impossibility of applying fairness strictly on the application at hand, or it is not a requirement to impose an exact constraint [128].

Fairness notions can be relaxed by considering a threshold on the ratio or difference between quantities. For instance, the requirement for statistical parity (Eq. (3.1)) can be relaxed in one of the two following ways:

- By allowing the ratio between the predicted acceptance rates of protected and unprotected groups to reach the threshold of τ (a.k.a $p\%$ rule defined as satisfying this inequality when $\tau = p/100$ [264]):

$$\frac{\mathbb{P}[\hat{Y} = 1 \mid A = 0]}{\mathbb{P}[\hat{Y} = 1 \mid A = 1]} \geq 1 - \tau \quad \forall\, \tau \in [0, 1] \tag{3.18}$$

  For $\tau = 0.2$, this condition relates to the 80% rule in disparate impact law [86, 25].

- By allowing the difference between the predicted acceptance rates of different groups to reach a threshold of τ [75]:

$$\Big|\mathbb{P}[\hat{Y} = 1 \mid A = 0] - \mathbb{P}[\hat{Y} = 1 \mid A = 1]\Big| \leq \tau \quad \forall\, \tau \in [0, 1] \tag{3.19}$$

A notable difference between the two types of relaxation is that the second one (Eq.(3.19)) is insensitive to which group/individual is the victim of discrimination as the formula uses the absolute value.

Fairness through awareness can be relaxed using three threshold values, α_1, α_2, and γ as follows [262]:

$$\mathbb{P}\Big[\mathbb{P}\Big[\big|M(v_i) - M(v_j)\big| > d(v_i, v_j) + \gamma \Big] > \alpha_2\Big] \leq \alpha_1. \tag{3.20}$$

The relaxation is allowing $M(v_i) - M(v_j)$ to exceed $d(v_i, v_j)$ by a margin of γ, but the fraction of individuals differing from them by γ should not exceed α_2. If the fraction exceeds α_2, the individual is said to be α_2-discriminated against.

Other relaxations can allow for more flexibility in applying fairness notions. For instance, Eq. (3.2) of conditional statistical parity can be modified by relaxing the strict equality $E = e$ as follows:

$$\mathbb{P}[\hat{Y} = 1 \mid e - \tau \leq E \leq e + \tau, A = 0] = \mathbb{P}[\hat{Y} = 1 \mid e - \tau \leq E \leq e + \tau, A = 1] \tag{3.21}$$

3.2.5 Group vs. individual fairness

Compared to individual fairness notions, the main concern for group fairness notions is that they are only suited to a limited number of coarse-grained, predetermined protected groups based on some sensitive attribute (e.g., gender, race, etc.). Hence, group fairness notions are unsuitable in the presence of intersectionality [61, 168], where individuals are often disadvantaged by multiple sources of discrimination: their race, class, gender, religion, and other inner traits. Typically, statistical fairness can only be applied across a small number of coarsely defined groups, and hence failing to identify discrimination on structured subgroups (e.g., single women), also known as "fairness gerrymandering" [123]. A simple alternative might be to apply statistical fairness across every possible combination of sensitive attributes. There are at least two problems with this approach. First, this can lead to an impossible statistical problem with a large number of sub-groups, which may lead, in turn, to overfitting. Second, groups that are not (yet) defined in anti-discrimination law may exist and may need protection [243]. Another issue with group fairness notions is their susceptibility to masking. Most group fairness notions can be gamed by adding arbitrarily selected samples to satisfy the fairness notion formula, that is, to "make up the numbers".

Compared to group fairness notions, individual fairness notions have the drawback that they can result in "unjust disparities in outcomes between groups" [35]. Another

critical issue for similarity-based individual fairness (e.g., fairness through awareness) is the difficulty of obtaining a similarity value between every pair of individuals. For example, even assuming that the similarity can be quantified between all individuals in the training data, it might be challenging to generalize to new individuals [35].

Several researchers assume that both group and individual fairness are prominent yet conflicting and suggest approaches to minimize the trade-offs between these notions [35]. For instance, [**?**] defines two different worldviews, WYSIWYG and WAE. The WYSIWYG (What you see is what you get) worldview assumes that the unobserved (construct) space and observed space are essentially the same. In contrast, the WAE (we're all equal) worldview implies no innate differences between groups of individuals based on specific potentially discriminatory characteristics. These two worldviews highlight the tension between group and individual fairness. For instance, in a job hiring example, the WYSIWYG might be the assumption that attributes like education level and job experience (which belong to the observed space) correlate well with the applicant's seriousness or hard work (properties of the construct space). This is to say there is some way to combine these two spaces to compare true applicant aptitude for the job correctly. On the other hand, the WAE claims that all groups will have almost the same distribution in the construct space of inherent abilities (here, seriousness and hard work), chosen as essential inputs to the decision-making process. The idea is that any difference in the groups' performance (e.g., academic achievement or education level) is due to factors outside their individual control (e.g., the quality of their neighborhood school) and should not be considered in the decision-making process. Thus, the choice between fairness notions must be based on an explicit worldview choice. Under a WYSIWYG worldview, only individual fairness notions achieve fairness (and group fairness notions are unfair). Under a WAE worldview, only group fairness notions achieve non-discrimination (and individual fairness notions are discriminatory)[9].

3.2.6 Conclusion

Implementing fairness is essential to guarantee that ML- based automated decision systems produce unbiased decisions and avoid unintentional discrimination against some sub-populations (typically minorities). This study discusses an important issue related to implementing fairness.

[9]The authors use the term *fairness* when discussing individual fairness and *non-discrimination* when discussing group fairness.

That is, several reasonable fairness requirements cannot be satisfied simultaneously. This means that fairness practitioners have to choose among them. In the following section, we present our work aimed at bridging the gap between fairness notions and real-world applications. Specifically, we address the challenge of the applicability of statistical fairness notions and the identification of fairness-relevant criteria to help select the most appropriate notion of fairness for a given scenario.

3.3 Applicability of Statistical Fairness Notions

In the context of automated decision-making, a consensual definition of fairness can be formulated as: "*absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits*" [164]. Mathematically, however, there is no consensual definition of fairness. Research papers often focus on a specific real-world scenario of an automated decision system and propose a fairness definition tailored to that scenario and its specificities. Consequently, several notions of fairness have been introduced in the literature (Section 3.2.2).

The very reason for having different flavors of fairness notions is how suitable each one is for specific real-world scenarios. Discussion about the suitability (and sometimes the applicability) of the fairness notions is very limited and scattered through several papers [166, 91, 264, 130, 58, 24]. In this work we show that each ML-based automated decision system can be different based on a set of criteria such as: whether the ground-truth exists, difference in base-rates between sub-groups, the cost of misclassification, the existence of a government regulation that needs to be enforced, etc. We then discuss the suitability and applicability of each fairness notion based on the list of criteria.

The results of this study are summarized in a decision diagram that hopefully can help researchers, practitioners, and policymakers to identify the subtleties of the ML-based automated decision system at hand and to choose the most appropriate fairness notion to use or at least rule out notions that can lead to wrong fairness/discrimination result.

Contributions. We propose a decision diagram (Fig. 3.3) integrating a set of fairness-related features of real-world scenarios that can help researchers, practitioners, and policymakers answer the question of "which notion of fairness is most appropriate to a given real-world scenario and why?".

Outline. The rest of the section is organized as follows. Section 3.3.1 discusses related work. Section 3.3.2 lists notable real-world ML systems where fairness is critical. Section 3.3.3 identifies a set of fairness-related characteristics of ML systems to recommend and/or discourage using fairness notions. The decision diagram is provided and described in Section 3.3.4, and Section 3.3.5 concludes the section.

3.3.1 Related work

In addition to the studies discussed in Section 3.2.1, we include further research on the applicability of fairness notions in real-world applications. In 2015, Zliobaite [281]

compiled a survey about fairness notions that have been introduced previously. He classified fairness notions into four categories: statistical tests, absolute measures, conditional measures, and structural measures. Statistical tests indicate only the presence or absence of discrimination. Absolute and conditional measures quantify the extent of discrimination, with the difference being that conditional measures consider legitimate explanations for the discrimination. These three categories correspond to notions of group fairness. Structural measures correspond to individual fairness notions[10]. Most of the fairness notions listed by Zliobaite are variants of the group fairness notions. For instance, the difference of means test (Section 4.1.2 in [281]) is a variant of balance for positive class (Eq. (3.10)). Although he dedicated one category to individual notions (structural measures). Regarding the applicability of notions, the only criterion considered was the type of variables (e.g., binary, categorical, numeric, etc.). The survey of Verma and Rubin [242] described a list of fairness notions similar to the list in Section 3.2.2. To illustrate how each notion can be computed in real scenarios, they used a loan granting real use case (German credit dataset [18]). Verma and Rubin did not address the applicability aspect in their work.

Brief discussions about the suitability of specific fairness notions can be found in a few papers. For instance, Zafar et al. [264] mentioned some application scenarios for statistical parity and equalized odds. Kleinberg et al. [130] discussed the applicability of calibration and balance notions. Through a discussion about the cost of unfair decisions on society, Corbett-Davies et al. [58] analyzed the impact of using statistical parity, predictive equality, and conditional statistical parity on public safety (criminal risk assessment). Unlike the scattered discussions about the applicability of fairness notions found in the literature, this study provides a complete reference to systemize the selection procedure of fairness notions.

3.3.2 Real-world Scenarios with Critical Fairness Requirements

As the work focuses on the applicability of fairness notions, we provide a list of notable real-world ML systems where fairness is critical. Failure to address the fairness requirement in these scenarios will lead to unacceptably biased decisions against individuals and/or sub-populations. These scenarios will provide concrete examples of situations where certain fairness notions are more suitable than others.

[10]Zliobaite does not use group vs individual notions, but indirect and direct discrimination.

Job Hiring. ML systems in hiring are increasingly used by employers to automatically screen candidates for job openings[11]. Commercial candidate, screening ML systems, include XING[12], Evolv [140], Entelo, Xor, EngageTalent, and GoHire. Typically, the input data used by the ML systems include affiliation, education level, job experience, IQ score, age, gender, marital status, address, etc. The ML systems output a decision and/or a score indicating how suitable/promising the application is for the job opening. A biased ML system leads to rejecting a candidate because of a trait that she cannot control (gender, race, sexual orientation, etc.). Such unfairness causes prejudice against the candidate but can also be damaging for the employer as excellent candidates might be missed.

Granting Loans. For decades, statistical and ML systems have been used to assess loan applications and determine which are approved and with which repayment plan and annual percentage rate (APR). The assessment proceeds by predicting the risk that the applicant will default on her repayment plan. Loan Granting ML systems currently in use include FICO, Equifax, Lenddo, Experian, TransUnion, etc. The common input data used for loan granting include credit history, loan purpose, loan amount requested, employment status, income, marital status, gender, age, address, housing status, and credit score. An unfair loan-granting ML system will either deny a deserving applicant a requested loan or give her an exorbitant APR, which, in the long run, will create a vicious cycle as the candidate will be very likely to default on her payments.

College Admission. Given the large number of admission applications, several colleges are now resorting to ML systems to reduce processing time and cut costs[13]. Existing college admission ML systems include GRADE [247], IBM Watson[14], Kira Talent[15]. The candidates' features typically include their previously attended institutions, SAT scores, extra-curricular activities, GPAs, test scores, interview scores, etc. The predicted outcome can be a simple decision (admit/reject) or a score indicating the candidate's potential performance in the requested field of study [**?**]. Unfair college admission ML systems may discriminate against a certain ethnic group (e.g., African-American [201]),

[11] In 2014, the automated job screening systems market was estimated at $500 million annual business and was growing at a rate of 10 to 15% per year [248].

[12] A job platform similar to LinkedIn. It was found that this platform ranked less qualified male candidates higher than more qualified female candidates [137].

[13] While the final acceptance decision is taken by humans, ML systems are typically used as a first filter to "clean-up" the list from clear rejection cases.

[14] A platform that uses natural language processing and personality traits to help students find the suitable college for them.

[15] A Canadian startup that sells a cloud-based admissions assessment platform to over 300 schools.

which could lead, in the long term, to economic inequalities and corrupt the role of higher education in society as a whole.

Criminal Risk Assessment. There is an increasing adoption of ML systems that predict risk scores based on historical data to guide human judges' decisions. The most common use case is to predict whether a defendant will re-offend (or recidivate). Examples of risk assessment ML systems include COMPAS [56], PSA [152], SAVRY [165], predPol [190]. Predicting risk and recidivism requires input information such as the number of arrests, type of crime, address, employment status, marital status, income, age, housing status, etc. Unfair risk assessment ML systems, as revealed by the highly publicized 2016 proPublica article [10], may result in biased treatment of individuals based solely on their race. In extreme cases, it may lead to wrongful imprisonment of innocent people, contributing to the cycle of violation and crime.

Teachers Evaluation and Promotion. ML systems are increasingly used by decision makers to decide which teachers to retain after a probationary period [44] and which tenured teachers to promote. An example of such ML systems is IMPACT [196]. Teacher evaluation ML systems take as input teacher-related features (age, education level, experience, surveys, classroom observations), student features (test scores, sociodemographics, surveys), and principals-related features (surveys about the school and teachers) to predict whether teachers are retained. A biased teacher evaluation ML system may lead to a systematic, unfair low evaluation for teachers in poor neighborhoods, which, very often, happen to be teachers belonging to minority groups [192]. In the long term, this may lead to a significant drop in students' performance and the compromise of overall school reputation [178].

Child Maltreatment Prediction. The objective of the ML systems in child maltreatment prediction is to estimate the likelihood of substantiated maltreatment (neglect, physical abuse, sexual abuse, or emotional maltreatment) among children. The system generates risk scores, triggering a targeted early intervention to prevent child maltreatment. Predictive risk model [238] has been developed to estimate the likelihood of substantiated maltreatment among children enrolled in New Zealand's public benefits system. AFST (Allegheny Family Screening Tool) [82] is designed to improve decision-making in Allegheny County's child welfare system. The features considered in this ML system include contemporaneous and historical information for children and caregivers. An unfair ML system may use a proxy variable to predict decisions based on the community rather than which child gets harmed. For example, a major cause of unfairness in AFST is the rate of referral calls; the community calls the child abuse hotline to report non-white families at a much higher rate than it does to report white

families [82]. In the long term, this creates a vicious cycle as families that have been reported will be the subject of more scrutiny and more requirements to satisfy, and eventually, will be more likely to fall short of these requirements and hence confirm the prediction of the system.

Health Care. For decades, ML algorithms have been processing anonymized electronic health records and flagging potential emergencies, to which clinicians are invited to respond promptly. Examples of features that might be used in disease (chronic conditions) prediction include vital signs, blood tests, socio-demographics, education, health insurance, home ownership, age, race, and address. The outcome of the ML system is typically an estimated likelihood of getting a disease. A biased disease prediction ML system can misclassify individuals in certain sub-populations disproportionately more than the dominant population. For instance, diabetic patients have known differences in associated complications across ethnicities [217]. Another example is described by Obermeyer et al. [176] where for the same prediction score, African-Americans were found to be sicker (more health issues) than whites because the ML system was relying on the cost of health services in the previous year (African-Americans were spending less on health services than whites) to predict the cost of health care in the coming years. Consequently, white patients were benefiting more from additional help programs than African Americans. More generally, because different subpopulations might have different characteristics, a single model to predict complications is unlikely to be best-suited for specific groups in the population even if they are equally represented in the training data [227]. Failure to predict disease likelihood promptly may, in extreme cases, impact people's lives.

Online Recommendation. Recommender systems are among the most widespread ML systems in the market, with many services to assist users in finding products or information that are of potential interest [113]. Such systems find applications on various online platforms such as Amazon, YouTube, Netflix, Linkedin, etc. An unfair recommender of ML systems can amplify gender bias in the data. For example, a recommender ML system called STEM, which aims to deliver advertisements promoting jobs in Science, Technology, Engineering, and Math fields, is deemed unfair as it has been shown that fewer women compared to men saw the advertisements due to gender imbalance [139]. Datta et al. [62] found that changing the gender bit in the Google Ad setting resulted in a significant difference in the type of job ads received: men received much more ads about high-paying jobs and career coaching services toward high-paying jobs compared to women.

Facial Analysis. Automated facial analysis systems are used to identify perpetrators from security video footage, to detect melanoma (skin cancer) from face images [81], to detect emotions [65, 84, 222], and to even determine individual's characteristics such as IQ, propensity towards terrorist crime, etc. based on their face images [252]. Flawed ML systems may lead to biased outcomes, such as wrongfully accusing individuals from specific ethnic groups (e.g., Asians, dark-skinned populations) for crimes (based on security video footage) at a much higher rate than the rest of the population. For instance, African Americans have been reported to be more likely to be stopped and investigated by law enforcement due to a flawed face recognition system [97]. An investigation of three commercial face-based gender classification systems found that the error rate for dark-skinned females can be as high as 34.7% while for light-skinned males, the maximum error rate is 0.8% [40].

Others. Other ML systems with fairness concerns include insurance policy prediction [213], income prediction [164, 276, 80, 204, 6], and university ranking [163, 178].

3.3.3 Fairness Notion Selection Criteria

To systemize the procedure for selecting the most suitable fairness notion for a specific ML system, we identify a set of criteria that can be used as a roadmap. We check whether each criterion holds in the problem at hand or not. Telling whether a criterion is satisfied does not typically require expertise in the problem domain. We note here that in some cases, these criteria can indicate whether a fairness notion is suitable and whether it is "acceptable" to use in the first place.

Ground Truth Availability. A ground truth value is the true and correct *observed* outcome corresponding to a given sample in the data. It should be distinguished from an *inferred* subjective outcome in historical data, which is decided by a human. An example of a scenario where ground truth is available is when predicting whether an individual has a disease. The ground truth value is observed by submitting the individual to a blood test[16] for example. An example of a scenario where ground truth is unavailable is predicting whether a job applicant is hired. The outcome in the training data is inferred by a human decision-maker, which is often a subjective decision, no matter how hard she tries to be objective. It is important to mention here that the availability of the ground truth depends on how the outcome is defined. Consider, for example, a college admission scenario. If the outcome in the training data is defined as whether the applicant is admitted or rejected, ground truth is not

[16] assuming the blood test is flawless.

available. If, however, the outcome is defined as whether the applicant will ultimately graduate from college with a high GPA, ground truth is available as it can be observed after a couple of years.

Base Rate is the Same Across Groups. The base rate is the proportion of positive outcomes in a population (Table 2.2). This rate can be the same or differs across sub-populations. For example, the base rates for diabetes disease occurrence for men and women are typically the same. But, for another disease such as prostate cancer, the base rates are different between men and women[17].

(Un)Reliable Outcome. In scenarios where ground truth is not available, the outcome (label) in the data is typically inferred by humans. In that case, the outcome of the training data can or cannot be reliable as it can encode human bias. The reliability of the outcome depends on the data collection procedure and how rigorous the data has been checked. Scenarios like job hiring and college admission may be more prone to unreliable outcome problems than recommender systems. A "one-size-fits-all" ML model in disease prediction that does not take into consideration the ethnic group of the individual may result in an unreliable outcome as well.

Presence of Explaining Variables. An explaining variable[18] is correlated with the sensitive attribute (e.g., race) legitimately. Any discrimination that can be explained using that variable is considered legitimate and acceptable. For instance, if all the discrepancies between male and female job hiring rates are explained by their education levels, discrimination can be deemed legitimate and acceptable.

Emphasis on Precision vs. Recall. Precision (the complement of target population error [68]) is defined as the fraction of positive instances among the predicted positive instances. In other words, how precise is that prediction if the system predicts an instance as positive? Recall (the complement of model error [68]) is defined as the fraction of the total number of positive instances that are correctly predicted positive. In other words, how many positive instances can the system identify? There is always a tradeoff between precision and recall (increasing one will lead, very often, to decreasing the other). Depending on the scenario, the fairness of the ML system may be more sensitive to one at the expense of the other. For example, granting loans to the maximum number of deserving applicants contributes more to fairness than making sure that an applicant who has been granted a loan really deserves it[19]. However, the

[17] While male prostate cancer is the second most common cancer in men, female prostate cancer is rare [71].

[18] Referred also as resolving variable.

[19] It is important to mention here that from the loan granting organization's point of view, the opposite is true. That is, it is more important to make sure that an applicant who has been granted a loan really deserves it and will not default in payments because the interest payments resulting from

opposite is true when firing employees: fairness is more sensitive to wrongly firing an employee than the maximum number of under-performing employees.

Emphasis on False Positive vs. False Negative. Fairness can be more sensitive to false positive misclassification (type I error) rather than false negative misclassification (type II error) or the opposite. For example, in a criminal risk assessment scenario, it is commonly accepted that incarcerating an innocent person (false positive) is more serious than letting a guilty person escape (false negative).

Cost of Misclassification. Depending on the scenario at hand, the cost of misclassification can be high (e.g., incarcerating an individual, firing an employee, rejecting a college application, etc.) or mild and without consequential impact (e.g., useless product recommendation, misleading income prediction, offensive online translation, abusive results in online autocomplete, etc.).

Prediction Threshold is Fixed or Floating. Decisions in ML systems are typically made based on predicted real-valued scores. In the case of a binary outcome, the score is turned into a binary value such as $\{0, 1\}$ by thresholding[20]. In some scenarios, it is desirable to interpret the real-value score as the probability of being accepted (predicted positive). The threshold used as a cutoff point where positive decisions are demarcated from negative decisions can be fixed or floating. A fixed threshold is set carefully and tends to be valid for different datasets and use cases. For instance, the high-risk threshold is typically fixed in recidivism risk assessment. Practitioners can arbitrarily select and fine-tune a floating threshold to accommodate a changing context. Acceptance score in loan granting scenarios is an example of a floating threshold, which can move up or down depending on the economic context.

Likelihood of Intersectionality. Intersectionality theory [61] focuses on a specific type of bias due to the combination of sensitive factors. An individual might not be discriminated against based on race or based on gender only, but she might be discriminated against because of a combination of both. Black women are particularly prone to this type of discrimination.

Likelihood of Masking. Masking is a form of intentional discrimination that allows decision-makers with prejudicial views to mask their intentions [25]. Masking is typically achieved by exploiting how fairness notions are defined. For example, suppose the notion of fairness requires an equal number of candidates to be accepted from two ethnic groups. In that case, the ML system can be designed to carefully select candidates

a loan are relatively small compared to the loan amount that could be lost. Here, we aim for fairness, while the loan granting organization's goal is a benefit.

[20]The threshold is defined by the decision makers depending on the context of interest.

from the first group (satisfying strict requirements) while selecting randomly from the second group just to "make the numbers".

Sources of Bias. Bias in the ML system outcome can arise from several possible sources at any stage in the data generation and ML pipeline. Framing sources of bias necessitates a deep understanding of the application at hand and can typically only be identified after a "post-mortem" analysis of the predicted outcome. However, in some real-world scenarios, one or more sources of bias may be more likely than others. In such cases, the suspected source of bias can be used as a criterion to select the most appropriate notion for fairness assessment. Sources of bias can be grouped broadly into six categories: historical, representation, measurement, aggregation, evaluation, and deployment [227]. Historical bias arises when the data reliably collected from the world leads to unwanted and socially unfavorable outcomes. For example, while data reliable collected indicates that only 5% of Fortune 500 CEOs are women [266], the resulting outcome of a prediction system based on this data is typically not wanted[21]. Representation bias arises when some non-protected populations are under-represented in the training data. Measurement bias arises when the features or label values are not measured accurately. For example, Street Bump is an application used in Boston City to detect when residents drive over potholes thanks to the accelerometers built into smartphones [60]. Collecting data using this application introduces a measurement bias due to the disparity in the distribution of smartphones according to the different districts in the city, which are often correlated with race or income level. Aggregation bias arises when sub-populations are aggregated while a single model is unlikely to fit all sub-populations. For instance, the genetic risk scores derived largely on European populations have been shown to perform very poorly in predicting osteoporotic fracture and bone mineral density in non-European populations, particularly in the Chinese population [143]. Evaluation bias arises when the training data differs significantly from the testing data. For instance, several ML systems are trained using benchmark datasets that may differ greatly from the target dataset. Deployment bias arises when there is a disparity between the initial purpose of an ML system and the way it is actually used. For instance, a child maltreatment ML system might be designed to predict the risk of child abuse after two years from the reception of a referral call, while in practice, it may be used to help social agents make decisions about an intervention. This can lead to a bias since the decision impacts the outcome [59].

[21] For this reason, Google has changed their image search result for CEO to return a higher proportion of women.

Legal Framework. Anti-discrimination regulations in several countries, particularly the US, distinguish between two legal frameworks: disparate treatment and disparate impact [25]. In the disparate treatment framework, a decision is considered unfair if it uses (directly or indirectly) the individual's sensitive attribute information. In the disparate impact framework, a decision is unfair if it results in an outcome that is disproportionately disadvantageous (or beneficial) to individuals according to their sensitive attribute information. Zafar et al. [264] formalized another fairness criterion, namely, disparate mistreatment, according to which a decision is unfair if it results in different misclassification rates for groups of people with different sensitive attribute information. Note that this criterion is currently not supported by a legal framework. ML fairness notions can be classified according to the type of fairness it evaluates. For instance, if a plaintiff is accusing an employer of intentional discrimination, she should consider the disparate treatment legal framework and, hence, a fairness notion that falls in that framework.

The Existence of Regulations and Standards. In some domains, laws and regulations might be imposed to avoid discrimination and bias. For instance, guidelines from the *U.S. Equal Employment Opportunity Commission* state that a difference in the probability of acceptance between two sub-populations exceeding 20% is illegal [24]. Another example might be an internal organizational policy imposing diversity among its employees.

3.3.4 Decision Diagram and Discussion

With the many fairness notions and the subtle resemblance between ML systems scenarios, deciding which fairness notion to use is not trivial. More importantly, selecting and using a fairness notion inappropriately in a scenario may detect unfairness in an otherwise fair scenario or the opposite, i.e., failing to identify unfairness in an unfair scenario.

One of the objectives of this study is to systemize the selection procedure of fairness notions. This is achieved by identifying a set of fairness-related characteristics (Section 3.3.3) of the scenario at hand and then using them to recommend the most suitable fairness notion for that specific scenario. The proposed systemized selection procedure is illustrated in the decision diagram of Fig. 3.3. The diagram is called a "decision diagram" and not a "decision tree" for the following reasons. In typical decision trees, every leaf corresponds to a single decision, which is a fairness notion that *should* be used. However, the diagram in Fig. 3.3 is designed such that every node indicates which notions are recommended, which notions to avoid, and which notions

must not be used. In addition, if a notion is not mentioned along the path, it can be safely used.

The diagram is composed of four types of nodes:

- **Decision node (diamond):** based on fairness-related characteristics (Section 3.3.3).
- **Recommended node (rectangle):** a leaf node indicating that the fairness notion is suitable for use given all fairness-related characteristics in the path to that node.
- **Warning node (triangle):** indicates that the fairness notion(s) is/are not recommended in the branch on the right of the node. This node can appear between two decision nodes in the middle of the edge.
- **Must-not node (circle):** the fairness notion must not be used.

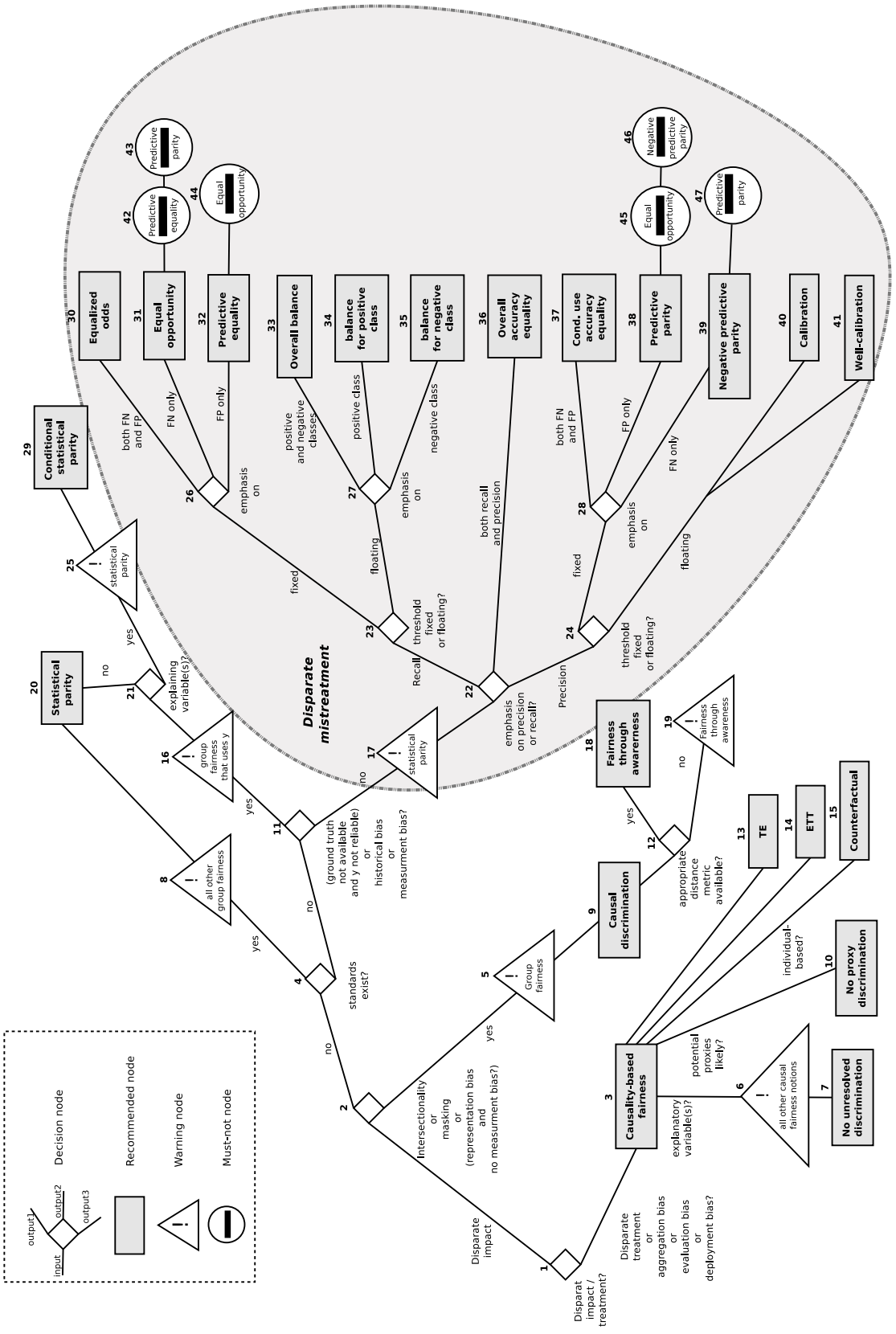


Fig. 3.3 Statistical fairness notions applicability decision diagram.

To illustrate how the diagram should be interpreted, consider the recommended node predictive parity (node 38). According to the diagram, predictive parity is recommended in the scenario where the legal framework is the disparate impact (decision node 1), intersectionality and/or masking are unlikely (decision node 2), there is no evidence that representation bias is likely (decision node 2), standards do not exist (decision node 4), ground-truth is available, or outcome Y is reliable (decision node 11), historical and measurement biases are unlikely (decision node 11), fairness is more sensitive to precision rather than recall (decision node 22), the prediction threshold is typically fixed (decision node 24) and the emphasis is on FPs (false positives) rather than FNs (false negatives) (decision node 28). In that particular scenario, equal opportunity must not be used (must-not node 45) because fairness in this scenario is particularly sensitive to FPs. In contrast, equal opportunity is completely insensitive to FPs. Similarly, negative predictive parity must not be used (must-not node 46) as fairness is sensitive to precision rather than recall. The warning node 17 along the same path indicates that statistical parity is unsuitable in this scenario. Finally, any fairness notion for which there is no warning node or a must-not node along the path of the scenario can be used in this scenario. For instance, all individual fairness notions can be used.

Consider the following scenario as a concrete example of situations where predictive parity (node 38) is recommended. When the outcome is influenced by the decision, some statistical quantities (e.g., FN, TN, etc.) are unlikely to be observed. Consequently, any fairness notion relying on these quantities is unsuitable in such cases. For example, in real-world cases of loan granting, a loan application predicted to default will not be approved. Consequently, neither negative statistics (TN nor FN) will be typically observed. Hence, fairness notions such as equalized odds and equality of opportunity cannot be used as they are defined in terms of TN and FN. In such cases, predictive parity (node 38) is recommended.

Node 1. Assessing fairness is very often performed in a legal case where a plaintiff is filing a claim against a party using an ML system. According to real-world legislation, particularly the American anti-discrimination law, this can fall into one of the two legal frameworks: disparate impact and disparate treatment. Suppose the plaintiff is filing the claim under the disparate impact framework. In that case, she can prove the liability of the defendant by using an observational group or individual fairness notion as the goal is to show that the practices and policies used by the defendant are facially neutral but have a disproportionately adverse impact on the protected class [25]. Suppose the plaintiff is filing a claim under the disparate treatment framework. In that case, observational fairness notions are often not enough to prove the defendant's

liability, as the goal is to show that the defendant has used the sensitive attribute to make the discriminatory decision. The recommended fairness notions, in that case, are causality-based[22] (recommended node 3) since all of them are expressed in terms of the causal effect of the sensitive attribute on the prediction.

Node 2. As explained above, any unintentional type of bias can also be "orchestrated" intentionally by decision-makers with prejudicial views. For instance, decision-makers can purposefully bias the data collection step to ensure that the ML system remains less favorable to protected classes. To reliably assess the bias in the presence of such masking attempts, all group fairness notions should be avoided as they are defined in terms of statistics about the different sub-populations and hence can more easily be gamed by prejudicial decision makers. Intersectionality is similar to masking as both lead to discrimination, which is difficult to detect using statistical measures and consequently requires more fine-grained measures. Therefore, individual fairness notions are recommended in the presence of both criteria (nodes 9 and 18).

Nodes 2, 3[23], and 11. In case one or more sources of bias are suspected ahead of time (before assessing fairness), the information can help warn against using some fairness notions. If representation bias is likely, the performance (accuracy) of the ML system on under-represented categories will often be worse. Such disparity in performance between groups may lead to unreliable fairness assessment if a group fairness notion is used, particularly disparate mistreatment notions (grayed section of the diagram). Individual fairness notions can assess fairness more reliably in such cases, provided that measurement bias is not likely (node 2). Suspecting historical or measurement bias means the features X and/or the label Y are unreliable. All group fairness notions using the label Y (disparate mistreatment) and individual notions are not recommended in that case. Statistical parity is recommended in such a situation. Finally, causality-based fairness notions are recommended in the presence of either aggregation, evaluation, or deployment bias. The reason is that the interventional and counterfactual quantities used in the definitions of these notions go beyond correlations and hence allow us to assess fairness more reliably in the presence of such bias. For instance, Coston et al. [59] propose counterfactual formulations of fairness metrics to properly account for the effect of an intervention (decision) on the outcome. Such an effect is a type of deployment bias.

Node 4. To reduce inequality and historical discrimination against sub-populations, particularly minorities, some states and organizations resort to equality standards

[22]We will focus on causality-based fairness notions and their applicability in Section 3.4.

[23]As node 3 involves the applicability of causality-based fairness notions, it will be thoroughly explained in Section 3.4.

and regulations such as the laws enforced by the US Equal Employment Opportunity Commission [1]. In the presence of such standards, an ML system should satisfy such standards to be deemed fair. Consequently, all that matters for fairness assessment is the proportion of positive prediction across all groups, corresponding to statistical parity.

Node 17. If no standards/regulations exist (node 4) and either the ground truth exists or the outcome label Y is available (node 11), statistical parity is not recommended (node 17) as it can lead to misleading results such as detecting unfairness in an otherwise fair scenario or failing to identify fairness in an unfair scenario. For instance, in a stop-and-frisk real-world scenario applied in New York City starting 1990 [28][24], the ground truth is available as by frisking an individual, a police officer can know with certainty the presence or no of illegal substance. In such cases, one or several disparate mistreatment notions (nodes 30-41) are more suitable to assess fairness.

Nodes 22-47. The bulk of Fig. 3.3 is dedicated to disparate mistreatment fairness notions and the criteria leading to each one of them. These notions define fairness in terms of the disparity of misclassification rates among the different groups in the population. Based on their definitions, selecting the most suitable notion to use depends on four criteria, namely, whether the emphasis is on precision or recall (node 22), whether the threshold is fixed or floating (nodes 23 and 24), whether the emphasis is on FNs or FPs (nodes 26 and 28), and finally, whether the emphasis is on the positive or negative class (node 27). As some notions focus only on FP or FN (nodes 31, 32, 38, and 39), any notion insensitive to either FP or FN must not be used (nodes 42 - 47).

The diagram may be misleading if it is interpreted very categorically. This occurs when a diagram user navigates it and ends up using the recommended fairness notion without considering other important elements specific to the scenario at hand. The diagram can also be misleading when it is unclear which branch to take in a decision node. For example, the question in decision node 22 (emphasis on precision or recall?) is difficult to answer categorically in several scenarios. The decision nodes 4, 21, 12, and even 2 are typically easier to navigate but can be challenging to settle in several scenarios. Moreover, in the presence of measurement bias, the values of some features and even the outcome label may not be reliable, making the diagram navigation more challenging. A potential solution would be to label one of the branches as default (to be followed when the answer is unclear), but this can often result in a sub-optimal decision. In summary, the diagram should be considered a guide and should never be used to supersede important elements specific to the scenario at hand.

[24] Assuming the absence of measurement bias.

Table 3.2 Correspondence between Fairness notions and the selection criteria: **C1**: disparate impact , **C2**: disparate treatment , **C3**: intersectionality/masking, **C4**: historical bias, **C5**: representational bias, **C6**: measurement bias, **C7**: aggregation/evaluation/deployment bias, **C8**: standards, **C9**: ground truth available, **C10**: Y not reliable, **C11**: explanatory variables, **C12**: precision, **C13**: recall, **C14**: FP, **C15**: FN, **C16**: threshold floating.
Notation: ✓: recommended, ⚠: warning, ✗: must not, −: insensitive.

	Legal Frame			**Suspected source of bias**								**Emphasis on**		**Emphasis on**		
Criterion / **Fairness notion**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**
Statistical parity	✓	⚠	⚠	✓	⚠	⚠	⚠	✓	⚠	✓	⚠	−	−	−	−	⚠
Conditional statistical parity	✓	⚠	⚠	✓	⚠	⚠	⚠	−	⚠	✓	✓	−	−	−	−	⚠
Equalized odds	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✓	✓	⚠
Equal opportunity	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✗	✓	⚠
Predictive equality	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✓	✓	⚠
Balance for positive class	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✗	✓	✓
Balance for negative class	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✓	✗	✓
Overall balance	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✓	✓	✓
Conditional use acc. equality	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	✓	✓	⚠
Predictive parity	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	✓	✗	⚠
Negative predictive parity	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	✗	✓	⚠
Calibration	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	−	−	✓
Well-calibration	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	−	−	✓
Overall accuracy equality	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	✓	✓	✓	⚠
Treatment equality	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	−	−	✓	✓	−
Total fairness	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	−	−	✓	✓	⚠
Causal discrimination	✓	⚠	✓	⚠	✓	⚠	−	−	✓	⚠	⚠	−	−	−	−	−
Fairness through awareness	✓	⚠	✓	⚠	✓	⚠	⚠	−	✓	⚠	−	−	−	−	−	−

Finally, Table 3.2 states explicitly the relationship between every selection criterion and every fairness notion. The table uses four symbols, namely, recommended (✓), warning (⚠), must-not (✗), and insensitive (−). Insensitive means that the choice of the fairness notion is independent of the selection criterion.

3.3.5 Conclusion

With the increasingly large number of fairness notions considered in the relatively new field of fairness in ML, selecting a suitable notion for a given ML system becomes a non-trivial task. There are two contributing factors. First, the boundaries between the defined notions are increasingly fuzzy. Second, applying inappropriately a fairness notion may report discrimination in an otherwise fair scenario, or vice versa, and fail to identify discrimination in an unfair scenario. This study addresses this problem by identifying fairness-related characteristics of the scenario at hand and then using them to recommend and/or discourage specific fairness notions. The main contribution of this work is to systemize the selection process based on a decision diagram. Navigating the diagram will result in recommending and/or discouraging using fairness notions.

One of the main objectives of this study is to bridge the gap between the real-world use case scenarios of automated (and generally unintentional) discrimination and the mostly technical tackling of the problem in the literature. Hence, the study can be particularly interesting to civil rights activists, civil rights associations, anti-discrimination law enforcement agencies, and practitioners in fields where automated decision-making systems are increasingly used.

More generally, in real scenarios, there are still two important obstacles to addressing the unfairness problem in automated decision systems. First, the victims of such systems are, very often, members of minority groups with limited influence in the public sphere. Second, automated decision systems are geared towards efficiency (typically money), and to optimize profit, they are designed to sacrifice the outliers as tolerable collateral damage. After all, the system benefits most of the population (employers finding ideal candidates, banks giving loans to minimum-risk borrowers, a society with recidivists locked in prisons, etc.).

The next section focuses on causality-based fairness notions and their applicability in real-world scenarios.

3.4 Applicability of Causality-Based Fairness Notions

The most commonly used fairness notions are observational (fairness notions presented in Section 3.2.2) and rely on mere correlation between variables. The main problem of correlation-based fairness notions is that they fail to detect discrimination in the presence of statistical anomalies such as Simpson's paradox [215]. A famous example of Simpson's paradox is the gender bias in 1973 Berkeley admission [33, 146]. That year, 44% of male applicants were admitted against only 34% of female applicants. While this looks like a bias against female candidates, when the same data was analyzed by the department, acceptance rates were approximately the same with a slight bias in favor of female candidates. In other words, the statistical conclusion drawn from the sub-populations differs from that of the entire population. Considering the fairness problem from the legal and philosophical point of view reveals another limitation of statistical fairness notions. In the disparate treatment liability framework [25], discrimination claims require plaintiffs to demonstrate a causal connection between the challenged decision (e.g., hiring, firing, admission) and the sensitive feature (e.g., gender, race). It is then necessary to investigate the causal relationship between the sensitive attribute and the decision rather than the associated relationship. Because of these two limitations, it is now widely accepted that causality is necessary to address the problem of fairness [146] appropriately.

Various causality-based fairness notions have been recently proposed to tackle the fairness problem through causal inference lenses. These include total effect [181], counterfactual fairness [136], counterfactual effects [268], interventional fairness [200], etc. These notions differ from statistical fairness approaches in that they are not totally based on data but consider additional knowledge about the structure of the world in the form of a causal model (Section 2.4). This additional knowledge helps to understand how data is generated in the first place and how changes in variables propagate in a system. Most of these fairness notions are defined in terms of non-observable quantities such as interventions (to simulate random experiments) and counterfactuals (which consider other hypothetical worlds in addition to the actual world). Such quantities cannot always be uniquely computed from observational data, which significantly hinders the applicability of causality-based notions in practical scenarios. Each one of the two main causal frameworks in the literature, namely, structural causal model (SCM) with causal graphs [181] and potential outcome [110] (both presented in Section 2.4), use a different approach to compute/estimate the

causal quantities using observational data. The SCM framework relies mainly on the identifiability criterion [210] to generate an expression for the causal quantity based only on observable probabilities. If the identifiability criterion is unsatisfied, the causal quantity can not be computed using the available observable data. In such case, as an alternative, if the complete structure of the causal model is available, it is possible to estimate the distribution of the latent variables U and consequently generate an estimation of the counterfactual outcomes [136]. The potential outcome framework approximates causal quantities using several estimation techniques (e.g., matching, re-weighting, etc.) [100].

Given a real-world scenario, selecting which fairness notion to use is a challenging and error-prone task, as using the wrong fairness notion may indicate unfairness in an otherwise fair scenario or the opposite (failing to detect unfairness in an unfair scenario). This study provides guidelines to help select a suitable causality-based fairness notion given a specific real-world scenario. The guidelines are summarized in a decision diagram (Fig. 3.11) that can be easily navigated using the characteristics of the real-world scenario at hand. On the other hand, according to Pearl's SCM framework, computing causal quantities (interventions and counterfactuals) depends on their identifiability. Hence, even if a fairness notion is appropriate in some setup, it might not be applicable because of identifiability issues. Placing the various causality-based fairness notions in Pearl's causation ladder with the three corresponding rungs (observation, intervention, and counterfactual) [185] (Fig 3.12) allows us to rank these notions and indicates how difficult it is to deploy each one of them in practice.

Contributions. The main contributions of this study are (1) a guideline to help select a suitable causality-based fairness notion given a specific real-world scenario and (2) a ranking of the fairness notions according to Pearl's causation ladder, indicating how difficult it is to deploy each notion in practice.

Outline. This section starts by illustrating the need for causality through a hypothetical example of teacher firing (Section 3.4.1). Section 3.4.2 examines a comprehensive list of causality-based fairness notions. A survey on the three approaches to computing causal quantities from observable data, namely, identifiability, estimation based on the full causal model, and potential outcome estimation, is provided in Section 3.4.3. The main contributions of the study, which are the suitability and applicability of causality-based fairness notions, are described in Section 3.4.4. Finally, Section 3.4.5 concludes.

3.4.1 The Need for Causality: an Example

Consider the hypothetical example[25] of an automated system for deciding whether to fire a teacher at the end of the academic year. Deployed teacher evaluation systems have been suspected of bias in the past. For example, IMPACT is a teacher evaluation system used in the city of Washington DC and has been found to be unfair against teachers from minority groups [196, 192, 178]. Assume that the system takes as input two features, namely, the location of the school where the teacher is working (C) and the initial[26] average level of the students in her class (A). The outcome is whether to fire the teacher (Y). Assume that all 3 variables are binary with the following values: if the school is located in a high-income neighborhood, $C = 1$. Otherwise (the school is located in a low-income neighborhood), $C = 0$. If the initial average score for the students assigned to the teacher is high, $A = 1$. Otherwise (initial level is low), $A = 0$. Firing a teacher corresponds to $Y = 1$ while retaining her corresponds to $Y = 0$. The level of students in a given class can be influenced by several variables, but in this example, assume that it is only influenced by the school's location; students in high-income neighborhoods are more advantaged and typically perform better in school.

Assume now that the automated decision system is suspected to be biased by the initial level of students assigned to the teacher. That is, it is claimed that the system will more likely fire teachers who have been assigned classes with low-level students at the beginning of the academic year, which is clearly unfair. In this case, the sensitive attribute is the initial level of students assigned to the teacher (A). For concreteness, consider the prediction system that yields the following conditional probabilities:

$$\begin{aligned} &\mathbb{P}[Y = 1 \mid A = 1, C = 0] = 0.02 && \mathbb{P}[A = 1 \mid C = 0] = 0.2 \\ &\mathbb{P}[Y = 1 \mid A = 1, C = 1] = 0.0675 && \mathbb{P}[A = 1 \mid C = 1] = 0.8 \\ &\mathbb{P}[Y = 1 \mid A = 0, C = 0] = 0.01 && \mathbb{P}[A = 0 \mid C = 0] = 0.8 \\ &\mathbb{P}[Y = 1 \mid A = 0, C = 1] = 0.25 && \mathbb{P}[A = 0 \mid C = 1] = 0.2 \end{aligned}$$

and that the dataset is collected from a population where schools are located with equal proportions in high-income and low-income neighborhoods, that is, $\mathbb{P}[C = 1] = \mathbb{P}[C = 0] = 0.5$. Assume also that the proportion of classes with a low initial average level of students is the same as the one with high average initial level of students, that is, $\mathbb{P}[A = 1] = \mathbb{P}[A = 0] = 0.5$. To keep the scenario simple, assume that the level of

[25] Inspired by the prior convictions example in [170].
[26] At the beginning of the academic year.

students A does not depend on any other feature except C and that the firing decision Y depends only on A and C.

A simple approach to check the fairness of the firing decision Y w.r.t the sensitive attribute A is to contrast the conditional probabilities: $\mathbb{P}[Y = 1 \mid A = 0]$ and $\mathbb{P}[Y = 1 \mid A = 1]$ which quantify, respectively, the likelihood of firing a teacher given that she is assigned students with an initial low level versus and the likelihood of firing a teacher given that she is assigned students with an initial high-level class. Such probabilities can be computed as follows:

$$\mathbb{P}[Y = 1 \mid A = a] = \sum_{c \in \{0,1\}} \mathbb{P}[Y = 1 \mid A = a, C = c] \times \mathbb{P}[C = c \mid A = a] \tag{3.22}$$

Hence,

$$\mathbb{P}[Y = 1 \mid A = 1] = 0.02 \times 0.2 + 0.0675 \times 0.8 = 0.058$$
$$\mathbb{P}[Y = 1 \mid A = 0] = 0.01 \times 0.8 + 0.25 \times 0.2 = 0.058$$

The firing rates between teachers assigned to low-level and high-level students appear equal, as the values are equal. Hence, no discrimination is detected[27]. This conclusion is flawed because it doesn't consider the mechanism by which the observed data is generated. In particular, the school's location where the teacher works influences both the initial level of students assigned to her and the decision to fire or retain her. The $\mathbb{P}[A|C]$ distribution indicates that 80% of classes in low-income neighborhoods have students with low initial levels ($\mathbb{P}[A = 0 \mid C = 0] = 0.8$) while 80% of classes in high-income neighborhoods have students with high initial levels ($\mathbb{P}[A = 1 \mid C = 1] = 0.8$). The automated decision system is biased in this case because $\mathbb{P}[Y = 1 \mid A = 0, C = 1]$, the probability of firing a teacher in high-income neighborhoods who is assigned a class with a low initial level, is exceptionally high (0.25). Using simple conditional probabilities on this collected dataset fails to appropriately account for that bias because very few teachers in high-income neighborhoods are assigned low-level classes in this particular dataset ($\mathbb{P}[A = 0 \mid C = 1] = 0.2$). Generally, any statistical fairness notion that relies solely on correlation between variables will fail to detect such bias.

The causal relationships between variables should be considered to avoid such misleading conclusions. Fig. 3.4 illustrates the causal relations between the three variables of the above example where the school's location C is a confounder. Based

[27]This corresponds to statistical parity (Eq. (3.1)).

on such a causal graph, a firing decision system is fair if it is as likely to fire teachers in the following two hypothetical cases: (1) when *all teachers in the population are assigned students of low level on average, and (2) when* all teachers in the population are assigned students of high level on average. This is achieved using intervention ($do()$ operator)[28] and allows to break the problematic dependence between A and C. The probabilities of firing a teacher in these two hypothetical cases are expressed as $\mathbb{P}[Y_{A=0} = 1] = \mathbb{P}[Y = 1 \mid do(A = 0)]$ and $\mathbb{P}[Y_{A=1} = 1] = \mathbb{P}[Y = 1 \mid do(A = 1)]$ respectively. In this simple graph, and assuming no other variable is used in the prediction, these probabilities can be computed as follows:

$$\mathbb{P}[Y_{A=a} = 1] = \sum_{c \in \{0,1\}} \mathbb{P}[Y = 1 \mid A = a, C = c] \times \mathbb{P}[C = c]$$

Hence,

$$\mathbb{P}[Y_{A=1} = 1] = 0.02 \times 0.5 + 0.0675 \times 0.5 = 0.0437$$
$$\mathbb{P}[Y_{A=0} = 1] = 0.01 \times 0.5 + 0.25 \times 0.5 = 0.13$$

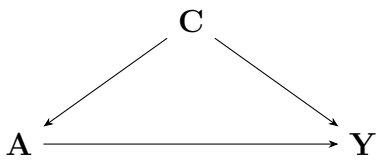


Fig. 3.4 Causal graph of the firing example. C: location of school, A: initial level of students, Y: firing.

The values confirm the existence of a bias against teachers who are assigned classes with initially low levels.

3.4.2 Causality-Based Fairness Notions

Assume that the sensitive attribute A and the outcome Y are binary variables where $A = a_1$ denotes the privileged group (e.g., male) and $A = a_0$ the disadvantaged group (e.g., female).

[28] Intervention and the $do()$ operator are already explained in Section 2.4.1.

You can refer to the simple job hiring example in Table A.11 in Appendix A.1.2 to understand better how the various causality-based fairness notions presented in this section are computed.

The most common non-causal fairness notion is total variation (TV), known as statistical disparity (Eq. (4.2)), demographic disparity, or risk difference. The total variation of $A = a_0$ on the outcome $Y = y$ with reference $A = a_1$ is defined using conditional probabilities as follows:

$$\mathrm{TV}_{a_1,a_0}(y) = \mathbb{P}[y \mid a_1] - \mathbb{P}[y \mid a_0] \tag{3.23}$$

Intuitively, $\mathrm{TV}_{a_1,a_0}(y)$ measures the difference between the conditional distributions of Y when we (passively) observe A changing from a_1 to a_0. The main limitation of TV is its purely statistical nature, which makes it unable to reflect the causal relationship between A and Y; that is, it is insensitive to the mechanism by which data is generated and collected. Total effect (TE) [181] is the causal version of TV and is defined in terms of experimental probabilities as follows:

$$\mathrm{TE}_{a_1,a_0}(y) = \mathbb{P}[y_{a_1}] - \mathbb{P}[y_{a_0}] \tag{3.24}$$

Recall that the notation $\mathbb{P}[y_{a_1}]$ is equivalent to $\mathbb{P}[Y = y|do(A = a_1)]$ (Section 2.4.1). TE measures the effect of the change of A from a_1 to a_0 on $Y = y$ along all the causal paths from A to Y. Intuitively, while TV reflects the difference in proportions of $Y = y$ in the current cohort, TE reflects the difference in proportions of $Y = y$ in the entire population. For the binary outcome case, TE is equivalent to the average treatment effect (ATE) [167] in the potential outcome framework, which is defined as follows:

$$\mathrm{ATE}_{a_1,a_0} = \mathbb{E}[Y^{a_1} - Y^{a_0}] \tag{3.25}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i^{a_1} - Y_i^{a_0}) \tag{3.26}$$

where n is the number of observed samples. ATE corresponds exactly to FACE in [124].

Computing exactly ATE requires the knowledge of both potential outcomes: the observed and the counterfactual. As the latter is almost impossible to observe, the exact computation of ATE is typically impossible. However, for the sake of illustration, we assume the counterfactual outcome is available and show how ATE is computed. Appendix A.1.2 shows how ATE and counterfactual outcomes can be estimated from observable data.

Computing the causal effect based only on the observed treatment group samples (e.g., female applicants only) corresponds to a variant of TE called the effect of treatment on the treated (ETT) [181] and is defined as:

$$\mathrm{ETT}_{a_1,a_0}(y) = \mathbb{P}[y_{a_1} \mid a_0] - \mathbb{P}[y_{a_0} \mid a_0] \tag{3.27}$$

In the binary outcome case, ETT corresponds to the average treatment effect on the treated ATT [167] in the potential outcome framework defined as:

$$\begin{aligned} \mathrm{ATT}_{a_1,a_0} &= \mathbb{E}[Y^{a_1}|A = a_0] - \mathbb{E}[Y^{a_0}|A = a_0] && (3.28)\\ &= \frac{1}{n_1} \sum_{i:A=a_0} (Y_i^{a_1} - Y_i^{a_0}) && (3.29) \end{aligned}$$

where n_1 is the number of samples in the treatment group. ATT is also called FACT in [124].

Average treatment effect on the control group (ATC) [167] is the same as ATT but focusing instead on the control group:

$$\begin{aligned} \mathrm{ATC}_{a_1,a_0} &= \mathbb{E}[Y^{a_1}|A = a_1] - \mathbb{E}[Y^{a_0}|A = a_1] && (3.30)\\ &= \frac{1}{n_2} \sum_{i:A=a_1} (Y_i^{a_1} - Y_i^{a_0}) && (3.31) \end{aligned}$$

where n_2 is the number of samples in the control group.

Conditional average treatment effect (CATE) [167] is defined similarly, but conditioning on some other covariate instead of the sensitive attribute A:

$$\begin{aligned} \mathrm{CATE}_{a_1,a_0}(X = x) &= \mathbb{E}[Y^{a_1}|X = x] - \mathbb{E}[Y^{a_0}|X = x] && (3.32)\\ &= \frac{1}{n_x} \sum_{i:X=x} (Y_i^{a_1} - Y_i^{a_0}) && (3.33) \end{aligned}$$

where n_x is the number of samples in the subgroup $X = x$.

Unlike the SCM framework, in the potential framework, it is possible to define individual treatment effect ITE [167], which is defined for every unit i as:

$$\mathrm{ITE}_{a_1,a_0}(i) = Y_i^{a_1} - Y_i^{a_0} \tag{3.34}$$

ATC, CATE, and ITE are defined and typically used in the potential outcome framework but have no equivalents in the SCM framework. However, although ATC and CATE

can be easily represented in the SCM formalism, ITE cannot be easily formalized in the SCM framework.

The job hiring example of Tables A.11 and A.12 in Appendix A.1.2 is interesting because it illustrates a statistical anomaly where some statistical notions such as TV fail to appropriately account for the bias between sub-populations (e.g., female vs. male). This simple job hiring scenario is similar to the Berkeley sex discrimination in college admission [33] where data showed a bias for male applicants overall, but when results were analyzed separately for each department, data showed a slight bias in favor of female candidates. The Berkeley scenario is typically used as an example of Simpson's paradox [215]. In both scenarios, by considering the outcome of the observable samples in the counterfactual setup, the above causality-based fairness notions could appropriately assess gender (A) discrimination on the outcome (Y). The job hiring example illustrating the statistical anomaly can be easily modified to reflect Simpson's paradox [215]. We provide such data in Table A.13 in Appendix A.1.2.

All the above causality-based fairness notions fall into the framework of disparate impact [25], which aims at ensuring the equality of outcomes among all groups (protected/treatment and unprotected/control). An alternative framework is the disparate treatment [25] which seeks equality of treatment achievable through prohibiting the use of the sensitive attribute in the decision process. The main idea is to split the causal effect between the sensitive attribute A and the outcome Y into several causal pathways, each of which is either fair, unfair, or spurious. Common fairness notions from the disparate treatment framework include direct effect, indirect effect, and path-specific effect [180]. An effect can be deemed fair, unfair, or spurious by an expert on the scenario at hand. An unfair effect is called discrimination. Direct discrimination is assessed using causal effect along the direct edge from A to Y. Indirect discrimination is measured using the causal effect along causal paths that pass through proxy attributes[29]. A fair or explainable discrimination is measured using causal pathways passing through explaining variables. The spurious effect corresponds to a pathway starting with an incident edge into the sensitive attribute A.

Fig. 3.5 presents a causal graph of the job hiring scenario involving an explaining variable E (e.g., education and academic degrees) and a proxy/redlining variable R (e.g., the hobby of the candidate). Hiring discrimination due to education level is legitimate and considered fair. In contrast, discrimination due to the hobby of the candidate is unfair as it is a proxy for gender (the type of hobby generally indicates the

[29]A proxy is an attribute that cannot be objectively justified if used in the decision-making process. It is also known as a redlining attribute.

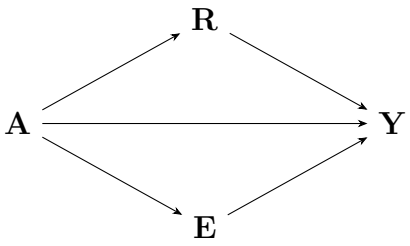


Fig. 3.5 Job hiring scenario where A is gender, Y: hiring decision, R: hobby of a candidate ($R = 1$ for mechanical hobby, $R = 0$ for non-mechanical hobby), and E: education level of the candidate ($E = 1$ for college degree, $E = 0$ for no college degree).

candidate's gender). Direct effect can be computed by simply "blocking" all indirect causal paths. An indirect causal path is a directed path from A to Y going through one or several mediator variables. For example, in Fig. 3.5, there are two indirect causal paths $A \rightarrow R \rightarrow Y$ and $A \rightarrow E \rightarrow Y$. To compute the direct causal effect ($A \rightarrow Y$), both indirect causal paths need to be blocked by adjusting on variables R and E. As there are no confounders, the direct effect can be computed as:

$$\begin{aligned} \text{DE}_{a_1,a_0}(y) &= \mathbb{P}[y \mid a_1, R, E] - \mathbb{P}[y \mid a_0, R, E] \\ &= \sum_r \sum_e \left(\mathbb{P}[y \mid a_1, r, e] - \mathbb{P}[y \mid a_0, r, e]\right) \end{aligned}$$

In the presence of confounders (between A and Y, between R and Y, etc.), natural direct effect (NDE) [180] is a more general notion that measures the direct causal effect and is defined as:

$$\text{NDE}_{a_1,a_0}(y) = \mathbb{P}[y_{a_0,Z_{a_1}}] - \mathbb{P}[y_{a_1}] \tag{3.35}$$

Where Z is the set of mediator variables and $\mathbb{P}[y_{a_0,Z_{a_1}}]$ is the probability of $Y = y$ had A been a_0 and had Z been the value it would naturally take if $A = a_1$. That is, A is set to a_0 in the single direct path $A \rightarrow Y$ and is set to a_1 in all other indirect paths ($A \rightarrow R \rightarrow Y$ and $A \rightarrow E \rightarrow Y$). Check Appendix A.1.2 to see how NDE is computed.

Natural indirect effect (NIE) [180] measures the indirect effect of A on Y and is defined as:

$$\text{NIE}_{a_1,a_0}(y) = \mathbb{P}[y_{a_1,Z_{a_0}}] - \mathbb{P}[y_{a_1}] \tag{3.36}$$

The problem with NIE is that it does not distinguish between the fair (explainable) and unfair (indirect discrimination) effects. Path-specific effect [181, 51, 255] is a more nuanced measure that characterizes the causal effect in terms of specific paths.

Given a path set π, the π-specific effect is defined as:

$$\text{PSE}^{\pi}_{a_1,a_0}(y) = \mathbb{P}[y_{a_0|_\pi, a_1|_{\overline{\pi}}}] - \mathbb{P}[y_{a_1}] \tag{3.37}$$

where $\mathbb{P}[y_{a_0|_\pi, a_1|_{\overline{\pi}}}]$ is the probability of $Y = y$ in the counterfactual situation where the effect of A on Y with the intervention (a_0) is transmitted along π, while the effect of A on Y without the intervention (a_1) is transmitted along paths not in π (denoted by: $\overline{\pi}$). Using the job hiring example of Fig. 3.5, Eq. (3.37) can be used to assess only unfair discrimination which is transmitted through the direct path $A \to Y$ and the indirect path $A \to R \to Y$. The third path $A \to E \to Y$ transmits explainable (fair) discrimination and hence, should not be considered.

No Unresolved Discrimination [127]. No unresolved discrimination is a fairness notion that falls into the disparate treatment framework and focuses on the indirect causal effects from A to Y. No unresolved discrimination is satisfied when no directed path from A to Y is allowed, except via a resolving (explaining) variable E. A resolving variable is any variable in a causal graph that is influenced by the sensitive attribute in a manner accepted as non-discriminatory. Fig. 3.6 presents two alternative causal graphs for the job hiring example. The graph at the left exhibits unresolved discrimination along the heavy paths: $A \to R \to Y$ and $A \to Y$. By contrast, the graph at the right does not exhibit any unresolved discrimination as the effect of A on Y is justified by the resolved variable E: $A \to E \to Y$.

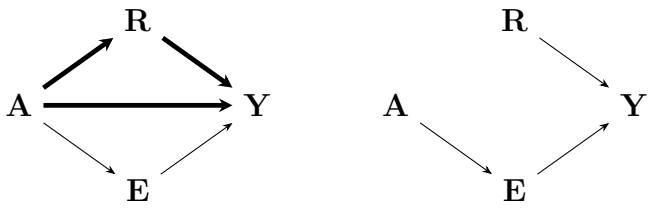


Fig. 3.6 Y exhibits unresolved discrimination in the left graph (along the heavy paths) but not the right one.

The assumption of valid causal graph availability limits the use of no unresolved discrimination in real scenarios. [127] provides formal proof that even with prior knowledge of resolving variables, it is not always possible to tell, based on observational data only, if Y satisfies no unresolved discrimination.

No Proxy Discrimination [127]. Similarly to no unresolved discrimination, no proxy discrimination focuses on indirect discrimination. A causal graph exhibits potential proxy discrimination if a path exists from the protected attribute A to the outcome Y that is blocked by a proxy/redlining variable R. It is called proxy because it is used to decide about the outcome Y while it is a descendent of A, which is significantly correlated with it in such a way that using the proxy in the decision has almost the same impact as using A directly. An outcome variable Y exhibits no proxy discrimination if the equality:

$$\mathbb{P}[Y \mid do(R = r)] = \mathbb{P}[Y \mid do(R = r')] \quad \forall\, r, r' \in dom(R) \tag{3.38}$$

holds for any potential proxy R.

Fig. 3.7 shows two similar causal graphs for the same job hiring example. The causal graph at the left presents potential proxy discrimination via the path: $A \rightarrow R \rightarrow Y$. However, the graph at the right is free of proxy discrimination as the edge between A and its proxy R has been removed due to the intervention on R $(R = r)$.

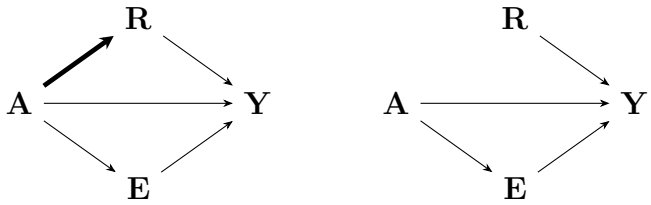


Fig. 3.7 The graph at the left exhibits potential proxy discrimination (along the heavy edge between A and R) but not in the right one.

Similarly to no unresolved discrimination, no proxy discrimination requires a valid causal graph. Hence, both fairness notions depend on the correct output of the causal discovery task (Chapter 5 introduces our work related to this problem).

X=x

Counterfactual Fairness [136]. Counterfactual fairness is a very strong fairness notion requiring equality between the observed and counterfactual outcomes for every individual. That is, an outcome Y is counterfactually fair if under any assignment of values $X = x$ and any individual in U,

$$\mathbb{P}[y_{a_1}(U) \mid X = x, A = a_1] = \mathbb{P}[y_{a_0}(U) \mid X = x, A = a_1] \tag{3.39}$$

where $X = V \backslash \{A, Y\}$ is the set of all remaining variables. As the latent variable U appears in Eq. (3.39), counterfactual fairness is an individual fairness notion. It is satisfied if the probability distribution of the outcome Y is the same for every possible individual in the actual and counterfactual worlds. Counterfactual fairness typically coincides with ITE (Eq. (3.34)).

[136] could test counterfactual fairness by making a very strong assumption. That is, they assumed the full structure of the causal model is available, including the latent variables U. They could then estimate the $\mathbb{P}[U]$ distribution using Markov chain Monte Carlo methods and the observed data. Thanks to the estimated distribution of $\mathbb{P}[U]$, they could compute counterfactuals using Pearl's three-step process: abduction, action, and prediction [181]. Hence, another sample with counterfactual sensitive value is generated for every individual in the population. Counterfactual fairness is finally assessed by comparing the density functions of the actual and counterfactual samples.

Counterfactual Effects [268]. By conditioning on the sensitive attribute $A = a$, Zhang and Bareinboim defined two variants of NDE (Eq. (3.35)) and NIE (Eq. (3.36)) which focus on the direct and indirect effect for a specific group. In addition, they characterize a third type of effect, spurious, which considers the back-door paths between A and Y, that is, paths with an arrow into A.

The three effects are defined as follows:

$$\mathrm{DE}_{a_1,a_0}(y|a) = \mathbb{P}[y_{a_0,Z_{a_1}}|a] - \mathbb{P}[y_{a_1}|a] \tag{3.40}$$

$$\mathrm{IE}_{a_1,a_0}(y|a) = \mathbb{P}[y_{a_1,Z_{a_0}}|a] - \mathbb{P}[y_{a_1}|a] \tag{3.41}$$

$$\mathrm{SE}_{a_1,a_0}(y) = \mathbb{P}[y_{a_1}|a_0] - \mathbb{P}[y|a_1] \tag{3.42}$$

where in Eq. (3.40) and (3.41), a can be a_0 or a_1. Considering the simple job hiring example and focusing on the female group ($A = 0$), $\mathrm{DE}_{1,0}(y|0)$ measures the change in the probability of Y (e.g., hiring) had A been 0 (female), while mediators E and R are kept at the level they would take had A been 1 (male). Appendix A.1.2 shows how counterfactual effects are computed in practice.

Compared to NDE and NIE, counterfactual effects focus only on individuals of a specific group (e.g., only female candidates) and characterize the causal effect through spurious (back-door) paths. This spurious effect is what makes causal relations different from mere statistical correlations. However, counterfactual indirect effect still does not distinguish between fair and unfair direct effects.

Counterfactual Error Rates [267]. Equalized odds (Eq. (3.3)) is an important statistical fairness notion that requires equality of error rates (TPR and FPR) across sub-populations. It can be re-written as follows:

$$\mathrm{ER}_{a_1,a_0}(\hat{y}|y) = \mathbb{P}[\hat{y} \mid a_1, y] - \mathbb{P}[\hat{y} \mid a_0, y] \tag{3.43}$$

where $\hat{y}$ denotes the prediction while y denotes the true outcome. The problem with this statistical notion is the difficulty in identifying the causes behind the discrimination, if any. [267] decomposes equalized odds(Eq. (3.43)) using three counterfactual measures corresponding to the direct, indirect, and spurious effects of A on $\hat{Y}$. The three measures are counterfactual direct error rate, counterfactual indirect error rate, and counterfactual spurious error rate. Let $\hat{y} = f(\hat{\mathbf{pa}})$ be a classifier where $\hat{\mathbf{PA}}$ is the set of input features (parent variables of $\hat{Y}$) for the classifier. The counterfactual error rates for a sub-population a, y (with prediction $\hat{y} \neq y$) are defined as:

$$\mathrm{ER}^{d}_{a_1,a_0}(\hat{y} \mid a, y) = \mathbb{P}[\hat{y}_{a_0,y,(\hat{\mathbf{PA}}\setminus A)_{a_1,y}} \mid a, y] - \mathbb{P}[\hat{y}_{a_1,y} \mid a, y] \tag{3.44}$$

$$\mathrm{ER}^{i}_{a_1,a_0}(\hat{y} \mid a, y) = \mathbb{P}[\hat{y}_{a_1,y,(\hat{\mathbf{PA}}\setminus A)_{a_0,y}} \mid a, y] - \mathbb{P}[\hat{y}_{a_1,y} \mid a, y] \tag{3.45}$$

$$\mathrm{ER}^{s}_{a_1,a_0}(\hat{y} \mid y) = \mathbb{P}[\hat{y}_{a_1,y} \mid a_0, y] - \mathbb{P}[\hat{y}_{a_1,y} \mid a_1, y] \tag{3.46}$$

For example, the counterfactual direct error rate (Eq. (3.44)) measures the error rate (disparity between the true and the predicted outcome) in terms of the direct effects of the sensitive attribute A on the prediction $\hat{Y}$. In the job hiring example, considering the female sub-population that *should* be hired ($A = 0$ and $Y = 1$), it reads: for a female candidate that should be hired, how would the prediction $\hat{Y}$ change had the candidate been a female (A been 0) while keeping all the other features $\hat{\mathbf{PA}} \setminus A$ at the level that they would attain had "she was male", compared to the prediction $\hat{Y}$ she would receive had "she was male" and should have been hired?

Appendix A.1.2 shows how counterfactual error rates can be computed in practice.

Interestingly, the statistical equalized odds error rate (Eq. (3.43)) can be decomposed in terms of the three above causality-based error rates:

$$\mathrm{ER}_{a_1,a_0}(\hat{y} \mid y) = \mathrm{ER}^{d}_{a_1,a_0}(\hat{y} \mid a_1, y) - \mathrm{ER}^{i}_{a_0,a_1}(\hat{y} \mid a_1, y) - \mathrm{ER}^{s}_{a_0,a_1}(\hat{y} \mid y) \tag{3.47}$$

Individual Direct Discrimination [273]. Individual direct discrimination aims to discover direct discrimination at the individual level. It is based on situation testing [30], a legally grounded technique for analyzing discrimination at an individual level. It compares the individual with similar individuals from both groups (protected and

unprotected). That is, for an individual i in question, find the k other individuals who are the most similar to i in the group $A = a_0$ and k similar individuals from the group $A = a_1$. The first set is denoted as S^+ while the second as S^-. The target individual is considered as discriminated if the difference observed between the rate of positive decisions in S^- and S^+ is higher than a predefined threshold τ (typically 5%).

Causal inference is used to define the distance function $d(i, i')$ required to select the elements of S^- and S^+. First, only attributes that are direct causes of the outcome should considered in the computation of the distance. That is, based on the causal graph, $Q = Pa(Y)\backslash\{A\}$ denotes the set of variables that should be used in the distance function. Second, the function definition should consider the causal effect of each of the selected attributes ($Q_k \in Q$) on the outcome. In particular, for each variable Q_k, $\text{CE}(q_k, q'_k)$ measures the causal effect on the outcome when the value of Q_k changes from q_k to q'_k and is defined as:

$$\text{CE}(q_k, q'_k) = \mathbb{P}[y_q] - \mathbb{P}[y_{q'_k, q\backslash\{q_k\}}] \tag{3.48}$$

where $\mathbb{P}[y_q]$ is the effect of the intervention that forces the set Q to take the set of values q, and $\mathbb{P}[y_{q'_k, q\backslash\{q_k\}}]$ is the effect of the intervention that forces Q_k to take the value q'_k and other attributes in Q to take the same values as q.

The two individual fairness notions mentioned above, namely, ITE (Eq. (3.34)) and counterfactual fairness (Eq. (3.39)), rely on the counterfactual outcome to assess fairness for every individual. Individual direct discrimination drops this requirement and uses instead the sets S^- and S^+ composed of similar individuals in both groups. Hence, it can be considered an estimation technique to circumvent counterfactual needs. However, the distance function between two individuals $d(i, i')$ is unnecessarily complex; it is defined in terms of the causal effects of every covariate X on the outcome Y. These causal effects are re-computed each time the distance between two individuals is needed. Matching techniques in the potential outcome framework use much simpler distance metrics. Matching techniques are discussed in Section 3.4.3.

Non-Discrimination Criterion [274]. Non-discrimination criterion is a group fairness notion that aims to discover and quantify direct discrimination through the direct causal effect of A on Y. Recall that, given a causal graph $\mathcal{G}$, a direct effect of A on Y is the causal effect through the edge $A \to Y$. The idea is to consider a modified graph $\mathcal{G}'$ where the edge in question ($A \to Y$) is discarded. A *block set* Q is a set of variables that blocks all causal effects from A to Y in the modified graph $\mathcal{G}'$. Hence, A and Y are independent conditioning on Q in $\mathcal{G}'$, that is, $(A \perp Y|Q)_{\mathcal{G}'}$.

Hence, conditioning on the same variables Q, any dependence between A and Y in $\mathcal{G}$ is due to the direct effect of A on Y, which indicates direct discrimination. This discrimination can be assessed using simply the risk difference [198]:

$$\mid \Delta P|_q \mid = \mid \mathbb{P}[y \mid a_1, q] - \mathbb{P}[y \mid a_0, q] \mid \tag{3.49}$$

where q is a value assignment for the block set Q and the absolute value to consider both positive and negative discriminations. No direct discrimination can be concluded if the risk difference is less than a threshold τ for all combinations of values of all block sets, that is, Eq. (3.49) holds for each value assignment q of each block set Q.

NDE (Eq. (3.35)) and counterfactual direct effect (Eq (3.40)) also focus on assessing the direct discrimination. Still, they both rely on nested counterfactual quantities that are not observable from the data. Non-discrimination criterion circumvents this difficulty by using block sets and considering all combinations of values of these block sets. Similarly to individual direct discrimination, it can be considered an estimation technique to avoid dealing with counterfactual quantities. This approach, however, does not work in semi-Markovian models as A and Y will never be independent in $\mathcal{G}'$ ($(A \perp Y|Q)_{\mathcal{G}'}$) because of hidden confounders.

Equality of Effort [107]. Equality of effort is a fairness notion that identifies discrimination by assessing how much effort the disadvantaged individual/group needs to reach a certain outcome level. A treatment variable T is selected to address the question: "To what extent should the treatment variable T change to make the individual (or a group of individuals) achieve a certain outcome level?". Hence, this notion focuses on whether the effort to reach a certain outcome level is the same for the protected and unprotected groups. Considering the simple job hiring example, the education level E is a good choice for the treatment variable. Two equality of effort notions are defined based on the potential outcome framework, individual γ-Equal effort and system γ-Equal effort. Let $Y_i^{(t)}$ be the potential outcome for individual i had T been t and $\mathbb{E}[Y_i^{(t)}]$ be the expected outcome for individual i. Situation testing [30] is used to estimate the counterfactual potential outcome in a similar way as individual direct discrimination (Eq. (3.48)). Let S^+ and S^- be the two sets of similar individuals with $A = a_0$ and $A = a_1$, respectively, and $\mathbb{E}[Y_{S^+}^{(t)}]$ be the expected outcome under treatment t for the subgroup of individuals S^+. The minimal effort needed to achieve γ-level of outcome variable within the subgroup S^+ is defined as:

$$\Psi_{S^+}(\gamma) = \underset{t \in T}{\operatorname{argmin}} \{\mathbb{E}[Y_{S^+}^{(t)}] \geq \gamma\} \tag{3.50}$$

Individual γ-Equal effort is satisfied for individual i if:

$$\Psi_{S^+}(\gamma) = \Psi_{S^-}(\gamma) \tag{3.51}$$

System γ-Equal effort is satisfied for a sub-population (e.g., $A = a_1$) if:

$$\Psi_{D^+}(\gamma) = \Psi_{D^-}(\gamma) \tag{3.52}$$

where D^+ and D^- are the subsets of the entire dataset with sensitive attributes a_1 and a_0, respectively. Both criteria can be used to measure the effort discrepancy between protected and unprotected groups by considering the difference $\Psi_{X^+}(\gamma) - \Psi_{X^-}(\gamma)$. Unlike most causality-based fairness notions that intervene (*do* operator) on the sensitive attribute A (y_a, Y_i^a, etc.), equality of effort intervenes instead on a treatment variable T ($Y_i^{(t)}$). The main limitation of the equality of effort notion is that a single treatment variable typically does not appropriately reflect the discrepancy between protected and unprotected groups.

Interventional and Justifiable Fairness [200]. Interventional and justifiable fairness is a group-level fairness that can be seen as a strong version of total effect (Eq. (3.24)). Instead of intervening only on the sensitive attribute A, interventional fairness intervenes on all remaining variables. Let K be a subset of V excluding A and Y, that is, $K \subseteq V - \{A, Y\}$. A predicting algorithm is K-fair if for any assignment of values $K = k$ and outcome $Y = y$:

$$\mathbb{P}[y_{a_1,k}] = \mathbb{P}[y_{a_0,k}] \tag{3.53}$$

A predicting algorithm is interventionally fair if it is K-fair for every set of variables K. Using the job hiring example of Fig. 3.5, interventional fairness holds between male and female groups if $\mathbb{P}[y_{1,E_u,R_v}] = \mathbb{P}[y_{0,E_u,R_v}]$, $\forall u, v \in \{0, 1\}$. The interventional fairness formula (Eq. (3.53)) is similar to the non-discrimination criterion formula (Eq. (3.49)). However, while Eq. (3.49) uses simple conditioning on A and covariates, Eq. (3.53) makes an intervention on A and all other covariates and hence works on Markovian as well as semi-Markovian models.

Justifiable fairness is a relaxation of interventional fairness achieved by classifying the variables as admissible (denoted as E) or inadmissible (denoted as R), which correspond, respectively, to explainable and proxy/redlining variables as defined previously. A predicting algorithm is justifiably fair if it is K-fair w.r.t only supersets of E, that is,

$K \supseteq E$. Hence, instead of intervening on all variables, it is enough to intervene on only admissible variables (or any superset of them). Graphically, suppose all directed paths from the sensitive attribute A to the outcome Y go through an admissible attribute in E. In that case, the algorithm is justifiably fair, which typically coincides with no-unresolved discrimination. Using the job hiring example (Fig. 3.5), justifiable fairness holds if $\mathbb{P}[y_{1,E_u}] = \mathbb{P}[y_{0,E_u}]$, $\forall u \in \{0, 1\}$. Notice that in case $E = \emptyset$, justifiable fairness coincides with interventional fairness. Interestingly, being based solely on interventions, interventional and justifiable notions of fairness do not require the presence of the underlying causal model. The only assumption is the ability to distinguish admissible and inadmissible variables.

Individual Equalized Counterfactual odds [188]. Individual equalized counterfactual odds is a stronger version of counterfactual fairness requiring, in addition, that the factual-counterfactual pair share the same value of the outcome Y. The aim is to have a counterfactual version of equalized odds (Eq. (3.3)). This is achieved by conditioning both sides of Eq. (3.39) on the same outcome $Y = y$. A predictor satisfies individual equalized counterfactual odds if:

$$\mathbb{P}[\hat{y}_{a_1} \mid X = x, y_{a_1}, A = a_1] = \mathbb{P}[\hat{y}_{a_0} \mid X = x, y_{a_0}, A = a_1] \tag{3.54}$$

The only difference with Eq. (3.39) is the additional conditioning $Y = y_{a_1}$ in the LHS and $Y = y_{a_0}$ in the RHS. Counterfactual error rates are the only other causality-based fairness notions considering the outcome Y (Eq. (3.43)). However, unlike counterfactual error rates, individual equalized counterfactual odds require intervention on Y. This is the only fairness criterion that requires intervention on the prediction $\hat{Y}$ and the actual outcome Y.

3.4.3 Computing Causal Quantities from Observable Data

Using causality-based fairness notions is challenging for two reasons. First, only the factual outcome can be observed among the two possible outcomes. The counterfactual outcome is usually impossible to observe (e.g., if the gender of a candidate is female (factual), it is impossible to observe the counterfactual outcome when the same candidate would have been male). Second, sensitive attribute values (e.g., male and female) are typically not randomly assigned in observational data. Hence, using observational data, the main difficulty in applying causality-based fairness notions is to compute and/or estimate the causal quantities (counterfactual outcomes, causal effects, counterfactual

effects, etc.). This includes all grayed columns in the simple toy datasets used in Appendix A.1.2 as well as all fairness notions such as ATE, ETT, counterfactual fairness, etc. Each causal framework, namely, SCM with causal graphs and potential outcomes, uses a different approach to compute/estimate the causal quantities using observational data. The SCM framework relies mainly on the identifiability criterion to generate an expression for the causal quantity based only on observable probabilities. If the identifiability criterion is not satisfied, the causal quantity cannot be computed using the available observable data. In such case, as an alternative, if the complete structure of the causal model is available, it is possible to estimate the distribution of the latent variables U and consequently generate an estimation of the counterfactual outcomes. The potential outcome framework approximates causal quantities using several estimation techniques (e.g., matching, re-weighting, etc.). The following subsections illustrate the above three approaches: identifiability, estimation based on full causal model, and potential outcome estimation techniques.

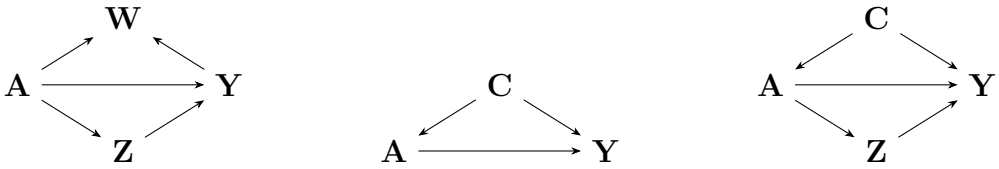
Identifiability. The identifiability of causal quantities has been extensively studied in the literature: causal effect (intervention) identifiability [93, 231, 232, 230, 210, 108, 212, 181], counterfactual identifiability [211, 212, 209, 254], direct/indirect effects [180] and path-specific effect identifiability [20, 209, 275, 272, 160]. This section summarizes the main identifiability conditions related to the specific problem of discrimination discovery (you can refer to our paper [157][30] for a more comprehensive study on identifiability).

- **Identifiability of Causal Effect (Intervention).** The causal effect of a cause variable X on an effect variable Y is computed using $\mathbb{P}[Y_x] = \mathbb{P}[Y|do(X = x)]$, the distribution of Y after the intervention $X = x$. In a discrimination setup, the cause is typically the sensitive attribute A. A basic case where identifiability can be avoided altogether is when it is possible to perform experiments by intervening on the sensitive attribute A. When this is possible, a randomized controlled trial (RCT) (Section 1.1.3) can be used to estimate the causal effect. RCT consists of randomly assigning subjects (e.g., individuals) to treatments (e.g., gender), then comparing the outcome Y of all treatment groups. However, in the context of ML fairness, RCT is often not an option as experiments can be too costly to implement, physically impossible, or ethically not acceptable (e.g., changing the gender of a job applicant). In Markovian models (no unobserved confounding), the causal effect is always identifiable (Corollary 3.2.6 in [181]). The simplest

[30] This paper is not included in the manuscript.

case is when there is no confounding between A and Y (Figure 3.8a). In that case, the causal effect matches the conditional probability regardless of any mediator:

$$\mathbb{P}[y_a] = \mathbb{P}[y|do(a)] = \mathbb{P}[y|a] \tag{3.55}$$



(a) Simple Markovian model with a collider (W) and a mediator (Z).

(b) Simple Markovian model with a confounder (C).

(c) Simple Markovian model with a confounder (C) and a mediator (Z).

Fig. 3.8 Simple Markovian causal graphs.

In presence of an observable confounder (Fig. 3.8b), $\mathbb{P}[y_a]$ is identifiable by adjusting on the confounder:

$$\mathbb{P}[y_a] = \sum_C \mathbb{P}[y|a,c]\ \mathbb{P}[c] \tag{3.56}$$

where the summation is on values c in the domain (sample space) of C denoted as $dom(C)$. Eq. (3.56) is called the back-door formula[31]. The backdoor adjusting formula is different from the joint probability

$$\mathbb{P}[y,a,c] = \mathbb{P}[y|a,c]\ \mathbb{P}[a|c]\ \mathbb{P}[c]$$

and the conditional probability

$$\mathbb{P}[y|a] = \sum_C \mathbb{P}[y|a,c]\ \mathbb{P}[c|a]$$

For semi-Markovian models, identifiability of $\mathbb{P}[y_a]$ is not guaranteed. In case it is identifiable, Pearl [181] proposes a *do*-calculus composed of three rules allowing the expression of interventional probabilities in terms of observational ones:

1. $\mathbb{P}[y_a|z,w] = \mathbb{P}[y_a|z]$ provided that the set of variables Z blocks all backdoor paths from W to Y after all arrows leading to A have been deleted.

[31] Called also adjustment formula or stratification.

2. $\mathbb{P}[y_a|z] = \mathbb{P}[y|a,z]$ provided that the set of variables Z blocks all backdoor paths from A to Y.
3. $\mathbb{P}[y_a] = \mathbb{P}[y]$ provided that there are no causal paths between A and Y.

do-calculus has been proven to be sound and complete in identifying interventional distributions [108]. For example, $\mathbb{P}[y_a]$ is identifiable in Fig. 3.9d. Appendix A.1.3 shows how this example is computed using the *do*-calculus rules.

As an alternative to using the *do*-calculus, several contributions in the identifiability literature focused on defining graphical patterns and mapping them to simple and concise intervention-free expressions [231, 232, 230, 210]. All

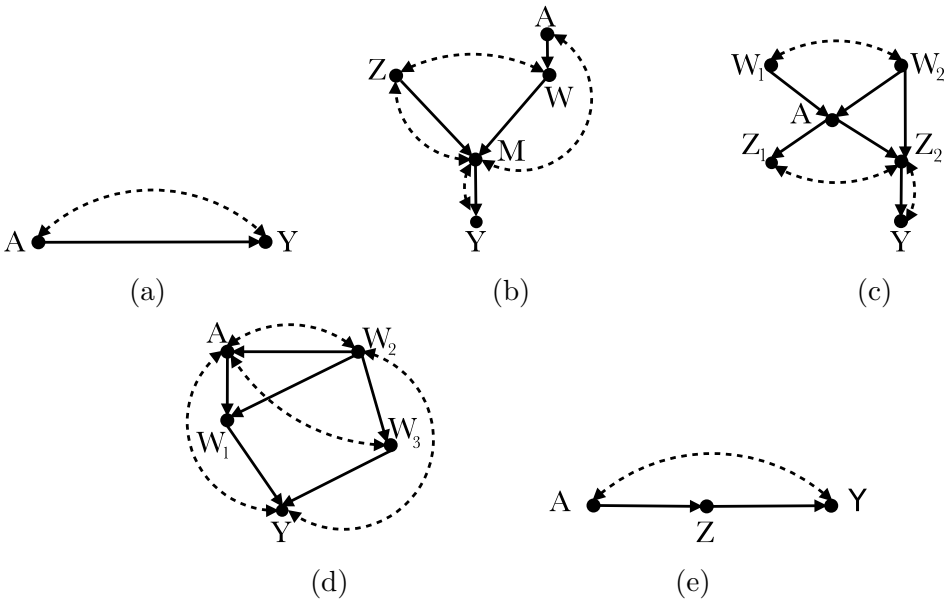


Fig. 3.9 Fig. 3.9a presents the "bow" graph, Fig. 3.9b illustrates the structure of a c-tree, Fig. 3.9c shows a semi-Markovian model where $\mathbb{P}[y_a]$ is observable, Fig. 3.9d presents a semi-Markovian model where $\mathbb{P}[y_a]$ is identifiable and Fig. 3.9e illustrates a simple example of the front-door criterion.

graphical criteria can be generalized to cases where the sensitive attribute is not connected to any of its children through a confounding path. In such cases, c-component factorization can be used. A c-component is a set of vertices in the graph such that every pair of vertices are connected by a confounding edge. The idea of c-component factorization is to decompose the identification problem into smaller sub-problems, that is, a disjoint set of c-components in order to calculate $\mathbb{P}[y_a]$. For example, in Fig. 3.9c, there are three c-components: $\{\{W1, W2\}, \{A\}, \{Z1, Z2, Y\}\}$. Hence, as long as no confounding path connects

A to any of its direct children, $\mathbb{P}[y_a]$ is identifiable. C-component factorization is used in the ID algorithm [212], which is proven complete for causal effect identification.

In case there is an unobservable confounding between the sensitive attribute A and the outcome Y, all the above criteria will fail. However, $\mathbb{P}[y_a]$ can still be identifiable using the front-door criterion. This criterion is satisfied in Fig. 3.9e and consists in having a mediator variable Z such that:

- there are no backdoor paths from A to Z,
- all backdoor paths from Z to Y are blocked by A.

A backdoor path from A to Z is any path starting at A with a backward edge $\leftarrow$ into A (e.g., $A \leftarrow \dots Z$). If such criterion is satisfied, $\mathbb{P}[y_a]$ can be computed as follows:

$$\begin{aligned}\mathbb{P}[y_a] &= \sum_Z \mathbb{P}[y|do(z)]\ \mathbb{P}[z|do(a)] \\ &= \sum_Z \mathbb{P}[y|z,a]\ \mathbb{P}[a]\ \mathbb{P}[z|a] \end{aligned} \tag{3.57}$$

Shpitser and Pearl [210] proved that all the unidentifiable cases of the causal effect $\mathbb{P}[y_a]$ boil down to a general graphical structure called the hedge criterion. Based on this criterion, they designed a complete identifiability algorithm called ID, which outputs the expression of $\mathbb{P}[y_a]$ if it is identifiable or the reason for the unidentifiability otherwise.

The simplest graph in which the causal effect between A and Y is not identifiable is the "bow" graph (Fig. 3.9a). This simple unidentifiability criterion can be generalized to a more complex graph called a c-tree. A c-tree is a graph that is at the same time a tree[32] and a c-component. Fig. 3.9b shows an example of a c-tree. If the causal graph is a c-tree rooted in the outcome variable Y, $\mathbb{P}[y_a]$ is unidentifiable [212].

- **Identifiability of Counterfactuals.** Most the causality-based fairness notions in the disparate treatment framework (NDE (Eq. (3.35)), path-specific effect (Eq. (3.37)), counterfactual effects (Eq. (3.43)), etc.) are defined in terms of counterfactual quantities. Hence, the applicability of those notions depends

[32]Notice that the direction of the arrows between nodes is reversed compared to the usual tree structure.

heavily on the identifiability of the counterfactuals composing them. In Markovian, as well as semi-Markovian models, if all parameters of the causal model are known (including $\mathbb{P}[u]$), any counterfactual is identifiable and can be computed using the three steps abduction, action, and prediction (Theorem 7.1.7 in [181]).

Let $P_* = \{P_x | X \subseteq V,\ x \textit{ a value assignment of } X\}$ be the set of all interventional distributions in a given causal model. While the identifiability of interventional probabilities $\mathbb{P}[y_a]$ is characterized based on observational probabilities $\mathbb{P}[v]$, the identifiability of counterfactuals is characterized in terms of interventional probabilities P_*. Then, combining the results of the identifiability of counterfactuals with the criteria of the identifiability of causal effect (intervention), a counterfactual can, in turn, be identified using observational probabilities $\mathbb{P}[v]$.

Given a causal graph $\mathcal{G}$ of a Markovian model and a counterfactual expression $\gamma = v_x|e$ with e some arbitrary set of evidence, identifying and computing $\mathbb{P}[\gamma]$ requires constructing a counterfactual graph which combines parallel worlds. Every world is represented by a model M_x corresponding to each subscript in the counterfactual expression. For example, given the causal graph in Fig. 3.4 and the counterfactual expression $y_{a_1}|a_0$, the resulting counterfactual graph is shown in Fig. 3.10d. The counterfactual graph should be "reduced" by merging together vertices that share the same causal mechanism (**make-cg** algorithm in [212] automates this procedure). The resulting counterfactual graph can be considered a typical causal graph for a larger causal model. Consequently, all the graphical criteria listed in the identifiability of causal effects above apply to the counterfactual graph to identify counterfactual quantities, in particular, the c-component factorization of the counterfactual graph [211]. ID* and IDC* algorithms [212] automate the identifiability and computation of counterfactuals based on all the above criteria. Note that ID* and IDC* output expressions in terms of interventional probabilities P_*. Then, the ID algorithm is used to express those interventional probabilities in terms of observational probabilities. The simplest unidentifiable counterfactual quantity is $\mathbb{P}[y_{a'}, y'_a]$, which is called the probability of necessity and sufficiency. The corresponding counterfactual graph is the W-graph with the same structure as Fig. 3.10a. This simple criterion can be generalized to the zig-zag graph (Fig. 3.10b) where the counterfactual $\mathbb{P}[y_a, w_1, w_2, z]$ is not identifiable. 9Pearl [181] proves two results about the identifiability of counterfactuals. First, for linear causal models (i.e., the functions $\mathbf{F}$ are linear), any counterfactual is experimentally (using P_*) identifiable whenever the model parameters are identified. Second, in linear causal models, if some

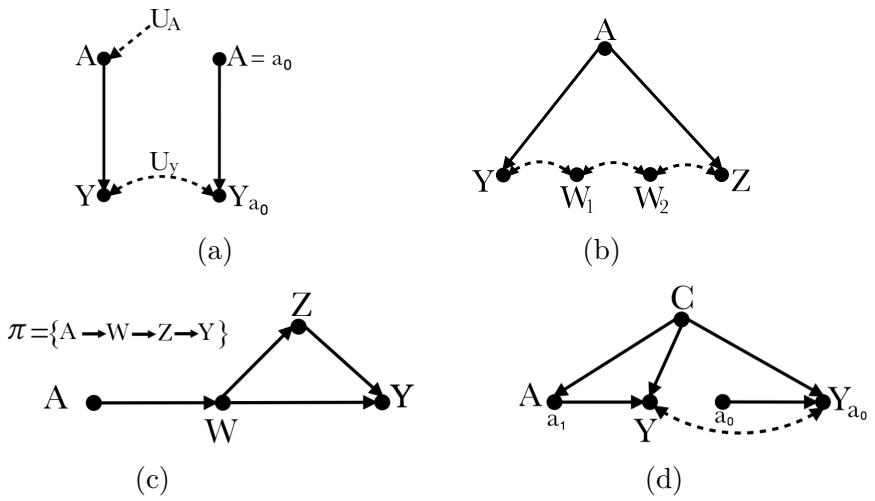


Fig. 3.10 Causal graphs.

of the model parameters are unknown, any counterfactual of the form $\mathbb{E}[Y_a|e]$ where e is some arbitrary set of evidence, is identifiable provided that $\mathbb{E}[y_a]$ is identifiable. Finally, there is no single necessary and sufficient criterion for the identifiability of counterfactuals in semi-Markovian models [20]. Appendix A.1.4 provides an example of the computation of a counterfactual probability of the teacher firing example of Fig. 3.4.

- **Identifiability of Direct and Indirect Effects.** In Markovian models, the average natural direct effect NDE and the average natural indirect effect NIE are always identifiable (from observational data) and can be computed as follows [180]:

$$\text{NDE}_{a_1,a_0}(Y) = \sum_s \sum_z \Big(\mathbb{E}[Y|a_0,z] - \mathbb{E}[Y|a_1,z]\Big)\mathbb{P}[z|a_1,s]\mathbb{P}[s] \tag{3.58}$$

$$\text{NIE}_{a_1,a_0}(Y) = \sum_s \sum_z \mathbb{E}[Y|a_1,z] \Big(\mathbb{P}[z|a_0,s] - \mathbb{P}[z|a_1,s]\Big)\ \mathbb{P}[s] \tag{3.59}$$

where Z is a set of mediator variables and S is any set of variables satisfying the back-door criterion between the sensitive variable A and the mediator variables Z, that is, *(i)* no variable in S is a descendant of A and *(ii)* S blocks all back-door paths between A and Z. A simpler formulation can be used in case there is no confounding between A and Z, where the need for S is dropped altogether:

$$\text{NDE}_{a_1,a_0}(Y) = \sum_z \Big(\mathbb{E}[Y|a_0,z] - \mathbb{E}[Y|a_1,z]\Big)\ \mathbb{P}[z|a_1] \tag{3.60}$$

$$\text{NIE}_{a_1,a_0}(Y) = \sum_z \mathbb{E}[Y|a_1, z] \left(\mathbb{P}[z|a_0] - \mathbb{P}[z|a_1]\right) \tag{3.61}$$

In semi-Markovian models, NDE and NIE are not generally identifiable, even if we have the luxury to perform any experiment using RCT, because of the nested counterfactuals $\mathbb{P}[Y_{a_0}, Z_{a_1}]$ and $\mathbb{P}[Y_{a_1}, Z_{a_0}]$ in Eq. (3.35) and Eq. (3.36), respectively. Nevertheless, these quantities are identifiable *from experimental data* provided that there is a set of variables W which are parents of the outcome variable Y but non-descendants of A and Z such that $Y_{a_1,z} \perp Z_{a_1}|W$ (reads: $Y_{a_1,z}$ and Z_{a_1} are independent conditional on W).

This condition can be easily checked from the causal graph as follows: W d-separates Y and Z in the graph formed by deleting all arrows emanating from A and Z, denoted simply as $(Y \perp Z|W)_{G_{\underline{AZ}}}$.

If such a graphical condition is satisfied, NDE and NIE can be computed from experimental quantities as follows:

$$\text{NDE}_{a_1,a_0}(Y) = \sum_{z,w} \left(\mathbb{E}[Y_{a_0,z}|w] - \mathbb{E}[Y_{a_1,z}|w]\right) \mathbb{P}[Z_{a_1} = z|w]\, \mathbb{P}[w] \tag{3.62}$$

$$\text{NIE}_{a_1,a_0}(Y) = \sum_{z,w} \mathbb{E}[Y_{a_1,z}|w] \left(\mathbb{P}[Z_{a_0} = z|w] - \mathbb{P}[Z_{a_1} = z|w]\right) \mathbb{P}[w] \tag{3.63}$$

- **Identifiability of Path-Specific Effects.** The identifiability of $\text{PSE}_\pi(a_1, a_0)$ in Markovian models depends on whether $\mathbb{P}[y|do(a_1|_\pi, a_0|_{\bar{\pi}})]$ is identifiable. Avin et al. [20] gave a single necessary and sufficient criterion for the identifiability of $\mathbb{P}[y|do(a_1|_\pi, a_0|_{\bar{\pi}})]$ in Markovian models called recanting witness criterion. This criterion holds when there is a vertex W along the causal path π that is connected to Y through another causal path not in π. For instance, Fig. 3.10c satisfies the recanting witness criterion when $\pi = A \to W \to Z \to Y$ with W as witness. The corresponding graph structure is called "kite" graph. When this criterion is satisfied, $\mathbb{P}[y|do(a_1|_\pi, a_0|_{\bar{\pi}})]$ is not identifiable, and consequently, $\text{PSE}_\pi(a_1, a_0)$ is not identifiable. Shpitser [209] generalizes this criterion to semi-Markovian models known as recanting district criterion.

Estimation Based on Full Knowledge of the Causal Model Parameters. The main reason behind the unidentifiability of causal quantities (causal effects, counterfactuals, etc.) is the presence of unobservable variables, namely, hidden latent variables. Some causality-based fairness notions, such as counterfactual fairness 3.39, can be

assessed in the presence of such unobservable latent variables. However, The only requirement is knowledge of the causal model structure (skeleton). Based on the causal model, the latent/background variables are estimated using observable data. Then, the predictor is trained using both observable (non-descendants of the sensitive attributes) as well as the estimated latent variables. Such predictors tend to be fairer than typical predictors (trained using only observable variables) since they consider hidden bias captured by latent variables. Given the full causal model, counterfactual fairness can be assessed by generating, for every observable data sample, a counterfactual data sample by simply changing the sensitive attribute value (e.g., turn male into female) and then using the three-step process (abduction, action, prediction) (Theorem 7.1.7 in [181]) to compute the outcome. The predictor is considered fair if the predicted outcome distributions of both groups (protected and unprotected) are similar.

Potential Outcome Estimation Techniques. Causal inference in the potential outcome framework focuses on estimating the causal effect of a treatment variable A (e.g., the sensitive attribute) on an effect variable Y (e.g., the decision outcome). As mentioned in Section 2.4.2, three assumptions are typically made for causal effect estimation: SUTVA, ignorability, and positivity. In line with the potential outcome framework literature, we focus on causal inference approaches that rely on the three assumptions [260, 100], namely, re-weighting [110], matching [167], and stratification [110].

- **Re-Weighting**: One of the main challenges of causal inference is that the sensitive attribute is not assigned at random in the observational data. That is, the distribution in the observed dataset does not reflect the true distribution. Sample re-weighting methods try to overcome this discrepancy by assigning appropriate weights to sample units in the observational data. The aim is to generate a pseudo-population on which the distributions of the protected (e.g., female) and unprotected (e.g., male) groups are the same as in the original total population. This is achieved by defining a balancing score $b(x)$ satisfying $A \perp x \mid b(x)$. The most common approach to balancing score is based on the propensity score [199], which is defined as the conditional probability of the sensitive attribute given background variables:

$$e(x) = \mathbb{P}[A = 1 \mid X = x] \tag{3.64}$$

Propensity scores can be used to equate groups based on covariates X. In inverse propensity weighting (IPW), the balancing score $b(x)$ for each sample is defined as:

$$b(x) = \frac{A}{e(x)} + \frac{1-A}{1-e(x)} \tag{3.65}$$

where $A = 1$ corresponds to the protected group and $A = 0$ corresponds to the unprotected group. The IPW estimator of ATE (Eq. (3.25)) is defined as:

$$\hat{ATE}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\hat{e}(x_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)\,Y_i}{1-\hat{e}(x_i)} \tag{3.66}$$

Notice that the estimation of ATE is based only on the observable outcome (no counterfactual outcomes) and on the estimation of $e(x_i)$, that is, $\hat{e}(x_i)$. Refer to Appendix A.1.2 for an example of how ATE is computed using the propensity score estimation technique.

When the propensity score is estimated, the normalized version of $\hat{\text{ATE}}_{IPW}$ is preferred:

$$\hat{\text{ATE}}_{IPW}^{norm} = \left[\sum_{i=1}^{n}\frac{A_i Y_i}{\hat{e}(x_i)}\Big/\sum_{i=1}^{n}\frac{A_i}{\hat{e}(x_i)} - \sum_{i=1}^{n}\frac{(1-A_i)\,Y_i}{1-\hat{e}(x_i)}\Big/\sum_{i=1}^{n}\frac{(1-A_i)}{1-\hat{e}(x_i)}\right] \tag{3.67}$$

The correctness of the IPW estimation relies heavily on the quality of the propensity score estimation ($\hat{e}(X)$). A slight misspecification of propensity scores may lead to a significant discrepancy in the ATE estimation. In such cases, doubly robust (DR) estimation is recommended [90]. DR combines IPW estimation with outcome regression so that the estimation remains valid even if one of the approaches is incorrect (but not both). Another limitation of IPW can be observed if the propensity score $e(X) = \mathbb{P}[A \mid X]$ for some value of X is small. In such a case, the estimation may suffer instability. To address this issue, trimming [141] is typically used. Trimming consists of removing the samples with a propensity score below a certain threshold.

- **Matching**: Matching techniques [167] focus on estimating the counterfactual outcome of units. The idea is to estimate the counterfactual outcomes $Y_i^1|A = 0$ and $Y_i^0|A = 1$ based on the matched neighbors of unit i in the opposite group. For example, given an observed female candidate f_k, estimating the counterfactual outcome (hiring decision) had she been a male is based on the units in the male group that are the most comparable to f_k. Hence, the first and main issue is to

define a similarity metric between two given units (e.g., x_i and x_j). The most common approach is to rely on the propensity scores of units:

$$D(i,j) = |e(x_i) - e(x_j)| \tag{3.68}$$

and its logit version:

$$D(i,j) = |logit(e(x_i)) - logit(e(x_j))| \tag{3.69}$$

which is preferred as it has been proven to reduce the bias [225]. The second issue is the matching algorithm: how many neighbors to consider and how these neighbors are weighted to obtain the estimation.

- **Stratification**: Stratification [110] uses the same principle underlying the identifiability approach: adjusting on confounders. The aim is to split the observed data into consistent groups so that the units in the same group can be considered sampled from data under RCT. The two ingredients of stratification are the splitting of groups and then the combination of the created groups. The stratification estimator of ATE can be defined generically as:

$$\hat{\text{ATE}}^{strat} = \sum_{k=1}^{K} m(k)[\overline{Y}_1(k) - \overline{Y}_0(k)] \tag{3.70}$$

where K is the number of stratification groups, $m(k)$ is the portion of units in group k to the total number of units N, $\overline{Y}_1(k)$ and $\overline{Y}_(k)$ are the CATE (Eq. (3.32)) for groups $A = 1$ and $A = 0$, respectively. $\hat{\text{ATE}}^{strat}$ expression has the same structure as the back-door formula (Eq. (3.56)).

If all variables needed for the stratification are observed and the available data is infinitely large, ATE^{strat} can lead to a consistent and unbiased estimator of ATE. However, in typical datasets, stratification may result in strata with few or no units. Consequently, some CATE estimates cannot be calculated using the available data. Propensity score can be used to address this data sparseness problem. The main idea is that "strata with identical propensity scores can be combined into more coarse strata" [167]. In other words, propensity score can be considered a single stratifying variable that usually results in larger strata. The same idea is used in the SCM framework to address the sparseness of data when computing identifiable expressions.

Other estimation methods in the potential outcome framework include tree-based methods [19], representation learning methods [29], and meta-learning methods [134].

3.4.4 Suitability and Applicability of Causality-based Fairness Notions

Suitability. In this section, we try to systemize the selection process by considering the subtleties of each causality-based fairness notion and defining 6 criteria that correspond to characteristics of the real-world scenario at hand. We check whether each criterion holds in the scenario at hand or not. Then, these answers will be used to recommend the most suitable causality-based fairness notion. The criteria are listed and briefly described as follows.

- **Presence of confounding**: A variable that is a common cause of two or more other variables.
- **Presence of explaining variable**: A variable correlated with the sensitive attribute such that any discrimination explained using that variable is considered legitimate and acceptable.
- **Likelihood of intersectionality**: A specific type of bias due to the combination of sensitive attributes. An individual might not be discriminated against based on race or gender only, but she might be discriminated against because of a combination of both.
- **Likelihood of masking**: A form of intentional discrimination that allows decision-makers with prejudicial views to discriminate against individuals or groups while masking their intentions.
- **Latent variables are known**: Latent (background) variables are not observable. However, they are identified in some scenarios, and their relationships with observable variables are known.
- **Ground truth or reliable outcome**: the label in the training data can or cannot be reliable. In several scenarios, the outcome is inferred by humans (job hiring, college admission, etc.) and hence can encode bias. The most reliable outcome is when the ground truth is available.

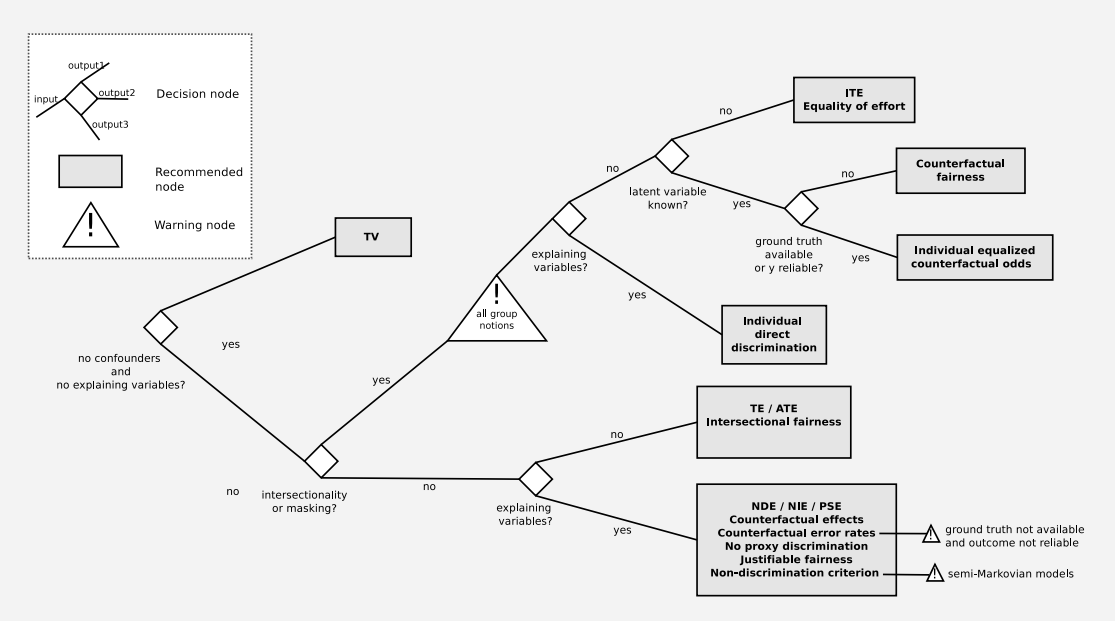


Fig. 3.11 Guideline for causality-based fairness notions selection.

The diagram in Fig. 3.11 can be used as a guideline to select an appropriate causality-based fairness notion given a real-world scenario.

Confounding variables result in backdoor paths between the sensitive attribute (A) and the outcome (Y). For example, the path $A \leftarrow C \rightarrow Y$ in Fig. 3.8 is a backdoor path. Backdoor paths are not causal but contribute to the association between the A and Y. Therefore, they are the reason why it is said that "correlation is different than causation". In the absence of confounding, the total causal effect (TE and ATE) coincides with the difference in conditional probabilities TV$= \mathbb{P}[y|a_1] - \mathbb{P}[y|a_0]$, which corresponds to statistical parity. On the other hand, if there are no explaining variables in the model representation of the world, both direct and indirect causal paths are discriminatory[33]. Consequently, assessing unfairness/bias due to the sensitive attribute does not require considering the different causal paths (direct, indirect, and path-specific) separately. In such cases (without confounding and explaining variables), causal inference is unnecessary to assess fairness appropriately.

Any unintentional bias can also be "orchestrated" intentionally by decision-makers with prejudicial views. To appropriately assess the bias in the presence of such masking attempts, it is recommended to avoid group-based notions as they can more easily be gamed by prejudicial decision-makers. Intersectionality is similar to masking as both lead to discrimination, which is difficult to detect at the group level and hence requires

[33]Indirect causal paths all go through proxy variables.

more fine-grained measures. Therefore, individual causality-based fairness notions are recommended in the presence of one of those criteria. For individual notions, in the presence of explaining variables, it is recommended to use individual direct discrimination (Eq. (3.48)) as it is the only individual notion listed in Section 3.4.2 that distinguishes direct from indirect discrimination. Counterfactual fairness (Eq. (3.39)) and individual equalized counterfactual odds (Eq. (3.54)) are recommended to be used in case the latent variables are known. If the ground truth is not available or the outcome Y is not reliable, individual equalized counterfactual odds is not recommended.

For the group causality-based fairness notions, if there are no explaining variables, there is no need to consider the different causal paths and hence TE, ATE, or interventional fairness can be safely used. In the presence of explaining variables, the remaining causality-based fairness notions are appropriate to use with two exceptions. First, the non-discrimination criterion is misleading if the causal model is semi-Markovian because the variable A can remain dependent even after conditioning on all observable variables because of the hidden confounders. Second, as counterfactual error rates (Eq. (3.43)) are expressed in terms of the true outcome Y, they are not recommended in case the ground truth is not available and the true outcome is not reliable.

Finally, note that ETT, ATT, and ATC are not generally used in fairness scenarios because, typically, the bias can be observed in both directions: when considering a disadvantaged group/individual as advantaged or the opposite. ETT is relevant when studying the effect of a treatment medicine on patients. For example, if a patient agrees to take the medicine and it turns out to be painful, she may be wondering about the chances of recovery if she did not take the treatment or if she took it with a lower dose. In this case, the opposite direction (the effect of treating an individual in the control group) is irrelevant.

In his book, *The Book of Why* [185], Pearl describes a causation ladder with three rungs: statistical observations (seeing), intervention (doing), and counterfactual (imagining). All causality-based fairness notions defined in Pearl's SCM framework (all notions in Section 3.4.2 except ATE, ATT, ATC, ITE, and equality of effort) are placed in the causation ladder, which will help us address the problem of their applicability in real-scenarios. The causation ladder is structured so that a quantity at a certain rung can be identified in terms of quantities at the rung just below it. Consequently, the higher the rung, the more challenging the problem of identifiability is, and hence, the less applicable a fairness notion is defined at that rung.

Fig. 3.12 shows the causation ladder and indicates at which rung every causality-based fairness notion stands. TV, the only non-causal fairness notion covered in this

study, is at rung 1. It is always applicable, provided a set of observations (dataset) is available. No unresolved and non-discrimination criteria are placed midway between rungs 1 and 2 as applicable, provided the causal graph is available along the dataset. The non-discrimination criterion, however, requires the Markov property to be applicable because causal dependence through unobservable paths cannot be blocked. It also has an exponential complexity since it considers all combinations of values of the parent variables of the outcome Y. A relaxation is described by the authors [274], but the notion remains computationally intractable.

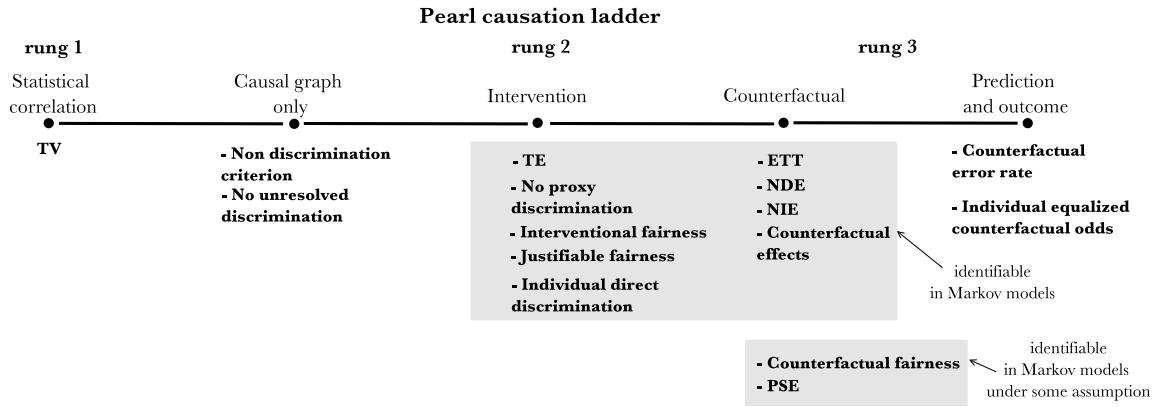


Fig. 3.12 Classification of causality-based fairness notions according to Pearl causation ladder [185].

Applicability. Fairness notions at rung 2 (TE, No-proxy discrimination, interventional and justifiable fairness, and individual direct discrimination) are applicable in any scenario where either experiments (RCT) are possible or hypothetical interventions are identifiable. As mentioned in Section 3.4.3, any intervention probability is identifiable from observational data in Markovian models. Hence, these fairness notions are always applicable to Markovian models. In semi-Markovian models, the applicability of these rung 2 notions depends on the identifiability of the intervention terms used in their respective definitions. For instance, for individual direct discrimination, the term in question is $\mathrm{CE}(q_k, q'_k)$ in Eq. (3.48).

The bulk of causality-based fairness notions are defined in terms of counterfactual quantities and hence are placed in rung 3 of the causation ladder. In Fig. 3.12, the counterfactual notions are ranked from top to bottom according to their degree of applicability. For instance, counterfactual effects are placed on top of counterfactual fairness to indicate that the former is applicable in more scenarios than the latter. In Markovian models, the top 4 notions (ETT, NDE, NIE, and counterfactual effects)

are always identifiable and hence applicable. That is, specific formulas are already available to compute each counterfactual term used in their definitions.

In Markovian models, the identifiability of counterfactual fairness depends on the identifiability of the term $\mathbb{P}[y_{a_1}|X = x, A = a_0]$ which is only identifiable if X does not contain any variable which is at the same time descendant of A and ancestor of Y, that is, $X \cap B = \emptyset$ where $B = An(Y) \cap De(A)$[254]. PSE is applicable provided that the model is Markovian and the recanting witness criterion is not satisfied. In semi-Markovian models, unless all model parameters are known (including $\mathbb{P}[u]$)[34], the identifiability of rung 3 fairness notions depends on the identifiability of counterfactuals, which rarely hold in practice.

Finally, counterfactual error rate and individual equalized counterfactual odds are special cases of rung 3 fairness notions as they are the only notions that condition on the true outcome Y to assess the fairness of the prediction $\hat{Y}$ (Eqs. (3.44),(3.45), (3.46), and (3.54)). Such conditioning has an important implication on identifiability since Y is a collider, and conditioning on a collider creates a dependence between the previous variables [181]. This leads to unobservable confounding between the causes of Y. Hence, even if the causal model is Markovian, applying both notions turns it into a semi-Markovian model. Zhang and Bareinboim [267] define an identifiability criterion for counterfactual error rate in Markovian models called the explanation criterion.

3.4.5 Conclusion

Notions of fairness inconsistent with the causal relationships in the data can lead to misleading conclusions about bias and discrimination of the outcomes. In particular, using causal reasoning to tackle the fairness problem in ML has at least three advantages. First, it appropriately measures discrimination in the presence of statistical anomalies (e.g., Simpson's paradox). Second, it provides a natural interpretation of causal relationships between variables supporting discrimination claims. This is particularly important in the disparate treatment legal framework. Third, it makes it possible to break down the dependence between the sensitive attribute and the outcome into different paths (direct, indirect, etc.), which allows us to assess fairness more accurately in the presence of acceptable and unacceptable discrimination.

Most of the causality-based notions of fairness examined in this study rely on the availability of the causal graph. The issue of generating causal graphs consistent with the observed data is a known problem in the causal inference literature. Studying it in

[34] In that case, it is possible to use the three steps abduction, action, and prediction [181].

the specific context of ML fairness is a relevant direction for future work. We tried tackling this problem in our studies in Chapter 5.

3.5 Conclusion

In this chapter, we have thoroughly examined the applicability of both statistical and causality-based fairness notions in real-life scenarios.

First, we explored statistical fairness notions, such as statistical parity and equal opportunity, and assessed their strengths and limitations when applied to real-world data. While these notions provide clear and measurable criteria for fairness, we found that their applicability can be context-dependent and sometimes challenging due to the complexity and variability of real-life situations.

Next, we delved into causality-based fairness notions, which aim to address the root causes of biased outcomes by leveraging causal inference techniques. These notions, such as counterfactual fairness, offer a deeper understanding of the mechanisms driving unfairness. Our analysis demonstrated that causality-based fairness notions are particularly suitable in scenarios where understanding and mitigating the underlying causes of bias is crucial. However, their application requires comprehensive data and sophisticated modeling, which can be resource-intensive.

Overall, our investigation reveals that both statistical and causality-based fairness notions have their respective advantages and limitations. The choice of which notion to apply depends heavily on the specific context, available data, and the goals of the fairness intervention.

With a solid understanding of how fairness notions can be applied, we now shift our focus to another ethical AI principle: privacy and how it interacts with fairness. Specifically, in the next chapter, we present our study on the impact of privacy on fairness.

Chapter 4

Impact of Privacy on Fairness

4.1 Introduction

This chapter investigates the intricate relationship between privacy and fairness in ML. As the adoption of ML models grows, so do concerns about their fairness and the privacy of the data they utilize. Understanding how privacy-preserving techniques impact fairness is crucial for developing ethical and effective AI systems.

The tension between fairness and DP is attracting increasing attention. For instance, some research works state that DP and fairness are at odds [85, 22, 5, 45, 191]. Conversely, in other lines of research, DP and fairness results align [148, 151, 42, 202]. However, the underlying reasons for this tension remain inadequately explored. *Therefore, a clear understanding of the relationship between privacy and fairness is highly needed.* This chapter presents three of our research works as a step in that direction.

Before delving into the three research studies, we present the generic framework used to assess the impact of privacy on LDP, along with the notation we use in this chapter.

Framework. Fig. 4.1 depicts the framework used in this chapter. Although the framework is consistently applied across all three studies, each study varies in terms of the specific attributes to which the obfuscation mechanism is applied and how it is implemented. We assume a given decision task, such as deciding whether to release a convict on parole or admit an applicant to a college program. We assume that we dispose of a set of data $S = (\mathbf{A}, X, Y)_{train} \bigcup (\mathbf{A}, X, Y)_{test}$ for building an ML model to help with the task, and for evaluating it. Specifically, $(\mathbf{A}, X, Y)_{train}$ is used for training the model, and $(\mathbf{A}, X, Y)_{test}$ to assess the fairness and utility of its predictions.

As shown in Fig. 4.1, in order to measure the impact of the LDP mechanism $\mathcal{L}$, we

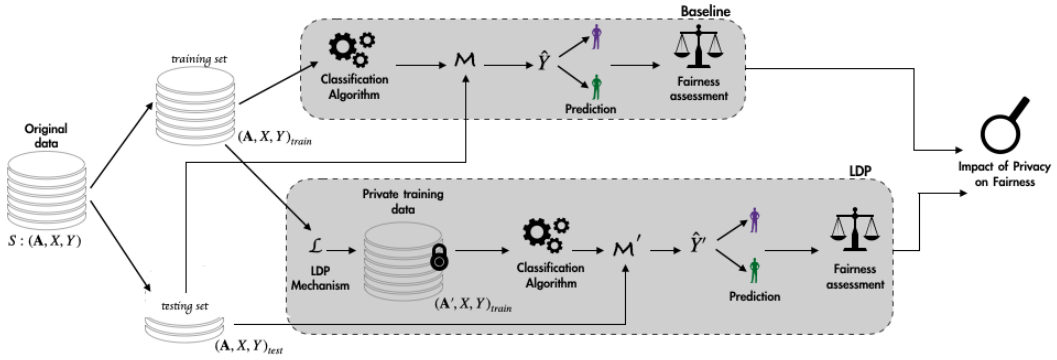


Fig. 4.1 Our framework to assess the impact of LDP on the fairness of an ML model.

train two models. The baseline model $\mathcal{M}$ (upper shaded box) is trained on the original data $(\mathbf{A}, X, Y)_{train}$, and we call its prediction $\hat{Y}$. Then, we obfuscate the training set by applying $\mathcal{L}$ to the $\mathbf{A}$ component of each sample in $(\mathbf{A}, X, Y)_{train}$. We recall that $\mathbf{A}$ designates the set of sensitive attributes in the data. The resulting data set $(\mathbf{A}', X, Y)_{train}$ is used to train (with the same classification algorithm and the same hyper-parameters) a second model $\mathcal{M}'$, whose prediction is called $\hat{Y}'$ (lower shaded box).

The difference between $\hat{Y}'$ and $\hat{Y}$ on the original testing data quantifies the impact of LDP on the fairness of the model. *It is important to emphasize that, in our framework, the individual predictions, both for $\mathcal{M}$ and $\mathcal{M}'$, are obtained by applying the models to the original testing data* $(\mathbf{A}, X, Y)_{test}$. Namely, in testing phase, $\hat{Y} = \mathcal{M}(\mathbf{A}, X)$ and $\hat{Y}' = \mathcal{M}(\mathbf{A}, X)$ (instead of $\hat{Y}' = \mathcal{M}(\mathbf{A}', X)$). This is because we argue that fairness must be evaluated on the true data. Indeed, even if a model was trained on obfuscated data, it is likely to receive the true data as input at the moment of its deployment. And in any case, the presence of proxies may reveal the true value of the sensitive variable anyway.

This chapter is organized as follows. First, we introduce our empirical study on the impact of collecting multidimensional data under ε-LDP guarantees on fairness and utility of seven state-of-the-art LDP protocols under a homogeneous encoding when training ML binary classifier (Section 4.2). Section 4.3 extends our empirical investigation to the impact of training a model with multiple sensitive attributes, obfuscated under LDP guarantees, using two variants (independent and combined) of the widely recognized k-ary randomized response mechanism. Finally, we shift our focus to a theoretical study of the interplay between privacy and fairness (Section 4.4). More specifically, we perform a quantitative study of how the fairness of the decisions

made by the ML model changes under LDP for different levels of privacy and data distributions. This study helps to identify conditions under which privacy measures can enhance or undermine fairness, offering a deeper understanding of the inherent trade-off between privacy and fairness.

4.2 Fairness Under Multidimensional Local Differential Privacy: Empirical Study 1

This study provides a comprehensive empirical analysis of how collecting multidimensional data under ε-LDP guarantees affects the fairness and utility in ML binary classification tasks. For fairness assessment, we employed various group fairness metrics, including disparate impact [25], equal opportunity [101], and overall accuracy [32]. The experimental evaluation encompasses seven state-of-the-art LDP protocols, namely, Generalized Randomized Response (k-RR) [115], Binary Local Hashing (BLH) [26], Optimal Local Hashing (OLH) [245], RAPPOR [79], Optimal Unary Encoding (OUE) [245], Subset Selection (SS) [244, 261], and Thresholding with Histogram Encoding (THE) [245] (All these LDP protocols have been introduced in Section 2.3.2). To broaden the scope of our study, we conducted our experiments on three benchmark datasets: Adult [70], ACSCoverage [70], and LSAC [250].

Moreover, since proxy variables can still introduce unintended biases and thus lead to unfair decisions [118], we consider the setting in which each sensitive attribute (sensitive and proxy attribute) is collected independently under ε-LDP guarantees. In other words, applying this independent setting automatically removes the correlation between the sensitive attributes. To this end, the privacy level ε should be divided among all sensitive attributes to ensure ε-LDP under sequential composition (Proposition 2). Let d_a be the total number of sensitive attributes, the LDP literature for multidimensional data [12, 195, 126, 145] considers a ***uniform*** solution that collects each sensitive attribute under $\frac{\epsilon}{d_a}$-LDP. In this work, we proposed a new ***k-based*** solution that considers the varying domain size k of different sensitive attributes. More precisely, for the i-th sensitive attribute, for $i \in [d_a]$, we allocate $\varepsilon_i = \frac{\varepsilon \cdot k_i}{\sum_{i=1}^{d_a} k_i}$.

In addition, this work explores a more dynamic scenario with a variable d_a. These attributes are selected randomly to underscore the generalization of our findings. Overall, this study challenges the common belief that using DP necessarily leads to worsened fairness in ML [22, 95]. More specifically, we show that training a classifier on LDP-based multidimensional data slightly improved fairness results without significantly

affecting classifier performance. We aim for this work to assist practitioners in collecting multidimensional user data in a privacy-preserving manner. By offering insights into the most suitable LDP protocols and privacy budget-splitting solutions, we hope to guide practitioners in making informed decisions tailored to their specific needs.

Contributions. We conducted a comprehensive empirical analysis on the impact of collecting (or pre-processing[1]) multidimensional data under ε-LDP guarantees on fairness and utility in ML binary classification. Moreover, we compared the impact on fairness and utility of seven state-of-the-art LDP protocols under a homogeneous encoding (see Fig. 4.2) when training ML binary classifiers. Additionally, we proposed a novel ***k-based*** solution for privacy budget allocation, which generally led to a better privacy-utility-fairness trade-off in our experiments. And finally, we open-sourced our codes in the following **GitHub repository** [11].

Outline. The rest of this section is organized as follows. Section 4.2.1 discusses related work. Next, Section 4.2.2 states the problem addressed, the fairness metrics considered, and the proposed ***k-based*** solution. Section 4.2.3 details the experimental setting and main results. Finally, we conclude this work indicating future perspectives in Section 4.2.4.

4.2.1 Related Work

Bagdasaryan, Poursaeed, and Shmatikov [22] studied the impact of training ε-DP deep learning (a.k.a. *gradient perturbation*) models on underrepresented groups. While maintaining the same hyperparameters as the non-private baseline model, the authors observed a more significant drop in accuracy for the underrepresented group. Similarly, Ganev et al. [95] observed disparities for the underrepresented group when generating ε-DP synthetic data for training ML models while keeping the default hyperparameters of differentially private generative models. In contrast, de Oliveira et al. [64] demonstrated that when searching for the best hyperparameters for both non-private and ϵ-DP models, the impact of DP on fairness is negligible. Recent works by Tran, Dinh, and Fioretto [235], by Emelianov and Perrot [78], and by Mangold et al. [161] investigated the reasons and contexts for this impact of central DP on fairness (discussed later in Section 4.4). Furthermore, Ficiu, Lawrence, and Paleyes [87] go beyond by proposing a framework to optimize a three-way objective for central DP ML models, namely, the fairness-privacy-utility trade-off.

[1]While the privacy-preserving mechanisms experimented within this study are specifically designed for a LDP setting, they can also be employed by a trusted server in a centralized DP setting.

In this work, our objective is to investigate the extent to which training an ML classifier on ε-LDP multidimensional data (*a.k.a. input perturbation*) while fixing the same set of hyperparameters has a detrimental effect on the disparities between the *privileged* (group receiving a favorable outcome) and the *unprivileged* (group receiving an unfavorable outcome) groups. Regarding the LDP setting, the work of Mozannar, Ohannessian, and Srebro [169] was the first to propose a fair classifier when obfuscating only the sensitive attribute with ε-LDP in both training and testing sets. More recently, the work of Chen et al. [47] considers a "semi-private" setting in which a small portion of users share their sensitive attribute with no obfuscation, while all other users apply an ε-LDP protocol.

While the two aforementioned research works [169, 47] answer interesting questions by collecting a single sensitive attribute using only the RR (Eq. (2.4)) protocol, we consider multiple sensitive attributes in this work, reflecting real-world data collections more accurately. Additionally, for a more comprehensive examination, we experimented with seven state-of-the-art ε-LDP protocols, as well as several fairness and utility metrics. Lastly, we propose a new privacy budget splitting solution named ***k-based***, which generally leads to better privacy-fairness-utility trade-offs in ML binary classification tasks than the commonly adopted ***uniform*** solution [12, 195, 126, 145].

4.2.2 Problem Setting and Methodology

Table 2.1 summarizes the notation used throughout this study. Note that in this work, we always consider a single sensitive attribute A to obfuscate and assess fairness w.r.t. that attribute. For LDP, we instead consider a set of sensitive attributes $\mathbf{A}$ where one of these attributes, namely A, is used to evaluate fairness.

Group Fairness Metrics Considered. In this work, we focus on group fairness metrics, which, as discussed in Section 3.2.2, evaluate the fairness of ML models across different demographic groups defined by sensitive attributes such as race, gender, and age. Let A be the sensitive attribute, $\hat{Y} \in \{0, 1\}$ be a predictor of a binary true decision $Y \in \{0, 1\}$. The metrics we use to evaluate fairness are listed below [2].

- **Disparate Impact (DI) [25].** DI is defined as the ratio of the proportion of positive predictions ($\hat{Y} = 1$) for the *unprivileged* group ($A = 0$) over the ratio

[2] In this work, rather than assessing fairness using equality-based metrics as originally defined, we evaluate fairness using difference-based metrics. For example, instead of using statistical parity (Eq. (3.1)), we use statistical disparity (SD) (Eq. (4.2)).

of the proportion of positive predictions for the *privileged* group ($A = 1$). The formula for DI is:

$$\text{DI} = \frac{\mathbb{P}[\hat{Y} = 1|A = 0]}{\mathbb{P}[\hat{Y} = 1|A = 1]}. \tag{4.1}$$

Note that a perfect DI value is equal to 1.

- **Statistical Disparity (SD) [3].** Instead of the ratio, SD computes the difference in the proportion of positive predictions for *privileged* and *unprivileged* groups and is defined as:

$$\text{SD} = \mathbb{P}[\hat{Y} = 1|A = 1] - \mathbb{P}[\hat{Y} = 1|A = 0]. \tag{4.2}$$

A perfect SD value is equal to 0.

- **Equal Opportunity Difference (EOD) [101].** EOD measures the difference in the true positive rates (i.e., recall) of the *privileged* and the *unprivileged* groups. Formally, EOD is defined as:

$$\text{EOD} = \mathbb{P}[\hat{Y} = 1|Y = 1, A = 1] - \mathbb{P}[\hat{Y} = 1|Y = 1, A = 0]. \tag{4.3}$$

A perfect EOD value is equal to 0.

- **Overall Accuracy Difference (OAD) [32].** OAD measures the difference in the overall accuracy rates between the *privileged* and the *unprivileged* groups. Formally, OAD is defined as:

$$\text{OAD} = \mathbb{P}[\hat{Y} = Y|A = 1] - \mathbb{P}[\hat{Y} = Y|A = 0]. \tag{4.4}$$

A perfect OAD value is equal to 0.

LDP Protocols Considered. All the LDP protocols presented in Section 2.3.2 are applied in this work, namely k-RR [115], BLH [26, 245], OLH [245], RAPPOR [79], OUE [245], SS [244, 261], and THE [245].

Considering the framework depicted in Fig. 4.1, our primary goal is to examine the impact of training an ML classifier on $S' = (X, \mathbf{A}', Y)$ compared to $S = (X, \mathbf{A}, Y)$ on fairness and utility, using different LDP protocols and privacy budget splitting solutions. Specifically, we focus on the scenario where each sensitive attribute in $\mathbf{A}$ is collected independently under LDP guarantees. In this case, to satisfy ε-LDP following

Proposition 2, the privacy budget ε must be split among the total number of sensitive attributes $d_a = |\mathbf{A}|$. The state-of-the-art [12, 195, 126, 145] solution, named ***uniform***, proposes to split the privacy budget ε evenly among all attributes, allocating $\frac{\varepsilon}{d_a}$ for each attribute. However, as different sensitive attributes have varying domain sizes k_i, for $i \in [d_a]$, we propose a new solution named ***k-based*** that splits the privacy budget ε proportionally to the domain size of the attribute. In other words, for the i-th attribute, we allocate $\varepsilon_i = \frac{\varepsilon \cdot k_i}{\sum_{i=1}^{d_a} k_i}$.

In addition, since each LDP protocol encodes and perturbs user data differently, we propose to compare all LDP protocols under the same encoding when training the ML classifier. More specifically, we used One-Hot Encoding (OHE) and Indicator Vector Encoding (IVE) [2] as all LDP protocols from Section 2.3.2 are designed for categorical data or discrete data with a known domain. For example, let Ω be the reported subset of a user after using SS as LDP protocol. Following IVE, we create a binary vector $\mathbf{z} = [b_1, \ldots, b_k] \in \{0, 1\}^k$ of length k, where the v-th entry is set to 1 if $v \in \Omega$, and 0, otherwise. In other words, $\mathbf{A}'$ represents the subset Ω in a binary format. Fig. 4.2 illustrates the LDP encoding and perturbation at the user side and how to achieve a "homogeneous encoding" for all the seven LDP protocols on the server side. Last, all non-sensitive attributes X are encoded using OHE.

4.2.3 Experimental Evaluation

In this section, we present the experimental setting and the results of our experiments. Our Research questions (RQ) are:

- **RQ1.** Overall, how does pre-processing multidimensional data with ε-LDP affect the fairness and utility of ML binary classifiers with the same hyperparameters used before and after obfuscation?
- **RQ2.** Which privacy budget-splitting solution has a better privacy-utility-fairness trade-off?
- **RQ3.** How do different LDP protocols affect the fairness and utility of an ML binary classifier, and which one is more suitable for the different real-world scenarios applied?

General setting. For all experiments, we consider the following setting:

- **Environment.** All algorithms are implemented in Python 3 with Numpy [240], Numba [138], and Multi-Freq-LDPy [13] libraries, and run on a local machine with

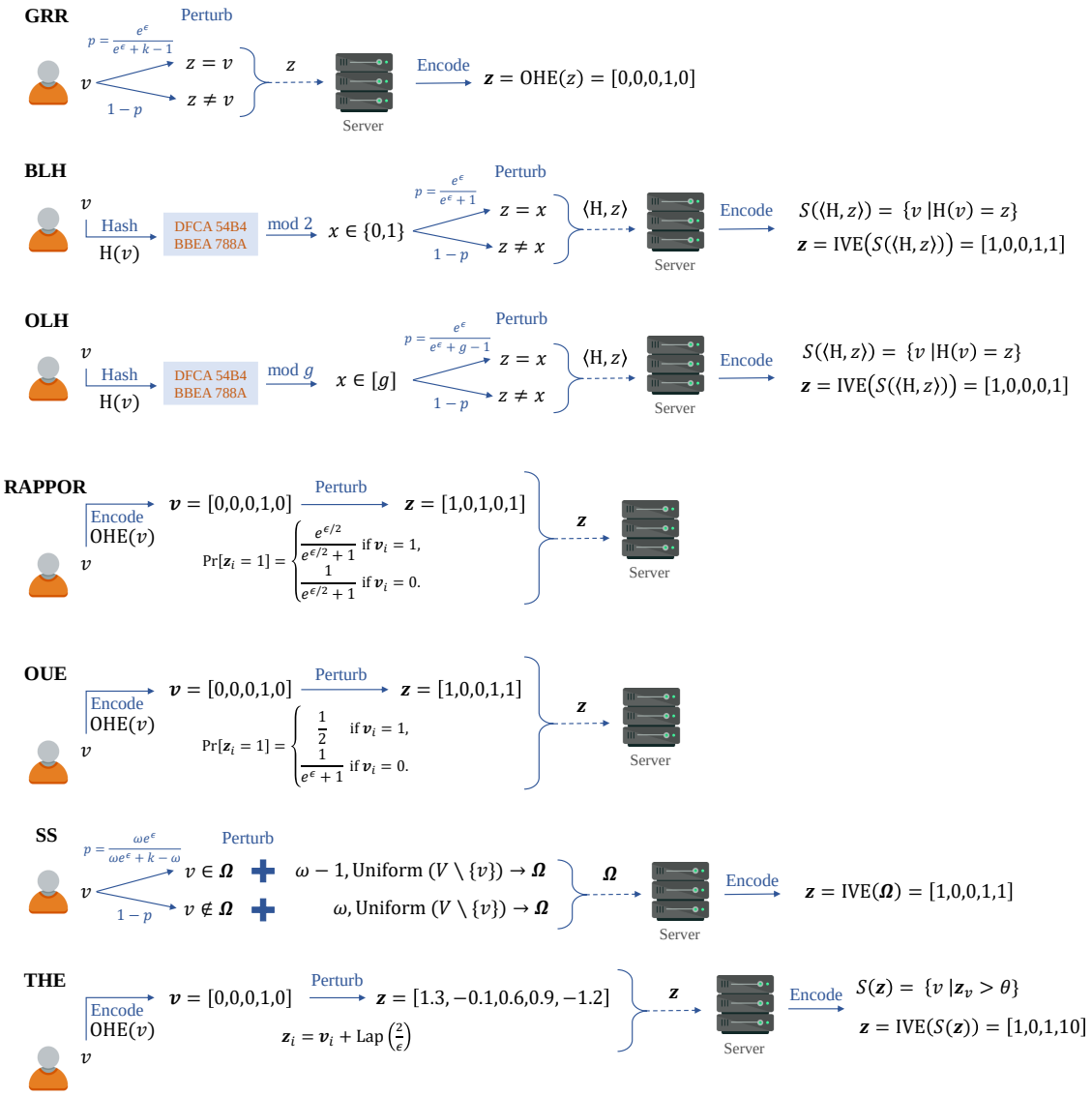


Fig. 4.2 Overview of client-side encoding and perturbation steps for the seven different LDP protocols applied. On the server side, there is also a post-processing step with one-hot encoding (OHE) or indicator vector encoding (IVE), if needed.

2.50GHz Intel Core i9 and 64GB RAM. The codes we develop for all experiments are available in a **GitHub repository** [11].

- **ML classifier.** We used the state-of-the-art[3] LGBM [122] as predictor $\hat{Y}$.
- **Encoding.** We only use discrete and categorical attributes, which are encoded using OHE or IVE (see Fig. 4.2), and the target is binary, i.e., $Y \in \{0, 1\}$.

[3]https://www.kaggle.com/kaggle-survey-2022.

- **Training and Testing Sets.** We randomly select 80% as training set and the remaining 20% as testing set. We apply LDP on the training set only. That is, the samples in the testing set are the original samples (i.e., no LDP).
- **Stability.** Since LDP protocols, train/test splitting, and ML algorithms are randomized, we report average results over 50 runs.

Datasets. Table 4.1 summarizes all datasets used in our experiments. For ease of reproducibility, we use real-world and open datasets.

Table 4.1 Description of the datasets used in the experiments of Section 4.2.

Dataset	n	A	$\mathbf{A}$, *domain size* k	Y
Adult_G	45849	gender	- gender, $k = 2$ - race, $k = 5$ - native country, $k = 41$ - age, $k = 74$ - hours per week, $k = 96$ - education, $k = 16$	income
Adult_R	45849	race	- race, $k = 2$ - gender, $k = 2$ - native country, $k = 41$ - age, $k = 74$ - hours per week, $k = 96$ - education, $k = 16$	income
ACSCoverage	98739	DIS	- DIS, $k = 2$ - AGEP, $k = 50$ - SEX, $k = 2$ - SCHL, $k = 24$ - RAC1P, $k = 2$ - NATIVITY, $k = 2$	PUBCOV
LSAC	20427	race	- race, $k = 2$ - gender, $k = 2$ - family income, $k = 5$ - full time, $k = 2$	pass bar

- **Adult.** This dataset contains information about individuals, including personal attributes such as gender, age, race, marital status, education, and occupation. The goal is to predict their income. We use 26000$ as a threshold to binarize the target variable income of the *reconstructed Adult* dataset [70]. After cleaning,

$n = 45849$ samples are kept. We excluded capital-gain and capital-loss and used the remaining 10 discrete and categorical attributes. We considered $\mathbf{A} = \{$gender, race, native country, age, hours per week, education$\}$ as the set of possible sensitive attributes for LDP obfuscation. We consider two scenarios for the Adult dataset depending on the sensitive attribute to assess fairness. *Gender* is the sensitive attribute for the first scenario, while *race* is considered as the sensitive attribute for the second scenario. We call the Adult dataset differently depending on the sensitive attribute at hand:

  - **Adult$_{\mathbf{G}}$:** With $A =$ gender.
  - **Adult$_{\mathbf{R}}$:** With $A =$ race[4].

- **ACSCoverage[5].** This dataset is retrieved with the `folktables` [70] Python package and the binary target PUBCO designates whether an individual is covered by public health insurance or not. We select 2018 and the "Texas" state, with $n = 98739$ samples. We removed DEAR, DEYE, DREM, and PINCP and used the remaining 15 discrete and categorical attributes. We considered $\mathbf{A} = \{$DIS, AGEP, SEX, SCHL, RAC1P, NATIVITY$\}$ as the set of sensitive attributes for LDP obfuscation and $A =$ DIS (i.e., disability) as the sensitive attribute to assess fairness.

- **LSAC.** This dataset is from the Law School Admissions Council (LSAC) National Bar Passage Study [250], and the binary target pass_bar indicates whether or not a candidate has passed the bar exam. After cleaning, $n = 20427$ samples are kept. We only consider as attributes gender, race, family income, full time, undergrad GPA score (discretized to $\{1.5, 2.0, ..., 4.5\}$), and LSAT score (rounded to the closest integer). The race attribute was binarized to {black, other}. We considered $\mathbf{A} = \{$gender, race, family income, full time$\}$ as the set of sensitive attributes for LDP obfuscation and $A =$ race as the sensitive attribute for fairness assessment.

Evaluated Methods. The methods we use and compare are:

- **(Baseline) NonDP.** This is our baseline with LGBM trained over original data (i.e., $S = (X, \mathbf{A}, Y)$). The non-sensitive X and sensitive $\mathbf{A}$ attributes

[4]We consider race as a binary attribute with the two categories: non-white and white.

[5]The full documentation for the description of all attributes is in https://www.census.gov/programs-surveys/acs/microdata/documentation.html.

are encoded with OHE. We searched for the best hyperparameters using Bayesian optimization [31] through 100 iterations varying: $max\_depth \in [3, 50]$, $n\_estimators \in [50, 2000]$, and $learning\_rate \in (0.01, 0.25)$.

- **LDP Protocols.** For all the four datasets of Table 4.1, we selected the number of sensitive attributes uniformly at random with $2 \leq d_a \leq |\mathbf{A}|$[6], in each of the 50 runs. We then pre-processed the sensitive attributes of the training sets (i.e., $\mathbf{A}' = \mathcal{L}(\mathbf{A})$) using each of the seven LDP protocols from Section 2.3.2 (i.e., k-RR, RAPPOR, OUE, SS, BLH, OLH, and THE). Next, we fixed the optimized hyperparameters found for the NonDP model and trained LGBM over $S' = (X, \mathbf{A}', Y)$ using these hyperparameters. To satisfy ε-LDP (Proposition 2), we split the privacy level following the two solutions described in Section 4.2.2, namely, the state-of-the-art ***uniform*** and our ***k-based*** solutions.

Metrics. We evaluate the performance of LGBM trained over the original data (i.e., NonDP baseline) and LDP-based data on privacy, utility, and fairness:

- **Privacy.** We vary the privacy parameter in the range of: $\varepsilon = \{0.25, 0.5, 1, 2, 4, 8, 10, 20, 50\}$. At $\varepsilon = 0.25$, the ratio of probabilities is bounded by $e^{0.25} \approx 1.3$, giving nearly indistinguishable distributions, whereas at $\varepsilon = 50$ almost no privacy is guaranteed.

- **Utility.** We use accuracy (acc), f1-score (f1), area under the receiver operating characteristic curve (auc), and recall as utility metrics;

- **Fairness.** We use the group fairness metrics presented earlier in this section (i.e., DI, SD, EOD, and OAD).

Main Results.

- **LDP Impact on Fairness.** Fig. 4.3 (Adult_G), Fig. 4.4 (Adult_R), Fig. 4.5 (ACSCoverage), and Fig. 4.6 (LSAC) illustrate the privacy-fairness trade-off for the NonDP baseline and all the seven LDP protocols, considering both the ***uniform*** and our ***k-based*** privacy budget splitting solutions. From these Figs, one can observe a general trend of slight improvement in fairness for all the seven LDP protocols under both the ***uniform*** and the ***k-based*** solutions. For instance, for the DI metric in Fig. 4.3, the NonDP data indicates a value of

[6]Once the total number of sensitive attributes d_a is established, the remaining $d_a - 1$ sensitive attributes are selected uniformly at random.

0.44 showing discrimination against women. Upon applying LDP protocols, DI tends to increase to approximately 0.48 (with $\epsilon \leq 2$), slightly improving fairness. Similarly, SD decreased from 0.37 to $\sim$ 0.34 after applying LDP protocols. Similar trends are observed for other fairness metrics such as EOD, indicating consistent improvements after applying LDP protocols.

The main exception occurred in Fig. 4.5 for the OAD metric, where the gap between the *privileged* and the *unprivileged* groups was accentuated (favoring the *unprivileged* group). Specifically, the OAD is -0.17 with the NonDP baseline. After applying all the LDP protocols using either the ***uniform*** or the ***k-based*** solutions, the gap between the *privileged* and the *unprivileged* groups increased to -0.25. In other words, we initially observed favoritism towards the *unprivileged* group (indicated by a negative value), which increased after applying LDP. It is noteworthy that even this exception contrasts with the findings of [22, 95] in central DP, where the underrepresented group is consistently negatively affected.

Additionally, it is noteworthy that when applying the ***uniform*** privacy budget splitting solution (refer to the left-side in the plots), all fairness metrics demonstrated less robustness to LDP than our ***k-based*** solution. Consequently, they reverted to the NonDP baseline value in low privacy regimes. On the other hand, when applying our ***k-based*** solution (refer to the right-side in the plots), all fairness metrics consistently performed better across all privacy regimes, especially for the Adult dataset, as depicted in Figs 4.3 and 4.4. For the ACSCoverage dataset, not all fairness metrics returned to the NonDP baseline value with our ***k-based*** solution. However, with the ***uniform*** solution, for $\varepsilon \geq 8$, all fairness metrics reverted to the NonDP baseline value. A similar behavior was noticed for the LSAC dataset, where our ***k-based*** solution exhibited more robustness to LDP than the ***uniform*** solution, reverting to the NonDP baseline values only when $\varepsilon \geq 20$ in contrast with $\varepsilon \geq 8$ for the ***uniform*** solution.

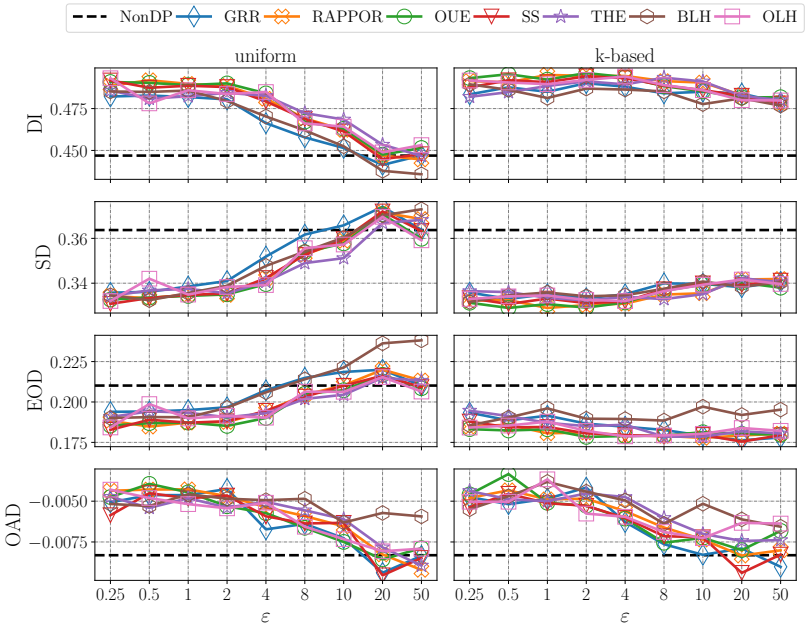


Fig. 4.3 Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ε-LDP protocol, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), on the Adult_{G} [70] dataset.

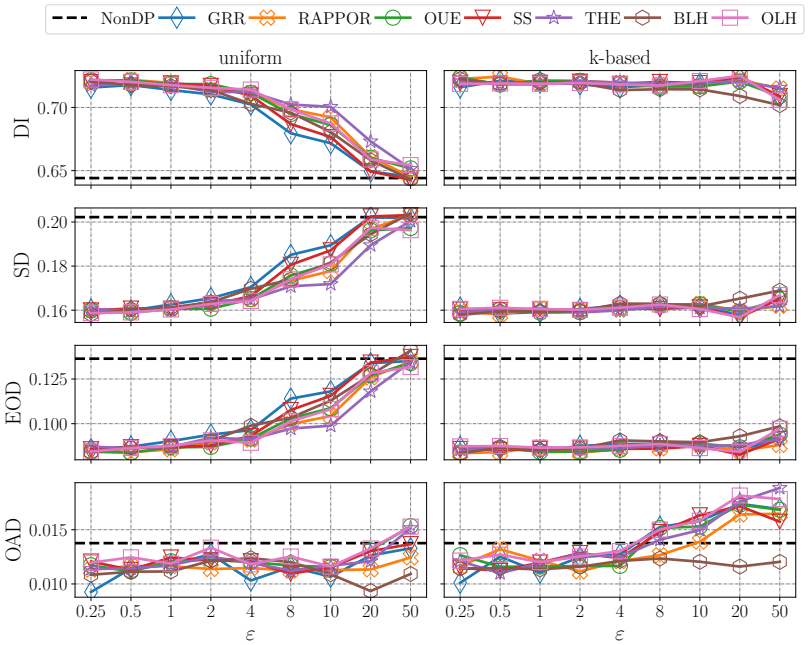


Fig. 4.4 Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ε-LDP protocol, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), on the Adult_{R} [70] dataset.

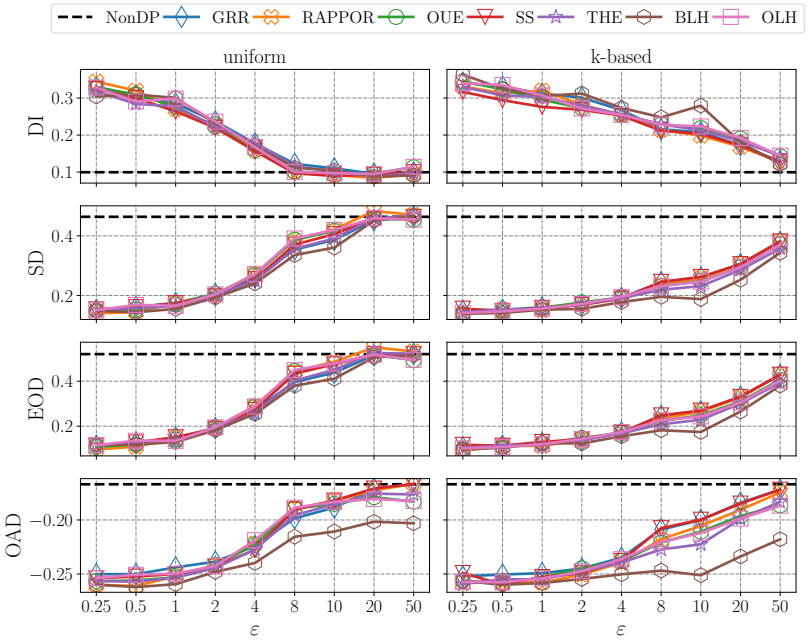


Fig. 4.5 Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ε-LDP protocol, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), on the ACSCoverage [70] dataset.

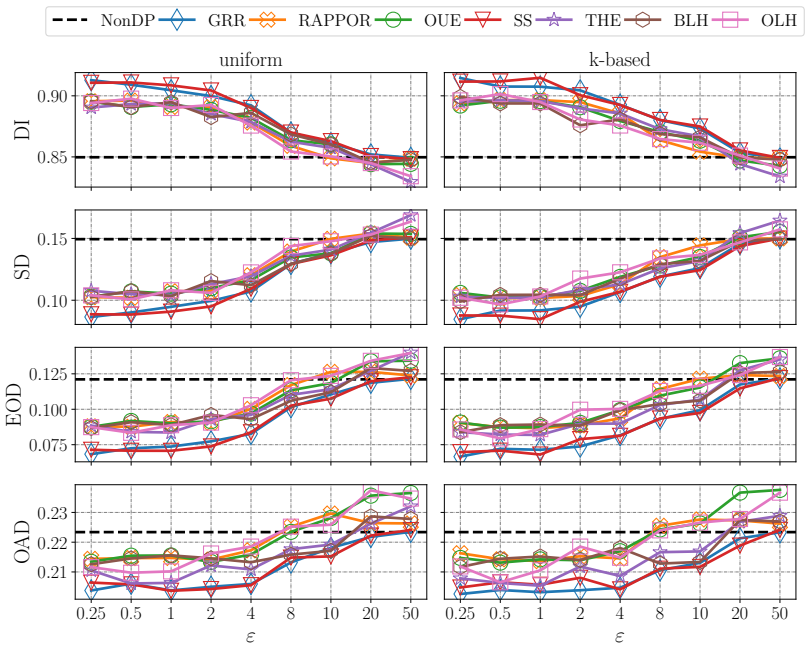


Fig. 4.6 Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ε-LDP protocol, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), on the LSAC [250] dataset.

- **LDP Impact on Utility.** Fig. 4.7 (Adult_G), Fig. 4.8 (Adult_R), Fig. 4.9 (ACSCoverage), and Fig. 4.10 (LSAC) illustrate the privacy-utility trade-off for the NonDP baseline and all the seven LDP protocols, considering both the ***uniform*** and our ***k-based*** privacy budget splitting solutions. From these Figures, one can note that, in general, the impact of ε-LDP on utility metrics is minor. For instance, for the Adult (see Figs 4.7 and 4.8) and LSAC (see Fig. 4.10) datasets, only $\sim 2\%$ of utility loss for all metrics is observed. Regarding privacy budget splitting, for the Adult_G and Adult_R datasets, our ***k-based*** solution was more robust to LDP as it only lost performance in high privacy regimes (i.e., smaller ε values). However, the ***uniform*** solution drops performance faster, i.e., even with $\varepsilon \leq 10$. One main explanation for this behavior is the high discrepancy in the domain size k of the sensitive attributes $\mathbf{A}$ and, consequently, more privacy level ε is allocated to those attributes with high k. For this reason, the ***uniform*** solution preserved more utility for the ACSCoverage dataset in Fig. 4.9, and both solutions had similar results for the LSAC dataset in Fig. 4.10 due to sensitive attributes with a small domain size k.

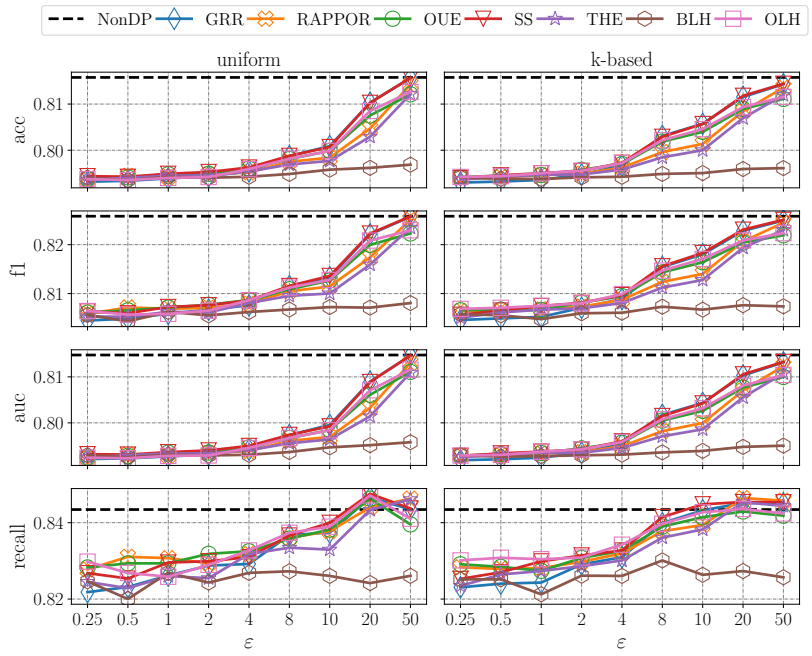


Fig. 4.7 Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ε-LDP protocol, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), on the Adult_G [70] dataset.

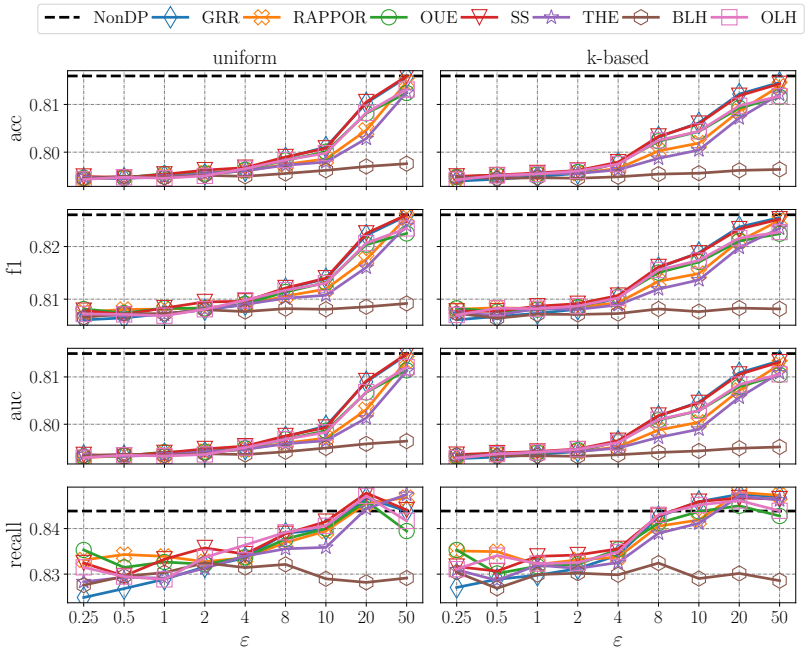


Fig. 4.8 Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ε-LDP protocol, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), on the Adult_{R} [70] dataset.

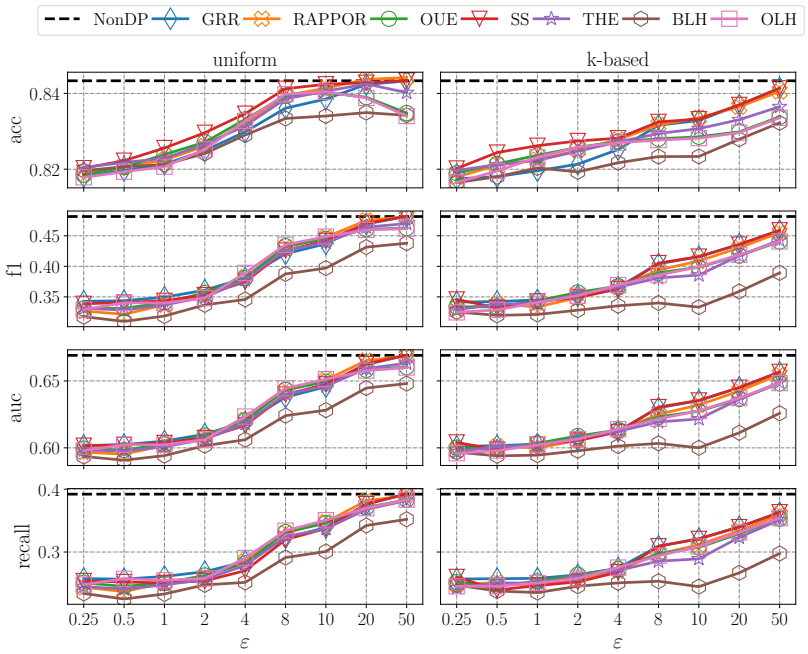


Fig. 4.9 Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ε-LDP protocol, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), on the ACSCoverage [70] dataset.

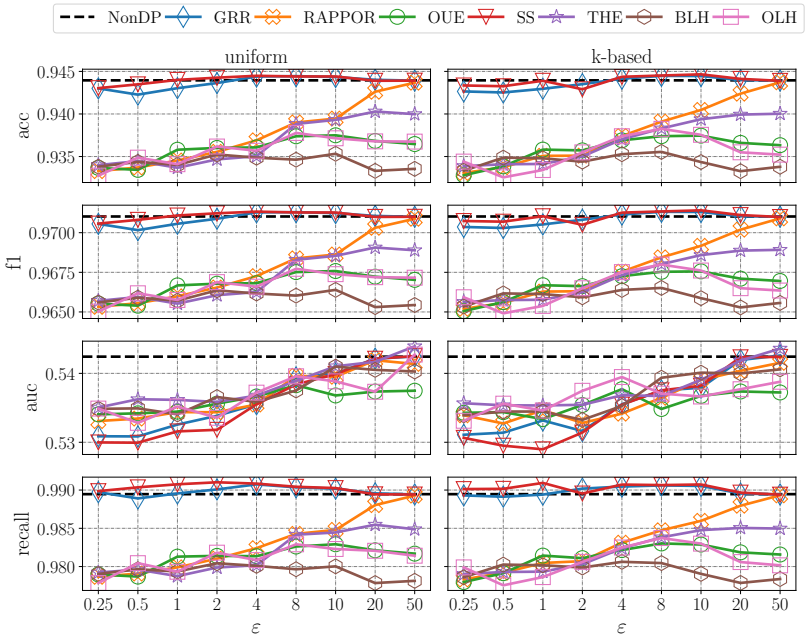


Fig. 4.10 Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ε-LDP protocol, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), on the LSAC [250] dataset.

- **Impact of the Number of Sensitive Attributes d_a.** Fig. 4.11 (fairness metrics) and Fig. 4.12 (utility metrics) illustrate the privacy-utility-fairness trade-off when varying the number of sensitive attributes d_a when applying the k-RR protocol on the Adult_G dataset. Naturally, as more sensitive attributes undergo obfuscation via ε-LDP, utility experiences a decline, albeit with a positive impact on fairness. Note that our experiments showed similar trend results for the other LDP protocols and datasets. Indeed, averaged results considering all the number of sensitive attributes d_a at once can be observed from Fig. 4.3 to Fig. 4.10.

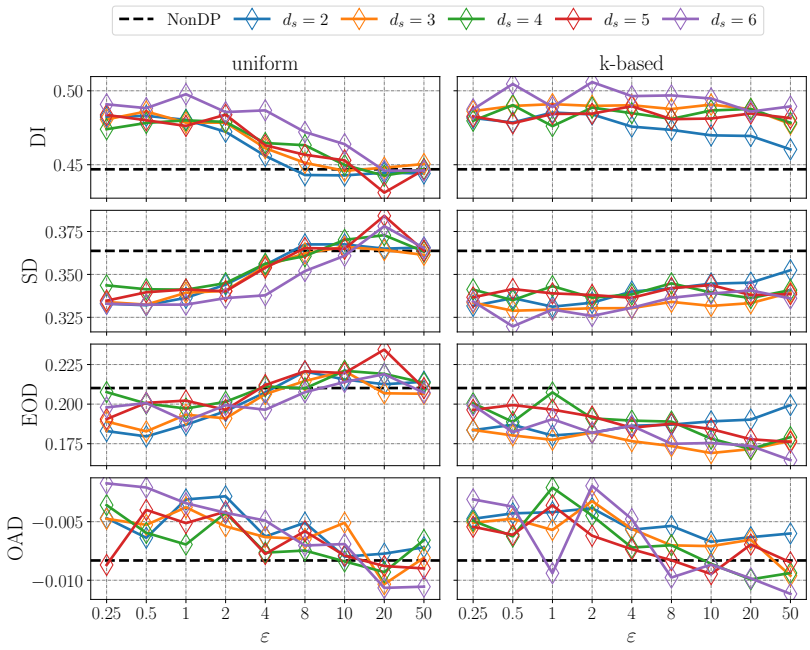


Fig. 4.11 Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the number of sensitive attributes d_s, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), with the k-RR protocol on the Adult_G [70] dataset.

Summary. We summarize our main findings for the three research questions formulated at the beginning of Section 4.2.3. **(RQ1)** Using the same hyperparameters configuration, ε-LDP positively affects fairness in ML (see Figs 4.3–4.6 and Fig. 4.11) while having a negligible impact on model's utility (see Figs 4.7–4.10 and Fig. 4.12). This contrasts the findings of [22, 95] that state that ε-DP negatively impacts fairness under the same hyperparameters configuration. While the aforementioned research works concern *gradient perturbation* in central DP, we focused on *input perturbation*, i.e., randomizing multiple sensitive attributes before training any ML algorithm, and discovered a positive impact of ε-LDP on fairness. **(RQ2)** Our ***k-based*** solution consistently led to better fairness improvement than the state-of-the-art ***uniform*** solution for all the four datasets. Regarding utility, ***k-based*** was better than ***uniform*** when sensitive attributes had higher domain sizes k (e.g., with the Adult_G and Adult_R datasets), which coincides with real-world data collections. Naturally, when all sensitive attributes have a binary domain, our ***k-based*** solution is equivalent to the ***uniform*** solution. For this reason, both state-of-the-art ***uniform*** and our ***k-based*** solution led to similar (favoring our ***k-based*** solution) privacy-utility-fairness trade-off for the LSAC dataset (see Figs 4.6 and 4.10). **(RQ3)** In general, k-RR and SS presented the best

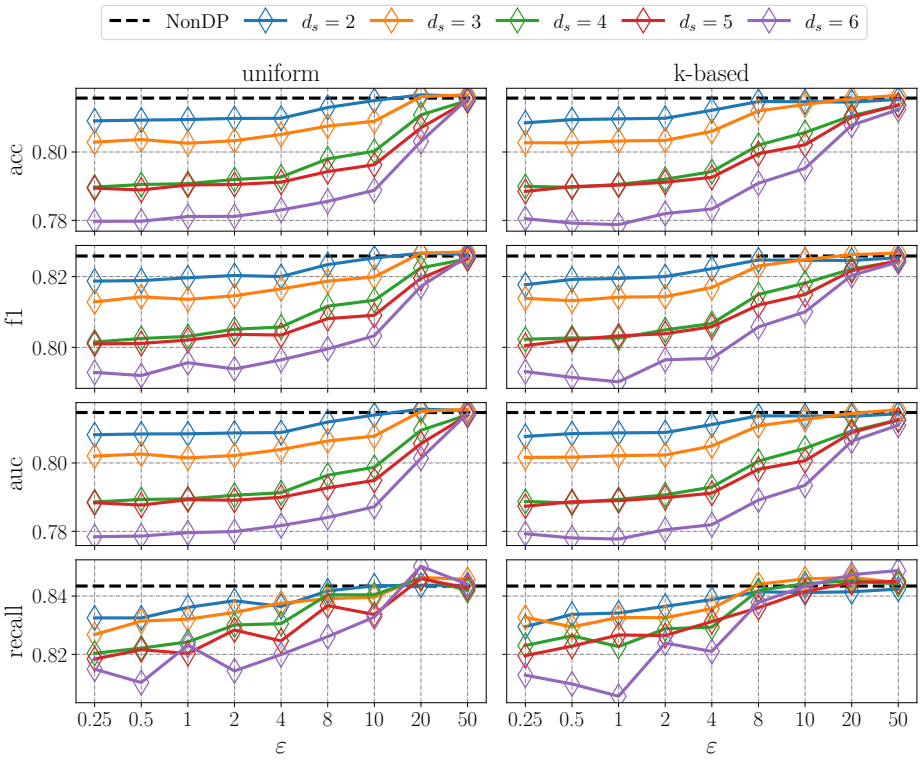


Fig. 4.12 Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the number of sensitive attributes d_s, and the privacy budget splitting solution (i.e., ***uniform*** on the left-side and our ***k-based*** on the right-side), with the k-RR protocol on the Adult_G [70] dataset.

privacy-utility-fairness trade-off for all the four datasets. This is because k-RR has only one perturbed output value and SS is equivalent to k-RR when $\omega = 1$, thus, not introducing *inconsistencies* for a user's profile. The term *inconsistency* refers to a user being in multiple categories in a given attribute, i.e., being both woman and man simultaneously. This can happen with all the other LDP protocols that utilize some specific encoding. For example, UE protocols perturb each bit independently, and when using LH protocols, many values can hash to the same perturbed value. In particular, since BLH hashes the input set $\mathbf{V} \rightarrow \{0, 1\}$, it consistently presented the worst utility results for all four datasets.

4.2.4 Conclusion

This work presented an in-depth empirical study of the impact of pre-processing multidimensional data with seven state-of-the-art ε-LDP protocols on fairness and utility in binary classification tasks. In our experiments, k-RR [115] and SS [244, 261]

presented the best privacy-utility-fairness trade-off in comparison with RAPPOR [79], OUE [245], THE [245], BLH [26], and OLH [245]. In addition, we proposed a new privacy budget splitting solution named ***k-based***, which generally led to better fairness and performance results than the state-of-the-art solution that splits ε uniformly (e.g., as in [12, 195, 126, 145]). Overall, while previous research [22, 95] has highlighted that central DP worsens fairness in ML under the same hyperparameter configuration, our study finds that LDP slightly improves fairness and does not significantly impair utility. As a perspective, we plan to investigate the impact of LDP pre-processing on different ML algorithms, such as deep neural networks, as well as different fairness metrics. Moreover, we plan to investigate the impact of Y distribution in a multidimensional LDP setting in Section 4.3. We also highlight that there is still much to explore in the area of privacy-fairness-aware ML, and this study's empirical results can serve as a basis for future research directions. For instance, in our work (presented later in Section 4.4), we formally investigated the privacy-fairness trade-off on binary classification when obfuscating the sensitive attribute A.

4.3 Fairness Under Multidimensional Local Differential Privacy: Empirical Study 2

This work can be considered an extension of the previous work presented in Section 4.2. In this work, we mainly focus on the well-known k-ary randomized response (k-RR) mechanism [115] (presented in Section 2.3.2) because all the seven LDP protocols presented similar trends in our study presented in Section 4.2. The choice of k-RR is motivated by its optimality for distribution estimation under several information theoretic utility functions [115], and also its design simplicity since k-RR does not require any particular encoding [245]. Specifically, since the output space is equal to the input space, k-RR provides optimal computational and communication costs for users. Moreover, no decoding step is needed on the server side. It also means that the server is free to use any post-processing coding techniques (e.g., one-hot encoding, mean encoding, binary encoding) to improve the usefulness of the ML model.

The k-RR mechanism has traditionally been mainly employed in the one-dimensional scenario in LDP and fairness literature [169, 48], where only one attribute is obfuscated. However, relying solely on LDP for a single sensitive attribute might be insufficient. This limitation stems from potential correlations that could allow attackers to reconstruct the privatized sensitive attribute. Hence, we specifically address scenarios involving multiple sensitive attributes, providing a more realistic representation of data collection in real-world contexts. Nevertheless, applying k-RR to multi-dimensional sensitive data presents greater challenges [72, 126]. For example, the naive approach of obfuscating each sensitive attribute independently results in the loss of any dependencies between sensitive attributes (our work presented in Section 4.2). In this study, in addition to this independent setting, we also explore a combined setting that transforms all sensitive attributes into a single attribute. Indeed, combined k-RR has not been extensively studied, and its impact on fairness remains unclear, a gap in understanding that we aim to address.

Contributions. The contributions of this study are threefold. First, we study the impact of LDP on fairness and utility by observing the behavior of sub-populations separately. This allows for a more complete understanding of how the fairness metrics behave under different LDP guarantees. Second, we compare both independent and combined settings for obfuscating multi-dimensional sensitive attributes under LDP guarantees. Third, we study how the Y distribution impacts the privacy-fairness-utility trade-off. The key findings of our empirical analysis are:

1. Generally, obfuscating data with LDP contributes to reducing disparity.

2. Obfuscating several sensitive attributes (multi-dimensional) reduces disparity more efficiently than obfuscating a single attribute (one-dimensional).

3. The multi-dimensional approaches of LDP (independent vs. combined) differ in their impact on fairness only at low privacy guarantees.

4. LDP obfuscation has, typically, a disproportionate impact on only one protected group, and this depends heavily on the distribution of the true decision Y.

Finally, to bridge the gap with practical applications, we frame the observations as concrete recommendations to practitioners considering both ethical concerns of privacy and fairness in ML applications.

Outline. The rest of this section is organized as follows. Section 4.3.1 discusses related work. Section 4.3.2 states the problem addressed in this study and provides some preliminaries. Section 4.3.3 details the experimental setting and discusses the main results. Finally, we conclude this work indicating future perspectives in Section 4.3.4.

4.3.1 Related Work

As this study closely relates to our work presented in Section 4.2, many of the research studies discussed in that section's related work also apply here. Therefore, we will highlight how this work differs from our previous research [16]. In particular, while [16] has only considered the *independent* setting for obfuscating the user's multi-dimensional data, for a more comprehensive examination, we considered in this study both *independent* and *combined* settings (defined below in Section 4.3.2). Another main difference with [16] is that we analyze the impact of LDP on fairness by varying the Y distribution.

4.3.2 Problem Setting and Methodology

Fig. 4.1 depicts the framework used in this work. Similar to the study described in Section 4.2, this work extends the obfuscation to multiple sensitive attributes $\mathbf{A}$, rather than focusing on a single attribute. Section 4.3.3 details all the k-RR settings considered in our experiments. Specifically, we assume there are $d_a \geq 2$ sensitive attributes $A_1, A_2, \ldots, A_{d_a}$, where the domain of each A_i is a discrete set of finite size $k_i = |dom(A_i)|$. We consider two state-of-the-art settings to apply k-RR on multi-dimensional data [126, 72]:

- **Independent k-RR (k-RR-Ind).** This naive approach applies k-RR independently on each attribute. In this study, we apply our *k-based* solution presented in Section 4.2. We recall that this approach consists of splitting ε among sensitive attributes based on their domain size.
- **Combined k-RR (k-RR-Comb).** This mechanism considers the Cartesian product $A_1 \times A_2 \times \ldots \times A_{d_a}$ as a single attribute and sanitizes it using k-RR parameterized with ε-LDP, where $k = k_1 \cdot k_2 \cdot \ldots \cdot k_{d_a}$.

Independent LDP on multi-dimensional data has been studied relatively well in the literature [195, 12, 126]. Moreover, its impact on fairness was the topic of our study [16]. Combined LDP, on the other hand, was not studied extensively. In particular, its impact on fairness is still unclear, a gap addressed in this study.

Group Fairness Metrics Considered. In this work, we focus on the following statistical group fairness metrics. These metrics are used to assess the impact of LDP on fairness.

- **Statistical Disparity (SD)** (Eq. 4.2).
- **Equal Opportunity Difference (EOD)** (Eq. 4.3).
- **Predictive Equality Disparity (PED)** [58] computes the difference in false positive rates (Table 2.2) between groups and it is formally defined as:

$$\text{PED} = \mathbb{P}[\hat{Y} = 1 \mid Y = 0, A = 1] - \mathbb{P}[\hat{Y} = 1 \mid Y = 0, A = 0]. \tag{4.5}$$

- **Overall accuracy difference (OAD)** (Eq. 4.4).
- **Predictive rate disparity (PRD)** [54] computes the difference in the positive predictive value (Table 2.2). PRD is formally defined as:

$$\text{PRD} = \mathbb{P}[Y = 1 \mid \hat{Y} = 1, A = 1] - \mathbb{P}[Y = 1 \mid \hat{Y} = 1, A = 0]. \tag{4.6}$$

Note that for all the fairness metrics mentioned above, the lower the values, the fairer the results.

4.3.3 Empirical Results and Analysis

To assess the impact of k-RR on fairness, two synthetic datasets and two real-world fairness benchmark datasets, namely *Adult* and *Compas*, are used. For each of these datasets, the fairness metrics presented above are applied.

General setting.

- **Environment.** All the experiments are implemented in Python 3. We use *Random Forest* model [38] for classification with its default hyper-parameters and we use the ten-fold cross-validation technique, both from Scikit-learn [186]. For k-RR mechanism, we use the implementation in Multi-Freq-LDPy [13]. The codes and datasets for all the experiments are available in a **GitHub repository** [153].
- **Stability.** Since LDP protocols, k-fold cross-validation, and ML algorithms are randomized, we report average results over 20 runs.

Datasets. A summary of all datasets used in this study is provided in Table 4.2.

Table 4.2 Metadata of the datasets used in the experiments of this study.

Dataset	n	A *(sensitive att.)*	$\mathbf{A}$ *(sensitive atts)*	Y	*Threshold*	$\mathbb{P}[Y=1]$
Synthetic	100K	A	- A	Y	$\tau_{Q1}=0.44$	0.75
			- C		$\tau_{Q2}=0.52$	0.5
			- M		$\tau_{Q3}=0.6$	0.25
Compas	5915	race	- race	risk score[7]	$\tau_{Q1}=1$	0.76
			- gender		$\tau_{Q2}=3$	0.47
			- age		$\tau_{Q3}=5$	0.26
Adult	32561	gender	- gender	income	$\tau_{Q1}=10K$	0.81
			- age		$\tau_{Q2}=27K$	0.5
			- race		$\tau_{Q3}=50K$	0.25
			- marital-status			
			- native-country			

- **Synthetic Dataset**: The causal model used to generate the synthetic dataset is depicted in Figure 4.13. A, C, and M are discrete variables[8], while Y is

[7] Unlike the synthetic and the *Adult* datasets, whose true decision is continuous, the true decision of the *Compas* dataset is discrete (score $\in \{0,1\}$). Thus, we use scores 1, 3, and 5 as thresholds for the Y distribution to be skewed to 0, balanced, and skewed to 1, respectively.

[8] C and A follow *Binomial* distributions while M follows *Multinomial* distribution.

a continuous variable that is a function of all the other variables such that: $Y = h(A, C, M)$. To study the impact of k-RR on fairness while varying the Y distribution, three thresholds are set for the true decision variable Y binarisation, resulting in three synthetic datasets differing solely by the distribution of Y. The thresholds and the resulting Y distribution for all datasets are shown in Table 4.2. Three scenarios are considered depending on the dataset, namely, Y distribution skewed to 1, balanced Y distribution, and Y distribution skewed to 0.

- **Benchmark Datasets**:

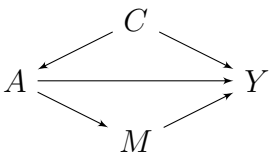
  - *Compas*: The *Compas* dataset includes data about defendants from Broward County, Florida, during 2013 and 2014 who were subject to *Compas* screening. Various information related to the defendants (e.g., race, gender, arrest date, etc.) were gathered by ProPublica [10], and the goal is to predict the two-year violent recidivism. Only black and white defendants assigned *Compas* risk scores within 30 days of their arrest are kept for analysis, leading to 5915 individuals in total. We consider race as the sensitive attribute. Five attributes are used in this study: race, sex, age, priors, and risk score. We use the *Compas* risk score as Y. The risk score consists of a rating of $1-10$; the higher the score, the more likely the defendant is to re-offend. Following the same reasoning as the synthetic datasets, we transform the risk score into a binary variable by choosing different thresholds to study the impact of Y distribution on the privacy-fairness trade-off. Three thresholds are used, leading to three different Y distributions: skewed to 1, almost balanced, and skewed to 0.

  - *Adult* (already defined in Section 4.2.3): The attributes considered in this work are age, gender, native country, education level, marital status, number of working hours per week, and income. Y is the income of an individual. Similarly to the other datasets, different thresholds are used to separate the positive true decision (high income) from the negative true decision (low income). Three thresholds are used in total, leading to three versions of the *Adult* dataset with skewed income distribution to 1 (threshold = 10K), balanced income distribution (threshold = 26K), and skewed income distribution to 0 (threshold = 50K[9]).

Applied Settings. Four settings (Table 4.3) are used to assess the impact of LDP on fairness. We vary the privacy level in the range of $\varepsilon = \{16, 8, 5, 3, 2, 1, 0.5, 0.1\}$.

- *noLDP* (Baseline): the model is trained using the original data (without privacy).

[9]The 50K threshold is used in the well-known Adult dataset mostly used in the literature [73].



(a) Causal Graph

$$C \sim \mathcal{B}(0.35),$$
$$A \sim \begin{cases} \mathcal{B}(0.55) \text{ if } C = 0, \\ \mathcal{B}(0.75) \text{ if } C = 1, \end{cases}$$
$$M \sim \begin{cases} \mathcal{M}(0.35, 0.4, 0.25) \text{ if } A = 0, \\ \mathcal{M}(0.5, 0.4, 0.1) \text{ if } A = 1, \end{cases}$$
$$Y = \alpha A + \beta M + \gamma C + \mathcal{U}_y, \quad \mathcal{U}_y \sim \mathcal{N}(0, 1)$$

(b) Structural equations

Fig. 4.13 Causal model of the synthetic datasets.

- *sLDP*: the model is trained using an obfuscated version of the data where only the sensitive attribute A is obfuscated using k-RR.
- *combLDP*: the model is trained using an obfuscated version of the data where a set of sensitive attributes $\mathbf{A}$, including the sensitive attribute A is obfuscated using k-RR-Comb (Section 4.3.2).
- *indLDP*: the model is trained using an obfuscated version of the data where the same set of sensitive attributes $\mathbf{A}$ is obfuscated using k-RR-Ind (Section 1.1.2). The privacy splitting solution used in the experiments is the *k-based* solution [16].

Table 4.3 Settings applied in this study.

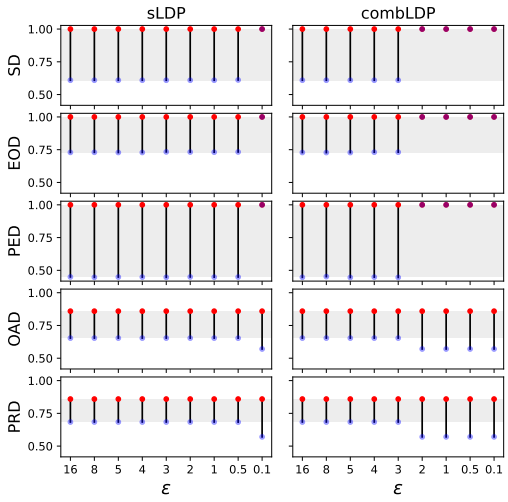
Settings applied	*k-RR applied to*
noLDP	no privacy
sLDP	A
combLDP	$\mathbf{A}$ using k-RR-Comb
indLDP	$\mathbf{A}$ using k-RR-Ind

Main Results.

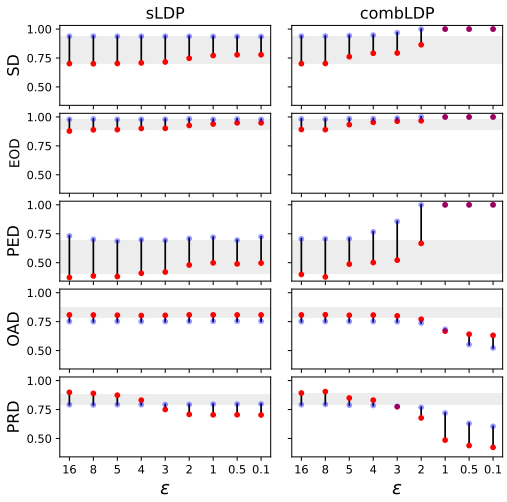
- **Impact of LDP on Fairness.** This set of experiments aims to study the effect of obfuscating data through LDP on the fairness of the model trained using that data. The experimental protocol consists of obfuscating data using either *sLDP* (one-dimensional) or *combLDP* (multi-dimensional) while decreasing the privacy budget ε toward more privacy requirements (small ε). Fairness is measured using

the various group metrics of Section 4.3.2, and the experiment is repeated for all three datasets (Synthetic, *Compas*, and *Adult*). Fig. 4.14 shows the obtained results. To better understand how LDP impacts fairness, the plots show the separate values for both groups: the *privileged* group ($A = 1$) in red dots and the *unprivileged* group ($A = 0$) in blue dots. Disparity between groups is then the difference between the two values (dots). In addition, for a better understanding of the trade-off, disparity in the baseline case (i.e., no obfuscation – *noLDP*) is shown using a gray shaded area. The following can be observed from the empirical results.

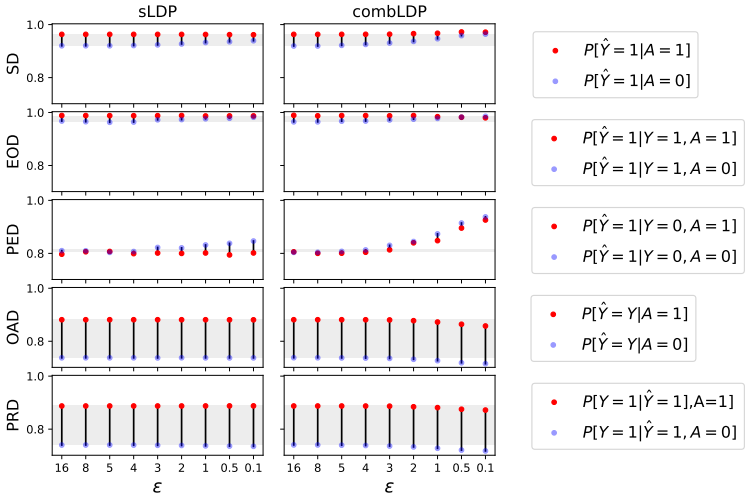
- **[Obs1]** *More privacy leads to less disparity.* For both *sLDP* and *combLDP*, the disparity decreases when imposing stronger privacy requirements (smaller ε). For example, in Fig. 4.14b, SD (first row) decreases from 0.23 to 0.15 (for *sLDP*) and to 0 (for *combLDP*). The same decreasing pattern can be observed for EOD (second row) and PED (third row). For OAD and PRD, however, disparity either stays unaffected (Fig. 4.14c) or increases (Figs 4.14a and 4.14b). These two fairness notions compare both groups' accuracy and precision (e.g., $Y = \hat{Y}$ for accuracy). Hence, the behavior is expected since imposing strong privacy guarantees typically leads to a decrease in the accuracy and precision of the classifier for one or both protected groups. But the drop is greater for one group than for the other. This is further detailed when studying the impact of the true decision distribution on the privacy-fairness-utility trade-off later in this section.



(a) Synthetic data



(b) Compas



(c) Adult

Fig. 4.14 Impact of LDP on disparity (y-axis) by varying the privacy level ε (x-axis). *sLDP* consists in obfuscating a single attribute (sensitive). *combLDP* consists in obfuscating all sensitive attributes. The gray shaded area represents the disparity results using the baseline model (*noLDP*). The formulas for the red and blue dots in each plot are shown on the right of Fig. 4.14(c) where red dots correspond to the fairness results for group $A = 1$ while the blue dots correspond to the fairness results for group $A = 0$.

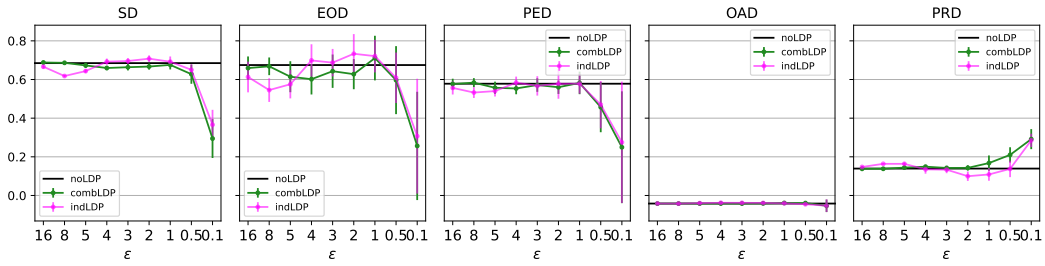
– **[Obs2]** *Multi-dimensional LDP reduces disparity more efficiently than one-dimensional LDP.* Both *sLDP* and *combLDP* lead to a decrease in disparity (previous observation). However, with *combLDP*, the reduction can be observed with weaker privacy guarantees (higher ε). In other words, the more attributes are obfuscated, the less privacy level ε is needed to improve

fairness. For instance, in Fig. 4.14a, the disparity disappears at $\varepsilon = 0.1$ for *sLDP*, but at $\varepsilon = 2$ for *combLDP*. This can be explained by the fact that obfuscating the sensitive attribute A (equivalent to removing that attribute from the training set when the privacy guarantees are strong enough) is insufficient to improve fairness due to proxies correlated with that attribute. Thus, by additionally obfuscating all attributes correlated with the sensitive attribute, weaker privacy guarantees are required to reduce the disparity between groups and, therefore, improve fairness.

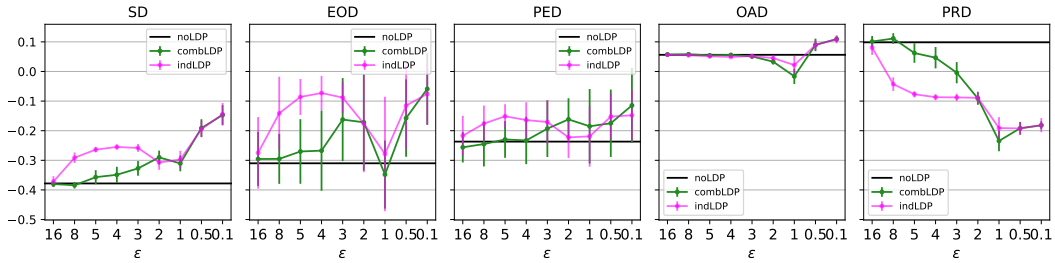
- [**Obs3**] *LDP has disproportionate impact on groups.* In most of the plots, one can observe that k-RR does not have an impact (or has a minor impact) on one group but a high impact on the other group. For instance, in the first three rows of Fig. 4.14a, the change in disparity is due to a significant change related to only the *unprivileged* group (blue dots). In other words, considering groups separately, k-RR impacts the fairness/utility of these groups differently.

- **k-RR-Ind vs. k-RR-Comb.** The impact of LDP on the fairness level of the obtained model depends on the multi-dimensional k-RR variant (Section 4.3.2) used for obfuscation. The following experiment is performed to compare the effects of k-RR-Ind and k-RR-Comb on the disparity between the *privileged* and *unprivileged* groups. Benchmark datasets (Synthetic, *Compas*, and *Adult*) are obfuscated using k-RR-Ind and k-RR-Comb while decreasing the privacy budget ε toward more strict privacy guarantees (very small ε). The obfuscated data is then used to train a predictor and the disparity of the model is then assessed using the fairness metrics of Section 4.3.2. Fig. 4.15 shows the result of the experiments.

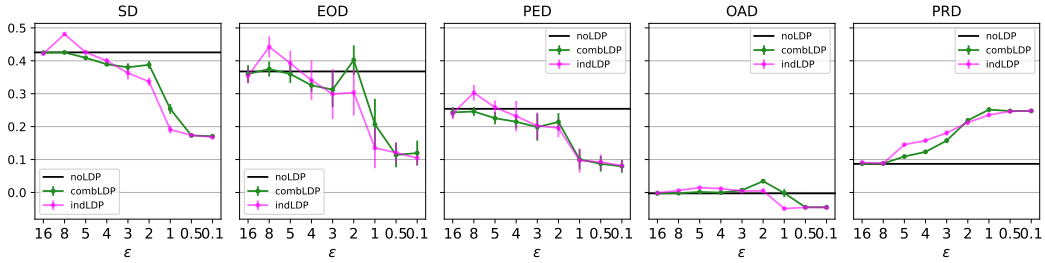
  - [**Obs4**] *For large ε, the efficiency to reduce disparity depends on the sensitive attributes inter-dependencies.* *Compas* and *Adult* experiments illustrate the two different behaviors. In *Compas* experiment (Figure 4.15b), at $\varepsilon = 4$, EOD for *indLDP* is -0.07 but -0.27 for *combLDP*. Recall, from Table 4.2, that in *Compas* dataset, three attributes are considered sensitive (race, gender, and age) with relatively low inter-dependencies between them. This explains why k-RR-Ind is more efficient in reducing disparity than k-RR-Comb for large ε values. In the *Adult* experiment result (Fig. 4.15c), k-RR-Comb is slightly more efficient than *indLDP* in reducing disparity. For instance, at $\varepsilon = 8$, EOD is at 0.44 for *indLDP* but at 0.37 for *combLDP*.



(a) Synthetic data



(b) Compas



(c) Adult

Fig. 4.15 Impact of *combLDP* and *indLDP* on disparity (y-axis) by varying the privacy level ε (x-axis) and obfuscating a set of sensitive attributes. The horizontal black line in each plot designates the fairness results when the baseline model (*noLDP*) is deployed.

This can be explained by the relatively high inter-dependencies of the five sensitive attributes (Table 4.2) considered in the *Adult* dataset.

– **[Obs5]** *For small ε, combLDP and indLDP have a similar impact on disparity.* In all plots of Fig. 4.15, for strict privacy guarantees (small ε), the disparity between sensitive groups converges to the same value whether the obfuscation was performed with *combLDP* or *indLDP*. In other words, by enforcing more privacy, both settings of k-RR improved fairness to the same extent.

- **The Effect of Changing the True Decision Distribution.** To assess disparity using the group fairness metrics (Section 4.3.2), the true decision variable Y must be binary. However, typically, the trained model predicts a continuous numerical value representing a score as a true decision. The score value needs to be thresholded to obtain a binary value. Consequently, the true decision variable Y distribution will depend on the threshold value. The following experiment is performed to study the effect of the true decision distribution on the disparity between protected groups while obfuscating data. Three different distributions are considered for each dataset (Synthetic, *Compas*, and *Adult*). The first distribution is obtained by considering a threshold value (τ_{Q1}) such that all instances in the three top quantiles have positive true decision ($Y = 1$). The threshold (τ_{Q2}) of the second distribution is selected such that the two top quantiles have positive true decision. The third threshold (τ_{Q3}) is selected such that only the instances in the top quantile have positive true decisions. Each of the obtained datasets is then obfuscated using *sLDP*, *combLDP*, and *indLDP*. Fig. 4.16 shows the experimental results for the *Adult* dataset (Results for Synthetic and *Compas* can be found in the appendix (Figs. B.2 and B.3). To better understand how fairness is impacted by the distribution of the true decision, the plots track the separate values for each protected group (dots on solid lines for the *privileged* group and dots on dashed lines for the *unprivileged* group). The difference between the two types of dots corresponds to the disparity. Finally, as previously mentioned, the gray area corresponds to the disparity of the baseline model (*noLDP*).

  - [**Obs6**] *When enforcing privacy, which group witnesses more accuracy drop depends on the true decision distribution.* Depending on the threshold for positive true decision (and hence the true decision distribution), the drop in accuracy[10] due to more tight privacy guarantees (smaller ε) is higher for one group than the other. In particular, the accuracy drops more for the *unprivileged* group $A = 0$ when the Y distribution is either skewed to 1 (τ_{Q1}) or balanced (τ_{Q2}), which corresponds to the first and second columns in Fig. 4.16. Whereas it drops more for the *privileged* group $A = 1$ when the Y distribution is skewed to 0 (τ_{Q1})[11].

[10] As this observation is about the accuracy, only the last two fairness metrics are concerned, that is, OAD and PRD corresponding to the two lower rows of Fig. 4.16.

[11] Note that this observation is also confirmed in the *Compas* dataset (Fig. B.3) but inverted since the *privileged* group in this dataset is the group $A = 0$. We generated a second synthetic dataset where the group $A = 0$ is *privileged* to confirm the inverted behavior. The plots are in Appendix B.1.1.

- **[Obs7]** *When enforcing privacy, which group contributes more to reduce the disparity depends on the true decision distribution.* Similarly to the above observation, the true decision distribution significantly impacts how each group (*privileged* vs. *unprivileged*) contributes to the disparity reduction while enforcing more privacy. In particular, the prediction rates per group (e.g., $\mathbb{P}[\hat{Y} = 1|A = 1]$ for SD) *increased more for the unprivileged group* $A = 0$ *when the true decision distribution is skewed to* 1 *(*τ_{Q1} *and* τ_{Q2}*) but* decreased more for the *privileged* group $A = 1$ when the true decision distribution is skewed to 0 (τ_{Q3})[12].
- **[Obs8]** *For a fair baseline model, enforcing privacy amplifies disparity.* The true decision distribution experiment exhibited an interesting behavior illustrated clearly in the *Adult* dataset results (Fig. 4.16). In particular, the disparity in the baseline predictor is relatively small for the PED metric with true decision distribution at threshold τ_{Q1}. However, training the predictor using obfuscated data resulted in disparity amplification. A similar behavior is observed for OAD with τ_{Q2}.

Based on the above observations, one can conclude the following statements:
Statement 1: *If* $A = a$ *is the privileged group (has a majority of* $Y = 1$*) then if* Y *is skewed to* 1*, adding noise affects more the accuracy of the unprivileged group* $A \neq a$ *else (*Y *is skewed to* 0*) adding noise affects more the accuracy of* $A = a$*. If* $A = a$ *is the privileged group (has a majority of* $Y = 1$*) then if* Y *is skewed to* 1*, adding noise affects more the accuracy of the unprivileged group* $A \neq a$ *else (*Y *is skewed to* 0*) adding noise affects more the accuracy of* $A = a$*.*
Statement 2: *If* $A = a$ *is the privileged group (has a majority of* $Y = 1$*), then if* Y *is skewed to* 1*, adding noise increases more the predicted rates for the unprivileged group* $A \neq a$ *else (*Y *is skewed to* 0*), adding noise decreases more the predicted rates for group* $A = a$*. If* $A = a$ *is the privileged group (has a majority of* $Y = 1$*), then if* Y *is skewed to* 1*, adding noise increases more the predicted rates for the unprivileged group* $A \neq a$ *else (*Y *is skewed to* 0*), adding noise decreases more the predicted rates for group* $A = a$*.*

Recommendations. Based on the observations obtained from the experimental analysis, one can propose the following recommendations for a practitioner considering

[12]Again, the behavior is reversed for the *Compas* dataset (Fig. B.3) for the same reason as the previous observation.

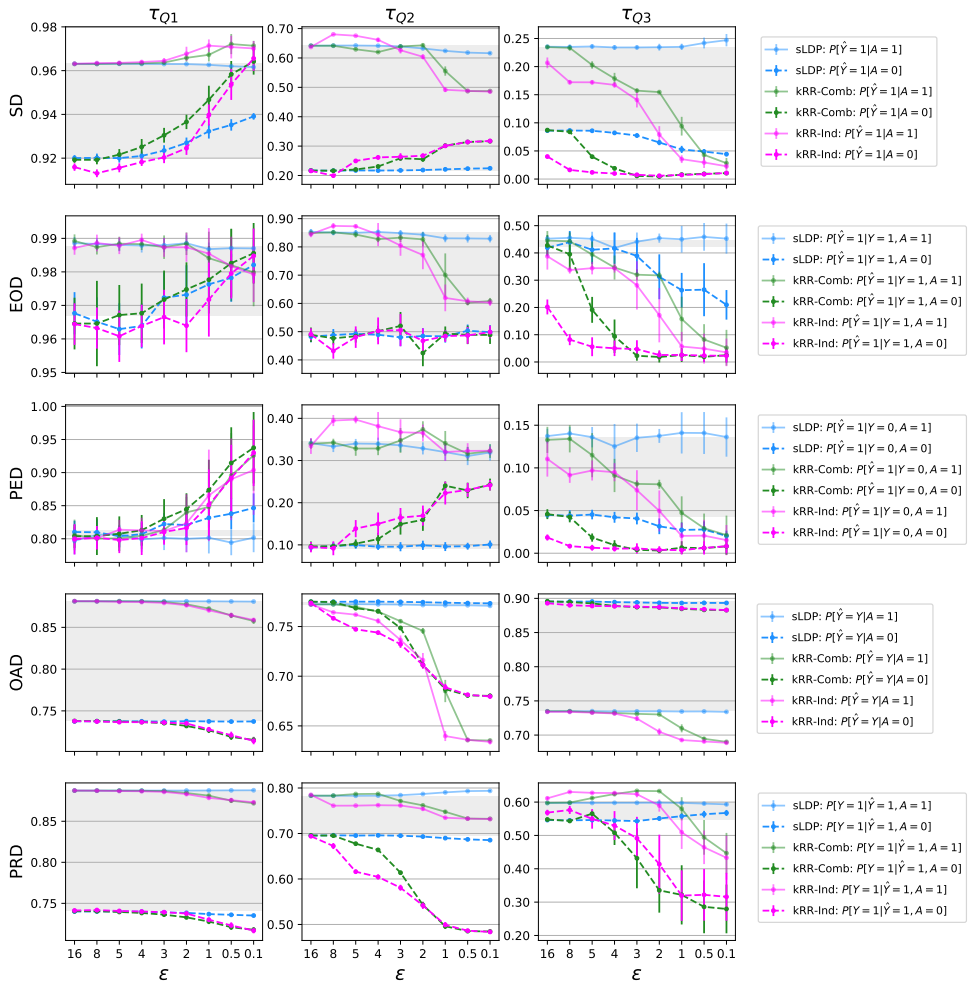


Fig. 4.16 Impact of Y distribution on the privacy-fairness trade-off. Columns 1, 2, and 3 illustrate the results for the *Adult* dataset when the Y distribution is skewed to 1, balanced, and skewed to 0, respectively. The gray shaded area represents the disparity results using the baseline model (*noLDP*).

a mechanism satisfying privacy and fairness guarantees. That is a mechanism allowing individual users to share their data while at the same time protecting their sensitive information and guaranteeing that the obtained model is fair w.r.t sub-populations and/or individuals.

A. LDP Data Obfuscation is an Efficient Mechanism for Reducing Disparity. Almost all observations from the experimental analysis confirm the conclusion that LDP obfuscation reduces disparity (**Obs1, Obs2, Obs4, Obs7**). The disparity reduction is often due to one group being more sensitive to the LDP obfuscation rather than the

other (**Obs2**). The only exception is when the predicted model using baseline (not obfuscated) data is already fair. In that case, LDP may create disparity (**Obs8**).

B. Obfuscating Several Sensitive Attributes Allows us to Reduce Disparity more Efficiently than a Single Attribute. Suppose a practitioner is interested in producing a fair model but with minimal privacy enforcement. In that case, it is recommended that she uses multi-dimensional LDP, obfuscating as many sensitive attributes as possible (**Obs2**).

C. Independent and Combined Variants of Multi-Dimensional LDP are Different Only with Weak Privacy Guarantees. The choice of the multi-dimensional approach of LDP (combined vs independent) matters only at low privacy guarantees (large ε) (**Obs4**). In that case, the practitioner's choice should depend on the level of interdependency between sensitive attributes. For high interdependency, a combined approach is more efficient in reducing disparity. For low or no interdependency, an independent approach is more efficient. At strict privacy guarantees (low ε), however, both approaches have a similar effect on disparity (**Obs5**).

D. Obfuscating Data Disproportionally Impacts Only one Group, Depending on the True Decision Distribution. A practitioner who obfuscates individual data with LDP should expect only one group to be significantly affected. And she can *guess* which group will be more affected by studying the true decision distribution. More precisely, if the true decision distribution is skewed towards the positive true decision (typically $Y = 1$), the *unprivileged* group will be more affected. Otherwise (if the true decision distribution is skewed to the negative true decision (typically $Y = 0$), it is the *privileged* group who will be more affected (**Obs7** and **Obs8**).

4.3.4 Conclusion

This work investigates how the accuracy and fairness of the decisions made by the model change under LDP, in particular, the k-RR mechanism, given different levels of privacy and different class distributions. To broaden the scope of our study, we employed various statistical group fairness metrics. We evaluated two settings for obfuscating multi-dimensional sensitive attributes under LDP, namely, independent and combined, on two synthetic and two benchmark datasets to substantiate our claims. We also investigated the impact of Y distribution on the impact of multidimensional LDP on fairness. The experimental analysis revealed very relevant observations that we framed as concrete

recommendations for ML practitioners aiming to guarantee both ethical privacy and fairness concerns. To the best of our knowledge, this is the first work that studies the effect of combined multi-dimensional LDP on fairness. In particular, we observed that independent and combined variants of multi-dimensional LDP are different only with weak privacy guarantees (high ε). For practitioners, the choice between the two variants of multi-dimensional LDP should be based on the extent of interdependence among sensitive attributes. In cases of high interdependency, employing a combined approach proves more effective in mitigating disparity. Conversely, opting for an independent approach is more efficient for low or negligible interdependency.

Although only associational fairness metrics are used, note that our work can be easily extended to causality-based fairness metrics. More specifically, given the causal graph, if some of the sensitive attributes happen to be in the back-door path[13] between the sensitive attribute A and the true decision Y, the most common causality-based fairness metric, namely, total effect (TE) (Eq. (3.24)), might behave differently than SD. However, TE will act precisely as SD if no sensitive attributes are in the back-door path between the sensitive attribute A and the true decision Y. Moreover, obfuscation might impact other causality-based fairness metrics (direct effect, indirect effect, etc.) depending on the location of the sensitive attributes on the causal graph, provided those metrics are identifiable [157].

[13]Recall that a back-door path is a path between A and Y with an edge into A.

4.4 A Systematic and Formal Study of the Impact of Local Differential Privacy on Fairness

The insights gained from our two empirical studies presented above highlight the complex interplay between privacy-preserving mechanisms and fairness in real-world applications. Despite the valuable findings, these empirical investigations revealed several underlying theoretical questions that remained unanswered. We were motivated to undertake a foundational study to address these gaps and build a deeper understanding. This study aims to systematically analyze the principles that govern the relationship between privacy and fairness, providing a theoretical framework that complements and enhances the empirical results.

Specifically, we formally study the impact of training a model with data obfuscated by the randomized response (RR) mechanism (Eq. (2.4)), a fundamental LDP protocol [115] that serves as a building block for more complex LDP mechanisms (Section 2.3.2).

Contributions. Our main contribution consists of a theoretical analysis of how the fairness of the prediction of an ML model is affected by the application of RR on the training data, depending on the level of privacy and the data distribution. In particular, we study three notions of fairness: SD (Eq. (4.2)), CSD (Eq. (4.7)), and EOD (Eq. (4.3)), and identify the conditions under which they are improved or reduced by RR. We then empirically validate our results by performing experiments on synthetic data and four real datasets, *Compas* [10], *Adult* [70], *German credit* [73], and *LSAC* [250]. Appendix B.2.1 contains all detailed proofs supporting our findings.

Outline. The rest of this section is organized as follows. Section 4.4.1 discusses related work. Section 4.4.2 states the problem addressed in this study. Then, section 4.4.3 presents our quantitative study of how the fairness of the prediction is affected by the application of the RR mechanism. Section 4.4.4 details the experimental setting and discusses the main results. Finally, we conclude this work indicating future perspectives in Section 4.4.5.

4.4.1 Related Work

This section provides additional existing research studies that explore the intriguing relationship between DP and fairness, specifically in the context of ML.

Central Differential Privacy. Sanyal et al. [202] present theoretical and experimental evidence demonstrating that private and accurate algorithms are inherently unfair.

They also illustrate that striving for both privacy and fairness results in inaccurate algorithms. A more recent research paper by Emelianov and Perrot [78] on the impact of output perturbation on individual and group fairness in binary linear classification shows that the impact of output perturbation on individual fairness, in general, depends on the dimension of the problem. They also derived a high probability bound on the group fairness of the private models compared to the non-private ones. They showed that the bound grows with the noise and depends on the non-private model's distribution of "angular margins". Tran, Dinh, and Fioretto [235] examined the accuracy disparities among different groups of individuals caused by output perturbation and differentially private stochastic gradient descent. They analyzed the data and model properties responsible for these disproportionate impacts, explored the reasons behind the unequal effects on various groups, and proposed guidelines to mitigate these issues.

Mangold et al. in [161] perform a theoretical analysis of the impact of central DP on fairness in classification. They prove that the difference in fairness levels between private and non-private models diminishes at a rate of $\tilde{O}\left(\sqrt{p}/n\right)$, where n represents the number of training records and p is the number of parameters. They also provide an empirical study using the central model with Gaussian noise for DP and $l2$-regularized logistic regression models for prediction.

Local Differential Privacy. In [169], Mozannar et al. show how to adapt non-discriminatory learners to work with privatized attributes, giving theoretical guarantees on performance. Our experimental analysis (discussed in Section 4.3) showed that obfuscating several sensitive attributes instead of obfuscating only the sensitive attribute used to assess fairness gives better results for fairness. Also, we observed that combined LDP, compared to independent LDP, reduces the disparity more efficiently at low privacy guarantees (high ε). Our empirical study presented in Section 4.2 also empirically deals with the impact on the fairness of applying LDP to multiple sensitive attributes. The analysis covers several fairness metrics and state-of-the-art LDP protocols. The results contrast with those obtained with central DP, as they show that LDP slightly improves fairness in learning tasks without significant loss of the model's accuracy.

These contrasting claims, most of which are backed only by experimental results, show that a systematic and foundational study of the relationship between privacy and fairness is highly needed. This work is a step in that direction.

4.4.2 Problem Setting and Methodology

Fig. 4.1 depicts the framework used in this work. Note that we solely obfuscate the sensitive attribute A in this study. We briefly recall the privacy setting and the fairness metrics applied in this study.

A predictor $\hat{Y}$ of an outcome Y is a function of a set of variables (A, X) where X[14] designates the set of non-sensitive attributes and $A \in \{0, 1\}$ represents the sensitive attribute. Note that X could include proxies to A, such as zip code, which could hint to race. We assume that $\hat{Y}$ and Y are binary random variables where $Y = 1$ (e.g., hiring a person) designates a positive outcome, and $Y = 0$ (e.g., not hiring a person) designates a negative outcome. For the remainder of this section, we assume that we have access to a (multi)set $S = \{(a_i, x_i, y_i)\}_{i=1}^{n}$ of n i.i.d samples from the distribution on $A \times X \times Y$.

We recall $A' = \mathcal{L}(A)$, the obfuscated version of the sensitive attribute A, where $\mathcal{L}$ is a certain randomized LDP mechanism. Thus, we denote an obfuscated version of S as $S' = (a'_i, x_i, y_i)_{i=1}^{n}$.

The LDP mechanism we consider here is the randomized response (RR) [246, 115] for a binary variable $a \in \{0, 1\}$, which is defined in Eq. (2.4).

Group Fairness Metrics Considered. In this study, we focus on the following statistical group fairness metrics. These metrics are used to assess the impact of LDP on fairness.

- **Statistical Disparity (SD)** (Eq. 4.2).
- **Conditional Statistical Disparity (CSD)** [58] computes the difference in predicted acceptance rates between groups conditioning on a set of explanatory attributes. In this work, we assume that all variables in X are (potential) explanatory variables, and we define conditional statistical disparity for each instance x of X as follows:

$$\text{CSD}_x = \mathbb{P}[\hat{Y} = 1 \mid X = x, A = 1] - \mathbb{P}[\hat{Y} = 1 \mid X = x, A = 0]. \tag{4.7}$$

  Note that, in general, x represents a tuple of values since X may contain more than one attribute.

- **Equal Opportunity Difference (EOD)** (Eq. (4.3)).

[14] X can be a vector of variables.

4.4.3 Quantitative Analysis of the Impact of Privacy on Fairness

In this section, we formally study the impact of LDP on fairness. Specifically, we perform a quantitative study of how the fairness of the prediction is affected by applying the RR mechanism to the sensitive values in the training data, depending on the level ε of privacy and the data distribution.

We briefly recall our setting. In addition to the sensitive attribute A and the true decision Y, which are binary, the data includes a set of non-sensitive attributes X with arbitrary values. We assume that the data model is *probabilistic*, in the sense that the data may contain tuples with the same values for X and A and different values for Y. $A' = \mathrm{RR}(A)$ is an obfuscated version of A obtained by applying the RR mechanism to A, and it is also binary. The prediction of the model trained on the original data is denoted by $\hat{Y}$, while that of the model trained on the obfuscated data, which we will call LDP model, is $\hat{Y}'$. Of course, $\hat{Y}$ and $\hat{Y}'$ are also binary. We assume that both models are deterministic. Namely, on a given input (x, a), $\mathcal{M}$ always outputs the same prediction. The same holds for $\mathcal{M}'$, although the prediction may be different from the one of $\mathcal{M}$.

Table 4.4 shows some abbreviations and definitions we use in this study. In particular, Δ_a^x denotes the difference between the frequency of the samples with the positive true decision ($Y = 1$) and those with the negative true decision ($Y = 0$), and have $A = a$ and $X = x$. On the other hand, Γ_a^x denotes the difference between the positive and negative decision rates *given* $A = a$ and $X = x$. $\Delta_a'^x$ and $\Gamma_a'^x$ denote the corresponding quantities in the obfuscated training data (i.e., on the samples with $A' = a$ and $X = x$).

In order to reason formally about the impact of privacy on fairness, we need to make a basic assumption about the training algorithm. Namely, we assume that the baseline model, in correspondence of the input (x, a), predicts $\hat{Y} = 1$ if $\Delta_a^x \geq 0$, namely the majority of the tuples in the training set with $X = x$ and $A = a$ have $Y = 1$, and predicts $\hat{Y} = 0$, otherwise. This assumption is quite natural, as, in general, an ML model should opt for the prevailing decision seen in training[15]. We make the same assumption for the LDP model $\mathcal{M}'$ (with A replaced by A'), which is reasonable since $\mathcal{M}$ and $\mathcal{M}'$ are trained with the same algorithm. Formally:

[15]Some learning algorithms like the Nearest Neighbours actually use a generalization of this criterion to produce the prediction.

Table 4.4 Abbreviations and definitions used in this study. $\hat{\mathbb{P}}$ denotes the empirical probability (frequency) on the training set.

Abbreviations
• $\hat{Y}_a^x \in \{0,1\}$: the prediction of the baseline model $\mathcal{M}$ on the input (x,a)
• $\text{CSD}_x = \mathbb{P}[\hat{Y}=1 \mid X=x, A=1] - \mathbb{P}[\hat{Y}=1 \mid X=x, A=0]$: Conditional statistical disparity in $\mathcal{M}$
• $\text{SD} = \mathbb{P}[\hat{Y}=1 \mid A=1] - \mathbb{P}[\hat{Y}=1 \mid A=0]$: statistical disparity in $\mathcal{M}$
• $\text{EOD} = \mathbb{P}[\hat{Y}=1 \mid Y=1, A=1] - \mathbb{P}[\hat{Y}=1 \mid Y=1, A=0]$: equal opportunity difference in $\mathcal{M}$
• $\hat{Y}_a'^x \in \{0,1\}$: the prediction of the LDP model $\mathcal{M}'$ on the input (x,a)
• $\text{CSD}'_x = \mathbb{P}[\hat{Y}'=1 \mid X=x, A=1] - \mathbb{P}[\hat{Y}'=1 \mid X=x, A=0]$: Conditional statistical disparity in $\mathcal{M}'$
• $\text{SD}' = \mathbb{P}[\hat{Y}'=1 \mid A=1] - \mathbb{P}[\hat{Y}'=1 \mid A=0]$: statistical disparity in $\mathcal{M}'$
• $\text{EOD}' = \mathbb{P}[\hat{Y}'=1 \mid Y=1, A=1] - \mathbb{P}[\hat{Y}'=1 \mid Y=1, A=0]$: equal opportunity difference in $\mathcal{M}'$
Definitions
• $\Delta_a^x = \hat{\mathbb{P}}[Y=1, X=x, A=a] - \hat{\mathbb{P}}[Y=0, X=x, A=a]$
• $\Gamma_a^x = \hat{\mathbb{P}}[Y=1 \mid X=x, A=a] - \hat{\mathbb{P}}[Y=0 \mid X=x, A=a]$
• $\Delta_a'^x = \hat{\mathbb{P}}[Y=1, X=x, A'=a] - \hat{\mathbb{P}}[Y=0, X=x, A'=a]$
• $\Gamma_a'^x = \hat{\mathbb{P}}[Y=1 \mid X=x, A'=a] - \hat{\mathbb{P}}[Y=0 \mid X=x, A'=a]$

Assumption 4.4.1 . *The prediction of $\mathcal{M}$ (baseline model) is:*

$$\hat{Y}_a^x = \begin{cases} 1 & \text{if} \quad \Delta_a^x \geq 0 \quad (\text{or, equivalently, } \Gamma_a^x \geq 0), \\ 0 & \text{otherwise.} \end{cases}$$

Assumption 4.4.2 . *The prediction of $\mathcal{M}'$ (LDP model) is:*

$$\hat{Y}_a'^x = \begin{cases} 1 & \text{if} \quad \Delta_a'^x \geq 0 \quad (\text{or, equivalently, } \Gamma_a'^x \geq 0), \\ 0 & \text{otherwise.} \end{cases}$$

The following Lemma relates the difference between the frequencies of positive and negative decisions in the obfuscated and original data. We recall that $p = {e^\varepsilon}/{(e^\varepsilon+1)}$ is the probability that the value reported by RR is the true value.

Lemma 4.4.1 . $\Delta_a'^x = p\,\Delta_a^x + (1-p)\,\Delta_{\bar{a}}^x$.

See proof on page 239.

The following Lemma relates the LDP model's prediction to the original data's statistics. It follows simply by case analysis from Lemma 4.4.1 and Assumption 4.4.2.

Lemma 4.4.2 .

$$\hat{Y}'^{x}_{a} = 1 \quad \textit{if} \quad \begin{cases} \Delta^x_a,\, \Delta^x_{\overline{a}} \geq 0, & \textit{or} \\ \Delta^x_a > 0 \quad \textit{and} \quad \Delta^x_{\overline{a}} < 0 \quad \textit{and} \quad e^\varepsilon \geq -\Delta^x_{\overline{a}}/\Delta^x_a, & \textit{or} \\ \Delta^x_a < 0 \quad \textit{and} \quad \Delta^x_{\overline{a}} > 0 \quad \textit{and} \quad e^\varepsilon \leq -\Delta^x_{\overline{a}}/\Delta^x_a\,. \end{cases}$$

$$\hat{Y}'^{x}_{a} = 0 \quad \textit{if} \quad \begin{cases} \Delta^x_a,\, \Delta^x_{\overline{a}} \leq 0 \quad \textit{and at least one of them is strictly negative}, & \textit{or} \\ \Delta^x_a > 0 \quad \textit{and} \quad \Delta^x_{\overline{a}} < 0 \quad \textit{and} \quad e^\varepsilon < -\Delta^x_{\overline{a}}/\Delta^x_a, & \textit{or} \\ \Delta^x_a < 0 \quad \textit{and} \quad \Delta^x_{\overline{a}} > 0 \quad \textit{and} \quad e^\varepsilon > -\Delta^x_{\overline{a}}/\Delta^x_a\,. \end{cases}$$

Impact of LDP on Conditional Statistical Disparity. In what follows, we analyze the effect of RR on conditional statistical disparity with respect to a specific tuple of values x of the explaining variables. To do so, we compare CSD'_x, which represents the conditional statistical disparity of prediction of the LDP model, with CSD_x which is the one of the baseline model. Following the principle that fairness should be assessed on the true inputs, we define CSD'_x as:

$$\text{CSD}'_x = \mathbb{P}[\hat{Y}' = 1 \mid X = x, A = 1] - \mathbb{P}[\hat{Y}' = 1 \mid X = x, A = 0].$$

Namely, the conditioning is on A and not on A'. Note that, since the models are deterministic, CSD_x and CSD'_x could equivalently be defined as:

$$\text{CSD}_x = \hat{Y}^x_1 - \hat{Y}^x_0 \qquad \text{and} \qquad \text{CSD}'_x = \hat{Y}'^x_1 - \hat{Y}'^x_0\,.$$

The following theorem states the relation between CSD_x and CSD'_x.

Theorem 4.4.1 Impact of LDP on CSD_x.

1. *if* $\text{CSD}_x > 0$ *then* $0 \leq \text{CSD}'_x \leq \text{CSD}_x$
2. *if* $\text{CSD}_x < 0$ *then* $\text{CSD}_x \leq \text{CSD}'_x \leq 0$
3. *if* $\text{CSD}_x = 0$ *then* $\text{CSD}'_x = \text{CSD}_x = 0$

See proof on page 239.

Essentially, the above theorem says that CSD'_x is always sandwiched between CSD_x and 0. Namely, if, in the baseline model, there is discrimination against one group, then obfuscating A tends to reduce the discrimination. It never introduces discrimination against the other group. In one extreme case, it may leave things unchanged, while, in the opposite extreme

case, it may remove the discrimination entirely. If, in the baseline model, we have conditional statistical parity ($\text{CSD}_x = 0$), then obfuscating A maintains this property.

It is important to note that Theorem 4.4.1 does not depend on whether the *unprivileged* group is the minority or the majority of the population.

Impact of LDP on Statistical Disparity. Using the results of CSD_x, in this section, we analyze the impact of privacy on SD by comparing SD' and SD, where SD' is the statistical disparity of the prediction of the LDP model, defined as:

$$\text{SD}' = \mathbb{P}[\hat{Y}' = 1 \mid A = 1] - \mathbb{P}[\hat{Y}' = 1 \mid A = 0]. \tag{4.8}$$

Again, note that we condition on A rather than A'.

We make the following assumption that we call the *uniform discrimination assumption*. Essentially, it says that if one group is discriminated against for some value $x*$ of X, then the other group cannot be discriminated against for other values x of X. This is a natural assumption in real-life scenarios. For example, consider an ML system that tries to predict whether to release an individual on parole, given the type of crime they have committed in the past. If the system (or the historical data in which it is trained) discriminates against an ethnic group in case of a minor crime, it would still discriminate against that same group in case of a major crime, or, at most, be fair. As another example, consider granting an application for a loan: If, for a certain amount of money requested, the applications from an ethnic group are accepted more frequently than those from the other group, it is unlikely that, for a different amount of money, the situation would be inverted.

Formally, the *uniform discrimination assumption* is stated as follows:

Assumption 4.4.3 . *Uniform discrimination assumption*

$$\textit{if } \exists x^* \ \Gamma_a^{x^*} > \Gamma_{\bar{a}}^{x^*} \textit{ then } \forall x \ \ \Gamma_a^x \geq \Gamma_{\bar{a}}^x$$

In the remainder of this section, we differentiate between two scenarios depending on whether X and A are independent. We will denote the case of independency by $X \perp A$, and the case of dependency by $X \not\perp A$[16].

- **First scenario: $\mathbf{X \perp A}$**
  We first consider the case of independency. We start by showing that we can quantitatively express SD in terms of the distribution of the data as follows:

[16] In real-life contexts, X and A are usually dependent.

Lemma 4.4.3 Quantification of SD.

$$\mathrm{SD} = \begin{cases} \mathbb{P}[\Delta_1^X \geq 0 \ \wedge \ \Delta_0^X < 0] & \textit{if } \exists x \ \Gamma_1^x > \Gamma_0^x \\ 0 & \textit{if } \forall x \ \Gamma_1^x = \Gamma_0^x \\ -\mathbb{P}[\Delta_1^X < 0 \ \wedge \ \Delta_0^X \geq 0] & \textit{if } \exists x \ \Gamma_1^x < \Gamma_0^x \end{cases}$$

See proof on page 240.

Analogously, we have:

Lemma 4.4.4 Quantification of SD′.

$$\mathrm{SD}' = \begin{cases} \mathbb{P}[\Delta_1'^X \geq 0 \ \wedge \ \Delta_0'^X < 0] & \textit{if } \exists x \ \Gamma_1'^x > \Gamma_0'^x \\ 0 & \textit{if } \forall x \ \Gamma_1'^x = \Gamma_0'^x \\ -\mathbb{P}[\Delta_1'^X < 0 \ \wedge \ \Delta_0'^X \geq 0] & \textit{if } \exists x \ \Gamma_1'^x < \Gamma_0'^x \end{cases}$$

See proof on page 241.

Using Lemma 4.4.1, by case analysis, the quantification of SD′ can be reformulated in terms of the distribution in the original data, as follows.

Lemma 4.4.5 Quantification of SD′ in terms of the distribution on the original data.

$$\mathrm{SD}' = \begin{cases} \mathbb{P}\left[\begin{array}{l} \Delta_1^X > 0 \ \wedge \ \Delta_0^X < 0 \ \wedge \\ e^\varepsilon \geq -\Delta_0^X/\Delta_1^X \ \wedge \ e^\varepsilon > -\Delta_1^X/\Delta_0^X \end{array}\right] & \textit{if } \exists x \ \Gamma_1^x > \Gamma_0^x \\ 0 & \textit{if } \forall x \ \Gamma_1^x = \Gamma_0^x \\ -\mathbb{P}\left[\begin{array}{l} \Delta_1^X < 0 \ \wedge \ \Delta_0^X > 0 \ \wedge \\ e^\varepsilon > -\Delta_0^X/\Delta_1^X \ \wedge \ e^\varepsilon \geq -\Delta_1^X/\Delta_0^X \end{array}\right] & \textit{if } \exists x \ \Gamma_1^x < \Gamma_0^x \end{cases}$$

We can now state the main result of this section: If $X \perp A$, then, like in the case of conditional statistical disparity, we have that SD′ is always sandwiched between SD and 0.

Theorem 4.4.2 Impact of LDP on SD. Case $X \perp A$.

1. *if* $\mathrm{SD} > 0$ *then* $0 \leq \mathrm{SD}' \leq \mathrm{SD}$
2. *if* $\mathrm{SD} < 0$ *then* $\mathrm{SD} \leq \mathrm{SD}' \leq 0$

3. if $\text{SD} = 0$ *then* $\text{SD}' = \text{SD} = 0$

The proof follows immediately from Lemmas 4.4.3 and 4.4.5 because the values of X that constitute the probability mass in the expression of SD' are a subset of those that constitute the probability mass in the expression of SD.

Discussion: Theorem 4.4.2 means that, from an unfair situation ($\text{SD} > 0$ or $\text{SD} < 0$), obfuscating the sensitive attribute A in general advantages the *unprivileged* group, but it never ends up discriminating the other group. (We will see in the next section that this is not always the case when some proxies to the sensitive attribute A exist in the data.)

In one extreme case, the situation does not change ($\text{SD}' = \text{SD}$). By looking at the expression quantifying SD and SD' in Lemmas 4.4.3 and 4.4.5, we can see that this happens when the noise we inject is small, i.e., for high values of ε, and, more precisely, when ε satisfies $\forall x\ \varepsilon \geq \max\{\ln(-\Delta^x_0/\Delta^x_1), \ln(-\Delta^x_1/\Delta^x_0)\}$.

In the opposite extreme case, the discrimination is totally eliminated ($\text{SD}' = 0$). This last case raises when we inject enough noise, and more precisely, when ε satisfies $\forall x\ \varepsilon < \max\{\ln(-\Delta^x_0/\Delta^x_1), \ln(-\Delta^x_1/\Delta^x_0)\}$.

In all the other cases, i.e., when for some x we have $\varepsilon \geq \max\{\ln(-\Delta^x_0/\Delta^x_1), \ln(-\Delta^x_1/\Delta^x_0)\}$ and for other x we have $\varepsilon < \max\{\ln(-\Delta^x_0/\Delta^x_1), \ln(-\Delta^x_1/\Delta^x_0)\}$, obfuscation removes some discrimination, but not entirely. Namely $0 < \text{SD}' < \text{SD}$ if SD is positive, or $\text{SD} < \text{SD}' < 0$ if SD is negative.

Note that the extreme case in which ε is 0 is equivalent to eliminating A entirely from the data. Hence, the takeout of this section is that the disparity between groups can be eliminated by removing the sensitive attribute, but it is important to remember that this is true only because there are no proxies to the sensitive attribute in the data ($X \perp A$).

Again, we note that Theorem 4.4.2 does not depend on whether the *unprivileged* group is the minority or the majority of the population.

- **Second scenario: $\mathbf{X \not\perp A}$**
  Usually, proxy attributes to the sensitive attribute A exist in the data. In other words, A and X are dependent ($X \not\perp A$). In this section, we study the impact of privacy on SD when $X \not\perp A$. Theorem 4.4.3 presents the results of the impact of privacy on SD in this scenario.

Theorem 4.4.3 Impact of LDP on SD. Case $X \not\perp A$.

1. *if* $\exists x\ \Gamma_1^x > \Gamma_0^x$ *then* $\mathrm{SD}' \leq \mathrm{SD}$
2. *if* $\exists x\ \Gamma_1^x < \Gamma_0^x$ *then* $\mathrm{SD} \leq \mathrm{SD}'$
3. *if* $\forall x\ \Gamma_1^x = \Gamma_0^x$ *then* $\mathrm{SD}' = \mathrm{SD}$

See proof on page 243.

Discussion: Theorem 4.4.3 confirms that also in the case $X \not\perp A$, in general, the *unprivileged* group benefits from privacy, and again, it does not depend on the *privileged* group being the majority or not. This finding is validated by our experiments on both synthetic and real-world datasets (cf. Figs 4.21 and 4.23 in Section 4.4.5).

Theorem 4.4.3 differs from Theorem 4.4.2 mainly on two points. First, SD and SD$'$ can have opposite signs. In other words, from a scenario where there is discrimination against one group, for instance, the group $A = 0$ (SD > 0), we can have, after obfuscation, a discrimination against the other group $A = 1$ (SD$' < 0$). We can even have scenarios in which, after obfuscation, the magnitude of unfairness against the other group is higher than the original one. This result is quite surprising. We simulated such a scenario using synthetic data (S5) and presented the results in Fig. 4.19 (Section 4.4.4).

Second, we note that in case 1 we can have SD < 0 despite the fact that $\exists x\ \Gamma_1^x > \Gamma_0^x$ (which, by the Assumption 4.4.3, implies that $\forall x\ \Gamma_1^x \geq \Gamma_0^x$), and similarly for case 3. From the proof of the above theorem, we can see that it is particularly likely to happen when $\mathbb{P}[X = x|A = 1] \ll \mathbb{P}[X = x|A = 0]$). This is a form of the *Simpson's paradox* called *Association Reversal* [184]: we have a scenario in which for all sub-populations (i.e., for all x) there is discrimination against one group, while when considering the whole population, the discrimination is against the other group. Note that privacy obfuscation does not break the paradox, because also SD$'$ is negative.

Another form of the *Simpson's paradox* called the *Yule's Association Paradox* [63] can happen when for all sub-populations, the model shows fair results (i.e., $\forall x\ \ \mathrm{CSD}_x = 0$), while for the whole population, it shows unfair results (SD$\neq 0$). In Section 4.4.4, we generated a synthetic dataset (S4) to illustrate such a paradox. Note that in this case, the privacy obfuscation has no effect on fairness: all the metrics remain the same. Indeed, if $\forall x\ \ \mathrm{CSD}_x = 0$, then $\forall x\ \ \mathrm{CSD}'_x = 0$, and all the metrics under consideration in this study are based on CSD'_x.

Impact of LDP on Equal Opportunity difference. In what follows, we consider the impact of privacy on EOD (Eq. (4.3)). This notion of fairness, by contrast to SD (Eq. (4.2)), considers, in addition to the prediction $\hat{Y}$, the true decision Y (cf. Eq. (4.3)).

The justification for the EOD as a notion of fairness is that Y is supposed to be reliable and not incorporate any bias (Hardt et al. [101]). Hence, if $\hat{Y}$ is consistent with Y, the prediction should be fair as well. Furthermore, thanks to this compatibility, and in contrast to other notions of fairness, EOD is, in general, going well along with accuracy (although there are exceptions: [189] has shown that, for certain distributions, Equal Opportunity implies trivial accuracy). We capture this principle in Assumption 4.4.4 here below, which states that the true decision Y is independent of the sensitive attribute A given X.

Assumption 4.4.4 . *Reliable Y. The decision Y is independent of the sensitive attribute for any value of the explaining variable. Namely:*

$$\mathbb{P}[Y = 1 \mid X = x, A = 1] = \mathbb{P}[Y = 1 \mid X = x, A = 0].$$

The limitation of EOD is that the "true" Y may not always be available. In its stead, the data may contain decisions that have been made in the past (which may not always have been fair), or decisions based on some proxy to the true Y. In any case, Assumption 4.4.4, may not always be satisfied in the data. When it is satisfied, however, we can obtain a strong result about the effect of privacy on EOD, similar to the one for SD. This is expressed by the theorem below.

Theorem 4.4.4 Impact of LDP on EOD.

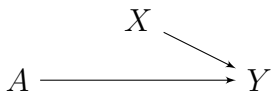
1. *if* $\text{EOD} > 0$ *then* $0 \leq \text{EOD}' \leq \text{EOD}$
2. *if* $\text{EOD} < 0$ *then* $\text{EOD} \leq \text{EOD}' \leq 0$
3. *if* $\text{EOD} = 0$ *then* $\text{EOD}' = \text{EOD} = 0$

See proof on page 244.

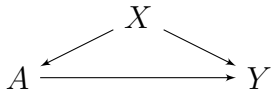
We recall that the above theorem holds under Assumption 4.4.4. On the other hand, it is valid regardless of whether X and A are independent.

4.4.4 Experimental Results and Discussion

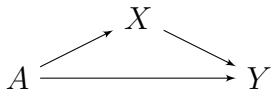
To validate our theoretical results, we have conducted a set of experiments on both synthetic and real-world fairness benchmark datasets. To each of these datasets, the fairness metrics presented in section 4.4.2 are applied to the baseline model $\mathcal{M}$ (model trained on the original samples) and to the LDP model $\mathcal{M}'$ (model trained on the obfuscated samples). Then, to assess the impact of privacy on fairness, in relation to Theorems 4.4.1, 4.4.2, 4.4.3, and



(a) Causal Graph for S1.



(b) Causal Graph for S2 and S6.



(c) Causal Graph for S3, S4 and S5.

Fig. 4.17 Causal graphs of the synthetic datasets.

4.4.4, the predictions of these two models are compared. Recall that the testing samples for both $\mathcal{M}$ and $\mathcal{M}'$ are always kept original without obfuscation. We vary the privacy parameter ε in the $\{16, 8, 2, 1, 0.85, 0.5, 0.4, 0.3, 0.2, 0.1\}$ for the synthetic datasets and in the $\varepsilon = \{16, 8, 5, 4, 3, 2, 1, 0.5\}$ for the real-world datasets. At $\varepsilon = 0.1$ (strong privacy), the ratio of probabilities is bounded by $\varepsilon^{0.1} \approx 1.05$, giving nearly indistinguishable distributions between the two groups, whereas at $\varepsilon = 16$ (weak privacy), the distributions are nearly the same as in the original data.

Data and Experiments.

- **Environment.** All the experiments are implemented in Python 3. We use *Random Forest* model [38] for classification with its default hyper-parameters and randomly select 80% as the training set and the remaining 20% as the testing set. For the RR mechanism, we use the implementation in Multi-Freq-LDPy [13].
- **Stability.** Since LDP protocols, train/test splitting, and ML algorithms are randomized, we report average results over 100 runs.
- **Datasets.** We validate our theoretical results with six synthetic datasets and four real-world datasets.

Synthetic Datasets. The causal graphs used to generate the synthetic datasets are depicted in Figure 4.17, and the joint empirical probabilities (frequencies) for the various

combinations of values are shown in Table 4.5. S1 differs from all other datasets in that X and A are independent, whereas in all other datasets, namely S2-S6, X and A are dependent. A and Y are binary variables while X is a discrete variable. In S1, S2, and S4, X is also a binary variable. S3 and S5 are generated to simulate the scenario where privacy shifts discrimination between groups. S5 shows an extreme scenario where $|\text{SD}'| > |\text{SD}|$, while $|\text{SD}'| < |\text{SD}|$ in S3. And finally, S4 includes a case of *Yule's Association Paradox* [63] (Section 4.4.3).

Table 4.5 Distributions of the synthetic datasets.

(a) S1.

$Y=1$	$X=0$	$X=1$
$A=1$	0.35	0.35
$A=0$	0	0.15
$Y=0$	$X=0$	$X=1$
$A=1$	0	0
$A=0$	0.15	0

(b) S2.

$Y=1$	$X=0$	$X=1$
$A=1$	0.28	0.38
$A=0$	0	0.12
$Y=0$	$X=0$	$X=1$
$A=1$	0	0
$A=0$	0.22	0

(c) S3.

$Y=1$	$X=0$	$X=1$	$X=2$
$A=1$	0.03	0.17	0.03
$A=0$	0	0.17	0.03
$Y=0$	$X=0$	$X=1$	$X=2$
$A=1$	0.24	0.03	0
$A=0$	0.1	0.2	0

(d) S4.

$Y=1$	$X=0$	$X=1$
$A=1$	0	0.4
$A=0$	0.03	0.34
$Y=0$	$X=0$	$X=1$
$A=1$	0.03	0.07
$A=0$	0.13	0

(e) S5.

$Y=1$	$X=0$	$X=1$	$X=2$
$A=1$	0.03	0.17	0.03
$A=0$	0	0.17	0.03
$Y=0$	$X=0$	$X=1$	$X=2$
$A=1$	0.24	0.03	0
$A=0$	0.03	0.27	0

(f) S6.

$Y=1$	$X=0$	$X=1$	$X=2$	$X=3$	$X=4$
$A=1$	0.05	0.08	0.09	0.13	0.14
$A=0$	0.02	0.03	0.06	0.03	0.04
$Y=0$	$X=0$	$X=1$	$X=2$	$X=3$	$X=4$
$A=1$	0.04	0.02	0.01	0.06	0
$A=0$	0.06	0.04	0.02	0.08	0

In the following plots, the ∞ symbol in the x-axis in each plot shows the fairness values when the model is trained on original samples (no privacy).

Fig. 4.18 shows the obtained results for S1-S4 presented above while S5 and S6 results are depicted in Fig. 4.19 and Fig. 4.20, respectively. For example, in S1, where some fairness measures show fair results in the baseline model $\mathcal{M}$, namely EOD and CSD_1, enforcing privacy helped maintain these fair results: $\text{SD}' = 0$ and $\text{CSD}'_1 = 0$. However, some fairness measures show unfair results against group $A = 0$ in the baseline model, namely SD, and

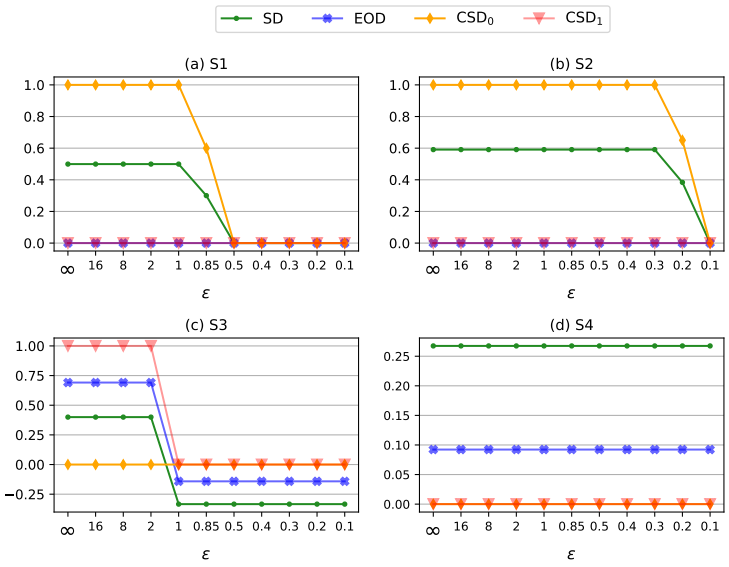


Fig. 4.18 Results for the synthetic dataset S1-S4, illustrating the impact of LDP on fairness (y-axis) for privacy level ε (x-axis). Note that in S3 we have $X \not\perp A$ and the fairness measure SD is inverted after obfuscation. Also, EOD is inverted after obfuscation. This is because Assumption 4.4.4 is not verified in this dataset. S4 illustrates Yule's Association Paradox, a variant of the *Simpson's paradox*. The fairness values on the original data (no privacy) are the values for $\varepsilon = \infty$.

CSD_0; thus, enforcing privacy removed discrimination when enough noise is added. In particular, at $\varepsilon = \ln(-\Delta_0^x/\Delta_1^x) = 0.85$, SD$'$ and CSD_0' values started to decrease and continued to decrease reaching full parity between groups.

As we proved theoretically in Theorem 4.4.3, and explained in Section 4.4.3, in S3 and S5 and from a scenario where SD and EOD show discrimination against the group $A = 0$, by applying privacy, the discrimination became against the other group $A = 1$. Note that this does not contradict Theorem 4.4.4, because S3 and S5 do not verify Assumption 4.4.4[17]. For S3, although this inversion of fairness conclusions (discrimination switching from one group to another when applying privacy), the disparity after obfuscation decreased: $|\text{SD}'| < |\text{SD}|$ ($|\text{SD}'| = 0.33$ and $|\text{SD}| = 0.39$). However, S5 (Fig. 4.19) shows an extreme case where the disparity between groups after obfuscation increased: $|\text{SD}'| > |\text{SD}|$ ($|\text{SD}'| = 0.46$ and $|\text{SD}| = 0.39$). In other words, not only has discrimination switched from one group to another after obfuscation, but the level of unfairness has also increased.

S4 shows a case of the *Yule's Association Paradox* [63], a variant of the *Simpson's paradox*. That is, the model $\mathcal{M}$ shows fair results for all sub-populations: $\text{CSD}_0 = \text{CSD}_1 = 0$. However, $\mathcal{M}$ shows unfair results for the whole population: SD$= 0.26$. As shown in Figure 4.18(d),

[17]We provide in Appendix B.2.2 a dataset called S7 that satisfies Assumption 4.4.4.

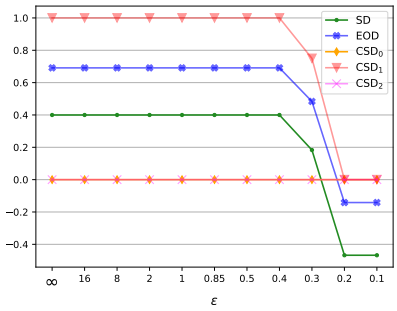


Fig. 4.19 Results for the synthetic dataset S5. Note that EOD is also inverted here after obfuscation. Again, this is because Assumption 4.4.4 is not verified in this dataset.

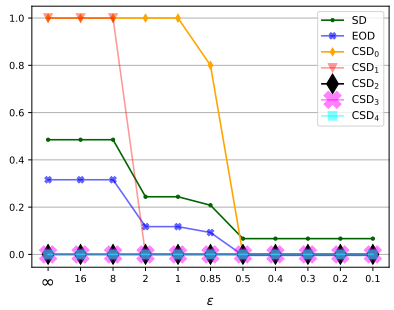


Fig. 4.20 Results for the synthetic dataset S6.

the paradox stayed even under a strong privacy regime ($\varepsilon = 0.1$) and hence obfuscating the sensitive attribute solely didn't remove the paradox from the data.

To better understand how privacy impacts fairness, the plots in Fig. 4.21 show the impact of privacy on $\mathbb{P}[\hat{Y} = 1 \mid A = a]$ and $\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = a]$ for both groups $A = 1$ and $A = 0$ while varying ε.

As mentioned in Section 4.4.3, the *unprivileged* group $A = 0$ benefits more from privacy. In other words, when obfuscating the sensitive attribute and aligning with our Theorems 4.4.1-4.4.4, the results of the acceptance rates and the true positive rates of the *unprivileged* group tend to increase. For instance, for all the synthetic datasets, it is clear that it is group $A = 0$ who advantages from privacy as shown in Fig. 4.21. In other words, there is an increasing trend of $\mathbb{P}[\hat{Y} = 1 \mid A = 0]$ and $\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 0]$.

Real-World Datasets. We consider the following four real-world datasets:

- *Compas* [10]: This dataset is already presented in Section 4.3.3. We recall that race ($A = 1$ for non-black individuals and $A = 0$ for black individuals) is considered as the sensitive attribute, and the risk of recidivism is the true decision. In this work, $Y = 1$ designates a low-risk recidivism score, while $Y = 0$ denotes a high-risk score. The number of priors of an individual is used as an explaining variable to compute CSD_x

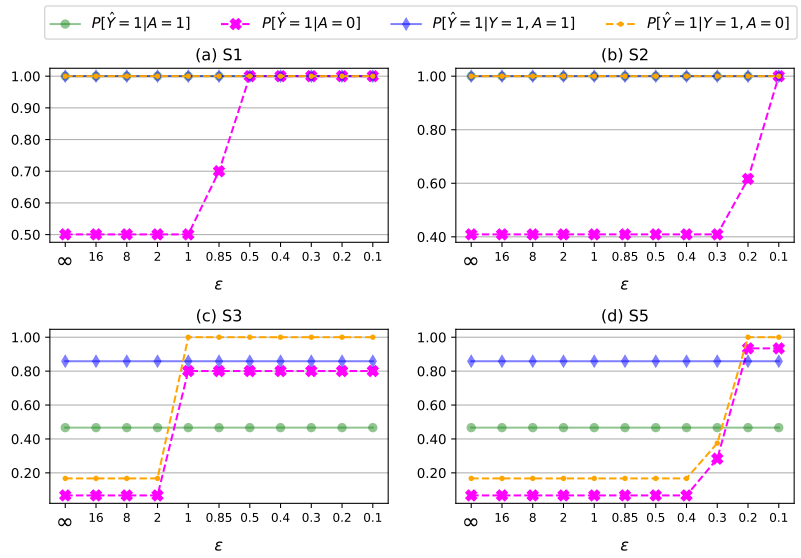


Fig. 4.21 Impact of LDP on disparity (y-axis) by varying the privacy level ε (x-axis) showing the behavior of fairness measures on groups separately when applying privacy. For readability, only SD and EOD are illustrated. Results for the synthetic datasets S1, S2, S3, and S5.

where $X = 1$ denotes a high number of priors, and $X = 0$ denotes a low number of priors.

- *Adult* [70]: This dataset is already presented in Section 4.3.3. We recall that gender is the sensitive attribute ($A = 1$ for men and $A = 0$ for women), and income is the true decision where $Y = 1$ designates a high income while $Y = 0$ denotes a low income. Education level is the attribute used as an explaining variable to compute CSD_x where $X = 1$ denotes a high education level, and $X = 0$ denotes a low education level.

- *German credit* [73]: This dataset includes data of 1000 individuals applying for loans. This dataset is designed for binary classification to predict whether an individual will default on the loan ($Y = 0$) or not ($Y = 1$) based on personal attributes such as gender, job, credit amount, credit history, etc. We consider gender the sensitive attribute where female applicants ($A = 0$) are compared to male applicants ($A = 1$). Credit history is the explaining attribute used to compute CSD_x where $X = 1$ denotes an applicant who has duly repaid in the past while $X = 0$ denotes a critical account for which the applicant has had late payments and/or defaults in the past.

- *LSAC* [250]: This dataset is already presented in Section 4.2.3. We recall that the sensitive attribute is race ($A = 0$ for blacks and $A = 1$ for other ethnic groups), and the true decision is "pass bar", which indicates whether a candidate has successfully passed the bar exam ($Y = 1$) or not ($Y = 0$). The explaining variable is the undergraduate

GPA score of an applicant where $X = 1$ indicates a high GPA and $X = 0$ denotes a low GPA.

The real-world datasets' distributions are shown in Table 4.6.

Table 4.6 Distributions of the real-world datasets.

(a) Compas.

$Y=1$	$X=0$	$X=1$
$A=1$	0.12	0.03
$A=0$	0.06	0.03
$Y=0$		
$A=1$	0.15	0.1
$A=0$	0.25	0.26

(b) Adult.

$Y=1$	$X=0$	$X=1$
$A=1$	0.06	0.53
$A=0$	0.02	0.21
$Y=0$		
$A=1$	0.03	0.06
$A=0$	0.02	0.07

(c) German credit.

$Y=1$	$X=0$	$X=1$
$A=1$	0.23	0.27
$A=0$	0.08	0.13
$Y=0$		
$A=1$	0.06	0.13
$A=0$	0.01	0.09

(d) LSAC.

$Y=1$	$X=0$	$X=1$
$A=1$	0.43	0.47
$A=0$	0.03	0.01
$Y=0$		
$A=1$	0.02	0.02
$A=0$	0.01	0.01

Fig. 4.22 shows the results of applying privacy on the four real-world datasets. As with the synthetic datasets and in alignment with our proofs, obfuscating the sensitive attribute tends to improve the fairness metrics considered in this study in all the datasets *except the German credit* one (we will discuss this latter case below). We believe that this is due to the fact that the real-world datasets do not always follow the "ideal" situation represented by our assumptions. In particular, the datasets we are considering contain other variables besides the X that we use as an explaining variable, which can influence the prediction.

For instance, in the *Compas* dataset, starting from discrimination against black individuals ($A = 0$), privacy reduced the disparity from SD = 0.21 to 0.15. Similarly, privacy decreased discrimination against black individuals from SD = 0.13 to SD = 0.09 in the *LSAC* dataset, and a similar decrease pattern is observed for all the other fairness measures. The *Adult* dataset also shows a slight disparity decrease caused by privacy. However, starting from a very high disparity between groups given a low level of education ($\text{CSD}_0 = 0.39$), the disparity is reduced to 0.22 at $\varepsilon = 0.1$.

Concerning the *German credit* dataset, the results show an unstable trend. This is because this data set does not satisfy the *uniform discrimination* assumption (Assumption 4.4.3).

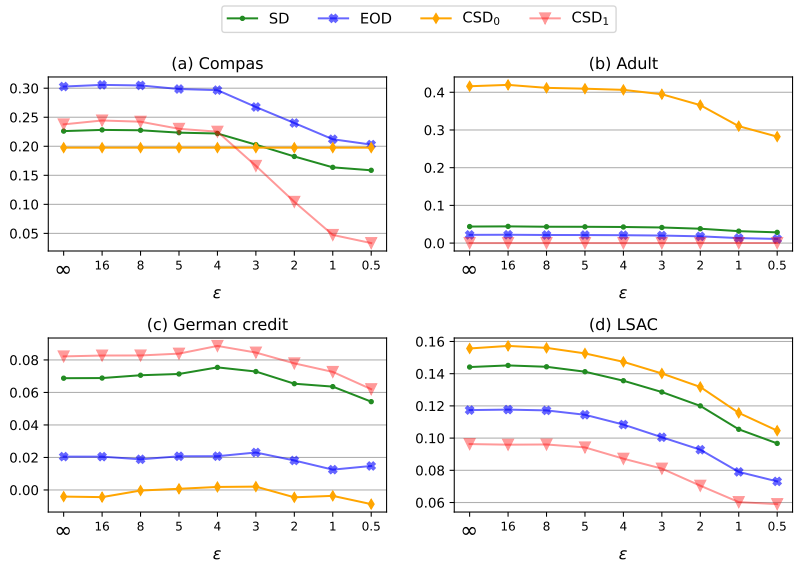


Fig. 4.22 Results for the real-world datasets. The *German credit* dataset does not satisfy Assumption 4.4.3, which explains its unstable behavior.

Indeed, we for $X = 0$, we have, for group $A = 1$:

$$\begin{aligned}\Gamma_1^0 &= \mathbb{P}[Y = 1 \mid X = 0, A = 1] - \mathbb{P}[Y = 0 \mid X = 0, A = 1] \\ &= \frac{0.23}{0.29} - \frac{0.06}{0.29} \\ &\approx 0.58\end{aligned}$$

while for the same $X = 0$, for group $A = 0$ we have:

$$\begin{aligned}\Gamma_0^0 &= \mathbb{P}[Y = 1 \mid X = 0, A = 0] - \mathbb{P}[Y = 0 \mid X = 0, A = 0] \\ &= \frac{0.08}{0.09} - \frac{0.01}{0.09} \\ &\approx 0.77\end{aligned}$$

Hence $\Gamma_1^0 < \Gamma_0^0$.
On the other hand, for $X = 1$ and group $A = 1$ we have:

$$\begin{aligned}\Gamma_1^1 &= \mathbb{P}[Y = 1 \mid X = 1, A = 1] - \mathbb{P}[Y = 0 \mid X = 1, A = 1] \\ &= \frac{0.27}{0.40} - \frac{0.13}{0.40} \\ &\approx 0.35\end{aligned}$$

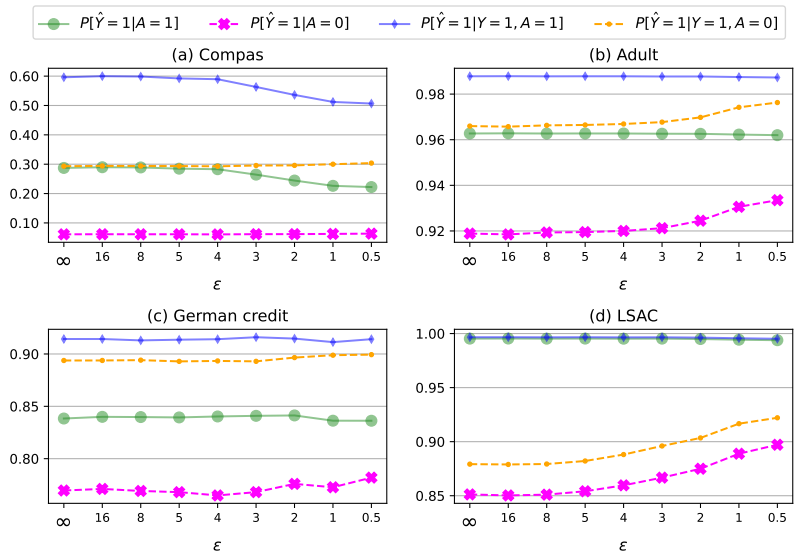


Fig. 4.23 Impact of LDP on disparity (y-axis) by varying the privacy level ε (x-axis) showing the behavior of fairness measures on groups separately when applying privacy. For readability, only SD and EOD are illustrated. Results for the real-world datasets.

while for the same $X = 1$, for group $A = 0$ we have:

$$\begin{aligned}\Gamma_0^1 &= \mathbb{P}[Y = 1 \mid X = 1, A = 0] - \mathbb{P}[Y = 0 \mid X = 1, A = 0] \\ &= \frac{0.13}{0.22} - \frac{0.09}{0.22} \\ &\approx 0.18\end{aligned}$$

Hence, $\Gamma_1^1 > \Gamma_0^1$, which means that the *German credit* dataset does not satisfy Assumption 4.4.3 It may also mean that the attribute "Credit history" is badly chosen as an explaining variable, and/or that it is not the main attribute influencing the decision.

To better understand how privacy impacts fairness, the plots in Fig. 4.23 show how the impact of privacy on $\mathbb{P}[\hat{Y} = 1 \mid A = a]$ and $\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = a]$ for both groups $A = 1$ and $A = 0$ while varying ε.

For instance, for the *Adult* dataset, we can observe that women's acceptance rate ($\mathbb{P}[\hat{Y} = 1 \mid A = 0]$) and true positive rate increased ($\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 0]$) from 0.91 to 0.93 and from 0.96 to 0.99, respectively. However, no change is observed for men ($A = 1$) even at strong privacy ($\varepsilon = 0.5$). A similar behavior is observed for the *LSAC* dataset. For the *Compas* dataset, while no change is observed for the black defendants' ($A = 0$) rates, a decrease is observed for the non-black defendants ($A = 1$). Similar behavior is also observed for the *German credit* dataset, where a slight increase in the acceptance rate and the true positive rate for women is observed while almost no change is observed for men.

LDP Impact on Model Accuracy. Figs 4.24 and 4.25 illustrate the impact of LDP on the accuracy of the model for the synthetic datasets and the real-world datasets, respectively. From these figures, one can note that, in general, the impact of obfuscating the sensitive attribute on model accuracy of the real-world datasets is minor. The drop in the utility is more apparent for the synthetic datasets but remains reasonable, with a maximum drop of 0.2 in S2 when $\varepsilon = 0.1$.

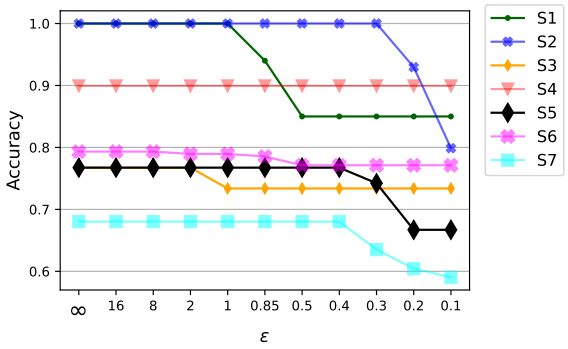


Fig. 4.24 Impact of LDP on the model accuracy for the synthetic datasets.

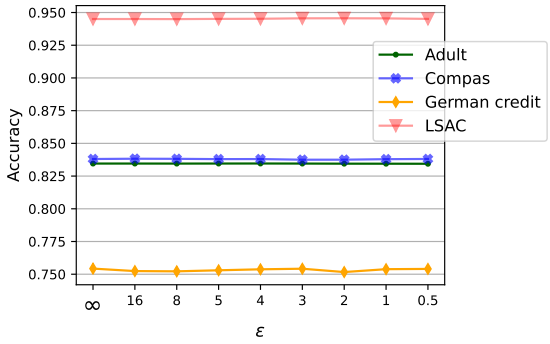


Fig. 4.25 Impact of LDP on the model accuracy for the real-world datasets.

4.4.5 Conclusion

This study formally examines how LDP affects fairness. More specifically, we provide bounds in terms of the joint distributions and the privacy level, delimiting the extent to which LDP can impact the fairness of the model. Our findings show that the *unprivileged* group benefits more than the *privileged* group when injecting enough noise into the sensitive attribute. Furthermore, for conditional statistical disparity and for equal opportunity difference, injecting noise, in general, improves fairness. This also holds for statistical disparity when the data contain no proxies to the sensitive attribute. However, when the data contains proxies, in

certain cases, by injecting enough noise, while the discrimination was originally against one group, it may be shifted to the other group after obfuscation, and the level of unfairness may be worse than before. Note that none of our results depend on whether the *unprivileged* group is the minority or the majority. Additionally, our work focuses on the RR mechanism, a fundamental LDP protocol [115] that serves as a building block for more complex LDP mechanisms (e.g., [26, 245, 79]).

In future work, we aim to extend our work to more fairness measures, particularly overall accuracy difference and others. We also believe that hiding only the sensitive attribute is crucial but not sufficient because proxies for this attribute may exist in the data and thus reveal sensitive information. Therefore, we plan to study formally the impact of LDP on multidimensional data.

4.5 Conclusion

This chapter explored the multifaceted relationship between privacy and fairness in ML through three research studies included in this dissertation. First, we presented two empirical studies (Sections 4.2 and 4.3) investigating the effects of applying LDP mechanisms to multidimensional data on fairness outcomes. These studies demonstrated that LDP slightly improves fairness and does not significantly impair utility. Moreover, our experimental analysis revealed very relevant observations that we framed as concrete recommendations for ML practitioners aiming to guarantee both ethical privacy and fairness concerns.

Next, we complemented our empirical findings with a foundational theoretical study (Section 4.4). This analysis delved into the underlying principles governing the interplay between LDP-preserving techniques and fairness metrics. Based on our theoretical study, we were able to elucidate the conditions under which the considered LDP mechanism, namely RR, either enhances or undermines fairness, thus bridging critical gaps identified in our empirical research.

Together, these studies contribute to a deeper understanding of the impact of LDP on fairness, providing valuable guidelines for practitioners and researchers alike. The next chapter shifts focus to another ethical AI principle: explainability. Specifically, we present our work on causal discovery from a fairness perspective. Causality, particularly causal discovery, is closely linked to explainability as it helps identify the root causes behind AI decisions or predictions. Therefore, our work on causality can be seen as a significant contribution to explainable AI. This upcoming chapter will first examine causal discovery approaches, their impact on fairness, and, most importantly, how slight differences between causal models can significantly impact fairness/discrimination conclusions (Section 5.2). Second, we will observe how causal discovery algorithms are impacted by the type of causal structures and the amount of injected bias in the data (Section 5.3).

Chapter 5

Causal Discovery Through the Lens of Fairness

5.1 Introduction

In recent years, the integration of causal discovery (CD) into various fields has gained significant attention, particularly in the context of ensuring fairness in decision-making processes. Understanding the causal relationships within data is crucial for identifying and mitigating bias, thereby enhancing the fairness of outcomes. This chapter delves into how CD algorithms impact fairness conclusions, examining the nuanced interplay between causal models and fairness assessments.

The first section explores various CD approaches and their implications for fairness in decision-making processes. CD methods, such as structural equation modeling and causal Bayesian networks, provide frameworks for uncovering the underlying causal structures within data. However, we highlight how slight differences between causal models can significantly impact conclusions regarding fairness and discrimination. By examining case studies and experimental results, we will demonstrate how these differences arise and discuss how they affect fairness assessments, thus emphasizing the critical role that precise causal modeling plays in determining fairness outcomes.

The second section focuses on how different causal structures and varying data bias levels impact the performance of CD algorithms and, consequently, fairness conclusions. We also examine the effect of the outcome variable binarization threshold on the discovered causal graph and, consequently, on fairness conclusions.

5.2 Causal Discovery for Fairness

The main impediment to causal inference is the unavailability of the true causal graph which indicates the causal relations between variables. Causal graphs can be set manually by field experts but are often generated using experiments (also called interventions). Identifying the causal graph is called CD or structural learning. As was stated in Section 1.1.3, RCTs are the gold standard of CD. However, RCTs are generally not feasible for practical, ethical, and scalability reasons. As an alternative to RCT, CD is typically done using statistical tests on observable data. However, even assuming the availability of an oracle that returns answers about conditional independencies in the data, a CD procedure can still be undecided about the causal graph. A large number of CD algorithms exist in the literature. Most of these algorithms fall into three categories: constraint-based, score-based, and procedures that exploit semi-parametric assumptions. We introduced some of them in Section 2.4.3. In the constraint-based category, algorithms rely mainly on the (conditional) independencies present in the data to discover causal relations between variables, as explained in the previous paragraph. Therefore, their efficiency depends on the reliability of the conditional independence test procedure. Score-based algorithms rely instead on goodness-of-fit tests. They learn causal graphs by maximizing a scoring criterion such as the Bayesian Information Criterion (BIC) [203], which trades off accuracy (fitness of graph to the data) with complexity (the number of parameters in the model). The most common assumptions relevant to the third category are the linearity of the model and the non-gaussianity of the regression residuals. As described, algorithms in the first two categories do not make strong assumptions about the parametric form or functions of causal connections. Therefore, they can be, theoretically, applied to many more scenarios than the third category. However, most available implementations of constraint-based and score-based CD algorithms model variables as multivariate Gaussian mixture, which implies linearity and Gaussianity of all continuous variables. Causal graphs returned by algorithms in the third category are more accurate than those of the two first categories, which are simply Markov equivalence classes.

This study studies the problem of discovering causal graphs to assess the fairness of ML-based decision systems.

Contributions. First, we discuss relevant details about applying the CD algorithms (that have been introduced in Section 2.4.3) in practice, including their assumptions, data types, and (conditional) independence tests. Second, we carry out an experimental analysis to illustrate the impact of the CD procedure on the structure of the causal graph and, consequently, on fairness conclusions.

Outline. The rest of this section is organized as follows. Section 5.2.1 discusses related work. Next, Section 5.2.2 discusses the applicability of CD algorithms in practice. Section 5.2.3 details the experimental setting and main results. Finally, we conclude this work indicating future perspectives in Section 5.2.4.

5.2.1 Related work

Several survey papers on CD can be found in the literature [98, 159, 220, 100, 263, 50, 174, 175]. Most of these surveys use the same classification that is used in this study, namely, constraint-based, score-based, and those relying on additional semi-parametric assumptions. Glymour et al. [98] review CD algorithms and their application in the biology and neurosciences fields. In particular, they provide general guidelines for their applicability in practice. Malinsky and Danks [159] focus more on their application for problems in Philosophy. The recent survey of Cheng et al. [50] mainly addresses CD approaches' evaluation procedures. Unlike existing surveys, this study provides a detailed but concise description of the major CD algorithms (i.e. PC, FCI, GES, and LiNGAM) and features a comparative empirical analysis. More importantly, it tackles the CD problem in the context of fairness.

Existing causality-based fairness approaches in the literature clear up the causal graph problem in two ways. Either they assume that the causal graph is known [170, 51, 136] or they use the available online implementations [194, 271, 278] of existing CD algorithms [275, 254, 253, 107, 274, 258]. Both ways are akin to skipping the important step of CD from observable data and its impact on the fairness conclusions.

This study considers major CD algorithms and illustrates the importance of the (different) graph structures on causality-based fairness notions (introduced in Section 3.4.2).

5.2.2 Applicability of Causal Discovery Algorithms

Not all algorithms can be used to discover causal relations in a given observed data. The type of the data variables (e.g., continuous vs. categorical), the type of the structural functions between variables (e.g., linear vs. non-linear), and the distribution of the noise (e.g., gaussian vs. uniform) are used to tell which algorithms can/cannot be used.

PC and FCI. PC algorithm requires three assumptions to hold: causal Markov condition (Eq. (2.6)), causal faithfulness (Section 2.4.2), and causal sufficiency (Section 2.4.2). Initially, PC was designed to take as input either entirely continuous or entirely discrete data. However, current implementations allow mixed data via the Conditional Gaussian test.

Because FCI is a PC variant, the same assumptions hold for FCI, except for causal sufficiency, which allows FCI to work in the presence of hidden confounders.

The conditional independence tests used to discover the skeleton of the graph for both PC and FCI have an α value for rejecting the null hypothesis, which is always a hypothesis of independence or conditional independence. For continuous variables, PC uses conditional Pearson correlation [236] (if the functional relations are linear and the data distribution is normal) or K-CI [270] (if no assumptions are made on the type of functions). For categorical

variables, PC uses either a chi-square or G likelihood ratio. The default value of α is 0.01. However, for categorical data, using a value of 0.05 is recommended[1].

The PC algorithm has been proven to be efficient for sparse graphs. Most of the PC processing time is spent in the first phase of skeleton identification. The obvious strategy is to test all possible conditional independence relations for each pair of variables (X and Y). A naive implementation will go through all possible subsets of variables, that is, $2^{|V|}$ subset, where $|V|$ is the number of variables. However, only subsets composed of adjacent variables to X and Y must be considered in practice. Overall, the complexity of PC is $\mathcal{O}(|V|^{d_{max}})$ where d_{max} denotes the maximal node degree in the graph [216]. The degree of a node X is the number of nodes adjacent to X. Hence, PC's efficiency depends heavily on the number of variables but, most importantly, on the sparsity of the causal relations.

GES. GES makes the same assumptions as PC and FCI: causal Markov condition, faithfulness, and sufficiency. For the type of data, the formulation of GES is very general; hence, it works for categorical, continuous, and mixed data. However, most of the theory about the statistical guarantees of the algorithm assumes joint gaussianity of the continuous variables. For instance, Chickering, in the original GES paper [52], defines GES for datasets in which all the variables are categorical (multinomial[2]) or all the variables are continuous and follow a joint Gaussian distribution. We consider the more general case of mixed data [9] because it captures the assumptions used in both continuous and discrete cases. The conditional Gaussian score calculates conditional Gaussian mixtures using the ratios of joint distributions. It makes the following assumptions: (A1) The continuous data were generated from a single joint (multivariate) Gaussian mixture where each Gaussian component exists for a particular setting of the discrete variables. (A2) The instances in the data are independent and identically distributed (iid). (A3) All Gaussian mixtures are approximately Gaussian.

It is also crucial in practice to have enough samples when conditioning on several categorical variables simultaneously, especially when these variables are parents of the same variable. This and A2 are the most relevant requirements in the all-discrete case. Regarding continuous variables, A1 implicitly implies linearity and Gaussianity of the residuals because of the nature of each Gaussian component; this is the most relevant assumption in the all-continuous case. It is important to mention that in the existing standard implementations of GES, such as Tetrad's fges[194] (written in Java), the pcalg[117, 103] library for R (written in C++ underneath), and other Python implementations [94, 116], the predictive models are pre-configured to linear regression for continuous variables. These assumptions may sound too strict, but the empirical evidence of several articles shows that GES performs well even when this assumptions do not hold exactly [9, 52, 103].

[1] https://cmu-phil.github.io/tetrad/manual/.

[2] The distribution of N i.i.d. categorical samples is multinomial.

In terms of complexity, the runtime of the worst-case scenario is upper bounded by $O(|V|^4 k \cdot \max(|V|\, k^2,\, n))$, where n is the number of samples, $|V|$ is the number of variables and k is the maximum number of parents a node can have (bounded by $|V|$ in general). This complexity can be decomposed as follows: (i) the algorithm visits at most $|V|(|V|-1)$ states because each transition adds or removes exactly one edge; (ii) each state has at most $|V|^2$ neighboring states because of the maximum number of edges that can be added or removed to a DAG; (iii) computing each neighboring state takes time $O(|E| \cdot k^2)$ [52] where E is the number of edges at the state (bounded by $|V| \cdot k$); and (iv) computing the BIC score difference between the state and one of its neighbors takes $O(n\, k)$ (assuming the continuous case). Therefore, the worst-case complexity is upper bounded by $O(|V|^4(|V|\, k^3 + n\, k)) = O(|V|^4 k \cdot \max(|V|\, k^2, n))$.

Consequently, GES should preferably be used with datasets consisting of few columns (its runtime grows with $|V|^5 k^3$) and many rows (linear w.r.t. n).

LiNGAM. Unlike PC, FCI, and GES, the assumption of causal faithfulness is not required in LiNGAM. Instead, directLiNGAM assumes that the data is continuous, the functional relations between variables are linear, and most importantly, with non-Gaussian noise terms [206]. The linearity assumption is important because the LiNGAM algorithm incapsulates and fits a linear regression model. It is possible to apply Direct LiNGAM despite the violations in linearity [208]. However, in such a case, the results should be interpreted cautiously, as the algorithm can fail to identify causal connections because of under-fitting. Linearity can be judged by simply eye-balling the pair-wise plots of the data.

The assumption of non-Gaussianity of the error terms is a crucial requirement of the algorithm, allowing it to determine causal directions. However, it cannot be tested *before* fitting the linear regression model and plotting out the error terms. An indication of the non-Gaussian distribution of the error terms in a linear model can be suspected if the distributions of the variables are strongly non-Gaussian. The distribution of the variables can be checked by plotting the histograms or applying Q-Q[3] tests. The exogenous variables can be Gaussian and have non-Gaussian error terms. However, they are discovered only *after* applying the model, so post-modeling testing of the compliance with the assumptions is recommended.

Most of the processing time of directLiNGAM is spent on the computation of residuals. The other heavy processing step is the regression to estimate the model parameters. According to Shimizu et al. [208], the total complexity of directLiNGAM algorithm is $\mathcal{O}(sn^3M^2 + n^4M^3)$, where s is the number of samples, n is the number of variables and $M(\ll s)$ is the maximal rank found by the low-rank decomposition used in the independence measure [208]. Alternatively, using prior knowledge can significantly reduce the complexity of residual computation.

[3] Q–Q (quantile-quantile) plot is a probability plot, which allows to graphically compare two probability distributions by plotting their quantiles against each other.

5.2.3 Experimental Analysis

To study the impact of the CD task on fairness, we apply the different CD algorithms on two synthetic datasets and six real-world fairness benchmark datasets. Table 5.1 provides a summary of all datasets. We use Tetrad [194] implementation of PC, FCI, and GES algorithms with a significance threshold (α) set to 0.01 for conditional independence testing. Different CI tests depend on the input data type and the search algorithm. For instance, *conditional Gaussian likelihood ratio test* and *conditional Gaussian score* are used for mixed data. For continuous data, *K-CI test* and *BIC score* are applied. For LiNGAM, we use the differences in Mutual Information for independence testing. Since the LiNGAM algorithm aims to establish causal order, it is determined by collecting an ordered ascending list of independence scores, the smallest corresponding to most exogenous variables. In the second phase of CD, where the graph is refined by estimating connection weights, we set a threshold (α) to 0.05 to exclude the connections with insignificant weights.

The only background knowledge we use in this study is temporal order using tiers. Variables are split into a set of ordered tiers (tier1, tier2, ... tier$n$) which imply the following constraints. A variable in tier$i$ can be the cause of variables in the same tier or in subsequent tiers ($i+1 \dots n$) but not of variables in previous tiers ($1 \dots i-1$).

With the presence of the causal graph, several causality-based fairness notions can be used to assess fairness [158]. Some qualitative notions can be applied by checking the structure of the graph. For instance, to tell if there is (or not) discrimination according to the "no unresolved discrimination" notion [127], one needs to check if there is a directed path from the sensitive attribute A to the outcome Y which does not go through a resolving (explaining) variable. Discrimination is concluded without further computation if such a path exists in the graph. A similar graph structure checking is needed for the "no proxy discrimination" [127]. For other quantitative fairness notions, the graph's structure is needed to distinguish between confounder, mediator, and collider variables. Quantitative fairness notions are typically computed by adjusting on variables. Adjusting on confounders allows the blocking of spurious/backdoor paths. Adjusting on mediators is needed for mediation analysis (direct vs. indirect vs. path-specific discrimination). Identifying colliders, however, allows us to avoid adjusting on them as this will introduce dependence that doesn't exist between variables.

We use five different causality-based fairness notions, namely, ATE_{IPW} (Eq. (3.66)), total effect (TE) (Eq. (3.24)), direct effect (DE) (Eq. (3.40)), indirect discrimination (ID) (Eq. (3.41)), and explainable discrimination (ED) (Eq. (3.41)). ID and ED compute both the indirect causal effect between the sensitive variable and the outcome. However, ID measures the path-specific effect with a proxy/redlining variable, while ED considers the path-specific effect with an explaining variable. Thus, while the first is discriminatory, the

Table 5.1 Characteristics of the datasets used for the structural learning.

Dataset	*Sample*	*Data type*	*Sensitive*	*Outcome*
Synthetic data (1)	10000	continuous	-	-
Synthetic data (2)	10000	continuous	-	-
Compas	5915	mixed	race	recidivism
Adult	32561	mixed	race	income
German credit	1000	mixed	sex	default
Dutch census	60420	mixed	sex	occupation
Boston housing	506	continuous	race	median price
Comm. & crime	1994	continuous	race	violent crime rate

second is legitimate and, hence, should be removed from the causal effect estimation. The *paths* package implementation [279] is used to estimate TE, DE, ID, and ED.

Computing (or estimating) discrimination using causality-based fairness notions consists of subtracting the probability of positive (desirable) output (e.g., hiring, granting a loan, etc.) for the unprivileged group (e.g., female) from the probability of positive output of the privileged group (e.g., male) as expressed in Eq. (3.24)). This leads to values in the range $[-1, +1]$. A value of 0 means the outcome is fair (no discrimination), a positive value indicates a discrimination *against* the protected group and a negative value indicates a discrimination *in favor* of the protected group.

Estimating discrimination using all the above measures requires the knowledge of the confounder and mediator variables. However, PC, FCI, and GES algorithms can output partially directed graphs (PDAG), which do not guarantee that a certain variable is a confounder or mediator since some edges are left undirected. In such cases, we consider all possible ways of directing the (typically few) undirected edges[4]. For instance, if there are two undirected edges $X - W$ and $Z - Y$, there are 4 ways of directing them: $X \to W$ and $Z \to Y$, $X \leftarrow W$ and $Z \to Y$, $X \to W$ and $Z \leftarrow Y$, and $X \leftarrow W$ and $Z \leftarrow Y$. For each combination, we compute the discrimination, and finally, we report the range of values. This can be seen as bounding the discrimination value.

Synthetic Linear Dataset. In general, synthetic datasets are crucial for testing CD algorithms systematically because, unlike real-world datasets, the ground truth graph is known and indisputable. Here, we use synthetic datasets to illustrate the main differences and characteristics of CD algorithms.

We generated two continuous linear datasets that have a very simple causal structure but are rich enough for analyzing and discussing the algorithms. Fig. 5.1 shows the six variables and their causal relationships. The first dataset uses Gaussian noise and the second

[4]As long as they don't introduce a v-structure.

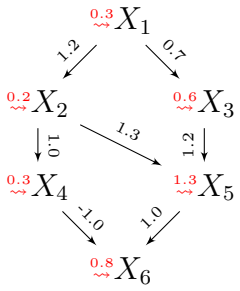
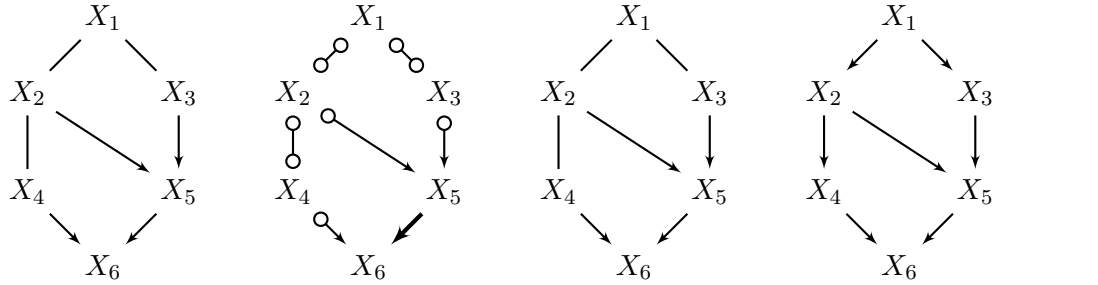


Fig. 5.1 Scheme description of the synthetic linear datasets used. Each edge has a weight, and the noise standard deviations are indicated in red. The value of a node is the weighted sum of the parents' values plus the noise.

uniform noise, both centered at zero and scaled to achieve the desired standard deviation (shown in red). For instance, values of variable X_5 are generated in the first dataset as $X_5 = 1.3X_2 + 1.2X_3 + \mathcal{N}(0, 1.3)$ while in the second dataset as $X_5 = 1.3X_2 + 1.2X_3 + \mathcal{U}(0, 1.3)$. Note that the weights were chosen randomly.



(a) PC (b) FCI (c) GES (d) DirectLiNGAM

Fig. 5.2 Generated causal graphs for the synthetic dataset with Uniform noise.

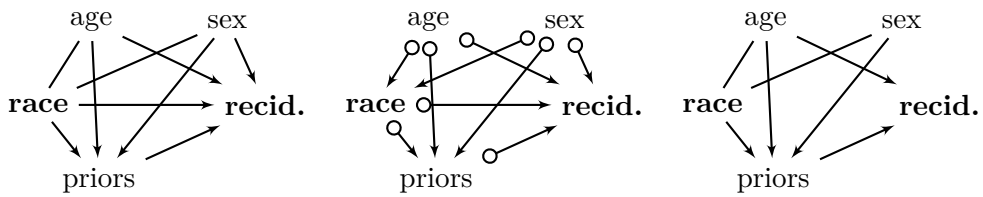
Fig. 5.2 shows the graphs generated based on the first dataset. PC, FCI, and GES generate the correct causal graph skeleton but fail to tell the direction of all edges. The structure corresponds to a Markov equivalence class (CPDAG) where 4 edges are (correctly) directed while the remaining 3 are left undirected. As expected, the constraint and score-based algorithms could identify the directions of all edges involved in v-structures. For the remaining edges ($X_1 - X_2, X_1 - X_3$, and $X_2 - X_4$), they couldn't identify the direction because all possible combinations of directions will lead to the same conditional independence relations between variables[5]. DirectLiNGAM, however, could generate the correct skeleton as well as the correct directions of the edges successfully. This is possible because the first dataset satisfies exactly the assumptions for the applicability of LiNGAM. That is, functional relations between variables are linear, values are continuous, and the noise distribution

[5]As long as the direction of edges do not introduce or remove a v-structure.

is non-Gaussian (uniform). It is important to mention that, for DirectLiNGAM, finding the correct causal structure also depends on setting the right threshold (α) for the linear regression step. For instance, the graph in Fig. 5.2d is obtained with a threshold $\alpha = 0.05$. Using a smaller value (e.g., $\alpha = 0.03$) leads to an extra false edge from X_3 to X_6.

We provide in Appendix C.1.1 the graphs generated from the second dataset following the same causal structure (Fig. 5.1) but with Gaussian noise.

Compas. The *Compas* dataset is already presented in Section 4.3.3. Five variables are used for structural learning, namely race, sex, age, priors, and recidivism. Age and priors are continuous, while the remaining variables are discrete. Three tiers in the partial order for temporal priority are used: race, sex, and age are defined in the first tier, priors are in the second tier, and recidivism is defined in the third tier. When found to be mediators, age and sex are considered as redlining variables, whereas priors are explaining variables. Since this dataset includes mixed data, the conditional Gaussian likelihood ratio test is used for PC and FCI, while the conditional Gaussian test is used for GES. For the same reason (mixed dataset), LiNGAM is not applied. Fig. 5.3 shows the generated causal graphs for PC, FCI, and GES. Note that, for clarity of illustration, in all subsequent causal graphs, the sensitive feature (on the left side) and the outcome (on the right side) are distinguished from the rest of the variables by highlighting them in bold.



(a) PC (b) FCI (c) GES

Fig. 5.3 Generated causal graph for the Compas dataset (**recid.** for recidivism.).

It is important to mention that the obtained graphs for the *Compas* dataset do not agree on the direct edge from the sensitive attribute (race) to the outcome variable (recidivism). There is such an edge according to PC and FCI but not according to GES. This is crucial to fairness as the direct effect is always discriminatory.

Fig. 5.4 shows the different discrimination measures using the different graphs. Both TE and ATE_{IPW} produce positive values, which indicate discrimination against non-white defendants.

Considering the PC CPDAG (Fig. 5.3a), the highest value of TE is obtained when there are no confounders (the two undirected edges are directed as $race \to age$ and $race \to sex$). In such a graph, TE coincides with TV, which equals 0.125. The same high value of TE is obtained with GES CPDAG (Fig. 5.3c) when the undirected edge is directed as $race \to age$.

In such no confounding case, the presence or absence of the direct edge $race \to recid.$ does not matter for TE. The smallest value for TE (0.050) is only obtained in FCI PAG (Fig. 5.3b), where both age and sex variables are confounders. This implies that the total effect goes through only two paths: $race \to recid.$ and $race \to priors \to recid.$. Such low TE value cannot be obtained in PC nor in GES CPDAGs because the edges $age \to race$ and $sex \to race$ will create a new v-structure and, hence, lead to a causal graph outside the Markov equivalence class.

The highest value (0.067) for DE is obtained in PC CPDAG when age is a mediator, but sex is a confounder ($race \to age$ and $sex \to race$). The smallest value (-0.012) is obtained when both variables are mediators. DE is naturally zero for GES and SBCN[6]. ID is highest (0.096) with PC when both age and sex are mediators ($race \to age$ and $race \to sex$). This is in line with GES as ID is highest (0.084) with the same directions of the edges ($race \to age$ and $race \to sex$). Surprisingly, when age is a confounder while sex remains a redlining, the indirect discrimination *against* blacks (0.096) becomes indirect discrimination *in favor* of blacks (-0.064). This is an example of Simpson's paradox [215, 33] when conditioning on a variable changes significantly the statistical conclusions.

When the edges are as directed as $race \to age$ and $sex \to race$, both PC and GES graphs produce the same ID value (-0.018). The case that leads to the highest discrepancy in ID values between PC and GES is $age \to race$ and $race \to sex$ (age is the confounder, and sex is a mediator). In such a setup, according to PC, ID is lowest (-0.064), while according to GES, ID is zero as there is a redlining path between race and recidivism. It is important to mention here that if a causal path is going through redlining and explaining variables (e.g., $race \to sex \to priors \to recid.$), it is considered part of explained discrimination. The rule of thumb is that any path containing at least one explaining variable is considered as part of explained discrimination[7]. ID is zero for FCI and SBCN for the same reason (without redlining paths).

According to all graphs, ED values are comparable as all explained discrimination goes through the single explaining variable (priors).

Overall, the *Compas* dataset shows that small variations in the graph structures can lead to significant differences in fairness conclusions. In particular, estimating TE using graphs generated by different CD algorithms can lead to a significant inconsistency ($0.125 - 0.050 = 0.075$) in assessing the amplitude of the discrimination against non-white defendants. Moreover, graphs generated by the same discovery algorithms (belong to the same Markov

[6]The results of SBCN (Suppes Bayes Causal Networks) [36] are excluded from the manuscript because SBCN is a specific type of causal graph for measuring fairness but which relies on a different interpretation of causal relationships (probabilistic causality [106]). However, you can consult the Appendix C.1.1 for the SBCN presentation.

[7]This interpretation can be justified by considering the simple path $race \to priors \to recid.$. Such a path is clearly part of explained discrimination as priors is an explaining variable. However, it also contains a "redlining" variable, which is the sensitive attribute race.

equivalence class) can lead to very different discrimination values (ID goes from a positive discrimination of 0.096 to a negative one (−0.064) due to reversing the direction of a single edge) which can be seen as a form of Simpson's paradox. Finally, the threshold value to decide about causal relations can also have important consequences on the fairness conclusion (missing *race* → *recid.* edge in GES and SBCN).

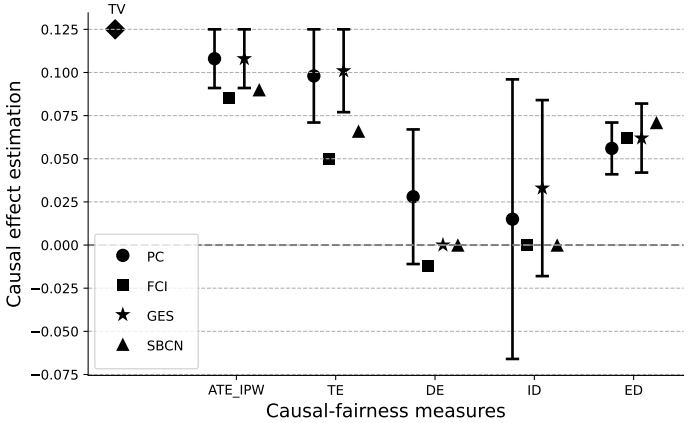
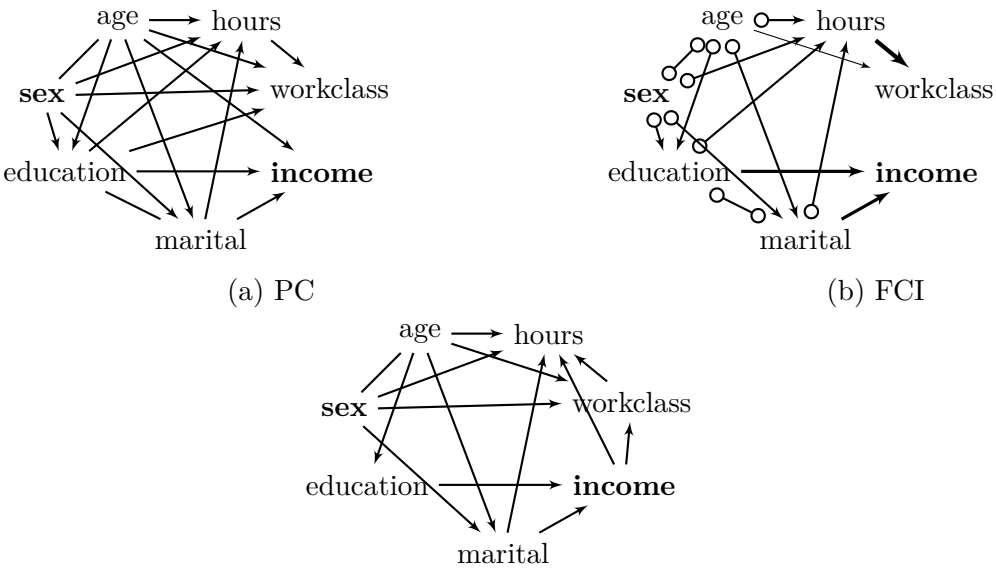


Fig. 5.4 Estimation of causal effects of the *Compas* dataset based on PC, FCI, GES, and SBCN.

Adult. The *Adult* dataset is already presented in Section 4.2.3. In this work, only 7 variables are used for structural learning: age, sex, education level, marital status, work class, and number of working hours per week. Age and number of working hours per week are continuous, while the remaining variables are discrete. Three tiers in the partial order for temporal priority are used: age and sex are defined in the first tier, education and marital status in the second tier, and work class, number of working hours per week, and income are defined in the last tier. When found to be mediators, variables age and marital status are considered as redlining, whereas education as explaining. The causal graphs generated by PC, FCI, and GES are shown in Fig. 5.5. As in the *Compas* dataset, LiNGAM cannot be used as data is mixed as well.



(a) PC (b) FCI

(c) GES

Fig. 5.5 Generated causal graph for the Adult dataset.

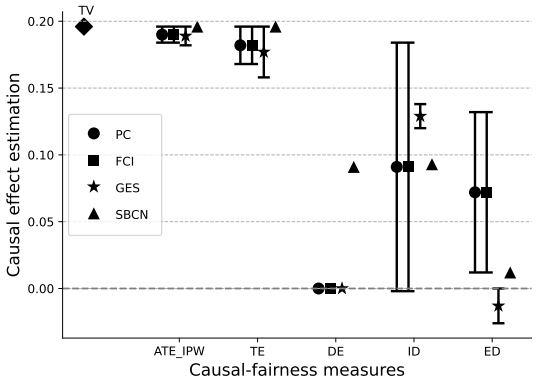


Fig. 5.6 Estimation of causal effects of the *Adult* dataset based on PC, FCI, GES, and SBCN.

There are two important notes about the generated graphs. First, only SBCN exhibits a direct edge between sex and income. Second, all remaining graphs have undirected edges (in particular, between sex and age). This leads to variability in the fairness measures as shown in Fig. 5.6. For instance, although all TE and ATE_{IPW} values are positive, which indicates discrimination against females, there is some variability in the extent of this discrimination. The highest discrimination can be seen in the GES CPDAG (Fig. 5.5c) where $sex \rightarrow age$ (age is a mediator) yields to TE= 0.196 whereas $age \rightarrow sex$ (age is a confounder) yields to TE = 0.157. DE is zero according to all graphs except for SBCN since it is the only one with a direct edge between sex and income. For PC and FCI graphs (having the same structure with two undecided edges), ID value ranges between -0.003 and 0.184 where the former is obtained with $age \rightarrow sex$ and $education \rightarrow marital$ and the

latter is obtained with $sex \to age$ and $education \to marital$. This is expected as $sex \to age$ opens an additional redlining path $sex \to age \to income$. In other words, having only one redlining path $sex \to marital \to income$ shows a very small indirect discrimination *in favor* of females. Opening the other redlining path (through age) turns that into a clear indirect discrimination *against* females. A possible explanation is that young married women tend to have low incomes due to motherhood responsibilities. In contrast, older married women pass that part of their lives and are more available for professional careers. Notice that the lowest ID value in GES (0.119 obtained with $age \to sex$) is significantly higher than the lowest ID value in PC and FCI (-0.003). The reason is that in GES, there is only one indirect (redlining and explained) path $sex \to marital \to recid.$ while in PC and FCI, there are three different paths ($sex \to marital \to income$, $sex \to education \to income$, and $sex \to education \to marital \to income$). Hence, the causal effect between sex and income in GES is only conveyed through the redlining path. In PC and FCI, the redlining path is split between the two other explained discrimination paths.

For ED, the highest value (0.132) is obtained in PC and FCI when age is confounder ($age \to sex$) and marital status is a mediator between education and income ($education \to marital$). The smallest value (-0.027) is obtained in GES when age is a mediator ($sex \to age$), which indicates a small explained discrimination *in favor* of females through the path $sex \to age \to education \to income$. This path is only possible as a single explaining path in GES CPDAG. In all the graphs obtained by the other algorithms, such a path is possible but along other explaining paths, particularly $sex \to education \to income$. This explains why the discrimination favoring females is only observable with GES. It is interesting to notice that in PC and FCI graphs, the explained discrimination through $sex \to education \to income$ is slightly positive (0.016), whereas in the GES graph, adding another mediator $sex \to age \to education \to income$ yields a slightly negative explained discrimination. As there is no overlap between the ranges of ED values in PC and FCI graphs on the one hand and GES on the other, and that values (although small) have different signs (positive vs. negative), the explained discrimination conclusions depend on which algorithm is used to discover causal relations.

Compared to the *Compas* dataset, the mediation analysis on the *Adult* dataset reveals two additional fairness-relevant observations. First, several CD algorithms can discover a specific causal path. However, the causal effect through that path may significantly differ depending on the presence of other causal paths that do not necessarily have the same interpretation (redlining or explaining paths). Second, even with the same causal path (e.g., $sex \to education \to income$), considering a mediator (e.g., age) can reverse the type of discrimination (e.g., $sex \to age \to education \to income$).

The experiments and the descriptions of the remaining datasets can be checked in Appendix C.1.2 (for *Dutch census*), Appendix C.1.3 (for *German credit*), Appendix C.1.4 (for *Boston housing*), and Appendix C.1.5 (for *Communities and crime*).

5.2.4 Conclusion

In this study, we provided a detailed and intuitive explanation of the major CD algorithms in the literature. Causal relations between variables are typically identified from observable data using CD algorithms as experiments and interventions (RCTs and A-B testing) are difficult to carry out in discrimination scenarios (requires changing inherent attributes of individuals such as gender or race). Constraint and score-based approaches to CD rely mainly on conditional independence tests and, hence, typically generate PDAGs with undirected edges. The third category relies rather on the independence between the cause variable and the residual of the regression to decide about the direction of the edges.

The main contributions of the study are two-fold. First, we show how the subtle differences between the CD algorithms can explain why they generate different causal graphs. Second and foremost, we demonstrate how slight differences between causal graphs may significantly impact fairness/discrimination conclusions.

Most causal approaches to fairness in the literature do not tackle the causal graph generation task. With this study, we hope to raise awareness about the importance of this step in the fairness assessment and enforcement pipeline, as any difference in the graph structure may lead to very different fairness conclusions. A natural follow-up work after this study is to design a new CD algorithm specifically tuned for fairness. This algorithm can be an adaptation of an existing algorithm but geared towards accurately discovering the sensitive attribute's causal effect on the outcome variable along the various directed paths. Another future direction would be to study the impact of pre-processing transformations on the structure of the generated graph and, consequently, on the fairness conclusions.

Having explored the impact of structural learning approaches on fairness assessments, we present in the next section our study on how CD algorithms perform when applied to biased data, investigating their robustness and accuracy under varying conditions of data bias.

5.3 Causal Discovery on Biased Data

In our study presented in the previous section, we showed how using a different CD algorithm may result in different causal graphs and, most importantly, how even slight differences between them can significantly impact fairness conclusions. This issue arises because current CD algorithms are not well-suited for handling the mixed data types commonly used in fairness assessments. Moreover, even simple experiments show that certain edges of the causal graph appear only when the bias level exceeds a threshold value [36].

The evaluation for CD algorithms typically follows a transductive approach, which involves comparing the structural difference between the discovered and ground-truth graphs. However, since obtaining datasets with ground truth is challenging, synthetic datasets are commonly used for benchmarking purposes [131, 205, 52]. Existing mechanisms allow the generation of datasets according to a desired causal graph [55]. Still, they cannot control the bias level in the generated data that can be exploited to develop and evaluate the effectiveness of bias mitigation approaches when developing ML models.

To address this issue, we propose a mechanism to generate synthetic datasets given causal graphs while allowing for adjustable bias levels. Our contribution is novel since it generates data according to the roles of variables within the causal structure (such as sensitive, confounding, mediator, and collider nodes) and facilitates examining how bias level affects the efficacy of CD methods.

Using the proposed synthetic data generation, we could study two fairness-related aspects of the CD procedure. First, we evaluated the reliability of CD algorithms in handling datasets with increasing bias levels while maintaining the same ground truth causal graph. Second, we examined the effects of different binarization thresholds for the outcome variable on these algorithms. Binarization is critical in this context, as it converts the continuous outcomes predicted by ML algorithms—such as job suitability or loan default risk—into binary decisions like approve or reject. Properly implementing this conversion is crucial for ensuring the algorithm's outputs effectively translate into the binary choices required in real-world scenarios [8].

The experimental analysis revealed that although some causal relations are present in the ground truth causal graph, CD algorithms could discover them only at a relatively high level of bias. Specifically, we demonstrate that these approaches can also suffer from instability in the presence of bias in the model. Therefore, there is a need for further discussion on the general pipeline used to assess fairness. In particular, tuning the degree of data bias allows us to understand how the system's misinterpretation depends on it. Related to outcome values binarization, experimental analysis revealed that the direct edge (sensitive to the outcome) is more affected by the binarization threshold in a collider structure than in a confounder or mediator structure. This holds significant importance within the context of

fairness, as a direct edge linking the sensitive variable to the outcome provides evidence of direct discrimination.

Contributions. The contributions of the study presented in this section are threefold: (1) a framework for generating synthetic datasets with a given causal graph and a desired bias level, (2) an empirical study of the bias level's impact on CD output, and (3) an empirical analysis of the effect of the outcome variable binarization threshold on the discovered causal graph.

Outline. The rest of this section is organized as follows. Section 5.3.1 discusses related work. Next, Section 5.3.2 presents our synthetic data generation mechanism. Section 5.3.3 details the experimental setting and main results. Finally, we conclude this section by indicating future perspectives in Section 5.3.4.

5.3.1 Related work

Identifying bias and mitigating it in real-world datasets usually relies on the statistical properties of the system. However, without knowledge of the underlying causal structure of the data, it is possible that non-existent biases are recognized or new ones are introduced through mitigation (e.g., the widely known Berkeley admissions case [33]). This section discusses relevant literature on each field, highlighting existing gaps and current drawbacks.

Causal Data Generation. Mechanisms for creating synthetic data based on causal frameworks are frequently used to test CD algorithms due to the lack of ground-truth causal structures in many benchmark datasets. Typically, these approaches rely on ML models [249, 55] or utilize existing network structures such as Erdős–Rényi or Scale-free to sample acyclic graphs [147, 277, 172]. However, these methods focus solely on modeling causal relations without considering fairness.

Fair Data Generation. Several studies propose synthetic data generation to ensure fairness in datasets [112, 46, 251, 265, 257]. A common approach to generating fairer synthetic data involves applying bias mitigation techniques during preprocessing and using GANs for data generation. Nevertheless, these methods do not account for causal frameworks, thus overlooking the necessity of maintaining causal relationships between variables to generate plausible data. Moreover, the option to control data bias is not considered.

Integrated Approaches for Causality and Fairness. Van Breugel et al. [239] present methods using various generators (one for each variable) that learn from the causal conditionals observed in the data. At inference time, variables are synthesized topologically, starting from the root nodes in the causal graph and sequentially synthesized, ending at the leaf nodes. While this approach is valuable for generating fair data considering

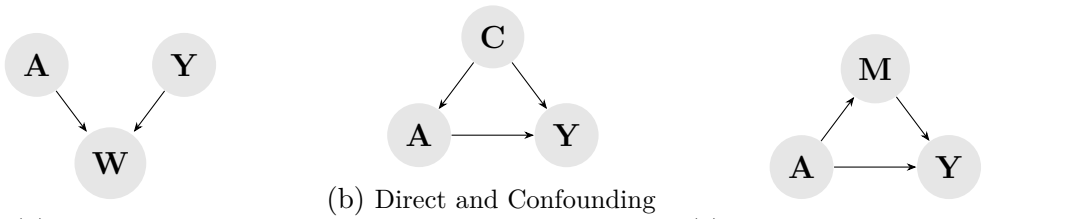
causal relationships, it does not allow for bias control. Another framework by Baumann et al. [27] aims to create synthetic data with various bias types (e.g., measurement, historical, representation, and omitted variable biases). Although this work does not explicitly reference causal models, it does consider generation following a graph structure. The limitation is that solely relying on a pre-defined graph structure may not accurately represent all possible scenarios. Our approach differs since it generates specific variable values depending on the variable's role (mediator, confounder, collider, sensitive, etc.) in the causal graph. This enables the generated data to include the ground truth of the causal model, which is crucial for benchmarking CD methods. Furthermore, unlike these previous approaches [239, 27], the proposed method allows for data generation based on a specific bias level. Controlling the bias level in synthetic datasets has several advantages; in particular, i) it aids in developing and evaluating bias mitigation techniques for ML models, and ii) it can be used to determine which CD algorithm best describes the data.

More related to CD, in our study [34], we highlighted the inconsistencies of CD algorithms when applied to benchmark datasets. The authors observed that state-of-the-art CD algorithms produce different causal graphs when applied to the same dataset. Most importantly, using causal fairness metrics [158], we observed that even slight deviations between the discovered graphs might lead to significant differences in fairness/discrimination conclusions. A possible explanation of these inconsistencies is that, since benchmark datasets are naturally biased, various CD algorithms are impacted differently by the level of bias in the data. This study is an attempt to understand better the impact of bias on CD.

5.3.2 Synthetic Data Generation

Fairness Terminology. In this work, we differentiate the terms *bias* and *discrimination*. Bias refers to the degree of imbalance in the generated data, specifically aimed at setting the degree to which a certain group is disadvantaged. The bias level is controlled through an input parameter in our data generation process (further details are provided later in this section). On the other hand, discrimination refers to the disparity between groups, which is quantitatively assessed using several fairness metrics (more details are provided in Section 5.3.3).

We propose a method for synthetic data generation with two primary features. First, leveraging the ground truth causal graph, our approach inherently incorporates the causal relationships among variables. Second, employing specific mechanisms tailored to each causal structure allows for setting a known bias level (`BL`), allowing the user to determine the extent of bias along different causal paths. We first recall the structures used because bias propagation is contingent upon the causal structure under analysis. Then, we provide



(a) Collider bias (b) Direct and Confounding biases (c) Direct and Mediation biases

Fig. 5.7 Causal graphs illustrating all types of causal structures used. Notably confounders (**C**), mediators (**M**) and colliders (**W**). Nodes are possibly multi-dimensional or empty.

an in-depth description of the data generation mechanism, highlighting how the bias is propagated in the data through the different causal structures.

Causal Structures. Considering a causal graph $\mathcal{G}$ in which at least one node represents a sensitive attribute, i.e., $A \in \mathbf{N}$, and one node represents the outcome attribute, i.e., $Y \in \mathbf{N}$, we consider three basic causal structures, namely, *confounding*, *collider*, and *mediation* [181]. By convention, we denote such structures using **C**, **W**, **M**, respectively. Fig. 5.7 depicts examples of causal graphs encompassing the different causal structures and types of biases. The rationale behind examining these structures is to accommodate the various forms of bias propagation through the causal graph while maintaining a simple and intuitive causal model to facilitate human readability and avoid unnecessary complexities that might hide the relevant patterns.

We also consider causal graphs incorporating various structure types simultaneously, specifically involving mediators and confounders. This combined consideration allows us to analyze the joint effects of these structures, providing insights into how their presence influences the overall bias propagation. Moreover, we typically consider two variants of each causal structure, one where the direct edge between the sensitive attribute and the outcome ($A \to Y$) is present (i.e., $A \in \mathbf{PA}_Y$) and one where it is absent (i.e., $A \notin \mathbf{PA}_Y$). The direct edge is critical when measuring discrimination because it always corresponds to unjustifiable bias.

Data Generation. Given $\mathcal{G}$ as input, we model the data generation using the following system of equations to formalize the relationships between variables. The equation of each variable depends on the corresponding node's role in the graph. The topological ordering of $\mathcal{G}$ dictates the sequence of variable generation, ensuring adherence to both causal dependencies and, if present, the accurate propagation of the bias along causal paths. Quantifying bias poses a significant challenge, mainly due to the complex task of defining bias within a mathematical framework [164]. As outlined at the beginning of this section, we denote the

term *bias* as advantage quantified by the gap between the distributions of the privileged and unprivileged groups. In what follows, we formulate how the bias propagates in our proposed data generation mechanism. For completeness, we consider different cases depending on the various causal structures in the causal graph. More specifically, for each node in $\mathcal{G}$, we formulate the data generation based on a specific structural equation that depends on the node's position in $\mathcal{G}$, which falls into two cases: i) the node is a root node; ii) the node is a child node.

Root Node Case. As the aim is to investigate the impact of bias, the generation mechanism depends on whether the node represents a sensitive attribute (A) or not (R). This distinction is essential because the sensitive attribute is assumed to be discrete, whereas any other root nodes could be continuous. Therefore, a root node A is distributed as a Bernoulli $\text{Bern}(p_A)$ variable where p_A controls the proportion of each sensitive group in the data. Whereas R is distributed according to a $\text{Beta}(\alpha_R, \beta_R)$ variable where α_R and β_R parameters that control the proportions of each value of R in the data. Formally,

$$\begin{aligned} A &:= D_A, \quad D_A \sim \text{Bern}(p_A) \\ R &:= D_R \quad D_R \sim \text{Beta}(\alpha_R, \beta_R) \end{aligned}$$

Notably, a root node can be a confounder ($R = C$) or other types with no implications on discrimination propagation, such as isolated nodes or other non-descendants of the sensitive attribute (e.g., $A \rightarrow Y \leftarrow R$).

Child Node Case. In the general case where a node X has at least one parent, the variable value is obtained as a combination of its parents in $\mathcal{G}$ plus independent noise U_X. Depending on the type of model used, this combination can be linear or nonlinear. As mentioned above, the data generation mechanism depends on the type of each parent node. We distinguish three types of parent nodes: a sensitive variable node (denoted $\mathbf{A}_X$), confounder nodes (denoted $\mathbf{C}_X$), and the remaining parent nodes (neither sensitive nor confounder, denoted $\mathbf{N}_X := \mathbf{PA}_X \setminus \{\mathbf{A}_X, \mathbf{C}_X\}$). The value of X is generated according to the following equation:

$$X := \sum_{i=1}^{|\mathbf{N}_X|} u_i N_i + \sum_{i=1}^{|\mathbf{C}_X|} v_i B_i^c + w_A B^A + U_X \tag{5.1}$$

where u, v, and w denote the weights that reflect the strength of the impact that each parent has on X, N_i is the value of i^{th} parent, and U_X represents a continuous independent random variable used to introduce variability into the generated dataset. B^A and B_i^c are the values through which the bias is propagated from the sensitive variable A to the child node X,

directly or through the confounder nodes, respectively. More formally, given a BL, B^A and B_i^c values are set as follows:

$$B = \begin{cases} \lambda Q_1 + \gamma Q_2 & \text{if } A = 0 \\ \gamma Q_1 + \lambda Q_2 & \text{otherwise} \end{cases} \tag{5.2}$$

where λ and γ are two positive numbers such that $0 > \lambda \geq \gamma$ and $\gamma/\lambda = \texttt{BL}$. Specifically, λ and γ are two parameters our generation mechanism inputs to quantify the disparity between privileged and non-privileged groups. Hence, $\texttt{BL} = 1$ (i.e., $\lambda = \gamma$) indicates a fair outcome, $\texttt{BL} > 1$ indicates a biased outcome against group $A = 0$, and $0 < \texttt{BL} < 1$ a biased outcome against group $A = 1$. Note that the difference between propagating the bias directly from the sensitive attribute to the outcome (B^A) or through the confounder (B_i^c) lies in the type of distributions of Q_1 and Q_2 values. This difference in the mechanism is critical for appropriately propagating the correct type of bias. Its significance lies in optimizing the incorporation of confounding factors, thereby improving the overall accuracy and reliability of the data generation process.

5.3.3 Experimental Analysis

This section presents the experimental setup, the main results related to the impact of varying data bias, and the effect of changing the outcome distribution on CD.

Experimental settings.

- **Causal Discovery Algorithms.** State-of-the-art CD algorithms are primarily designed for continuous data, presenting a challenge for our study involving mixed data types. Our datasets include binary sensitive and outcome attributes alongside other variable types. Therefore, employing algorithms explicitly tailored to handle mixed data is imperative. This approach guarantees the accuracy and reliability of our causal analysis across diverse data types. Accordingly, we use PC [117] from the `g-castle` library [271]. Among the reasons for choosing the `g-castle` implementation is that it identifies all causal relationships in the graph, typically not leaving indirect edges. To ensure the reliability and robustness of our findings, we conduct 10 runs of each CD approach on every synthetically generated dataset.

- **Synthetic Data.** We use the data generation mechanism described in Section 5.3.2 to evaluate how datasets with known apriori bias levels impact CD experimentally. For each experiment, we generate 5000 samples. As mentioned above, the sensitive attribute A follows Bernoulli(p_A) where $p_A = 0.55$. Then, we set $\alpha_R = 3$ and $\beta_R = 3$ for the $Beta$ distribution of R nodes. Concerning the BL, we set $\gamma = [1, 1.2, 1.4, 1.6, 1.8, 2, 2.5, 3, 3.5]$

while $\lambda = 1$. For B^A, we set $Q_1 \sim \mathcal{N}(0.25, 0.1)$ and $Q_2 \sim \mathcal{N}(0.75, 0.1)$, while for B_i^c we use $Q_1 \sim \text{Uniform}(C, 1)$ and $Q_2 \sim \text{Uniform}(0, C)$.

- **Real-World Datasets.** The real-world datasets used in our experiments are: *Adult*, *Compas*, and *Communities and crime*. All three datasets have already been introduced in Section 4.2.3, Section 4.3.3, and Section C.1.5, respectively.

- **Results Visualization.** To provide a clear and intuitive visualization of the results, we use the causal graphs directly, labeling each edge with either the level of discrimination (Figs 5.8, 5.9, and 5.10) or the range of outcome thresholds (Figs 5.12, 5.11, and 5.13) at which the CD algorithm discovers the edge. In addition, we define a standardized way of naming the causal graphs for ease of recall and readability of the results. In all the experiments, the estimated causal graphs are named as the causal structure in question (e.g., **M**, **C**, **W**) with the number of paths as superscript and the number of nodes as a subscript. For instance, $\mathbf{M}_2^1$ represents a causal graph with two mediators on a single path, and $\mathbf{C}_2^2$ represents a causal graph with two confounders, each one on a path. $\mathbf{M}_2^1\mathbf{C}_1^1$ implies a causal graph with one mediator path with two nodes and one confounder path with one confounder.

- **Evaluation Measures.** As stated in Section 5.3.2, discrimination between groups is quantified using selected fairness notions. We opt to use the Statistical Disparity (SD) (Eq. (4.2)) and the Disparate Impact (DI) (Eq. (4.1)).

  We validate our data generation mechanism by applying causality-based fairness notions, namely the natural direct effect (NDE) (Eq. (3.35)) and the natural indirect effect (NIE) (Eq. (3.36)), which quantify the direct and indirect impact of the sensitive attribute on the outcome, respectively.

Causal Discovery Using Biased Data. In this section, we analyze how altering the discrimination level within a dataset affects CD. The results are presented through causal graphs, where each edge is labeled with the `BL` value at which the PC algorithm first discovers that specific edge. Solid lines indicate edges in both the ground-truth graph and the one discovered by the CD algorithm, while dashed lines represent inferred edges not present in the ground-truth.

Fig. 5.8 illustrates the results for the mediator structures when $A \notin \mathbf{PA}_Y$ (no $A \to Y$ edge in the ground-truth graph). In a fair setting ($\texttt{BL} = 1$), the PC algorithm detects only direct links between mediators and the outcome as it prioritizes immediate impact on Y. As the data becomes more discriminatory ($\texttt{BL} > 1$), PC can uncover more complex relationships. Specifically, at $\texttt{BL} = 1.2$, causal links between the sensitive attribute and mediators are discovered. Despite the absence of $A \to Y$ in the ground truth, PC identifies it beyond a certain `BL`. Notably, identifying $A \to Y$ edge in terms of `BL` does not depend on the number

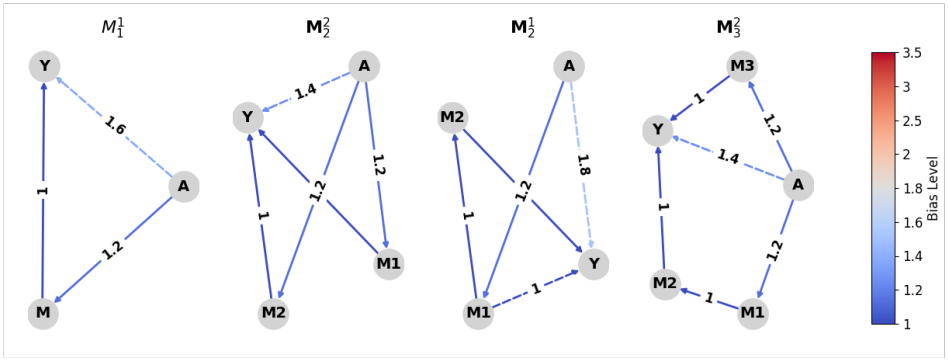


Fig. 5.8 Mediators with $A \notin \mathbf{PA}_Y$ using PC. The color bar shows the discrimination level; 1 indicates edge detection under fair data. Solid lines represent inferred edges present in the ground truth, while dashed lines indicate absent edges in the ground truth.

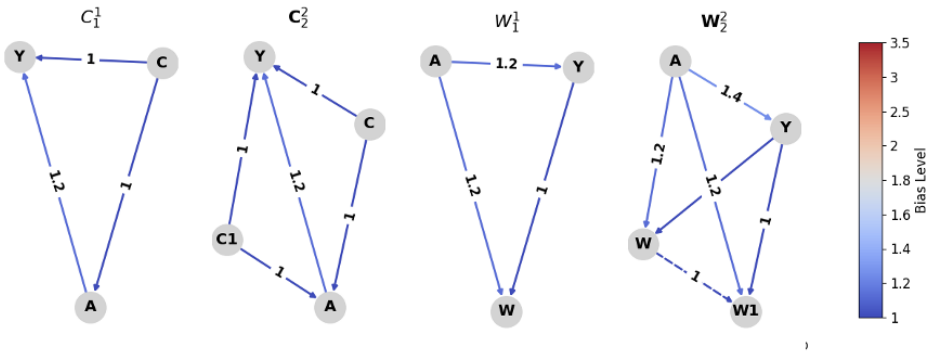


Fig. 5.9 Confounders and Colliders with $A \in \mathbf{PA}_Y$.

of mediation paths in the causal graph. Concerning **C** structures (Fig. 5.9 - Left), when $\texttt{BL} = 1$ and the link $A \to Y$ is absent from in ground truth, PC exhibits a robust capability in capturing the underlying causal structure. In contrast, when the ground truth includes the $A \to Y$ edge, PC identifies it when $\texttt{BL} \geq 1.2$. This behavior does not change even in the presence of multiple confounders. Fig. 5.9 on the right shows the results for the **W** structure when the direct edge $A \to Y$ is present in the ground-truth graph. Regardless of the number of colliders in the graph, $Y \to W$ is inferred when the data is fair. At this BL, in the case where two colliders are present ($\mathbf{W}_2^2$), PC erroneously identifies the causal relationship between the two collider nodes. Moreover, the sensitive attribute's causal links are discovered when $\texttt{BL} \geq 1.2$.

Finally, we explore causal structures incorporating multiple relationship types (i.e., mediators and confounders), aiming to simulate the complexity of real situations (Fig. 5.10).

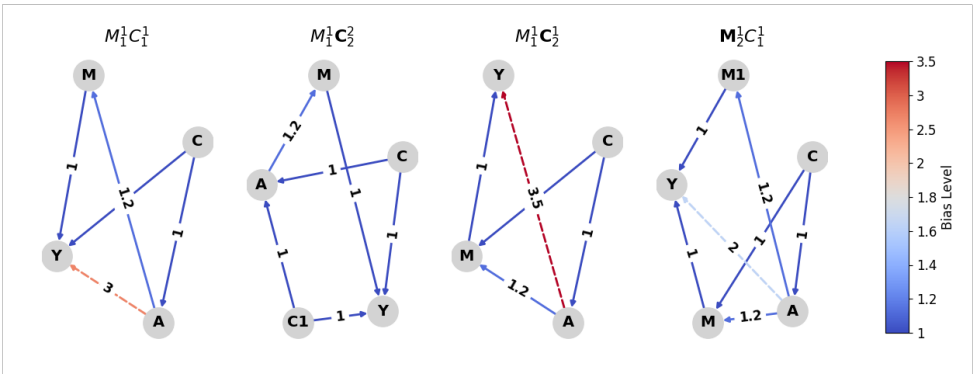


Fig. 5.10 Mediators and Confounders with $A \notin \mathbf{PA}_Y$.

We specifically vary the number of structures for each type across different graphs. We observe that up to $\texttt{BL} \leq 1.2$, the observations remain consistent with those from the structures considered individually. However, PC erroneously identifies a dependency between the sensitive attribute and the outcome at a higher level of discrimination (e.g., $\texttt{BL} = 3.5$). In the case of two mediators and one confounder ($\mathbf{M_1^2C_1^1}$), this misidentification occurs at $\texttt{BL} = 1.8$.

In summary, observing results across diverse causal structures could offer various insights. The main observation is that when data is unbiased w.r.t. a sensitive attribute, CD methods fail to identify some relevant edges in the causal graph. In particular, the first edge in mediated (indirect) paths, and most importantly, the direct edge $A \to Y$ in confounder and collider structures. As the presence/absence of the direct edge is crucial for measuring discrimination (NDE, NIE, and PSE [158]), it is highly recommended to rely on experts of the field (or the case study at hand) specifically when data is unbiased. Relying only on the output of CD algorithms may produce misleading results that can have significant implications when applied in real-world scenarios.

Impact of the Outcome Distribution on Causal Discovery. In this section, we study the impact of binarizing the outcome Y on CD. In particular, we observe how varying the outcome distribution by changing the threshold used for binarization affects the estimated causal graph. As stated at the beginning of Section 5.3 and discussed later in this section, changing the distribution of the outcome implicitly impacts the level of bias in the data and, therefore, on the CD results. We validate our findings on both synthetic and real-world datasets.

The causal graphs depicted in Figs 5.11, 5.12, and 5.13 represent the estimated causal graphs by PC of the synthetic datasets when the outcome Y is not binarized.

To examine how the distribution of the outcome affects CD, we compare the causal graphs obtained when Y is not binary with those obtained when Y is binary across various thresholds, ranging from 0.1 to 0.8. If the causal graphs produced when binarizing Y closely align with the graph obtained when Y is not binarized, this indicates the robustness of the CD algorithm to the binarization process.

The black solid and dotted edges represent the edges present and absent, respectively, when Y is not binarized. The red dashed edge indicates that an edge is absent when Y is not binarized and stays absent for all thresholds used to binarize Y. The intervals on some edges represent the threshold range at which the CD algorithm discovers the edge. Thus, the wider the range, the more robust the CD algorithm is to the binarization of the outcome. For example, a black solid edge with an interval $[0.1, 0.8]$ implies that the PC algorithm could discover the edge before and after binarisation for all the chosen thresholds. However, a black solid edge with an interval $[0.3, 0.5]$ implies that the PC algorithm could discover the edge before binarization and only for the thresholds ranging from 0.3 to 0.5 after binarization.

Below are the main observations depending on the causal structure studied: **M**, **C**, **W**. We also study the impact of varying the Y distribution when the different causal structures coexist (**MC**). Since we noticed that the `BL` plays a crucial role in the impact of Y distribution on CD, we provide the results for two different `BL`s, namely 10 (biased data) and 1 (fair data). These values are indicated in the title of each structure next to its name (e.g., M_1^1 - `BL` : 10 or M_1^1 - `BL` : 1).

Fig. 5.11 illustrates the impact of the outcome distribution on the mediator structure (**M**) discovery. Firstly, notice that when data is biased (`BL` = 10), PC could identify much more causal relations[8] than when data is fair (`BL` = 1), which corroborates the findings of the previous section. We observe that the higher the `BL`, the more binarization impacts the CD algorithm[9]. In other words, when the data is fair (`BL`=1), the interval of the thresholds on all edges is almost the full interval: $[0.1, 0.8]$ on black edges or stays absent in all thresholds in case of absence of the edge (dashed red edges in the causal graphs) when Y is not binarized. However, PC becomes less robust to binarization when the data is biased. This observation becomes more apparent when the number of mediators or the number of mediator paths increases ($\mathbf{M}_2^2$ - `BL` : 10, $\mathbf{M}_3^2$ - `BL` : 10). For example, some edges ($A \to M2$ in $\mathbf{M}_2^2$ - `BL` : 10) are not discovered when the outcome is not binarized (depicted in black dashed edges in the graphs); however, they are discovered when the outcome Y is binarized for all thresholds range. Moreover, some edges ($M2 \to Y$ in $\mathbf{M}_3^2 -$ `BL` : 10) while present in the graph when Y is not binary, they disappeared in almost all the thresholds range after Y binarization.

[8]Recall that red edges mean that the causal relation was not determined when the outcome Y was not binarized.

[9]Note that we run experiments with other `BL`s and this observation is confirmed, namely the higher the `BL`, the more binarization of the outcome impacts the CD algorithm.

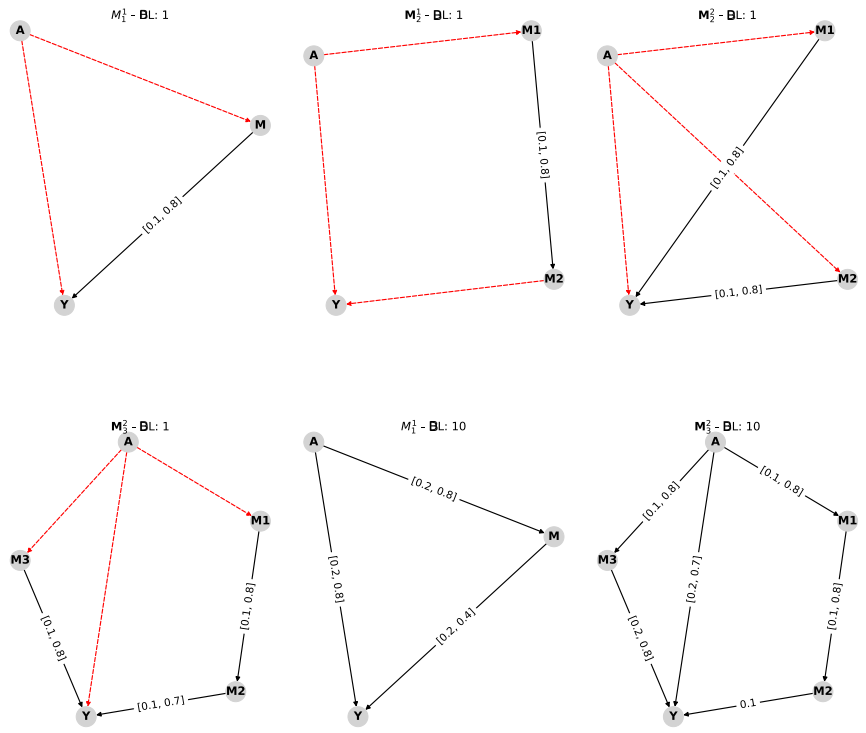


Fig. 5.11 Estimated causal graphs for the mediator structure. The black solid and dotted edges represent the edges present and absent, respectively, when Y is not binarized. The red dashed edge indicates that an edge is absent when Y is not binarized and stays absent for all thresholds used to binarize Y. The intervals on some edges represent the threshold range at which the CD algorithm discovers the edge.

Regarding the confounder structure (**C** in Fig. 5.12), $A \leftarrow C$ is highly impacted by Y binarization. In other words, while absent when Y is not binarized, it appears in the full range of thresholds when only one confounder exists ($\mathbf{C}_1^1$ - BL : 10). Moreover, $C1 \rightarrow Y$ is only impacted when the data is fair (BL $= 1$). This happens when the number of confounders increases ($\mathbf{C}_2^2$ - BL : 1).

A crucial observation of the impact of Y distribution is related to the direct edge $A \rightarrow Y$, which implies direct discrimination in a fairness context. That is, $A \rightarrow Y$ is generally unaffected by Y binarization for both the mediator and confounder structures. However, this is not true with the collider structure (**W**) where the direct edge $A \rightarrow Y$ is impacted by binarization, and this is more apparent with high BL ($\mathbf{W}_1^1$ - BL : 10 and $\mathbf{W}_2^2$ - BL : 10).

So, in general, the impact of Y distribution affects the direct edge $A \rightarrow Y$ when **W** exist between A and Y but not when only **M** or only **C** exist between A and Y. However, combining both **M** and **C** structures in the same graph (Fig. 5.13), the impact of Y distribution becomes

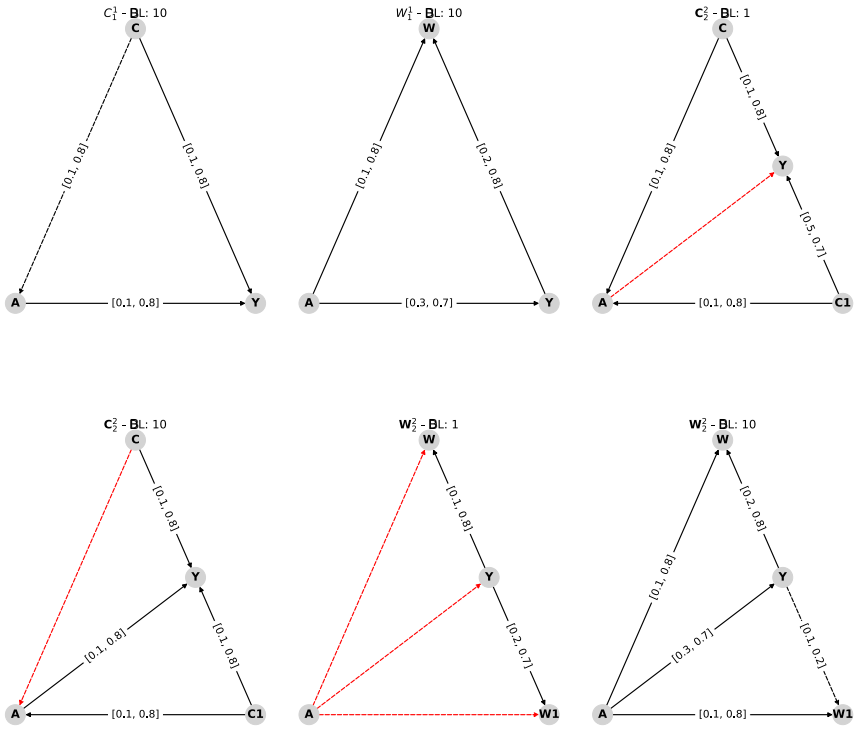


Fig. 5.12 Estimated causal graphs for the confounder and the collider structures.

apparent in particular when BL is high. Moreover, $A \leftarrow C \rightarrow Y$ and $A \rightarrow M \rightarrow Y$ are highly impacted by binarization when the data is biased in the **CM** structure. On the other hand, when the data is fair (BL $= 1$), all the estimated causal graphs when Y is binarized coincide with the one when Y is kept unbinarized in the mediator and the collider structures. This is not always the case for the confounder structure.

Another notable observation is that the binarization threshold highly impacts discovering the confounding path $A \leftarrow C \rightarrow Y$. In particular, when data is biased, discovering the edge $C \rightarrow A$ depends on the binarization threshold, while for unbiased data, discovering the edge $C \rightarrow Y$ depends on the binarization threshold. Identifying all confounding paths is crucial to measure discrimination[10], it is recommended to rely on experts to identify the graph correctly. Moreover, in the presence of a collider structure or a coexistence of confounders/mediators structures, although present when Y is not binary, the direct edge $A \rightarrow Y$ is only identified for a small range of threshold values when Y is binary. In the following section, we examine the impact of the Y distribution on real data sets.

[10]Total effect (TE) measure appropriately adjusts on all confounders. Hence, if CD fails to identify a node as a confounder, TE will yield misleading results as it will not adjust on confounder variables.

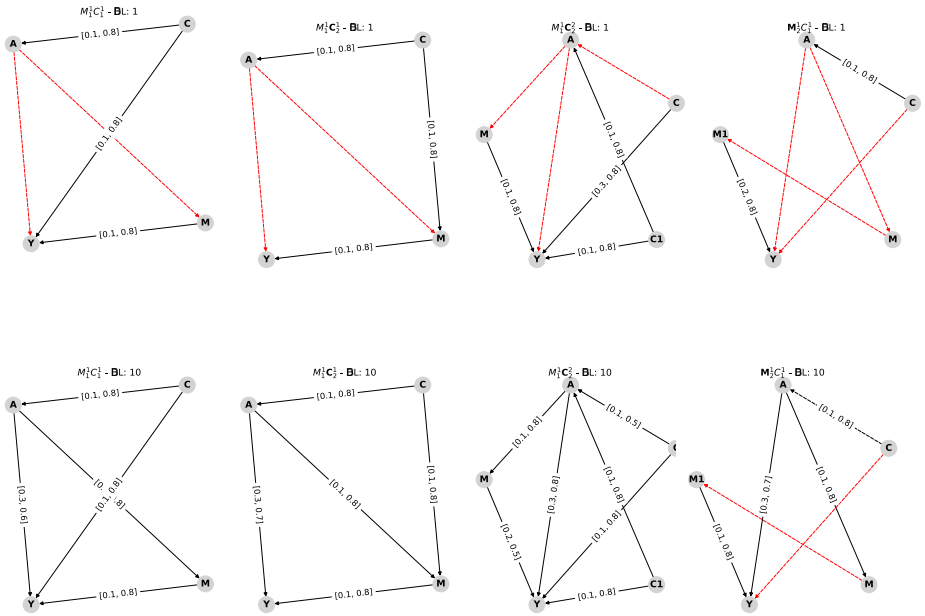


Fig. 5.13 Estimated causal graphs for the combined structures.

Experiments on Real-World Datasets. To validate the crucial impact of binarization of the outcome on CD, we run our experiments on three real-world fairness benchmark datasets, namely *Adult*, *Compas*, and *Communities and crime* (All datasets are already presented in the manuscript). The causal graphs of *Adult* and *Compas* are provided in Section 5.2.3 while that of *Communities and crime* is provided in Appendix C.1.5. As for the synthetic datasets, the outcome distribution is tightly related to the BL. For instance, for the *Compas* dataset, although the direct edge *race* $\rightarrow$ *risk_score* is present when the risk score is not binary, it disappeared in the two last thresholds 8 and 9 (high-risk scores). At these two thresholds, the disparity between groups is almost equal to 0 (0.05 and 0.01 for threshold = 8 and threshold = 9, respectively).

For the *Adult* dataset, we note that the direct edge *gender* $\rightarrow$ *income* appears only when the threshold for binarising an individual's income is $10K$ or $50K$. In all the other thresholds, no direct edge is discovered. This is crucial because, in the literature, the version of the *Adult* dataset primarily used is when the income is binarized to $50K$ [73]. However, simply changing the threshold used to binarize the outcome significantly impacts CD, and therefore, fairness conclusions. The same behavior is also observed in the *Communities and crime* dataset where PC is proven to be not robust to the violence rate binarization. In particular, the impact is depicted on the direct edge *race* $\rightarrow$ *violence_rate*, appearing in most thresholds but disappearing in three (0.1, 0.4, and 0.6).

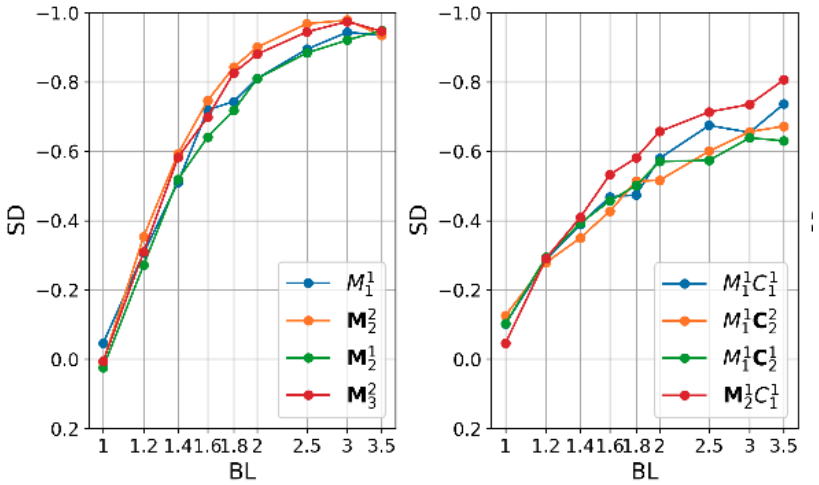


Fig. 5.14 Statistical Disparity for mediators and confounders where the edge $A \to Y$ is present in the ground truth causal graph.

5.3.4 Conclusion

Assessing discrimination in ML automated decision systems increasingly relies on the causal relations between variables rather than mere correlations. Causal relations are captured through causal graphs, typically identified using CD algorithms. This study aims to study the behavior of CD algorithms in the presence of biased data. To this end, we proposed a mechanism that takes a causal graph and a bias level as input and generates a biased synthetic dataset satisfying the causal structure of the graph with a desired bias level. Regarding the impact of the bias level on CD, the main observation is that when data is unbiased w.r.t a sensitive attribute, CD algorithms fail to identify important edges of the graph. In particular, the first edge in mediated (indirect) paths, and most importantly, the direct edge ($A \to Y$) in confounder and collider structures. Consequently, relying on experts in the field (or the case study at hand) is highly recommended specifically when data is unbiased. Relying only on the output of CD algorithms may produce misleading results. The most notable observation regarding the impact of the binarization threshold on CD is the importance of the threshold value in discovering confounding paths.

5.4 Conclusion

In this chapter, we have explored the intricate relationship between CD and fairness, focusing on two primary areas of investigation.

We began by examining various CD approaches and their implications for fairness. Our analysis highlighted that even slight differences between causal models could lead

to significantly different conclusions regarding fairness and discrimination. This section underscored the critical role of precise causal modeling in fairness assessments, demonstrating that the choice of the CD method can profoundly influence the outcomes of such evaluations. By presenting case studies and experimental results, we illustrated the variability in fairness conclusions stemming from different causal interpretations, emphasizing the necessity for careful selection and validation of CD approaches in fairness studies.

In the second section, we investigated how different types of causal structures and varying levels of injected bias in data impact the performance of CD algorithms. Our findings revealed that the robustness and reliability of these algorithms can be significantly affected by the complexity of causal structures and the presence of bias.

Building on the insights gained from examining the impact of CD methods on fairness, we plan to design a CD algorithm specifically tuned for fairness in future work.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis has comprehensively explored three pivotal research areas in the domain of fairness in algorithmic decision-making systems. Our investigations provided critical insights into the interplay between fairness, privacy, and causal discovery, offering valuable contributions to the field of ML and artificial intelligence. Here, we summarize the key findings from each research work and discuss their broader implications.

1. **Applicability of Statistical and Causality-Based Fairness Notions: Chapter 3**

   Our first study delved into the applicability of both statistical and causality-based fairness notions across diverse application domains. We evaluated how well these notions align with stakeholder preferences and societal norms. This work underscored the complexity of fairness in algorithmic decision-making, revealing that different fairness criteria can lead to varied outcomes depending on the context and stakeholder values. In particular, we proposed two decision diagrams (one focusing more on the statistical fairness notions and one on the causality-based fairness notions) integrating a set of fairness-related features of real-world scenarios that can help researchers, practitioners, and policymakers answer the question of "which notion of fairness is most appropriate to a given real-world scenario and why?".

   Our systematic evaluation highlighted the necessity of a nuanced approach to fairness, advocating for context-aware fairness assessments that consider the unique characteristics of each application domain.

2. **Impact of Local Differential Privacy (LDP) on Fairness: Chapter 4**

   The second research focus examined the intersection of privacy and fairness. In particular, the impact of LDP on group fairness. Our empirical studies found that obfuscating multi-dimensional sensitive attributes via LDP mechanisms in general,

improves fairness. In addition, we proposed a new privacy budget splitting solution named ***k-based***, which generally led to better fairness and performance results than the state-of-the-art solution that splits ε uniformly. Moreover, we observed that the true decision distribution has an important effect on which group is more sensitive to privacy and we summarized our findings as recommendations to guide practitioners in adopting effective privacy-preserving practices while maintaining fairness and utility in ML applications.

Motivated by our empirical studies, we conducted a rigorous quantitative assessment of how different levels of privacy and data distributions influence ML model decisions. Our findings demonstrated that while LDP can protect individual privacy, it also has significant implications for fairness. More specifically, our findings show that the *unprivileged* group benefits more than the *privileged* group when injecting enough noise into the sensitive attribute. Furthermore, for conditional statistical disparity and for equal opportunity difference, injecting noise, in general, improves fairness. This also holds for statistical disparity when the data contain no proxies to the sensitive attribute. However, when the data contains proxies, in certain cases, by injecting enough noise, while the discrimination was originally against one group, it may be shifted to the other group after obfuscation, and the level of unfairness may be worse than before.

3. **Causal Discovery and Algorithmic Fairness: Chapter 5**

   In our third study, we explored causal discovery in relation to algorithmic fairness, investigating how the process of uncovering causal relationships impacts the structure of causal graphs and subsequent fairness conclusions. In particular, we demonstrated how slight differences between causal graphs may significantly impact the conclusion on fairness/discrimination. In addition, we proposed a novel data generation mechanism that creates biased synthetic datasets based on causal graphs and specified bias levels. This approach enables us to systematically analyze the influence of various causal discovery algorithms on different causal structures and the extent of introduced bias. Our results indicate that the choice of causal discovery method can significantly affect the fairness of the resulting models, highlighting the importance of selecting appropriate algorithms and understanding their implications for bias and fairness.

In conclusion, this thesis contributes to the growing body of knowledge on fairness in machine learning and artificial intelligence, providing practical insights and methodological advancements that pave the way for more equitable algorithmic decision-making systems.

6.2 Future Work

The findings from these three research areas collectively advance our understanding of fairness in algorithmic decision-making systems. They emphasize the need for a multidisciplinary approach integrating statistical, causal, and privacy perspectives to develop fair and equitable algorithms. Our studies also point to several key directions for future research:

1. **Contextual Fairness Assessments:** Future work should develop frameworks that tailor fairness evaluations to specific application contexts, incorporating stakeholder input and societal norms. In particular, and for a short-term future plan, we believe that validating our proposed decision diagrams for the applicability of fairness notions by experts in real-world contexts will enhance their credibility and help us improve further and enrich these diagrams. As a long-term objective, we aim to create a robust, scalable framework that integrates statistical and causality-based fairness notions into AI systems across various sectors.

2. **Relationship Between Privacy and Fairness:** Investigating advanced privacy-preserving techniques that minimize trade-offs with fairness and utility of the ML model and exploring their applicability in real-world scenarios will be crucial. In the short term, we plan to extend our empirical studies by investigating the impact of LDP pre-processing on different ML algorithms, such as deep neural networks, as well as different fairness measures. As a medium-term future work, we aim to extend our systematic and formal study to more fairness measures, particularly overall accuracy equality. We also believe that hiding only the sensitive attribute is crucial but not sufficient because proxies for this attribute may exist in the data and thus reveal sensitive information. Therefore, we plan to study formally the impact of LDP on multidimensional data.

3. **Causal Discovery and Fairness:** Further research is needed to refine causal discovery algorithms and data generation methods, ensuring reliable and fair outcomes across diverse settings. A natural short-term follow-up work after this study is to conduct additional experiments by testing with synthetic datasets designed to simulate real-world biases to improve the performance of existing causal discovery algorithms. Our medium-term objective is to design a new causal discovery algorithm specifically tuned for fairness. This algorithm can adapt to an existing algorithm but is geared towards accurately discovering the sensitive attribute's causal effect on the outcome variable along the various directed paths. Another future direction would be to study the impact of pre-processing transformations on the structure of the generated graph and, consequently, on the fairness conclusions.

Bibliography

[1] (1965). US Equal Employment Opportunity Commission. https://www.eeoc.gov/.

[2] (2023). Indicator vector. Available online: https://en.wikipedia.org/wiki/Indicator_vector (accessed on 04 April 2023).

[3] Agarwal, A., Agarwal, H., and Agarwal, N. (2022). Fairness score and process standardization: framework for fairness certification in artificial intelligence systems. *AI and Ethics*, pages 1–13.

[4] Agarwal, S. (2020). Trade-offs between fairness, interpretability, and privacy in machine learning. Master's thesis, University of Waterloo.

[5] Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.

[6] Alao, D. and Adeyemo, A. (2013). Analyzing employee attrition using decision tree algorithms. *Computing, Information Systems, Development Informatics and Allied Research Journal*, 4.

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

[8] Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., and Zhioua, S. (2023). Survey on fairness notions and related tensions. *EURO Journal on Decision Processes*, 11:100033.

[9] Andrews, B., Ramsey, J., and Cooper, G. F. (2018). Scoring bayesian networks of mixed variables. *International journal of data science and analytics*, 6(1):3–18.

[10] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. propublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*.

[11] Arcolezi, H. H. (2023). LDP impact on fairness repository. https://github.com/hharcolezi/ldp-fairness-impact.

[12] Arcolezi, H. H., Couchot, J.-F., Al Bouna, B., and Xiao, X. (2021). Random sampling plus fake data: Multidimensional frequency estimates with local differential privacy. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 47–57, New York, NY, USA. Association for Computing Machinery.

[13] Arcolezi, H. H., Couchot, J.-F., Gambs, S., Palamidessi, C., and Zolfaghari, M. (2022). Multi-freq-ldpy: Multiple frequency estimation under local differential privacy in python. In Atluri, V., Di Pietro, R., Jensen, C. D., and Meng, W., editors, *Computer Security – ESORICS 2022*, pages 770–775, Cham. Springer Nature Switzerland.

[14] Arcolezi, H. H. and Gambs, S. (2023). Revealing the true cost of local privacy: An auditing perspective. *arXiv preprint arXiv:2309.01597*.

[15] Arcolezi, H. H., Gambs, S., Couchot, J.-F., and Palamidessi, C. (2023a). On the risks of collecting multidimensional data under local differential privacy. *Proceedings of the VLDB Endowment*, 16(5):1126–1139.

[16] Arcolezi, H. H., Makhlouf, K., and Palamidessi, C. (2023b). (local) differential privacy has no disparate impact on fairness. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 3–21. Springer.

[17] Arcolezi, H. H., Pinzón, C. A., Palamidessi, C., and Gambs, S. (2023c). Frequency estimation of evolving data under local differential privacy. In *Proceedings of the 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28 - March 31, 2023*, pages 512–525. OpenProceedings.org.

[18] Asuncion, A. and Newman, D. (2007). Uci machine learning repository.

[19] Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

[20] Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the 19th international joint conference on Artificial Intelligence*, pages 357–363.

[21] Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, 61(6):54–61.

[22] Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[23] Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *NIPS Tutorial*.

[24] Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[25] Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.

[26] Bassily, R. and Smith, A. (2015). Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 127–135, New York, NY, USA. Association for Computing Machinery.

[27] Baumann, J., Castelnovo, A., Crupi, R., Inverardi, N., and Regoli, D. (2023). Bias on demand: A modelling framework that generates synthetic data with bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1002–1013.

[28] Bellin, J. (2014). The inverse relationship between the constitutionality and effectiveness of new york city stop and frisk. *BUL Rev.*, 94:1495.

[29] Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. (2007). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137.

[30] Bendick, M. (2007). Situation testing for employment discrimination in the united states of america. *Horizons stratégiques*, (3):17–39.

[31] Bergstra, J., Yamins, D., and Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, ICML'13, page I–115–I–123. JMLR.

[32] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

[33] Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404.

[34] Binkytė, R., Makhlouf, K., Pinzón, C., Zhioua, S., and Palamidessi, C. (2023). Causal discovery for fairness. In Dieng, A., Rateike, M., Farnadi, G., Fioretto, F., Kusner, M., and Schrouff, J., editors, *Proceedings of the Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*, volume 214 of *Proceedings of Machine Learning Research*, pages 7–22. PMLR.

[35] Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 514–524.

[36] Bonchi, F., Hajian, S., Mishra, B., and Ramazzotti, D. (2017). Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 3(1):1–21.

[37] Bowen, N. K. and Guo, S. (2011). *Structural equation modeling*. Oxford University Press.

[38] Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

[39] Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877.

[40] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91.

[41] Cai, R., Qiao, J., Zhang, K., Zhang, Z., and Hao, Z. (2018). Causal discovery from discrete data using hidden compact representation. *Advances in neural information processing systems*, 31.

[42] Carey, A. N., Bhaila, K., and Wu, X. (2023). Randomized response has no disparate impact on model accuracy. In *2023 IEEE International Conference on Big Data (BigData)*, pages 5460–5465. IEEE.

[43] Celis, D. and Rao, M. (2019). Learning facial recognition biases through vae latent representations. In *proceedings of the 1st international workshop on fairness, accountability, and transparency in MultiMedia*, pages 26–32.

[44] Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., and Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5):124–27.

[45] Chang, H. and Shokri, R. (2021). On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE.

[46] Chaudhari, B., Chaudhary, H., Agarwal, A., Meena, K., and Bhowmik, T. (2022). Fairgen: Fair synthetic data generation. *arXiv preprint arXiv:2210.13023*.

[47] Chen, C., Liang, Y., Xu, X., Xie, S., Hong, Y., and Shu, K. (2022a). On fair classification with mostly private sensitive attributes. *arXiv preprint arXiv:2207.08336*.

[48] Chen, C., Liang, Y., Xu, X., Xie, S., Hong, Y., and Shu, K. (2022b). When fairness meets privacy: Fair classification with semi-private sensitive attributes. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

[49] Cheng, J. and Zeng, J. (2023). Shaping ai's future? China in global AI governance. *Journal of Contemporary China*, 32(143):794–810.

[50] Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K. S., and Liu, H. (2022). Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*.

[51] Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808. PKP Publishing Services Network.

[52] Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

[53] Chizhikova, A., Billinghurst, H., Elizabeth, M., Hossain, S., Kulkarni, A., Guibon, G., and Couceiro, M. (2024). Factorizing gender bias in automatic speech recognition for mexican spanish.

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

[55] Cinquini, M., Giannotti, F., and Guidotti, R. (2021). Boosting synthetic data generation with effective nonlinear causal discovery. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, pages 54–63.

[56] Compas (2020). Compas. https://www.equivant.com/northpointe-risk-need-asses sments/.

[57] Corbett-Davies, S., Gaebler, J., Nilforoshan, H., Shroff, R., and Goel, S. (2023). The measure and mismeasure of fairness. *J. Mach. Learn. Res*, 24:1–117.

[58] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806.

[59] Coston, A., Mishler, A., Kennedy, E. H., and Chouldechova, A. (2020). Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 582–593.

[60] Crawford, K. (2013). Think again: big data. why the rise of machines isn't all it's cracked up to be. *Foreign Policy*, 10.

[61] Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43:1241.

[62] Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, 2015(1):92–112.

[63] David, H. A. and Edwards, A. W. F. (2001). *Yule's Paradox ("Simpson's Paradox")*, pages 137–143. Springer New York, New York, NY.

[64] de Oliveira, A. S., Kaplan, C., Mallat, K., and Chakraborty, T. (2023). An empirical analysis of fairness notions under differential privacy. *arXiv preprint arXiv:2302.02910*.

[65] Dehghan, A., Ortiz, E. G., Shu, G., and Masood, S. Z. (2017). Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*.

[66] del Barrio, E., Gordaliza, P., and Loubes, J. (2020). Review of mathematical frameworks for fairness in machine learning. *stat*, 1050:26.

[67] Derous, E. and Ryan, A. M. (2019). When your resume is (not) turning you down: Modelling ethnic bias in resume screening. *Human Resource Management Journal*, 29(2):113–130.

[68] Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc.*

[69] Ding, B., Kulkarni, J., and Yekhanin, S. (2017). Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30.

[70] Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

[71] Dodson, M. K., Cliby, W. A., Keeney, G. L., Peterson, M. F., and Podritz, K. C. (1994). Skene's gland adenocarcinoma with increased serum level of prostate-specific antigen. *Gynecologic oncology*, 55(2):304–307.

[72] Domingo-Ferrer, J. and Soria-Comas, J. (2022). Multi-dimensional randomized response. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4933–4946.

[73] Dua, D. and Graff, C. (2017). UCI machine learning repository.

[74] Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

[75] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

[76] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg.

[77] Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

[78] Emelianov, V. and Perrot, M. (2024). On the impact of output perturbation on fairness in binary linear classification. *arXiv preprint arXiv:2402.03011*.

[79] Erlingsson, U., Pihur, V., and Korolova, A. (2014). RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067, New York, NY, USA. ACM.

[80] Esmaieeli Sikaroudi, A. M., Ghousi, R., and Sikaroudi, A. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *Journal of Industrial and Systems Engineering*, 8(4):106–121.

[81] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

[82] Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[83] European Commission (2021). Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (accessed on 13 March 2024).

[84] Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570.

[85] Farrand, T., Mireshghallah, F., Singh, S., and Trask, A. (2020). Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pages 15–19.

[86] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

[87] Ficiu, B., Lawrence, N. D., and Paleyes, A. (2023). Automated discovery of trade-off between utility, privacy and fairness in machine learning models. *arXiv preprint arXiv:2311.15691*.

[88] Forré, P. and Mooij, J. M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv preprint arXiv:1807.03024*.

[89] Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143.

[90] Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767.

[91] Gajane, P. and Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

[92] Galhotra, S., Brun, Y., and Meliou, A. (2017). Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510.

[93] Galles, D. and Pearl, J. (1995). Testing identifiability of causal effects. In *Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence*, pages 185–195. ACM.

[94] Gamella, J. (2021). Greedy equivalence search (GES) algorithm for causal discovery. https://github.com/juangamella/ges. Accessed: 2022-03-16.

[95] Ganev, G., Oprisanu, B., and De Cristofaro, E. (2022). Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6944–6959. PMLR.

[96] Garg, P., Villasenor, J., and Foggo, V. (2020). Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, pages 3662–3666. IEEE.

[97] Garvie, C. (2016). *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology.

[98] Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524.

[99] Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International conference on machine learning*, pages 2357–2365. PMLR.

[100] Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37.

[101] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

[102] Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102.

[103] Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464.

[104] Heckerman, D. (2008). A tutorial on learning with bayesian networks. *Innovations in Bayesian networks: Theory and applications*, pages 33–82.

[105] Hern, A. (2016). Partnership on ai'formed by google, facebook, amazon, ibm and microsoft. *The Guardian*, 28:2016.

[106] Hitchcock, C. (2002). Probabilistic causation. *Stanford Encylopedia of Philosophy (archive)*.

[107] Huan, W., Wu, Y., Zhang, L., and Wu, X. (2020). Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*, pages 743–751, USA. ACM.

[108] Huang, Y. and Valtorta, M. (2006). Identifiability in causal bayesian networks: A sound and complete algorithm. In *Proceedings of the national conference on Artificial Intelligence*, volume 21, page 1149, London. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, AAAI Press.

[109] IEEE (2016). (the ieee global initiative on ethics of autonomus and intelligent systems. Available online: https://standards.ieee.org/industry-connections/ec/autonomous-systems/ (accessed on 14 March 2024).

[110] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[111] Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

[112] Jang, T., Zheng, F., and Wang, X. (2021). Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7908–7916.

[113] Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender systems: an introduction*. Cambridge University Press.

[114] Johndrow, J. E., Lum, K., et al. (2019). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220.

[115] Kairouz, P., Bonawitz, K., and Ramage, D. (2016). Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444.

[116] Kalainathan, D. and Goudet, O. (2019). Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*.

[117] Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

[118] Kallus, N., Mao, X., and Zhou, A. (2022). Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981.

[119] Kamiran, F. and Žliobaitė, I. (2013). Explainable and non-explainable discrimination in classification. In *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases*, pages 155–170. Springer.

[120] KAMIRAN, F., ZLIOBAITE, I., and CALDERS, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems (Print)*, 35(3):613–644.

[121] Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. (2011). What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826.

[122] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[123] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR.

[124] Khademi, A., Lee, S., Foley, D., and Honavar, V. (2019). Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, USA. ACM.

[125] Khalil, A., Ahmed, S. G., Khattak, A. M., and Al-Qirim, N. (2020). Investigating bias in facial analysis systems: A systematic review. *IEEE Access*, 8:130751–130761.

[126] Kikuchi, H. (2022). Castell: Scalable joint probability estimation of multi-dimensional data randomized with local differential privacy. *arXiv preprint arXiv:2212.01627*.

[127] Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666.

[128] Kim, J. S., Chen, J., and Talwalkar, A. (2020). Model-agnostic characterization of fairness trade-offs. *arXiv preprint arXiv:2004.03424*.

[129] Kim, M., Reingold, O., and Rothblum, G. (2018). Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, pages 4842–4852.

[130] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H., editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[131] Kocaoglu, M., Dimakis, A., and Vishwanath, S. (2017). Cost-optimal learning of causal graphs. In *International Conference on Machine Learning*, pages 1875–1884. PMLR.

[132] Krco, N., Laugel, T., Loubes, J.-M., and Detyniecki, M. (2023). When mitigating bias is unfair: A comprehensive study on the impact of bias mitigation algorithms. *arXiv preprint arXiv:2302.07185*.

[133] Kulkarni, A., Tokareva, A., Qureshi, M. R. R., and Couceiro, M. (2024). The balancing act: Unmasking and alleviating asr biases in portuguese. In *EACL 2024 LT-EDI WorkShop*.

[134] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

[135] Kurz, V., Orland, A., and Posadzy, K. (2018). Fairness versus efficiency: how procedural fairness concerns affect coordination. *Experimental economics*, 21:601–626.

[136] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

[137] Lahoti, P., Gummadi, K. P., and Weikum, G. (2019). ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE.

[138] Lam, S. K., Pitrou, A., and Seibert, S. (2015). Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15, New York, NY, USA. Association for Computing Machinery.

[139] Lambrecht, A. and Tucker, C. E. (2018). Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. *An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads (March 9, 2018)*.

[140] Leber, J. (2013). The machine-readable workforce. *MIT Technology Review*. https://www.technologyreview.com/2013/05/27/178320/the-machine-readable-workforce.

[141] Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174.

[142] Li, N., Li, T., and Venkatasubramanian, S. (2006). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE.

[143] Li, Y.-M., Peng, C., Zhang, J.-G., Zhu, W., Xu, C., Lin, Y., Fu, X.-Y., Tian, Q., Zhang, L., Xiang, Y., et al. (2019). Genetic risk factors identified in populations of european descent do not improve the prediction of osteoporotic fracture and bone mineral density in chinese populations. *Scientific reports*, 9(1):1–9.

[144] Lipton, Z., McAuley, J., and Chouldechova, A. (2018). Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135.

[145] Liu, G., Tang, P., Hu, C., Jin, C., and Guo, S. (2023). Multi-dimensional data publishing with local differential privacy. In *Proceedings of the 26th International Conference on Extending Database Technology, EDBT 2023*, pages 183–194.

[146] Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. (2018). Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*.

[147] Lopez, R., Hütter, J.-C., Pritchard, J., and Regev, A. (2022). Large-scale differentiable causal discovery of factor graphs. *Advances in Neural Information Processing Systems*, 35:19290–19303.

[148] Lyu, L., He, X., and Li, Y. (2020). Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2355–2365.

[149] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1):3–es.

[150] Mahawaga Arachchige, P. C., Bertok, P., Khalil, I., Liu, D., Camtepe, S., and Atiquzzaman, M. (2020). Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7):5827–5842.

[151] Maheshwari, G., Denis, P., Keller, M., and Bellet, A. (2022). Fair nlp models with differentially private text encoders. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6913–6930.

[152] Majdara, A. and Nematollahi, M. R. (2008). Development and application of a risk assessment tool. *Reliability Engineering & System Safety*, 93(8):1130–1137.

[153] Makhlouf, K. (2023). Impact ldp on fairness repository. https://github.com/KarimaMakhlouf/Impact_of_LDP_on_Fairness.

[154] Makhlouf, K., Arcolezi, H. H., Zhioua, S., Brahim, G. B., and Palamidessi, C. (2023). On the impact of multi-dimensional local differential privacy on fairness. *European Conference of Machine Learning (ECML)- journal track*.

[155] Makhlouf, K., Stefanović, T., Arcolezi, H. H., and Palamidessi, C. (2024a). A systematic and formal study of the impact of local differential privacy on fairness: Preliminary results. In *2024 IEEE 37th Computer Security Foundations Symposium (CSF)*, pages 1–16. IEEE.

[156] Makhlouf, K., Zhioua, S., and Palamidessi, C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642.

[157] Makhlouf, K., Zhioua, S., and Palamidessi, C. (2022). Identifiability of causal-based ml fairness notions. In *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 1–8.

[158] Makhlouf, K., Zhioua, S., and Palamidessi, C. (2024b). When causality meets fairness: A survey. *Journal of Logical and Algebraic Methods in Programming*, page 101000.

[159] Malinsky, D. and Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470.

[160] Malinsky, D., Shpitser, I., and Richardson, T. (2019). A potential outcomes calculus for identifying conditional path-specific effects. *Proceedings of machine learning research*, 89:3080.

[161] Mangold, P., Perrot, M., Bellet, A., and Tommasi, M. (2023). Differential privacy has bounded impact on fairness in classification. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23681–23705. PMLR.

[162] Mani, S., Spirtes, P. L., and Cooper, G. F. (2012). A theoretical study of y structures for causal discovery. *arXiv preprint arXiv:1206.6853*.

[163] Marope, P. T. M., Wells, P. J., and Hazelkorn, E. (2013). *Rankings and accountability in higher education: Uses and misuses*. Unesco.

[164] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

[165] Meyers, J. R. and Schmidt, F. (2008). Predictive validity of the structured assessment for violence risk in youth (savry) with juvenile offenders. *Criminal Justice and Behavior*, 35(3):344–355.

[166] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2020). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.

[167] Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

[168] Mougan, C., Álvarez, J. M., Ruggieri, S., and Staab, S. (2023). Fairness implications of encoding protected categorical attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 454–465.

[169] Mozannar, H., Ohannessian, M., and Srebro, N. (2020). Fair learning with private demographic data. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7066–7075. PMLR.

[170] Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access.

[171] Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE.

[172] Ng, I., Zhu, S., Chen, Z., and Fang, Z. (2019). A graph autoencoder approach to causal structure learning. *arXiv preprint arXiv:1911.07420*.

[173] Nikulin, M. S. (2001). Hellinger distance. *Encyclopedia of mathematics*, 78.

[174] Nogueira, A. R., Gama, J., and Ferreira, C. A. (2021). Causal discovery in machine learning: Theories and applications. *Journal of Dynamics & Games*, 8(3):203.

[175] Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., and Gama, J. (2022). Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1449.

[176] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

[177] Oja, E. and Hyvarinen, A. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

[178] O'Neill, C. (2016). Weapons of math destruction. *How Big Data Increases Inequality and Threatens Democracy*.

[179] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

[180] Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence*, pages 411–420.

[181] Pearl, J. (2009). *Causality*. Cambridge university press.

[182] Pearl, J. (2010). Causal inference. *Causality: objectives and assessment*, pages 39–58.

[183] Pearl, J. (2012). Judea pearl on potential outcomes. http://causality.cs.ucla.edu/blog/index.php/2012/12/03/judea-pearl-on-potentialoutcomes/.

[184] Pearl, J. (2022). Comment: understanding Simpson's paradox. In *Probabilistic and causal inference: The works of judea Pearl*, pages 399–412. The American Statistician.

[185] Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.

[186] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[187] Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568.

[188] Pfohl, S. R., Duan, T., Ding, D. Y., and Shah, N. H. (2019). Counterfactual reasoning for fair clinical risk prediction. In *Machine Learning for Healthcare Conference*, pages 325–358.

[189] Pinzón, C., Palamidessi, C., Piantanida, P., and Valencia, F. (2022). On the impossibility of non-trivial accuracy in presence of fairness constraints. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 7993–8000. AAAI Press.

[190] predPol (2020). predpol. https://www.predpol.com.

[191] Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., and Miklau, G. (2020). Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199.

[192] Quick, K. (2015). The unfair effects of impact on teachers with the toughest jobs. *The Century Foundation*. https://tcf.org/content/commentary/the-unfair-effects-of-impact-on-teachers-with-the-toughest-jobs/?agreed=1.

[193] Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. (2017). A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3:121–129.

[194] Ramsey, J. D., Zhang, K., Glymour, M., Romero, R. S., Huang, B., Ebert-Uphoff, I., Samarasinghe, S., Barnes, E. A., and Glymour, C. (2018). Tetrad—a toolbox for causal discovery. In *8th International Workshop on Climate Informatics*.

[195] Ren, X., Yu, C.-M., Yu, W., Yang, S., Yang, X., McCann, J. A., and Philip, S. Y. (2018). Lopub: high-dimensional crowdsourced data publication with local differential privacy. *IEEE Transactions on Information Forensics and Security*, 13(9):2151–2166.

[196] Rhee, M. (2019). Impact: The dcps evaluation and feedback system for school-based personnel. https://dcps.dc.gov/page/impact-dcps-evaluation-and-feedback-system-school-based-personnel.

[197] Rice, W. E. (1996). Race, gender, redlining, and the discriminatory access to loans, credit, and insurance: An historical and empirical analysis of consumers who sued lenders and insurers in federal and state courts, 1950-1995. *San Diego L. Rev.*, 33:583.

[198] Romei, A. and Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638.

[199] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[200] Salimi, B., Rodriguez, L., Howe, B., and Suciu, D. (2019). Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810.

[201] Santelices, M. V. and Wilson, M. (2010). Unfair treatment? the case of freedle, the sat, and the standardization approach to differential item functioning. *Harvard Educational Review*, 80(1):106–134.

[202] Sanyal, A., Hu, Y., and Yang, F. (2022). How unfair is private learning? In *Uncertainty in Artificial Intelligence*, pages 1738–1748. PMLR.

[203] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

[204] Sexton, R. S., McMurtrey, S., Michalopoulos, J. O., and Smith, A. M. (2005). Employee turnover: a neural network solution. *Computers & Operations Research*, 32(10):2635–2651.

[205] Shanmugam, K., Kocaoglu, M., Dimakis, A. G., and Vishwanath, S. (2015). Learning causal graphs with small interventions. *Advances in Neural Information Processing Systems*, 28.

[206] Shimizu, S. (2014). Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98.

[207] Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

[208] Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248.

[209] Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035.

[210] Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444.

[211] Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. In *23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007*, pages 352–359.

[212] Shpitser, I. and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979.

[213] Shrestha, Y. R. and Yang, Y. (2019). Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms*, 12(9):199.

[214] Simoiu, C., Corbett-Davies, S., Goel, S., et al. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216.

[215] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

[216] Sondhi, A. and Shojaie, A. (2019). The reduced pc-algorithm: Improved causal structure learning in large random networks. *J. Mach. Learn. Res.*, 20(164):1–31.

[217] Spanakis, E. K. and Golden, S. H. (2013). Race/ethnic difference in diabetes and diabetic complications. *Current diabetes reports*, 13(6):814–823.

[218] Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72.

[219] Spirtes, P., Meek, C., and Richardson, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21:211–252.

[220] Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen.

[221] Spirtes, P. L., Meek, C., and Richardson, T. S. (2013). Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*.

[222] Srinivasan, R., Golomb, J. D., and Martinez, A. M. (2016). A neural basis of facial action recognition in humans. *Journal of Neuroscience*, 36(16):4434–4442.

[223] Stel, V. S., Dekker, F. W., Zoccali, C., and Jager, K. J. (2013). Instrumental variable analysis. *Nephrology Dialysis Transplantation*, 28(7):1694–1699.

[224] Stolberg, H. O., Norman, G., and Trop, I. (2004). Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544.

[225] Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

[226] Suppes, P. (1973). A probabilistic theory of causality. *British Journal for the Philosophy of Science*, 24(4).

[227] Suresh, H. and Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8):73.

[228] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.

[229] Team, D. P. (2017). Learning with privacy at scale. https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

[230] Tian, J. (2004). Identifying linear causal effects. In *AAAI*, pages 104–111.

[231] Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573.

[232] Tian, J. and Pearl, J. (2015). On the identification of causal effects.

[233] Tikka, S. and Karvanen, J. (2017a). Enhancing identification of causal effects by pruning. *The Journal of Machine Learning Research*, 18(1):7072–7094.

[234] Tikka, S. and Karvanen, J. (2017b). Simplifying probabilistic expressions in causal inference. *The Journal of Machine Learning Research*, 18(1):1203–1232.

[235] Tran, C., Dinh, M., and Fioretto, F. (2021). Differentially private empirical risk minimization under the fairness lens. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27555–27565. Curran Associates, Inc.

[236] Tsagris, M. (2019). Bayesian network learning with the pc algorithm: an improved and correct variation. *Applied Artificial Intelligence*, 33(2):101–123.

[237] Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., and Floridi, L. (2022). The ethics of algorithms: key problems and solutions. *AI & SOCIETY*, 37(1):215–230.

[238] Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., and Jiang, N. (2013). Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American journal of preventive medicine*, 45(3):354–359.

[239] Van Breugel, B., Kyono, T., Berrevoets, J., and Van der Schaar, M. (2021). Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233.

[240] Van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30.

[241] VanderWeele, T. J. and Hernán, M. A. (2012). Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *American journal of epidemiology*, 175(12):1303–1310.

[242] Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE.

[243] Wachter, S. and Mittelstadt, B. (2019). A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. *Colum. Bus. L. Rev.*, page 494.

[244] Wang, S., Huang, L., Wang, P., Nie, Y., Xu, H., Yang, W., Li, X.-Y., and Qiao, C. (2016). Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*.

[245] Wang, T., Blocki, J., Li, N., and Jha, S. (2017). Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 729–745, Vancouver, BC. USENIX Association.

[246] Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

[247] Waters, A. and Miikkulainen, R. (2014). Grade: Machine learning support for graduate admissions. *AI Magazine*, 35(1):64–64.

[248] Weber, L. and Dwoskin, E. (2014). Are workplace personality tests fair? *The Wall Street Journal*. https://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257.

[249] Wen, B., Colon, L. O., Subbalakshmi, K., and Chandramouli, R. (2021). Causal-tgan: Generating tabular data using causal generative adversarial networks. *arXiv preprint arXiv:2104.10680*.

[250] Wightman, L. F. (1998). Lsac national longitudinal bar passage study. lsac research report series.

[251] Wu, X., Xu, D., Yuan, S., and Zhang, L. (2022). Fair data generation and machine learning through generative adversarial networks. In *Generative Adversarial Learning: Architectures and Applications*, pages 31–55. Springer.

[252] Wu, X. and Zhang, X. (2016). Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, pages 4038–4052.

[253] Wu, Y., Zhang, L., and Wu, X. (2018). On discrimination discovery and removal in ranked data using causal graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2536–2544.

[254] Wu, Y., Zhang, L., and Wu, X. (2019a). Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

[255] Wu, Y., Zhang, L., Wu, X., and Tong, H. (2019b). Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, pages 3404–3414.

[256] Xiong, X., Liu, S., Li, D., Cai, Z., and Niu, X. (2020). A comprehensive survey on local differential privacy. *Security and Communication Networks*, 2020:1–29.

[257] Xu, D., Yuan, S., Zhang, L., and Wu, X. (2018). Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE.

[258] Yan, J. N., Gu, Z., Lin, H., and Rzeszotarski, J. M. (2020). Silva: Interactively assessing machine learning fairness using causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

[259] Yang, M., Guo, T., Zhu, T., Tjuawinata, I., Zhao, J., and Lam, K.-Y. (2024). Local differential privacy and its applications: A comprehensive survey. *Computer Standards & Interfaces*, 89:103827.

[260] Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

[261] Ye, M. and Barg, A. (2018). Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676.

[262] Yona, G. and Rothblum, G. (2018). Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pages 5680–5688. PMLR.

[263] Yu, K., Li, J., and Liu, L. (2016). A review on algorithms for constraint-based causal discovery. *arXiv preprint arXiv:1611.03977*.

[264] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

[265] Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778.

[266] Zarya, V. (2018). The share of female ceos in the fortune 500 dropped by 25% in 2018. https://fortune.com/2018/05/21/women-fortune-500-2018/.

[267] Zhang, J. and Bareinboim, E. (2018a). Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pages 3671–3681.

[268] Zhang, J. and Bareinboim, E. (2018b). Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[269] Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655.

[270] Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press.

[271] Zhang, K., Zhu, S., Kalander, M., Ng, I., Ye, J., Chen, Z., and Pan, L. (2021). gcastle: A python toolbox for causal discovery.

[272] Zhang, L. and Wu, X. (2017). Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics*, 4(1):1–16.

[273] Zhang, L., Wu, Y., and Wu, X. (2016). Situation testing-based discrimination discovery: A causal inference approach. In *IJCAI*, volume 16, pages 2718–2724.

[274] Zhang, L., Wu, Y., and Wu, X. (2017a). Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1335–1344.

[275] Zhang, L., Wu, Y., and Wu, X. (2017b). A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3929–3935.

[276] Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., and Zhu, X. (2018). Employee turnover prediction with machine learning: A reliable approach. In *Proceedings of SAI intelligent systems conference*, pages 737–758. Springer.

[277] Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.

[278] Zheng, Y., Huang, B., Chen, W., Ramsey, J., Gong, M., Cai, R., Shimizu, S., Spirtes, P., and Zhang, K. (2024). Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8.

[279] Zhou, X. and Yamamoto, T. (2023). Tracing causal paths from experimental and observational data. *The Journal of Politics*, 85(1):250–265.

[280] Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. In *The 2nd workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) at ICML'15*.

[281] Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089.

Appendix A

Chapter 3: Fairness Notions and their Applicability

A.1 Examples for Fairness Notions Computation

A.1.1 Statistical Fairness Notions

Table A.1 illustrates a simple job hiring scenario. Each sample in the dataset has the following attributes: education level (numerical), job experience (numerical), age (numerical), marital status (categorical), gender (binary), and a label (binary). The sensitive attribute is the applicant's gender; that is, we are focusing on whether male and female applicants are treated equally. Table A.1(b) presents the predicted decision (first column) and the predicted score value (second column) for each sample. The threshold value is set to 0.5.

Statistical Parity

In the ML system of Table A.1, it means that one should not hire proportionally more applicants from one group than the other. The calculated predicted acceptance rate of hiring male and female applicants is 0.57 (4 out of 7) and 0.4 (2 out of 5), respectively. Thus, the ML system of Table A.1 does not satisfy statistical parity.

Conditional Statistical Parity

Table A.2 shows two possible combination values for E. The first combination (education level=8 and job experience=2) includes samples Female 1, Female 2, Male 4, and Male 5 for which the prediction is clearly discriminative against women as the predicted acceptance rates for men and women are 1 and 0.5, respectively. The second combination (education level=12 and job experience=8) includes Female 3 and Male 6 in which the prediction is fair

Table A.1 A simple job hiring example. Y represents the data label indicating whether the applicant is hired (1) or rejected (0). $\hat{Y}$ is the prediction which is based on the score S. A threshold of 0.5 is used.

(a) Dataset

Gender	Education Level	Job Experience	Age	Marital Status	Y
Female 1	8	2	39	single	0
Female 2	8	2	26	married	1
Female 3	12	8	32	married	1
Female 4	11	3	35	single	0
Female 5	9	5	29	married	1
Male 1	11	3	34	single	1
Male 2	8	0	48	married	0
Male 3	7	3	43	single	1
Male 4	8	2	26	married	1
Male 5	8	2	41	single	0
Male 6	12	8	30	single	1
Male 7	10	2	28	married	1

(b) Prediction

$\hat{Y}$	S
1	0.5
0	0.1
1	0.5
0	0.2
0	0.3
1	0.8
0	0.1
0	0.1
1	0.5
1	0.5
1	0.8
0	0.3

Table A.2 Application of conditional statistical parity by controlling on education level and job experience.

(a) Dataset

Gender	Education Level	Job Experience	Age	Marital Status	Y
Female 1	8	2	39	single	0
Female 2	8	2	26	married	1
Female 3	12	8	32	married	1
Male 4	8	2	26	married	1
Male 5	8	2	41	single	0
Male 6	12	8	30	single	1

(b) Prediction

$\hat{Y}$	S
1	0.5
0	0.1
1	0.5
1	0.5
1	0.5
1	0.8

(predicted acceptance rate is 1 for both applicants). Overall, the prediction is not fair as it does not hold for one combination of values of E.

Equalized Odds

In Table A.1, the TPR for male and female groups is 0.6 and 0.33, respectively, while the FPR is exactly the same (0.5) for both groups. Consequently, the equalized odds does not hold.

The scenario in Table A.3 shows an extreme case of a job hiring dataset where the male group has a large number of false positives (Male $7-100$) while equal opportunity is satisfied.

Table A.3 An extreme job hiring scenario satisfying equal opportunity. All Male $7-100$ samples are false positives (label Y is 0 and prediction $\hat{Y}$ is 1).

(a) Dataset

Gender	Education Level	Job Experience	Age	Marital Status	Y
Female 1	8	2	39	single	1
Female 2	8	2	26	married	0
Female 3	12	8	32	married	1
Male 4	8	2	26	married	1
Male 5	8	2	41	single	0
Male 6	12	8	30	single	1
Male 7	10	5	32	married	0
...	...	...	...	...	0
Male 100	8	10	27	single	0

(b) Prediction

$\hat{Y}$	S
1	0.5
0	0.1
0	0.3
1	0.5
0	0.2
0	0.4
1	0.8
1	...
1	0.7

Conditional Use Accuracy Equality

The calculated PPVs for male and female applicants in our hiring example (Table A.1) are 0.75 and 0.5, respectively. NPVs for male and female applicants are both equal to 0.33. Overall the dataset in Table A.1 does not satisfy conditional use accuracy equality.

An example Proving Proposition 4

Table A.4 illustrates an example that satisfies overall accuracy but not conditional use accuracy equality.

Treatment Equality

Table A.5 shows a dataset that fails to satisfy all previous notions, yet treatment equality is satisfied. Treatment equality can be used in real-world scenarios where only the type of misclassification rate matters for fairness.

Table A.4 A job hiring scenario satisfying overall accuracy but not conditional use accuracy equality.

	Group 1 (Female)			Group 2 (Male)			
	Gender	Y	$\hat{Y}$	Gender	Y	$\hat{Y}$	
OA = 0.625	F1	1	1	M1	1	1	OA = 0.625
PPV = 1	F2	1	0	M2	0	1	PPV = 0.4
NPV = 0.25	F3	1	0	M3	0	1	NPV = 1
	F4	0	0	M4	0	0	
	F5	1	1	M5	0	0	
	F6	1	1	M6	0	0	
	F7	1	0	M7	0	1	
	F8	1	1	M8	1	1	

Table A.5 A job hiring scenario satisfying treatment equality but not satisfying all of the previous notions.

	Group 1 (Female)			Group 2 (Male)			
	Gender	Y	$\hat{Y}$	Gender	Y	$\hat{Y}$	
TPR = 0.33	F1	1	1	M1	1	1	TPR = 0.8
FPR = 0.8	F2	0	0	M2	1	1	FPR = 0.66
PPV = 0.2	F3	0	1	M3	1	1	PPV = 0.66
NPV = 0.33	F4	0	1	M4	1	1	NPV = 0.5
OA = 0.25	F5	0	1	M5	0	0	OA = 0.625
FN/FP = 0.5	F6	0	1	M6	0	1	FN/FP = 0.5
	F7	1	0	M7	0	1	
	F8	1	0	M8	1	0	

Total Fairness

Table A.6 shows a scenario where total fairness holds. More generally, total fairness is satisfied in the very uncommon situation where the proportions of TPs, TNs, FPs, and FNs are the same in all groups.

Balance

Table A.7 shows a job hiring scenario where the average score for female candidates that should be hired ($Y = 1$) is 7.1 while it is 4.7 for male candidates. The scenario is not balanced for positive class. Note that, despite the significant difference between these two average

Table A.6 A job hiring scenario satisfying total fairness.

	Group 1 (Female)			Group 2 (Male)			
	Gender	Y	$\hat{Y}$	Gender	Y	$\hat{Y}$	
TPR = 0.5	F1	1	1	M1	1	1	TPR = 0.5
FPR = 0.66	F2	0	0	M2	1	1	FPR = 0.66
PPV = 0.33	F3	0	1	M3	0	0	PPV = 0.33
NPV = 0.5	F4	0	1	M4	0	0	NPV = 0.5
OA = 0.4	F5	1	0	M5	0	1	OA = 0.4
FN/FP = 0.5				M6	0	1	FN/FP = 0.5
				M7	0	1	
				M8	0	1	
				M9	1	0	
				M10	1	0	

values, for a score threshold value of 5, the scenario of Table A.7 satisfies both statistical parity (Eq. 3.1) and equal opportunity (Eq. 3.4).

Table A.7 A job hiring scenario satisfying statistical parity and equal opportunity (for a score threshold value of 5) but neither balance for positive class nor balance for negative class.

(a) Group 1 (Female)

Gender	Y	S
F1	1	9
F2	1	8
F3	0	8
F4	1	4.5
F5	0	4.5
F6	0	3.5

(b) Group 2 (Male)

Gender	Y	S
M1	1	6.2
M2	1	6
M3	0	5.5
M4	0	1
M5	1	2
M6	0	2

Calibration

the scenario of Table A.8 does not satisfy calibration.

Table A.9 shows a job hiring scenario satisfying calibration, but not predictive parity.

Well-Calibration Table A.10 (a) is a job hiring scenario that is calibrated (the proportion of applicants that should be hired for every score value is the same for male and female

Table A.8 A job hiring scenario satisfying predictive parity (for any threshold smaller than 0.7 or larger than 0.8) but not calibration.

(a) Group 1 (Female)

Gender	Y	S
F1	1	0.85
F2	1	0.8
F3	0	0.8
F4	1	0.7
F5	0	0.7
F6	0	0.4
F7	1	0.4
F8	0	0.4

(b) Group 2 (Male)

Gender	Y	S
M1	1	0.85
M2	1	0.8
M3	1	0.8
M4	0	0.7
M5	0	0.7
M6	1	0.4
M7	0	0.4
M8	0	0.4

Table A.9 A job hiring scenario satisfying calibration but not predictive parity (for any threshold).

(a) Group 1 (Female)

Gender	Y	S
F1	1	0.8
F2	1	0.8
F3	1	0.7
F4	1	0.7
F5	0	0.7
F6	0	0.7
F7	0	0.3
F8	0	0.3

(b) Group 2 (Male)

Gender	Y	S
M1	1	0.8
M2	1	0.8
M3	1	0.7
M4	0	0.7
M5	0	0.3
M6	0	0.3

groups) but not well-calibrated (the score value does not coincide with the proportion of applicants that should be hired). Table A.10 (b) is both calibrated and well-calibrated. Garg et al. [96] show that the difference between calibration and well-calibration is a simple difference in mapping. That is, "the scores of a calibrated predictor can, using a suitable transformation, be converted to scores satisfying well-calibration".

A.1.2 Causality-Based Fairness Notions

Table A.11 shows a simple job hiring dataset where A is the sensitive attribute corresponding to the gender ($A = 1$ for male and $A = 0$ for female), C is a covariate corresponding to the job type ($C = 0$ for flexible schedule job and $C = 1$ for non-flexible job schedule), and Y is the outcome corresponding to the hiring decision ($Y = 0$ for not-hired and $Y = 1$ for hired).

Table A.10 Calibration vs well-calibration.

(a) Calibrated but not well-calibrated

s	0.4	0.7	0.8	0.85
Female	0.33	0.5	0.6	0.6
Male	0.33	0.5	0.6	0.6

(b) Calibrated and well-calibrated

s	0.4	0.7	0.8	0.85
Female	0.4	0.7	0.8	0.85
Male	0.4	0.7	0.8	0.85

Table A.11 A job hiring example with 24 applications. A is the gender (sensitive attribute) where $A = 1$: male, $A = 0$: female. C is the job type where $C = 0$: flexible time job, $C = 1$: non-flexible time job. Y is the hiring decision (outcome) where $Y = 0$: not-hired, $Y = 0$: hired.

Female applicants (Treatment group)				Male applicants (Control Group)			
i	A	C	Y	i	A	C	Y
1:	0	0	1	13:	1	0	1
2:	0	0	1	14:	1	0	0
3:	0	0	0	15:	1	0	0
4:	0	0	0	16:	1	0	0
5:	0	0	0	17:	1	1	1
6:	0	0	0	18:	1	1	1
7:	0	0	0	19:	1	1	1
8:	0	0	0	20:	1	1	1
9:	0	1	1	21:	1	1	0
10:	0	1	1	22:	1	1	0
11:	0	1	1	23:	1	1	0
12:	0	1	0	24:	1	1	0

Total Variation (TV)

In the example of Table A.11:

$$\text{TV} = \mathbb{P}[Y = 1 \mid A = 1] - \mathbb{P}[Y = 1 \mid A = 0] = \frac{5}{12} - \frac{5}{12} = 0.$$

So according to TV, the predicted hiring decision is fair.

Later sections will show how ATE and counterfactual outcomes can be estimated from observable data. Table A.12 shows the same job hiring dataset but with counterfactual outcomes.

Table A.12 The job hiring example with counterfactual outcomes. A^{cf} denotes the candidate's gender in the counterfactual world. Y^{cf} denotes the counterfactual potential outcome.

	Female applicants (Treatment group)						Male applicants (Control Group)					
i	A	C	Y	A^{cf}	Y^{cf}	i	A	C	Y	A^{cf}	Y^{cf}
1:	0	0	1	1	1	13:	1	0	1	0	1
2:	0	0	1	1	0	14:	1	0	0	0	1
3:	0	0	0	1	1	15:	1	0	0	0	0
4:	0	0	0	1	0	16:	1	0	0	0	0
5:	0	0	0	1	0	17:	1	1	1	0	1
6:	0	0	0	1	0	18:	1	1	1	0	1
7:	0	0	0	1	0	19:	1	1	1	0	1
8:	0	0	0	1	0	20:	1	1	1	0	1
9:	0	1	1	1	1	21:	1	1	0	0	1
10:	0	1	1	1	1	22:	1	1	0	0	0
11:	0	1	1	1	0	23:	1	1	0	0	0
12:	0	1	0	1	0	24:	1	1	0	0	0

Average Total Effect (ATE)

ATE is computed by considering the average potential outcome if the gender is female $A = 0$, that is, $\frac{1}{n}\sum_{i=1}^{n}(Y_i^0)$ and the same if the gender is male $A = 1$, $\frac{1}{n}\sum_{i=1}^{n}(Y_i^1)$. The former $(\sum_{i=1}^{n}(Y_i^0))$ corresponds to the average of the observed outcomes (Y) of samples 1 to 12 and counterfactual outcomes (Y^{cf}) of samples 13 to 24, which gives $\frac{12}{24} = \frac{1}{2}$. Similarly, the average potential outcome if the gender is male corresponds to the counterfactual outcomes of samples 13 to 24 and the observed outcomes of samples 1 to 12, which gives $\frac{9}{24} = \frac{3}{8}$. Hence, ATE= $\frac{3}{8} - \frac{1}{2} = -\frac{1}{8}$ which indicates a positive bias for female.

Average Treatment Effect on the Treated (ATT)

In the example of Table A.12, ATT corresponds to the difference between the average observable outcome (Y) and the average counterfactual outcome (Y^{cf}) in samples 1 to 12, that is, ATT= $\frac{5}{12} - \frac{4}{12} = \frac{1}{12}$, which confirms the positive bias for female.

Average Treatment Effect on the Control Group (ATC)

Using the example of Table A.12, ATC $= \frac{5}{12} - \frac{7}{12} = -\frac{1}{6}$.

Conditional Average Treatment Effect (CATE)

Using the covariate $C = 0$ (flexible schedule jobs) in the hiring example of Tabel A.12, $\text{CATE}(C = 0) = \frac{3}{12} - \frac{4}{12} = -\frac{1}{12}$, which is again confirming hiring decisions in favor of female.

Individual Treatment Effect (ITE)

In Table A.12, $\text{ITE}(i = 3) = 0 - 1 = -1$ which indicates a discrimination against the female applicant $i = 3$.

An Example of Data Exhibiting the Simpson's Paradox

According to the job hiring example of Tables A.11 and A.12, there exists a statistical anomaly where some statistical notions such as TV fail to appropriately account for the bias between sub-populations (e.g., female vs. male). Notice first that, according to the collected data, both female and male candidates are hired at the same rate $\frac{5}{12}$. Notice also that if the hiring rates are adjusted according to the job type, female candidates are hired at an equal or higher rate for both types of jobs: for flexible schedule jobs ($C = 0$), the hiring rates are the same $\frac{1}{4}$ and for non-flexible jobs ($C = 1$), the hiring rates are $\frac{3}{4}$ for female and $\frac{4}{8} = \frac{1}{2}$ for male. The explanation for such a counter-intuitive result is that most female candidates (8 out of 12) are applying for flexible schedule jobs (for family reasons), in which hiring is more difficult. On the other hand, few male candidates (4 out of 12) are applying for flexible schedule jobs and instead massively applying for the more accessible non-flexible jobs (8 out of 12 applicants). To appropriately assess discrimination in this case, there is a need to adjust on the job type variable C, that is, assessing discrimination for each job type separately. This simple job hiring scenario is similar to the Berkeley sex discrimination in college admission [33] where data showed a bias for male applicants overall, but when results were analyzed separately for each department, data showed a slight bias in favor of female candidates.

Table A.13 shows an example with 30 observed samples. In such cohort, TV= $\frac{1}{15}$ indicates a discrimination against female applicants. However, all causal notions ($\text{TE} = \text{ATE} = -\frac{2}{15}$, $\text{ATT} = -\frac{2}{15}$, and $\text{ATC} = -\frac{2}{15}$, $\text{CATE}(C = 0) = -\frac{2}{15}$, and $\text{CATE}(C = 1) = -\frac{2}{15}$ are indicating a bias in favor of female.

Natural Direct Effect (NDE)

To see how NDE is computed, consider the sample dataset in Table A.14 corresponding to the causal graph in Fig. 3.5. Similarly to the previous examples, we assume the counterfactual values are available (grayed columns). The cohort consists of 6 female candidates and 6 male candidates. Y^{cf} is the counterfactual potential outcome (the gender differs from the observed sample). E_1 is the education level had the gender was male. R_1 is the hobby of the candidate had the gender was male. Y_{0,E_1,R_1} is the hiring decision had (1) the gender was female and

Table A.13 The job hiring example with a Simpson's paradox.

	Female applicants (Treatment group)						Male applicants (Control Group)					
i	A	C	Y	A^{cf}	Y^{cf}	i	A	C	Y	A^{cf}	Y^{cf}
1:	0	0	1	0	1	16:	1	0	1	1	1
2:	0	0	1	0	1	17:	1	0	0	1	1
3:	0	0	1	0	0	18:	1	0	0	1	0
4:	0	0	0	0	0	19:	1	0	0	1	0
5:	0	0	0	0	0	20:	1	0	0	1	0
6:	0	0	0	0	0	21:	1	1	1	1	1
7:	0	0	0	0	0	22:	1	1	1	1	1
8:	0	0	0	0	0	23:	1	1	1	1	1
9:	0	0	0	0	0	24:	1	1	1	1	1
10:	0	0	0	0	0	25:	1	1	1	1	1
11:	0	1	1	0	1	26:	1	1	1	1	1
12:	0	1	1	0	1	27:	1	1	1	1	1
13:	0	1	1	0	1	28:	1	1	0	1	1
14:	0	1	1	0	0	29:	1	1	0	1	0
15:	0	1	0	0	0	30:	1	1	0	1	0

Table A.14 A job hiring scenario corresponding to the causal graph in Fig. 3.5 (Section 3.4.2).

i	A	E	R	Y	Y^{cf}	E_1	R_1	Y_{0,E_1,R_1}	E_0	R_0	Y_{1,E_0,R_0}	Y_{0,E_1,R_0}
1:	0	1	0	1	1	1	1	1	1	0	1	1
2:	0	1	0	1	1	1	1	1	1	0	1	1
3:	0	1	1	0	1	0	1	1	1	1	1	0
4:	0	0	0	1	1	1	0	1	0	0	1	1
5:	0	0	0	0	1	0	1	0	0	0	1	1
6:	0	0	0	0	0	0	0	0	0	0	0	0
7:	1	1	1	1	1	1	1	1	1	0	1	1
8:	1	1	0	1	1	1	0	1	1	0	1	1
9:	1	1	1	1	0	1	1	0	1	1	0	0
10:1	0	1	1	1	0	1	1	1	0	1	1	1
11:1	0	1	0	1	0	1	1	0	0	1	0	1
12:1	0	1	0	0	0	1	0	0	0	1	1	1

(2) the education and hobby were set to the values if the candidate was male. According to Eq. (3.35), $\text{NDE}_{1,0}(y=1) = \mathbb{P}[y_{0,E_1,R_1}] - \mathbb{P}[y_1] = \frac{8}{12} - \frac{9}{12} = -\frac{1}{12}$ which indicates a direct discrimination against female candidates.

Natural Indirect Effect (NIE)

In the example of Table A.14, $\text{NIE}_{1,0}(y=1) = \mathbb{P}[y_{1,E_0,R_0}] - \mathbb{P}[y_1] = \frac{10}{12} - \frac{9}{12} = \frac{1}{12}$

Path Specific Effect (PSE)

Given $\pi = \{A \to Y, \quad A \to R \to Y\}$, $\text{PSE}^{\pi}_{1,0} = \mathbb{P}[Y_{0,E_1,R_0}] - \mathbb{P}[y_1] = \frac{8}{12} - \frac{9}{12} = -\frac{1}{12}$ which indicates a discrimination against female candidates.

Counterfactual Effects

Considering the simple job hiring example and focusing on the female group ($A = 0$), $\text{DE}_{1,0}(y|1)$ measures the change in the probability of Y (e.g., hiring) had A been 1 (female), while mediators E and R are kept at the level they would take had A been 1 (male).

Using the values in Table A.14, $\text{DE}_{1,0} = \mathbb{P}[y_{0,E_1,R_1}|0] - \mathbb{P}[y_1|0] = \frac{4}{6} - \frac{5}{6} = -\frac{1}{6}$ which indicates a direct counterfactual discrimination against female. Similarly, $\text{IE}_{1,0} = \mathbb{P}[y_{1,E_0,R_0}|0] - \mathbb{P}[y_1|0] = \frac{5}{6} - \frac{5}{6} = 0$ which indicates the absence of counterfactual indirect discrimination. $\text{SE}_{1,0}(y)$ reads the change in the probability of hiring Y had A been 0 (male) for the female candidates w.r.t the probability of hiring of male candidates. Using Table A.14, $\text{SE}_{1,0}(y) = \mathbb{P}[y_1|0] - \mathbb{P}[y|1] = \frac{5}{6} - \frac{4}{6} = \frac{1}{6}$ which indicates a spurious effect in favor of female.

Counterfactual Error Rates

Table A.15 A job hiring scenario for counterfactual direct error rate ER^d computation. $E_{1,1}$ is a short version of $E_{A=1,Y=1}$. $R_{1,1}$ means $R_{A=1,Y=1}$. $\hat{Y}_{0,1,E_{1,1},R_{1,1}}$ means $\hat{Y}_{A=0,Y=1,E_{1,1},R_{1,1}}$. $Y_{1,1}$ means $Y_{A=1,Y=1}$.

i	A	E	R	$\hat{Y}$	Y	$E_{1,1}$	$R_{1,1}$	$\hat{Y}_{0,1,E_{1,1},R_{1,1}}$	$\hat{Y}_{1,1}$
1:	0	1	0	1	1	1	0	1	1
2:	0	1	0	1	1	1	0	1	1
3:	0	1	1	0	1	0	1	1	1
4:	0	0	0	1	1	1	0	1	0
5:	0	0	0	0	1	0	0	0	0
6:	0	0	0	0	0	0	0	0	0
7:	1	1	1	1	1	1	1	1	1
8:	1	1	0	1	1	1	0	1	1
9:	1	1	1	1	0	0	1	0	1
10:	1	0	1	1	1	0	1	0	0
11:	1	0	1	0	1	0	1	1	0
12:	1	0	1	0	0	0	1	0	0

Table A.15 shows the values (observed and counterfactual) needed to compute counterfactual direct error rate ER^d for the female candidates that should be hired ($A = 1$ and

$Y = 1$).

$$\text{ER}^d(\hat{Y} = 1|A = 0, Y = 1) = \mathbb{P}[\hat{Y}_{A=0,Y=1,E_{1,1},R_{1,1}}|A = 0, Y = 1] - \mathbb{P}[\hat{Y}_{A=1,Y=1}|A = 0, Y = 1]$$

where $E_{1,1}$ is a short version of $E_{A=1,Y=1}$ which refers to the education level of the candidate had "she" been male and hired. $R_{1,1}$ means $R_{A=1,Y=1}$ and indicates the hobby of the candidate had "she" been male and hired. $\hat{Y}_{A=0,Y=1,E_{1,1},R_{1,1}}$ reads the hiring decision had the candidate was female, hired, with education $E_{1,1}$, and hobby $R_{1,1}$. $Y_{A=1,Y=1}$ reads the hiring decision had the candidate was male and hired. Using the values in Table A.15 (rows 1 to 5 in the last two columns), $\text{ER}^d(\hat{Y} = 1|A = 0, Y = 1) = \frac{4}{5} - \frac{3}{5} = \frac{1}{5}$ which indicates a higher direct error rate for the female group.

Potential Outcome Estimation Techniques

Re-weighting. Table A.16 shows the values of propensity ($e(c_i)$) as well as balance ($b(c_i)$) scores for each unit i in the simple job hiring example. Using Eq. (3.66), the $\hat{\text{ATE}}_{IPW}$ estimation of ATE is 0.25 indicating discrimination in favor of the female group.

Table A.16 Estimation of ATE using inverse propensity weighting (IPW) on the job hiring example with propensity score $e(c_i)$ and balancing score $b(c_i)$.

	Female applicants (Treatment group)						Male applicants (Control Group)					
i	A	C	Y	$e(c_i)$	$b(c_i)$	i	A	C	Y	$e(c_i)$	$b(c_i)$
1:	1	0	1	2/3	3/2	13:	0	0	1	2/3	3
2:	1	0	1	2/3	3/2	14:	0	0	0	2/3	3
3:	1	0	0	2/3	3/2	15:	0	0	0	2/3	3
4:	1	0	0	2/3	3/2	16:	0	0	0	2/3	3
5:	1	0	0	2/3	3/2	17:	0	1	1	1/3	3/2
6:	1	0	0	2/3	3/2	18:	0	1	1	1/3	3/2
7:	1	0	0	2/3	3/2	19:	0	1	1	1/3	3/2
8:	1	0	0	2/3	3/2	20:	0	1	1	1/3	3/2
9:	1	1	1	1/3	3	21:	0	1	0	1/3	3/2
10:	1	1	1	1/3	3	22:	0	1	0	1/3	3/2
11:	1	1	1	1/3	3	23:	0	1	0	1/3	3/2
12:	1	1	0	1/3	3	24:	0	1	0	1/3	3/2

$\hat{\text{ATE}}^{norm}_{IPW}$ for the same example equals 0.125, which is a perfect estimation of ATE in this case as both values coincide (ATE = 0.125).

A.1.3 An Example of Computing $\mathbb{P}[y_a]$ by Applying *do*-calculus

$\mathbb{P}[y_a]$ is identifiable in Fig. 3.9d. By applying the chain rule following the topological order: $W_2 < A < W_1 < W_3 < Y$, we get:

$$\begin{aligned}
\mathbb{P}[y_a] &= \sum_{w_1 w_2 w_3} \mathbb{P}[y|do(a), w_1, w_2, w_3]\ \mathbb{P}[w_1|do(a), w_2]\ \mathbb{P}[w_2] \\
&\qquad \times \mathbb{P}[w_3|w_2, w_1, do(a)] && \text{(A.1)} \\
&= \sum_{w_1 w_2} \mathbb{P}[y|do(a), w_1, w_2]\ \mathbb{P}[w_1|do(a), w_2]\ \mathbb{P}[w_2] && \text{(A.2)} \\
&= \sum_{w_1 w_2} \mathbb{P}[y|do(a), w_2]\ \mathbb{P}[w_1|a, w_2]\ \mathbb{P}[w_2] && \text{(A.3)} \\
&= \sum_{w_1 w_2} \sum_{a'} \mathbb{P}[y|a', w_2, do(w_1)]\ \mathbb{P}[a'|do(w_1), w_2] \\
&\qquad \times \mathbb{P}[w_1|a, w_2]\ \mathbb{P}[w_2] && \text{(A.4)} \\
&= \sum_{w_1'} \sum_{w_2'} \sum_{a'} \mathbb{P}[y|w_1', w_2', a']\ \mathbb{P}[a'|w_2']\ \mathbb{P}[w_1'|w_2', a]\ \mathbb{P}[w_2'] && \text{(A.5)}
\end{aligned}$$

Note that w_3 is omitted from (A.2) since it is considered latent [233]. Applying Rule 2 followed by Rule 3 to the first term in (A.2) yields $\mathbb{P}[y|do(a), w_2]$ (A.3). Likewise, applying Rule 2 to the second term in (A.2) leads to $\mathbb{P}[w_1|a, w_2]$. Thus, the original problem reduces to identifying the term $\mathbb{P}[y|do(a), w_2]$ in (A.3). Here, we cannot apply Rule 2 to exchange $do(a)$ with a because $G_{\underline{A}}$ (graph obtained by removing all emanating arrows from A) contains a backdoor path from A to Y. Thus, to block that path, we need to condition and sum over all values of A as shown in Eq. (A.4) ($\sum_{a'}\ \mathbb{P}[y|a', w_2, do(w_1)]\ \mathbb{P}[a'|do(w_1), w_2]$. Now, applying Rule 2 to $\mathbb{P}[y|a', w_2, do(w_1)]$ and Rule 3 to $\mathbb{P}[a'|do(w_1), w_2]$ and adding the other terms results in the final expression in (A.5). The problem of *do*-calculus is the difficulty of determining the correct order of application of the rules. Using the wrong order may hinder the identifiability or produce a very complex expression [234].

A.1.4 Computation of the counterfactual probability of the teacher firing example

To illustrate the computation of a counterfactual probability, consider the teacher firing example of Fig. 3.4 and the counterfactual probability $\mathbb{P}[y_{a_0}|a_1]$ which reads the probability of firing a teacher who is assigned a class with a high initial level of students (a_1) had she been assigned a class with a low initial level of students (a_0). Applying **make-cg** algorithm based on this counterfactual quantity produces the counterfactual graph in Fig. 3.10d, which combines two worlds: the actual world where the teacher has normally $A = a_1$ and the counterfactual world where *the same* teacher is assigned $A = a_0$. Both variables C are reduced to a single variable, and Y and Y_{a_0} are connected by an unobservable confounder.

The counterfactual graph is composed of three c-components $\{C\}, \{A\}, \{Y, Y_{a_1}\}$. Applying algorithm IDC* [212] results in:

$$\mathbb{P}[y_{a_0}|a_1] = \frac{\sum_{y,c} Q(c)\, Q(a_1)\, Q(y, y_{a_0})}{\mathbb{P}[a_1]} \tag{A.6}$$

where $Q(\mathbf{v}) = \mathbb{P}[\mathbf{v}|pa(\mathbf{V})]$ in the counterfactual graph. Hence,

$$\begin{aligned}
\mathbb{P}[y_{a_0}|a_1] &= \frac{\sum_{y,c} \mathbb{P}[c]\, \mathbb{P}[a_1|c]\, \mathbb{P}[y, y_{a_0}|c]}{\mathbb{P}[a_1]} \\
&= \frac{\sum_{c} \mathbb{P}[c]\, \mathbb{P}[a_1|c]\, \mathbb{P}[y_{a_0}|c]}{\mathbb{P}[a_1]} && \text{(A.7)} \\
&= \frac{\sum_{c} \mathbb{P}[c]\, \mathbb{P}[a_1|c]\, \mathbb{P}[y|a_0, c]}{\mathbb{P}[a_1]} && \text{(A.8)} \\
&= \frac{0.5 \times 0.2 \times 0.01 + 0.5 \times 0.8 \times 0.25}{0.5} \\
&= 0.202
\end{aligned}$$

y in Eq. (A.7) is cancelled by summation while $\mathbb{P}[y_{a_0}|c]$ in the same equation is transformed into $\mathbb{P}[y|a_0, c]$ in Eq. (A.8) using Rule 2 of the *do*-calculus.

Appendix B

Chapter 4: Impact of Privacy on Fairness

B.1 Fairness Under Multidimensional Local Differential Privacy: Empirical Study 2

B.1.1 Results of the Synthetic Dataset 2

The synthetic dataset 2 follows the exact same causal model depicted in Fig. 4.13. The data distribution is the only difference between the Synthetic datasets 1 and 2. More specifically, synthetic data 2 differs from synthetic data 1 solely by Y distribution.

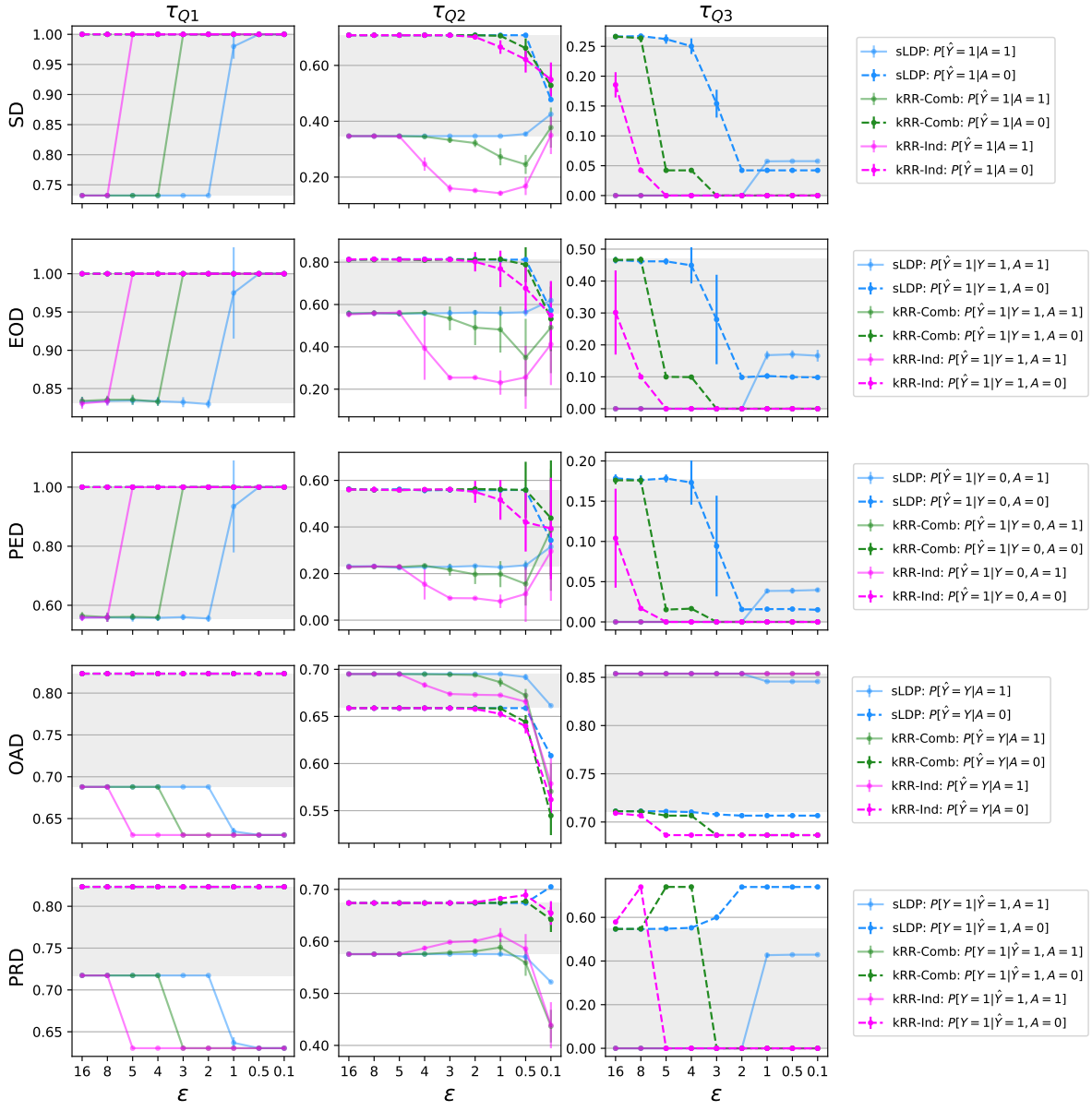


Fig. B.1 Impact of k-RR on fairness for the Synthetic datasets 2 generated with three different thresholds leading to different Y distributions. The gray shaded area represents the disparity results using the baseline model (*noLDP*).

B.1.2 Results of the Synthetic Dataset 1 and the Compas Datasets

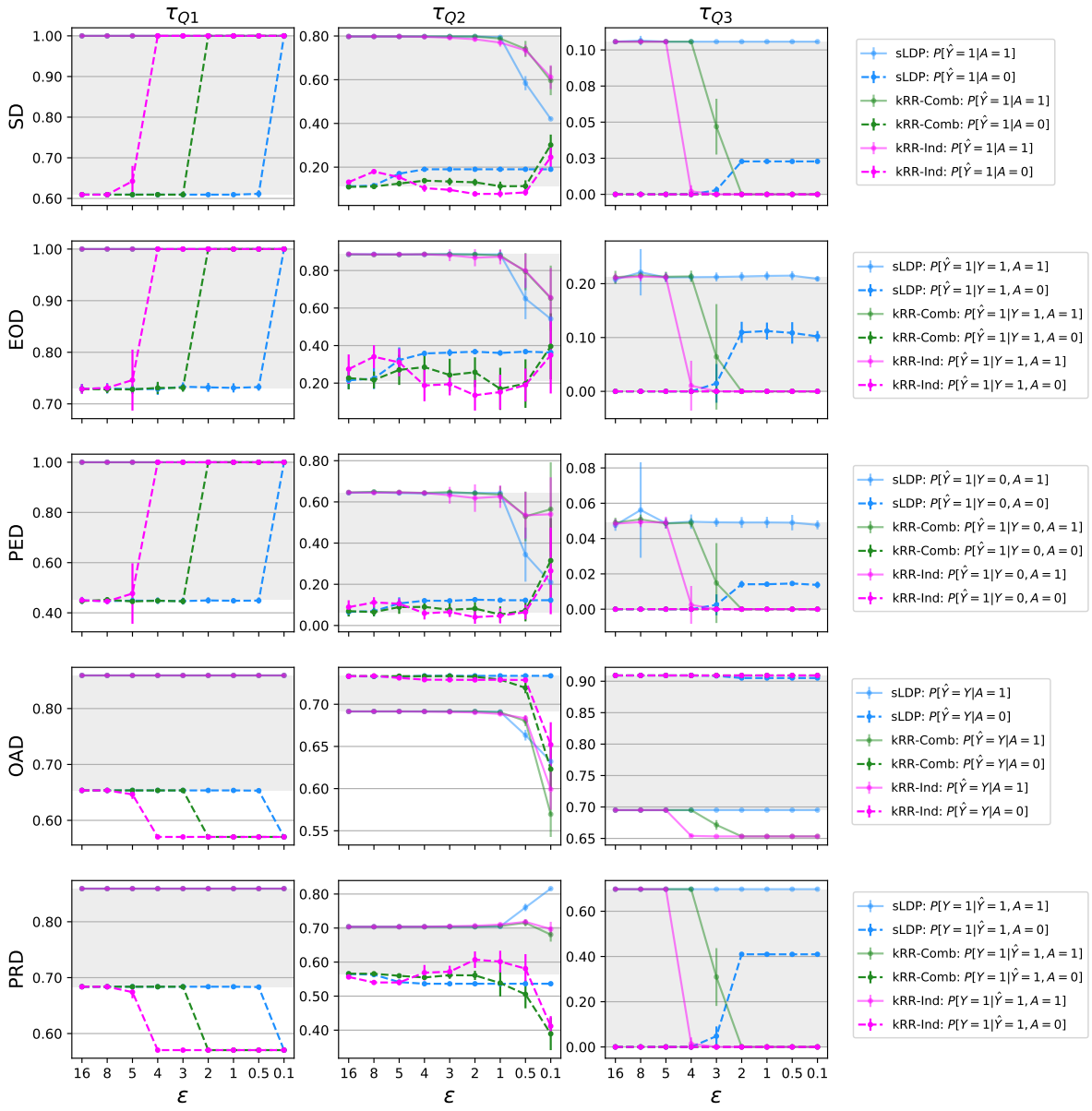


Fig. B.2 Impact of k-RR on fairness for the synthetic dataset 1 generated with three different thresholds leading to different Y distributions. The gray shaded area represents the disparity results using the baseline model (*noLDP*).

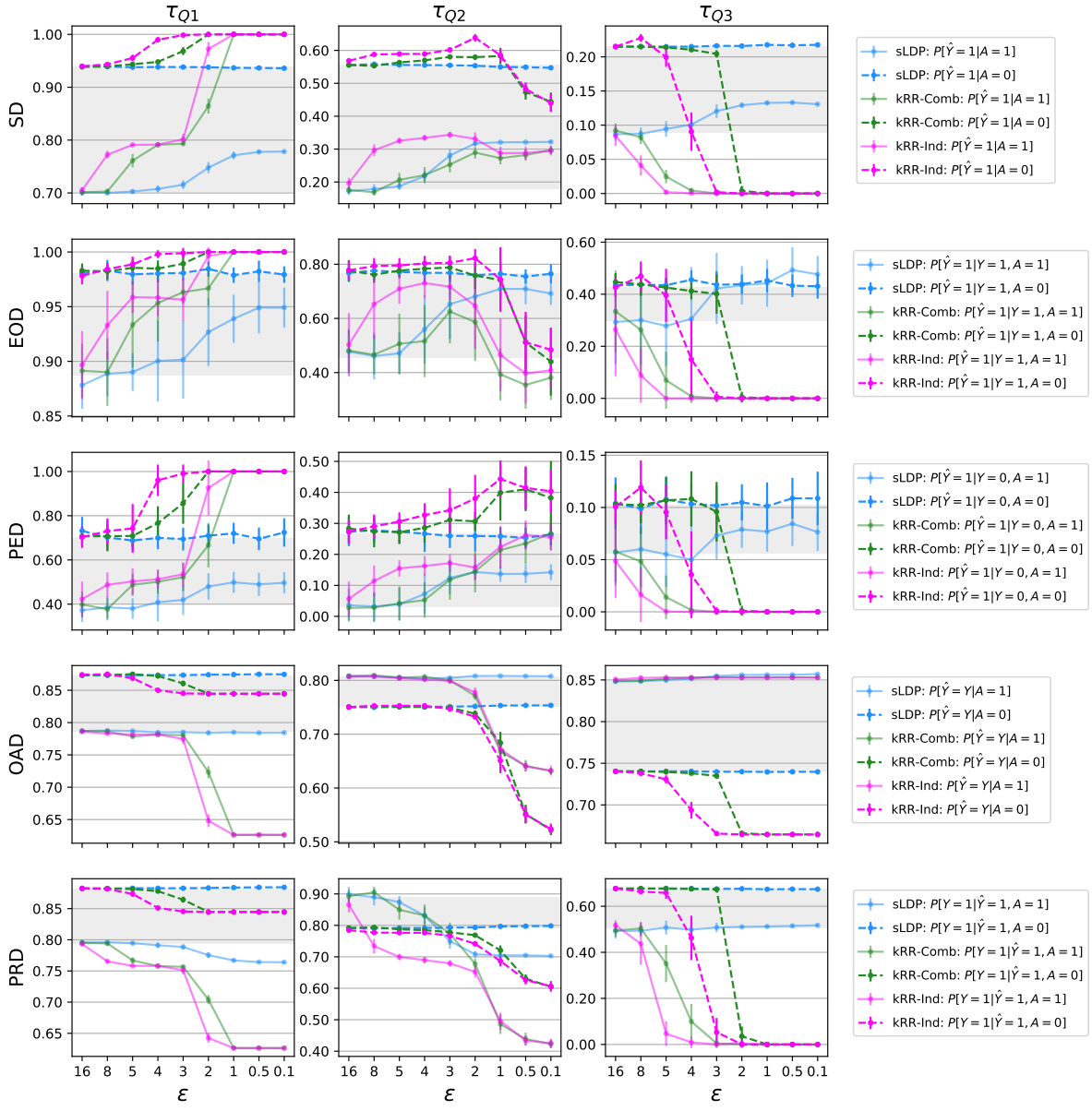


Fig. B.3 Impact of Y distribution on the privacy-fairness trade-off. Columns 1, 2, and 3 illustrate the results for the *Compas* dataset when the Y distribution is skewed to 1, balanced, and skewed to 0, respectively. The gray shaded area represents the disparity results using the baseline model (*noLDP*).

B.2 A Systematic and Formal Study of the Impact of Local Differential Privacy on Fairness

B.2.1 Proofs

Lemma 4.4.1 . $\Delta_a'^x = p\,\Delta_a^x + (1-p)\,\Delta_{\overline{a}}^x$.

Proof of Lemma 4.4.1.

$$\begin{aligned}
\Delta_a'^x &= \hat{\mathbb{P}}[Y=1, X=x, A'=a] - \hat{\mathbb{P}}[Y=0, X=x, A'=a] \\
&= p\,\hat{\mathbb{P}}[Y=1, X=x, A=a] + (1-p)\hat{\mathbb{P}}[Y=1, X=x, A=\overline{a}] - \left(p\,\hat{\mathbb{P}}[Y=0, X=x, A=a] + (1-p)\hat{\mathbb{P}}[Y=0, X=x, A=\overline{a}]\right) \\
&= p\left(\hat{\mathbb{P}}[Y=1, X=x, A=a] - \hat{\mathbb{P}}[Y=0, X=x, A=a]\right) + (1-p)\left(\hat{\mathbb{P}}[Y=1, X=x, A=\overline{a}] - \hat{\mathbb{P}}[Y=0, X=x, A=\overline{a}]\right) \\
&= p\,\Delta_a^x + (1-p)\,\Delta_{\overline{a}}^x
\end{aligned}$$

□

Theorem 4.4.1 Impact of LDP on CSD_x.

1. *if* $\mathrm{CSD}_x > 0$ *then* $0 \leq \mathrm{CSD}'_x \leq \mathrm{CSD}_x$
2. *if* $\mathrm{CSD}_x < 0$ *then* $\mathrm{CSD}_x \leq \mathrm{CSD}'_x \leq 0$
3. *if* $\mathrm{CSD}_x = 0$ *then* $\mathrm{CSD}'_x = \mathrm{CSD}_x = 0$

Proof of Theorem 4.4.1.

1. if $\mathrm{CSD}_x > 0$ then, according to Assumption 4.4.1, $\hat{Y}_1 = 1$ and $\hat{Y}_0 = 0$.
   Hence, $\Delta_1^x \geq 0$ and $\Delta_0^x < 0$.
   Using Lemma 4.4.2, we have:

$$\hat{Y}_1'^x = \begin{cases} 1 \text{ if } \Delta_1^x > 0 \text{ and } e^\varepsilon \geq -\Delta_0^x/\Delta_1^x \\ 0 \text{ if } (\Delta_1^x > 0 \text{ and } e^\varepsilon < -\Delta_0^x/\Delta_1^x) \text{ or } \Delta_1^x = 0 \end{cases}$$

   and

$$\hat{Y}_0'^x = \begin{cases} 1 \text{ if } \Delta_1^x > 0 \text{ and } e^\varepsilon \leq -\Delta_1^x/\Delta_0^x \\ 0 \text{ if } (\Delta_1^x > 0 \text{ and } e^\varepsilon > -\Delta_1^x/\Delta_0^x) \text{ or } \Delta_1^x = 0 \end{cases}$$

   Consequently, three scenarios are possible:

- $\hat{Y}_1'^x = 0 \wedge \hat{Y}_0'^x = 0$ if $\Delta_1^x = 0$
  or $\Delta_1^x > 0$ and $e^\varepsilon < -\Delta_0^x/\Delta_1^x$ and $e^\varepsilon > -\Delta_1^x/\Delta_0^x$
- $\hat{Y}_1'^x = 1 \wedge \hat{Y}_0'^x = 0$ if $\Delta_1^x > 0$ and $e^\varepsilon \geq -\Delta_0^x/\Delta_1^x$ and $e^\varepsilon > -\Delta_1^x/\Delta_0^x$
- $\hat{Y}_1'^x = 1 \wedge \hat{Y}_0'^x = 1$ if $\Delta_1^x > 0$ and $e^\varepsilon \geq -\Delta_0^x/\Delta_1^x$ and $e^\varepsilon \leq -\Delta_1^x/\Delta_0^x$

Note that the case $\hat{Y}_1'^x = 0 \wedge \hat{Y}_0'^x = 1$ is not possible. Indeed, $\hat{Y}_1'^x = 0 \wedge \hat{Y}_0'^x = 1$ implies $e^\varepsilon < -\Delta_0^x/\Delta_1^x$ and $e^\varepsilon \leq -\Delta_1^x/\Delta_0^x$. Note that the two fractions are one the inverse of the other. Hence, one of them is smaller than 1, or both are 1. Therefore, we would have $e^\varepsilon < 1$, which is not possible because $\varepsilon \geq 0$.
Hence we have $\text{CSD}'_x = 0$ or $\text{CSD}'_x = 1$, i.e., $0 \leq \text{CSD}'_x \leq \text{CSD}_x$.

2. Case 2 ($\text{CSD}_x < 0$) is analogous to case 1. That is, proving this case amounts to replacing 0 by 1 and 1 by 0 in case 1 proof.

3. if $\text{CSD}_x = 0$, two cases are possibles:

   - $\hat{Y}_1^x = 0 \wedge \hat{Y}_0^x = 0$. This means that $\Delta_1^x < 0 \wedge \Delta_0^x < 0$. By Lemma 4.4.2, we derive $\hat{Y}_1'^x = 0 \wedge \hat{Y}_0'^x = 0$. Hence, $\text{CSD}'_x = 0$.

   - $\hat{Y}_1^x = 1 \wedge \hat{Y}_0^x = 1$. This means that $\Delta_1^x \geq 0 \wedge \Delta_0^x \geq 0$. By Lemma 4.4.2, we derive $\hat{Y}_1'^x = 1 \wedge \hat{Y}_0'^x = 1$. Hence, $\text{CSD}'_x = 0$.

□

Lemma 4.4.3 Quantification of SD.

$$\text{SD} = \begin{cases} \mathbb{P}[\Delta_1^X \geq 0 \ \wedge \ \Delta_0^X < 0] & \textit{if } \exists x \ \Gamma_1^x > \Gamma_0^x \\ 0 & \textit{if } \forall x \ \Gamma_1^x = \Gamma_0^x \\ -\mathbb{P}[\Delta_1^X < 0 \ \wedge \ \Delta_0^X \geq 0] & \textit{if } \exists x \ \Gamma_1^x < \Gamma_0^x \end{cases}$$

Proof of Lemma 4.4.3.

$$
\begin{aligned}
\mathrm{SD} &\overset{\text{def}}{=} \mathbb{P}[\hat{Y}=1|A=1] - \mathbb{P}[\hat{Y}=1|A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}=1, X=x|A=1] - \sum_x \mathbb{P}[\hat{Y}=1, X=x|A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}=1|X=x, A=1]\cdot\mathbb{P}[X=x|A=1] - \sum_x \mathbb{P}[\hat{Y}=1|X=x, A=0]\cdot\mathbb{P}[X=x|A=0] \\
&\overset{(a)}{=} \sum_x \hat{Y}_1^x\ \mathbb{P}[X=x|A=1] - \sum_x \hat{Y}_0^x\ \mathbb{P}[X=x|A=0] \\
&\overset{(b)}{=} \sum_x \hat{Y}_1^x\ \mathbb{P}[X=x] - \sum_x \hat{Y}_0^x\ \mathbb{P}[X=x] \\
&\overset{(c)}{=} \sum_{\substack{x:\\ \Delta_1^x\geq 0}} \mathbb{P}[X=x] - \sum_{\substack{x:\\ \Delta_0^x\geq 0}} \mathbb{P}[X=x]
\end{aligned}
\tag{B.1}
$$

In step (a), we replace $\mathbb{P}[\hat{Y}=1|X=x, A=1]$ and $\mathbb{P}[\hat{Y}=1|X=x, A=0]$ by their corresponding abbreviated forms $\hat{Y}_1^x$ and $\hat{Y}_0^x$. Step (b) follows from $X \perp A$. Step (c) follows because $\hat{Y}_1^x = 1$ when $\Delta_1^x \geq 0$, and $\hat{Y}_1^x = 0$, otherwise. Similarly, $\hat{Y}_0^x = 1$ when $\Delta_0^x \geq 0$, and $\hat{Y}_0^x = 0$, otherwise.

Then, we consider three cases:

- Case $\exists x\ \Gamma_1^x > \Gamma_0^x$. By Assumption 4.4.3 (uniform discrimination) we have that $\forall x\ \Gamma_1^x \geq \Gamma_0^x$. Also, recall that $\Gamma_a^x \geq 0$ if and only if $\Delta_a^x \geq 0$. Therefore, in the expression (B.1), for each x such that $\Delta_0^x \geq 0$, we also have $\Delta_1^x \geq 0$, which concludes the proof for this case.
- Case $\forall x\ \Gamma_1^x = \Gamma_0^x$. We have that $\Delta_0^x \geq 0$ if and only if $\Delta_1^x \geq 0$, hence the two terms in the expression (B.1) are equal.
- Case $\exists x\ \Gamma_1^x < \Gamma_0^x$. This is the symmetric of the first case. Following the same reasoning (with 0 and 1 exchanged), we have that, in the expression (B.1), for each x such that $\Delta_1^x \geq 0$, we also have $\Delta_0^x \geq 0$.

□

Lemma 4.4.4 Quantification of SD′.

$$
\mathrm{SD}' = \begin{cases}
\mathbb{P}[\Delta_1'^X \geq 0\ \wedge\ \Delta_0'^X < 0] & \text{if } \exists x\ \Gamma_1'^x > \Gamma_0'^x \\
0 & \text{if } \forall x\ \Gamma_1'^x = \Gamma_0'^x \\
-\mathbb{P}[\Delta_1'^X < 0\ \wedge\ \Delta_0'^X \geq 0] & \text{if } \exists x\ \Gamma_1'^x < \Gamma_0'^x
\end{cases}
$$

Proof of Lemma 4.4.4.

$$
\begin{aligned}
\mathrm{SD}' &\stackrel{\text{def}}{=} \mathbb{P}[\hat{Y}'=1|A=1] - \mathbb{P}[\hat{Y}'=1|A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}'=1, X=x|A=1] - \sum_x \mathbb{P}[\hat{Y}'=1, X=x|A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}'=1|X=x, A=1]\cdot\mathbb{P}[X=x, A=1] - \sum_x \mathbb{P}[\hat{Y}'=1|X=x, A=0]\cdot\mathbb{P}[X=x, A=0] \\
&= \sum_x \hat{Y}'^{x}_{1}\, \mathbb{P}[X=x|A=1] - \sum_x \hat{Y}'^{x}_{0}\, \mathbb{P}[X=x|A=0] \\
&= \sum_{\substack{x:\\ \Delta'^{x}_{1}\geq 0}} \mathbb{P}[X=x] - \sum_{\substack{x:\\ \Delta'^{x}_{0}\geq 0}} \mathbb{P}[X=x]
\end{aligned}
$$

The proof proceeds like the one in Lemma 4.4.3. We need, however, the following result, which states that LDP obfuscation preserves *uniform discrimination assumption* (Assumption 4.4.3).

Lemma B.2.1 . *If* $\exists x^*\ \Gamma'^{x^*}_{a} > \Gamma'^{x^*}_{\overline{a}}$ *then* $\forall x\ \Gamma'^{x}_{a} \geq \Gamma'^{x}_{\overline{a}}$

Proof

We prove the property by showing that $\Gamma'^{x}_{a} > \Gamma'^{x}_{\overline{a}}$ if and only if $\Gamma^{x}_{a} > \Gamma^{x}_{\overline{a}}$, and that $\Gamma'^{x}_{a} = \Gamma'^{x}_{\overline{a}}$ if and only if $\Gamma^{x}_{a} = \Gamma^{x}_{\overline{a}}$. Then, clearly, the statement of the theorem derives from the assumption of *uniform discrimination* for the original data (before obfuscation).

It is easy to see that

$$
\Gamma'^{x}_{a} = \frac{p\Delta^{x}_{a} + (1-p)\Delta^{x}_{\overline{a}}}{p\mathbb{P}[X=x, A=a] + (1-p)\mathbb{P}[X=x, A=\overline{a}]}
$$

Let us prove that $\Gamma'^{x}_{a} > \Gamma'^{x}_{\overline{a}}$ if and only if $\Gamma^{x}_{a} > \Gamma^{x}_{\overline{a}}$:

$$
\begin{aligned}
\Gamma'^{x}_{a} &> \Gamma'^{x}_{\overline{a}} \\
\Leftrightarrow & \\
\frac{p\Delta^{x}_{a} + (1-p)\Delta^{x}_{\overline{a}}}{p\mathbb{P}[X=x, A=a] + (1-p)\mathbb{P}[X=x, A=\overline{a}]} &> \frac{p\Delta^{x}_{\overline{a}} + (1-p)\Delta^{x}_{a}}{p\mathbb{P}[X=x, A=\overline{a}] + (1-p)\mathbb{P}[X=x, A=a]} \\
\Leftrightarrow & \\
p^2\Delta^{x}_{a}\mathbb{P}[X=x, A=\overline{a}] + (1-p)^2\Delta^{x}_{\overline{a}}\mathbb{P}[X=x, A=a] &> p^2\Delta^{x}_{\overline{a}}\mathbb{P}[X=x, A=a] + (1-p)^2\Delta^{x}_{a}\mathbb{P}[X=x, A=\overline{a}] \\
\Leftrightarrow & \\
\left.\begin{array}{r} p^2\Gamma^{x}_{a}\mathbb{P}[A=a, X=x]\mathbb{P}[A=\overline{a}, X=x] \\ + \\ (1-p)^2\Gamma^{x}_{\overline{a}}\mathbb{P}[A=\overline{a}, X=x]\mathbb{P}[A=a, X=x] \end{array}\right\} &> \left\{\begin{array}{l} p^2\Gamma^{x}_{\overline{a}}\mathbb{P}[A=\overline{a}, X=x]\mathbb{P}[A=a, X=x] \\ + \\ (1-p)^2\Gamma^{x}_{a}\mathbb{P}[A=\overline{a}, X=x]\mathbb{P}[A=a, X=x] \end{array}\right. \\
\Leftrightarrow & \\
\mathbb{P}[A=\overline{a}, X=x]\mathbb{P}[A=a, X=x]\left(p^2\Gamma^{x}_{a} + (1-p)^2\Gamma^{x}_{\overline{a}}\right) &> \mathbb{P}[A=\overline{a}, X=x]\mathbb{P}[A=a, X=x]\left(p^2\Gamma^{x}_{\overline{a}} + (1-p)^2\Gamma^{x}_{a}\right) \\
\Leftrightarrow & \\
p^2\left(\Gamma^{x}_{a} - \Gamma^{x}_{\overline{a}}\right) &> (1-p)^2\left(\Gamma^{x}_{a} - \Gamma^{x}_{\overline{a}}\right) \\
\Leftrightarrow & \\
\Gamma^{x}_{a} &> \Gamma^{x}_{\overline{a}}
\end{aligned}
$$

The property $\Gamma'^x_a = \Gamma'^x_{\bar{a}}$ if and only if $\Gamma^x_a = \Gamma^x_{\bar{a}}$ can be proved similarly, just replace the ">" symbol by "=". □

Theorem 4.4.3 Impact of LDP on SD. Case $X \not\perp A$.

1. *if* $\exists x\ \Gamma^x_1 > \Gamma^x_0$ *then* $\mathrm{SD}' \leq \mathrm{SD}$
2. *if* $\exists x\ \Gamma^x_1 < \Gamma^x_0$ *then* $\mathrm{SD} \leq \mathrm{SD}'$
3. *if* $\forall x\ \Gamma^x_1 = \Gamma^x_0$ *then* $\mathrm{SD}' = \mathrm{SD}$

Proof of Theorem 4.4.3. We prove the result for the case $\exists x\ \Gamma^x_1 > \Gamma^x_0$, the other two cases can be proven similarly.
Recall that:

$$\mathrm{SD} = \sum_{\substack{x:\\ \Delta^x_1 \geq 0}} \mathbb{P}[X = x|A = 1] - \sum_{\substack{x:\\ \Delta^x_0 \geq 0}} \mathbb{P}[X = x|A = 0] \quad \hat{Y}^x_1, \hat{Y}^x_0 = 1$$

Since we are considering the case $\exists x\ \Gamma^x_1 > \Gamma^x_0$, from Assumption 4.4.3 (*uniform discrimination*) we derive $\forall x\ \Gamma^x_1 \geq \Gamma^x_0$, hence:

$$\mathrm{SD} = \sum_{\substack{x:\\ \Delta^x_1, \Delta^x_0 \geq 0}} \mathbb{P}[X = x|A = 1] - \mathbb{P}[X = x|A = 0] \ + \sum_{\substack{x:\\ \Delta^x_1 \geq 0, \Delta^x_0 < 0}} \mathbb{P}[X = x|A = 1]$$

After obfuscation, from Lemma B.2.1 we have that $\forall x\ \Gamma'^x_1 > \Gamma'^x_0$. Hence:

$$\begin{aligned}
\mathrm{SD}' &= \sum_{\substack{x:\\ \Delta'^x_1, \Delta'^x_0 \geq 0}} \mathbb{P}[X = x|A = 1] - \mathbb{P}[X = x|A = 0] \ + \sum_{\substack{x:\\ \Delta'^x_1 \geq 0, \Delta'^x_0 < 0}} \mathbb{P}[X = x|A = 1] \\
&= \sum_{\substack{x:\\ \Delta^x_1, \Delta^x_0 \geq 0}} \mathbb{P}[X = x|A = 1] - \mathbb{P}[X = x|A = 0] \ + \sum_{\substack{x:\\ \Delta^x_1 > 0, \Delta^x_0 < 0,\\ -\frac{\Delta^x_0}{\Delta^x_1} \leq e^\varepsilon \leq -\frac{\Delta^x_1}{\Delta^x_0}}} \mathbb{P}[X = x|A = 1] - \mathbb{P}[X = x|A = 0] \\
&\quad + \sum_{\substack{x:\\ \Delta^x_1 \geq 0, \Delta^x_0 < 0\\ e^\varepsilon \geq -\frac{\Delta^x_0}{\Delta^x_1},\\ e^\varepsilon > -\frac{\Delta^x_1}{\Delta^x_0}}} \mathbb{P}[X = x|A = 1]
\end{aligned}$$

By case analysis, and similar to the proof of Theorem 4.4.2, we can conclude Theorem 4.4.3. The main difference with Theorem 4.4.2, is that SD' contains the additional term

$$\sum_{\substack{x:\\ \Delta^x_1 > 0, \Delta^x_0 < 0,\\ -\frac{\Delta^x_0}{\Delta^x_1} \leq e^\varepsilon \leq -\frac{\Delta^x_1}{\Delta^x_0}}} \mathbb{P}[X = x|A = 1] - \mathbb{P}[X = x|A = 0]$$

which can be negative and large enough to cause SD′ to go below 0. Hence, SD′ and SD can be opposite signs. □

Theorem 4.4.4 Impact of LDP on EOD.

1. *if* $\text{EOD} > 0$ *then* $0 \leq \text{EOD}' \leq \text{EOD}$
2. *if* $\text{EOD} < 0$ *then* $\text{EOD} \leq \text{EOD}' \leq 0$
3. *if* $\text{EOD} = 0$ *then* $\text{EOD}' = \text{EOD} = 0$

Proof of Theorem 4.4.4.

$$
\begin{aligned}
\text{EOD} &\overset{\text{def}}{=} \mathbb{P}[\hat{Y}=1|Y=1,A=1] - \mathbb{P}[\hat{Y}=1|Y=1,A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}=1, X=x|Y=1,A=1] - \sum_x \mathbb{P}[\hat{Y}=1, X=x|Y=1,A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}=1|X=x,Y=1,A=1]\cdot\mathbb{P}[X=x|Y=1,A=1] - \sum_x \mathbb{P}[\hat{Y}=1|X=x,Y=1,A=0]\cdot\mathbb{P}[X=x|Y=1,A=0] \\
&= \sum_x \frac{\mathbb{P}[\hat{Y}=1,Y=1|X=x,A=1]}{\mathbb{P}[Y=1|X=x,A=1]}\cdot\mathbb{P}[X=x|Y=1,A=1] - \sum_x \frac{\mathbb{P}[\hat{Y}=1,Y=1|X=x,A=0]}{\mathbb{P}[Y=1|X=x,A=0]}\cdot\mathbb{P}[X=x|Y=1,A=0] \\
&\overset{(a)}{=} \sum_{\substack{x:\\ \Delta_1^x\geq 0}} \mathbb{P}[X=x|Y=1,A=1] - \sum_{\substack{x:\\ \Delta_0^x\geq 0}} \mathbb{P}[X=x|Y=1,A=0] \\
&\overset{(b)}{=} \sum_{\substack{x:\\ \Delta_1^x\geq 0}} \mathbb{P}[X=x|Y=1] - \sum_{\substack{x:\\ \Delta_0^x\geq 0}} \mathbb{P}[X=x|Y=1]
\end{aligned}
$$

(a) follows from the fact that both $\frac{\mathbb{P}[\hat{Y}=1,Y=1|X=x,A=1]}{\mathbb{P}[Y=1|X=x,A=1]}$ and $\frac{\mathbb{P}[\hat{Y}=1,Y=1|X=x,A=0]}{\mathbb{P}[Y=1|X=x,A=0]}$ are equal to 1 for $x:\Delta_1^x \geq 0$ and $x:\Delta_0^x \geq 0$, respectively. And (b) follows because of the reliability assumption 4.4.4.

Now, after obfuscation and following the same reasoning as in the proofs of Theorems 4.4.2 and 4.4.3, we have:

$$
\begin{aligned}
\text{EOD}' = &\sum_{\substack{x:\\ \Delta_1^x,\Delta_0^x\geq 0}} \mathbb{P}[X=x|Y=1] - \mathbb{P}[X=x|Y=1] \quad + \sum_{\substack{x:\\ \Delta_1^x>0,\Delta_0^x<0,\\ -\frac{\Delta_0^x}{\Delta_1^x}\leq e^\varepsilon\leq -\frac{\Delta_1^x}{\Delta_0^x}}} \mathbb{P}[X=x|Y=1] - \mathbb{P}[X=x|Y=1] \\
&+ \sum_{\substack{x:\\ \Delta_1^x,\Delta_0^x\geq 0,\\ e^\varepsilon\geq -\frac{\Delta_0^x}{\Delta_1^x},\\ e^\varepsilon> -\frac{\Delta_1^x}{\Delta_0^x}}} \mathbb{P}[X=x|Y=1] - \mathbb{P}[X=x|Y=1]
\end{aligned}
$$

The rest is deduced by case analysis.

□

B.2.2 Results for S7

Below are the data distribution (Table B.1) and the results of the dataset S7. The data was generated following the causal graph depicted in Fig. 4.17(c). The results of applying privacy on fairness are illustrated in Fig. B.4. Note that in this dataset, the Assumption 4.4.4 is satisfied.

Table B.1 Distributions of the synthetic dataset S7.

$Y=1$		$X=0$	$X=1$	$X=2$	$X=3$	$X=4$
	$A=1$	0.05	0.07	0.04	0.06	0.05
	$A=0$	0.05	0.07	0.04	0.06	0.05
$Y=0$		$X=0$	$X=1$	$X=2$	$X=3$	$X=4$
	$A=1$	0	0.06	0.05	0.02	0
	$A=0$	0.09	0.04	0.06	0.02	0.12

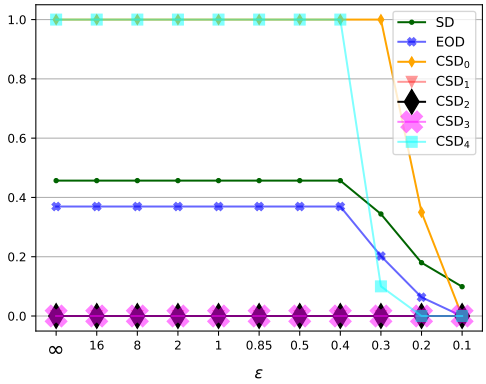


Fig. B.4 Results for the synthetic dataset S7.

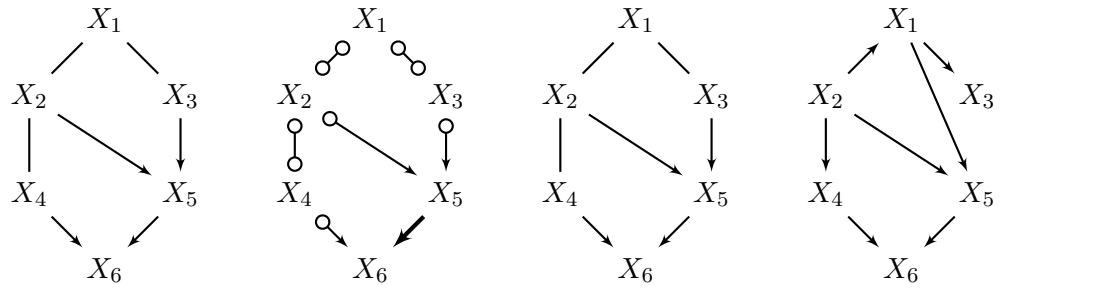
Appendix C

Chapter 5: Causal Discovery Through the Lens of Fairness

C.1 Additional Experiments for Section 5.2

C.1.1 Results for the Second Synthetic Dataset with Gaussian Noise

Fig. C.1 shows the graphs generated from the second dataset following the same causal structure (Fig. 5.1) but with Gaussian noise. PC, FCI, and GES generate the same graph structure as the first dataset. The only additional detail is that FCI is very confident about the $X_5 \to X_6$ edge (highlighted with a thicker edge). DirectLiNGAM, however, generates a graph with several discrepancies compared with the correct graph. Both graphs generated by LiNGAM fail even to identify v-structures correctly. This shows that LiNGAM is not reliable when the non-Gaussianity assumption of the noise does not hold.



(a) PC (b) FCI (c) GES (d) DirectLiNGAM

Fig. C.1 Generated causal graphs for the synthetic dataset with Gaussian noise.

SBCN. A Suppes-Bayes Causal Network (SBCN) [36] is a different type of causal graph that is used specifically for fairness assessment purposes. SBCN deviates from the causal graphs used above in three aspects. First, vertices in an SBCN correspond to Bernoulli variables with binary values. For example, $\langle$ Gender = female $\rangle$ and $\langle$ Gender = male $\rangle$ correspond to two different vertices. Second, causal relations between vertices follow Suppes's definition of causality [106, 226] (different from the typical definition of causality [181]) which requires temporal priority and probability raising. For example, a node a is a cause of a node y ($a \rightarrow y$) if and only if a occurs before y (temporal priority) and the cause a raises the probability of the effect y, that is, $\mathbb{P}[y|a] > \mathbb{P}[y|\neg a]$ (probability raising). Third, every edge (causal relation) is assigned a weight corresponding to the confidence score. The weight is simply the extent of the probability raising ($W(a, y) = \mathbb{P}[y|a] - \mathbb{P}[y|\neg a]$).

Discovering the SBCN structure from the data is a hybrid approach using constraint-based as well as score-based ideas.

Measuring fairness/discrimination using the generated SBCN is based on random walks. That is, based on the weighted edges between vertices, it is possible to measure several types of fairness notions (e.g., group and individual discrimination, direct and indirect discrimination, etc.). For instance, group discrimination is measured using a number n of random walks that begin from a node v (e.g., $\langle \mathsf{gender} = \mathsf{female} \rangle$) and reach the node corresponding to the negative decision (e.g., $\langle \mathsf{decision} = \mathsf{not\ hired} \rangle$). This corresponds to the discrimination score ds^-:

$$ds^-(v) = \frac{rw_{v \rightarrow \delta^-}}{n} \tag{C.1}$$

where $\delta^- \in V$ represents the node of the negative decision (e.g., *not hired*) and $rw_{v \rightarrow \delta^-}$ represents the number of random walks starting at vertex v and reaching δ^- earlier than δ^+ (node of the positive decision e.g. *hired*). Note that the choice of a path in a random walk is based on the weights of the out-goings edges. Being at node x, the probability of moving to node y rather than another neighbor node is:

$$p(x, y) = \frac{W(x, y)}{\sum_{z \in \mathsf{outgoing}(x)} W(x, z)} \tag{C.2}$$

$\mathsf{outgoing}(x)$ represents the set of outgoing edges from x. In case a random walk reaches a node with no outgoing edges before attaining the decision node, it is restarted from the starting node.

SBCN is used similarly to compute favoritism (positive discrimination), indirect, genuine, individual, and sub-group discrimination.

C.1.2 Results for the Dutch Census Dataset

The *Dutch Census* dataset consists of 60,420 tuples where the sensitive attribute is the sex of an individual and the outcome is her occupation (job). Six attributes are used for structural learning, namely age, sex, economic status, education, marital status, and occupation. Age is continuous, while the remaining variables are discrete. Three tiers in the partial order for temporal priority are used: age and sex are defined in the first tier, education is in the second tier, and marital status, economic status, and occupation are defined in the third tier. When found to be mediators, only education considered explaining variable. The rest (age, employment status, and marital status) are considered redlining variables.

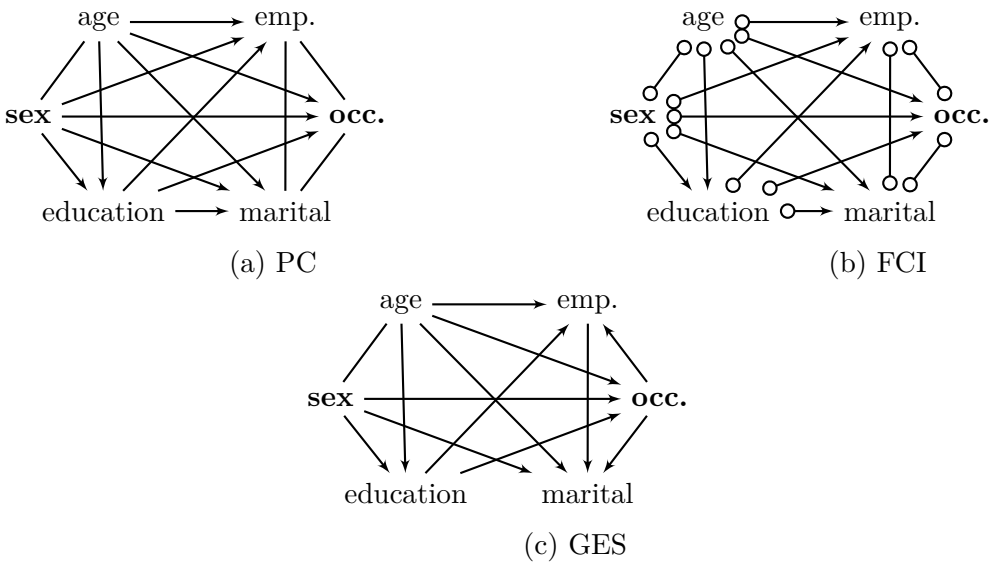


Fig. C.2 Generated causal graph for the Dutch census dataset. Occ. stands for occupation and emp. stands for whether an individual is employed or not.

The obtained graph in Fig. C.2 shows that PC and FCI produce very similar structures, which are significantly different from the GES graph. In particular, the $status - occupation$, $marital - occupation$, and $status - marital$ edges are undirected in PC and FCI, but directed in GES. This has a significant consequence on the set of possible causal paths between the sensitive attribute and the outcome.

As shown in Fig. C.3, total effect measures (as high as 0.3 obtained when age is a confounder $age \rightarrow sex$)) indicate a significant discrimination against females. The highest variability is observed for DE values. When age is a confounder and employment status and marital status variables are mediators, PC and FCI graphs exhibit 6 causal paths

For PC and FCI, there are in total 17 causal paths between *sex* and *occupation*, whereas in GES, there are only 4.

Note that the edge *emp.* − *marital* should be left unoriented as orienting it in either way will create a collider. The edges *emp.* − **occ.** and *marital.* − **occ.** should be oriented as *emp.* → **occ.** and *marital.* → **occ.** for the same reason (i.e.; not creating colliders).

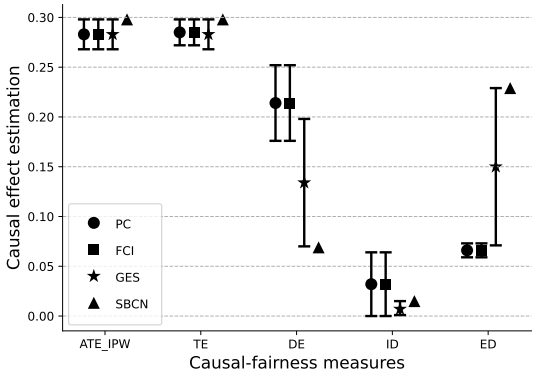


Fig. C.3 Estimation of causal effects of the Dutch census dataset based on PC, FCI, GES, and SBCN.

As shown in Figure C.3, all measures (except ID) are negative and hence indicate a discrimination *against* female individuals. In particular, total effect values are around −0.3 according to all graphs. According to PC and FCI graphs, there is a high variability in the value of ED depending on whether *age* variable is a confounder or a mediator (redlining) and whether *status* and *marital* variables are mediators or colliders. For instance, if *age* is a confounder and *status* and *marital* are mediators, apart from the direct causal link (*sex* → *occupation*), there will be only one causal path, namely, *sex* → *education* → *occupation*, and hence the DE will be at its lowest (−0.25). The only measure that might return positive values is ID for PC and FCI where there are 7 possible indirect causal paths going through redlining variables. The total indirect discrimination can be slightly positive, indicating a discrimination *in favor* of females. If taken separately, such values are misleading because they should be considered along with direct discrimination (DE). For GES, there is only one possible indirect discrimination path (*sex* → *age* → *occupation*) and two possible explaining discrimination paths (*sex* → *education* → *occuptation* and *sex* → *age* → *education* → *occupation*).

C.1.3 Results for the German Credit Dataset

The *German credit* dataset[1] contains data of 1000 individuals applying for loans. The variables used for causal graph generation are sex, age, credit amount, employment length, and default. Age and credit amount are continuous, while the remaining variables are discrete. This dataset is designed for binary classification to predict whether an individual will default on the loan (1) or not (0). We consider sex as a sensitive feature where female applicants are compared to male applicants. Three tiers in the partial order for temporal priority are

[1]https://archive-beta.ics.uci.edu/ml/datasets/statlog+german+credit+data

used: age and sex are defined in the first tier, credit amount and employment length in the second tier, and default is defined in the third tier. If found as a mediator, the age variable is considered as redlining.

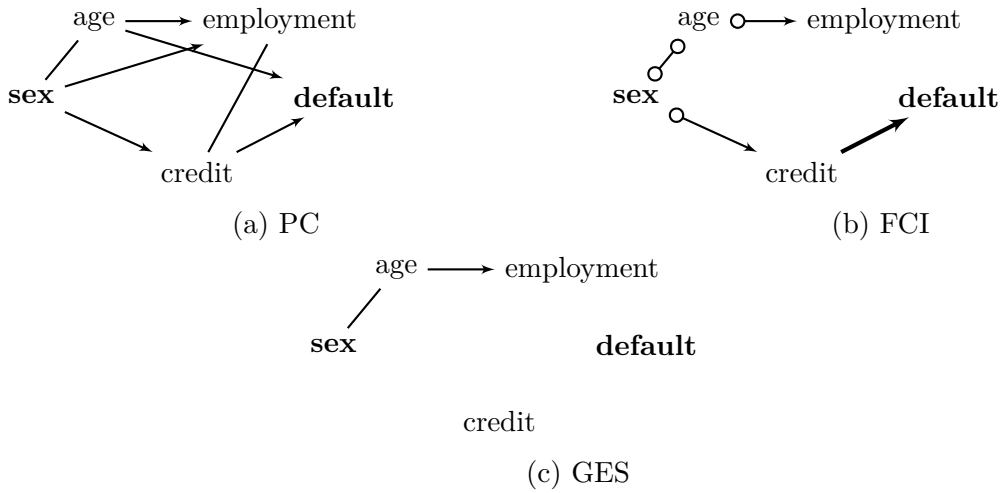


Fig. C.4 Generated causal graph for the German credit dataset.

Compared to previous datasets, German credit leads to sparser causal graphs (Fig.s C.4). The most extreme case is GES, which could not identify any dependence between sex and default variables. Besides, no algorithm could detect a direct dependence between sex and default variables. Interestingly, all graphs (except GES) show a causal relation from credit amount to default, with FCI very confident about it. Discrimination values in Fig. C.5 show that all discrimination measures are either zero or slightly positive, indicating small discrimination against females. For GES, all causal effects are equal to zero due to the absence of any causal path from sex to default. The range of TE *and* ATE_{IPW} values for PC is relatively wide. The lowest value (0.021) is obtained when age is a confounder ($age \rightarrow sex$). The highest value (0.074) is obtained when age is a mediator (redlining). In total, there are 4 possible causal paths from sex to default, according to PC. There is only one causal path for FCI: $sex \rightarrow credit \rightarrow default$. Therefore, ID is different than zero in PC since the same path is the only possible indirect discrimination, and ED is different than zero in FCI.

C.1.4 Results for the Boston Housing Dataset

The Boston housing dataset holds the statistics on 506 cases of Boston areas including diverse variables used to predict median real estate value in the district. The data has a sensitive predictor variable - the proportion of black people living in the area. The data is collected by the U.S Census Service and can be found in StatLib archive [2] and was originally published

[2] http://lib.stat.cmu.edu/datasets/boston

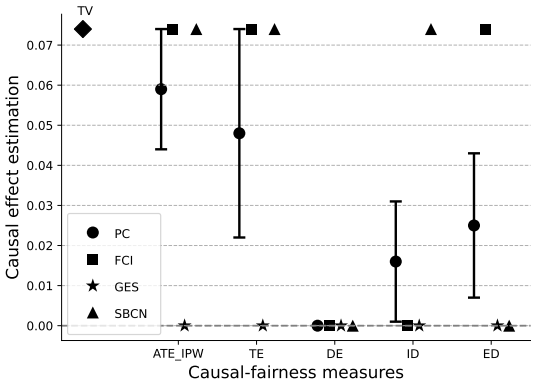


Fig. C.5 Estimation of causal effects of the German credit dataset based on PC, FCI, GES, and SBCN.

by Harrison et al. [102]. The dataset has been used extensively to benchmark machine learning algorithms, however its use for fairness in machine learning is very limited[3]. The dataset originally contains 14 variables, but only 7 are used for empirical experiments. We removed two variables because of missing values and another 5 to avoid multicollinearity and simplify the graphs. All the variables in the data are continuous, mostly following non-Gaussian distribution (as found by the quantiles tests (QQ)). By contrast to the above datasets, LiNGAM is applied along with all the other search algorithms since the data is totally continuous.

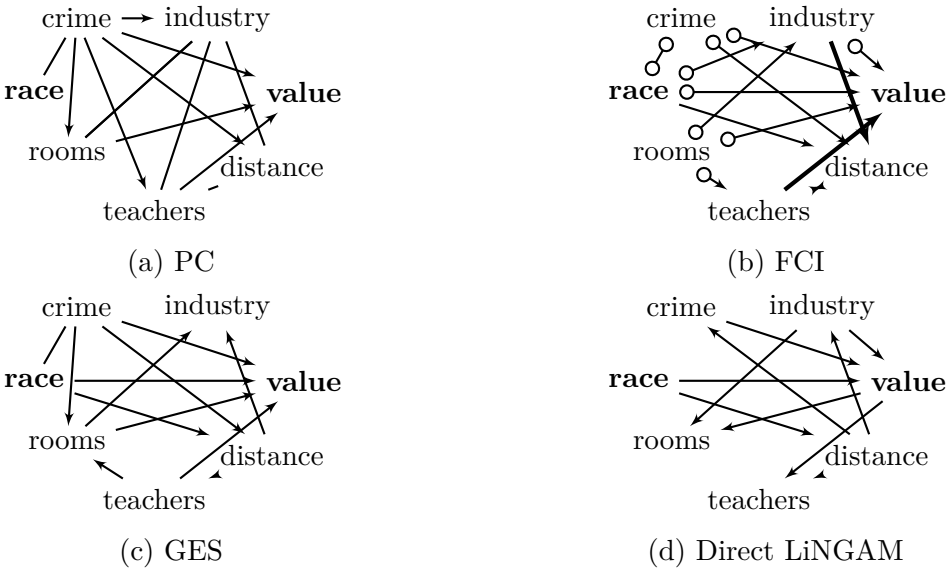


Fig. C.6 Generated causal graph for the Boston Housing dataset.

[3]To the best of our knowledge, it has been only used to illustrate fairness preprocessing tools in SciKitLearn https://scikit-fairness.readthedocs.io/en/latest/fairness_boston_housing.html

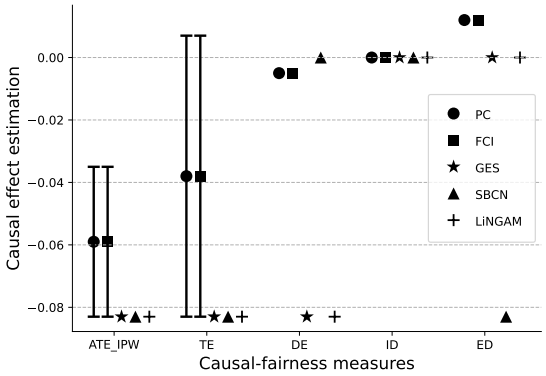


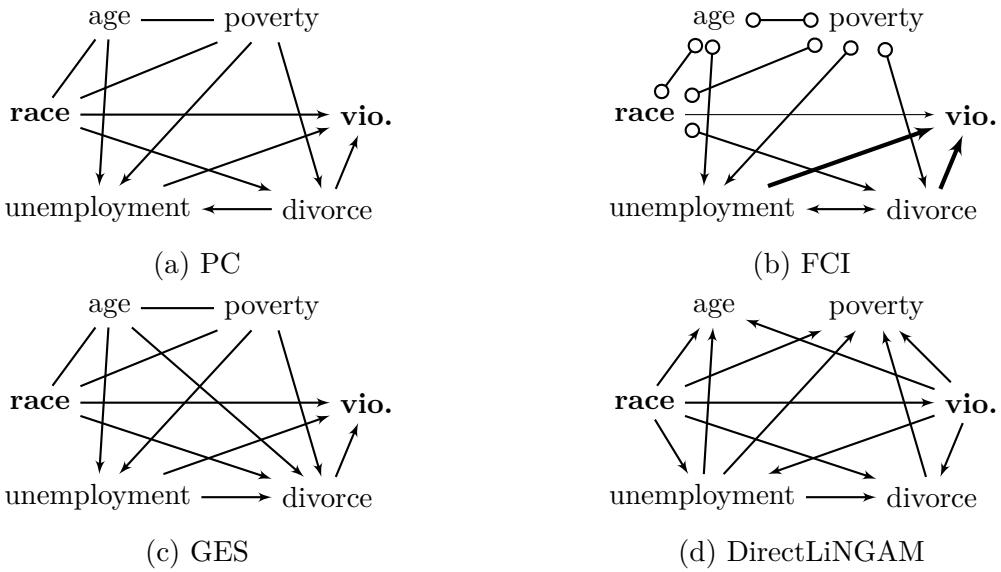
Fig. C.7 Estimation of causal effects of the Boston housing dataset based on PC, FCI, GES, and SBCN.

The generated graphs are shown in Fig. C.6. The direct effect appears in all graphs. The number of causal paths between race and value varies greatly between graphs. These are 13, 6, 2, 3, and 3 according to PC, FCI, GES, and LiNGAM, respectively. For instance, the two possible paths according to GES are $race \to value$ and $race \to distance \to industry \to value$. The most notable feature of the discrimination values in Fig. C.7 is that ID is zero according to all graphs. This is due to the fact that all mediator variables are explaining; crime rate, distance to employment centers, number of rooms in houses, etc. These can be clearly used to justify discrimination legitimately. All measures return either zero or some slightly positive values, which, surprisingly, indicate a slight (≤ 0.08) discrimination *in favor* of areas with higher proportions of black individuals.

C.1.5 Results for the Communities and Crime Dataset

The communities and crime dataset[4] contains data relevant to per capita violent crime rates in several communities in the United States, and the outcome is this crime rate. The variables used for causal graph generation are continuous: race, age, poverty rate, unemployment rate, divorce rate, and violent crime rate. Race is considered a sensitive variable. Three tiers are used: race, age, and poverty rate, which are defined in the first tier. Divorce and unemployment rates are defined in the second tier and violent crime rate in the last tier. No variable can be considered as explaining; hence, we treat them as redlining if found as mediators.

[4]https://archive.ics.uci.edu/ml/datasets/communities+and+crime



(a) PC (b) FCI

(c) GES (d) DirectLiNGAM

Fig. C.8 Generated causal graph for the communities and crime census dataset. Vio. stands for violence.

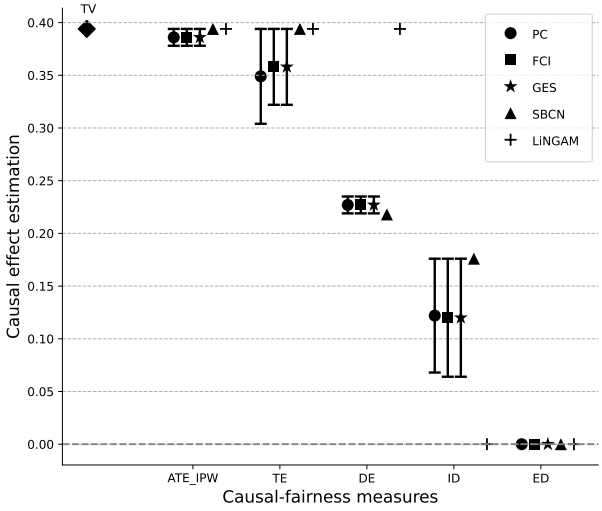


Fig. C.9 Estimation of causal effects of the communities and crime dataset based on PC, FCI, GES, and SBCN.

Fig. C.8 shows the generated graphs. The direct edge between race and violence is identified by all algorithms. PC and FCI generated graphs differ only in the direction of the unemployment-divorce edge. The numbers of causal paths possible in all graphs are similar with the striking exception of LiNGAM. These numbers are 7, 9, 9, and 9 for PC, FCI, and GES, respectively, but only 1 (the direct path *race* → *violence*) for LiNGAM. This can be a strong indicator that the dataset does not satisfy the LiNGAM assumption. In particular, the non-Gaussianity of the noise terms. Fig. C.9 shows the values of the discrimination measures. Both measures of total effect indicate significant discrimination (almost 0.4) against blacks.

TE ranges from 0.303 when age is a confounder and 0.394 when age is a mediator. There are no confounders according to LiNGAM and SBCN graphs, hence, TE coincides with total variation (Eq. (3.23)). DE values are comparable except for LiNGAM. In the latter, since the direct edge is the only causal path between race and violence, DE also coincides with TE, and all indirect effects (ID and ED) are equal to zero. The high variability in ID values is directly linked to the role of the age variable (whether it is a confounder or a mediator). ED is zero according to all graphs since there are no explaining variables.

It is important to mention that despite the flawed graph returned by LiNGAM, the total effect is similar to the values computed based on other graphs because all discrimination is considered direct, and hence, the indirect discrimination is zero. This does not reflect the correct mediation analysis; the other more reliable graphs were returned. In other words, the total effect value according to LiNGAM is correct, but the direct and indirect discrimination values are flawed. More generally, in case of the absence of confounding between the sensitive attribute and the outcome variable, a flawed causal model does not impact the reliability of the total effect as the latter coincides with total variation, which can be computed independently of the causal graph. However, splitting the causal effect between direct, indirect, and explained types of discrimination depends heavily on the mediation structure of the graph.

ECOLE
DOCTORALE

Titre : Promouvoir l'IA éthique et responsable : explorer l'équité, la confidentialité et l'explicabilité à travers des perspectives causales

Mots clés : équité, confidentialité, explicabilité, causalité

Résumé : Cette thèse explore l'intersection de la confidentialité, de l'équité et de la causalité dans le domaine de l'apprentissage automatique et de la prise de décision basée sur les données. Les principales contributions de la thèse peuvent être résumées comme suit : (1) nous étudions l'applicabilité des notions d'équité statistiques et basées sur la causalité dans divers domaines d'application, en évaluant leur alignement avec les préférences des parties prenantes et les normes sociétales dans les systèmes de prise de décision algorithmiques ; (2) nous menons une étude systématique et formelle sur l'impact de la confidentialité différentielle locale sur l'équité. Nous évaluons quantitativement la façon dont les décisions du modèle d'apprentissage automatique changent en fonction de différents niveaux de confidentialité et de distribution des données. De plus, nous examinons empiriquement les implications en matière d'équité de la collecte de plusieurs attributs sensibles dans le cadre du confidentialité différentielle locale ; (3) nous étudions la découverte causale à travers le prisme de l'équité algorithmique, en analysant comment le processus de découverte causale influence la structure des graphiques causals et, par conséquent, les conclusions sur l'équité. En outre, nous proposons un nouveau mécanisme de génération de données pour produire des ensembles de données synthétiques biaisés basés sur des graphiques causals et des niveaux de biais spécifiés, explorant l'influence de différents algorithmes de découverte causale sur diverses structures causales et le degré de biais introduit. Dans l'ensemble, cette thèse contribue au corpus croissant de littérature sur l'intelligence artificielle éthique et responsable en offrant des perspectives théoriques complétées par des considérations pratiques pour les décideurs politiques, les praticiens et les chercheurs cherchant à développer des systèmes algorithmiques plus justes qui adhèrent aux principes de confidentialité et d'explicabilité.

Title : Advancing Ethical and Responsible AI: Exploring Fairness, Privacy, and Explainability through Causal Perspectives

Keywords : fairness, privacy, explainability, causality

Abstract : This dissertation explores the intersection of privacy, fairness, and causality within the realm of machine learning (ML) and data-driven decision-making. The dissertation's main contributions can be summarized as follows: (1) we investigate the applicability of statistical and causality-based fairness notions in diverse application domains, evaluating their alignment with stakeholder preferences and societal norms in algorithmic decision-making systems; (2) we conduct a systematic and formal study on the impact of local differential privacy (LDP) on fairness. We quantitatively assess how ML model decisions change under varying levels of privacy and data distributions. Additionally, we empirically examine the fairness implications of collecting multiple sensitive attributes under LDP; (3) we study causal discovery through the lens of algorithmic fairness, analyzing how the causal discovery process influences the structure of causal graphs and, consequently, fairness conclusions. Furthermore, we propose a novel data generation mechanism to produce biased synthetic datasets based on causal graphs and specified bias levels, exploring the influence of different causal discovery algorithms on various causal structures and the degree of introduced bias. Overall, this thesis contributes to the growing body of literature on ethical and responsible artificial intelligence by offering theoretical insights complemented by practical considerations for policymakers, practitioners, and researchers seeking to develop fairer algorithmic systems that adhere to privacy and explainability principles.

Institut Polytechnique de Paris
91120 Palaiseau, France