Department of Electrical and Information Engineering

Electrical and Information Engineering Ph.D. Program

SSD: ING-INF/05 - Information Processing Systems

Final Dissertation

Fostering Responsible Artificial Intelligence: An Evaluation Approach Grounded in Counterfactual Reasoning

by

Giandomenico Cornacchia

Supervisors

Prof. Tommaso Di Noia

Prof. Fedelucio Narducci

Prof. Azzurra Ragone

Coordinator of the Ph.D. Program

Prof. Mario Carpentieri

Course XXXVI, 01/11/2020 - 31/10/2023

Abstract

The increasing application of Artificial Intelligence and Machine Learning models poses potential risks of unethical behaviour and, in light of recent regulations, has attracted the attention of the research community. Current AI regulations require discarding sensitive features (e.g., gender, race, religion) in the algorithm's decision-making process to prevent unfair outcomes. However, even without sensitive features in the training set, algorithms can persist in discrimination. Indeed, when sensitive features are omitted (*fairness under unawareness*), they could be inferred through non-linear relations with the so-called proxy features. Several researchers focused on seeking new fairness definitions or developing approaches to identify biased predictions without helping to answer the following question: Which fairness definition should be used and satisfied in a deployed model? Consequently, what metric should we satisfy? However, what metrics can better quantify the unfair behavior of a model? These questions remain open challenges in the field. Furthermore, a limitation of the proposed approaches is that they focus solely on a discrete and limited space; only a few analyze the minimum variations required in the user characteristics to ensure a positive outcome for the individuals (counterfactuals).

This dissertation aims to bridge the gap in the fairness domain by proposing a new fairness perspective. Starting from the recent academic literature in the area, this thesis will intertwine with the issues close to the field of responsible AI by offering insights in the following directions: (i) we propose a framework grounded in counterfactual reasoning to reveal the potential hidden bias of a machine learning model that can persist even when sensitive features are discarded, (ii) we propose a simple procedure to identify and quantify the relationship between sensitive characteristics and proxy features. (iii) we leverage counterfactual reasoning to explain the model decision building a responsible pipeline for the credit score domain.

Publications

Some ideas and figures have appeared previously in other publications. A complete list of my publications is available in the following (the symbol * denotes papers where I contributed as main author).

[1] *Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio. "Auditing fairness under unawareness through counterfactual reasoning." In: *Information Processing & Management* 60.2 (2023), p. 103224. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2022.103224.

[2] *Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "A General Architecture for a Trustworthy Creditworthiness-Assessment Platform in the Financial Domain." In: *Annals of Emerging Technologies in Computing (AETiC)* 7.2 (2023).

[3] *Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Reasoning for Decision Model Fairness Assessment." In: *Companion Proceedings of the ACM Web Conference 2023*. WWW '23 Companion. Austin, TX, USA: Association for Computing Machinery, 2023, pp. 229–233. ISBN: 9781450394192. DOI: 10.1145/3543873.3587354.

[4] *Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Reasoning for Bias Evaluation and Detection in a Fairness under Unawareness setting." In: *ECAI*. Vol. 372. Frontiers in Artificial Intelligence and Applications. IOS Press, 2023, pp. 477–484. DOI: 10.3233/FAIA230306.

[5] *Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Reasoning for Responsible AI Assessment." In: *Ital-IA*. Vol. 3486. CEUR Workshop Proceedings. CEUR-WS.org, 2023, pp. 347–352.

[6] *Giandomenico Cornacchia, Francesco M. Donini, Fedelucio Narducci, Claudio Pomo, and Azzurra Ragone. "Explanation in Multi-Stakeholder Recommendation for Enterprise Decision Support Systems." In: *Advanced Information Systems Engineering Workshops - CAiSE 2021 International Workshops, Melbourne, VIC, Australia, June 28 - July 2, 2021, Proceedings*. Ed. by Artem Polyvyanyy and Stefanie Rinderle-Ma. Vol. 423. Lecture Notes in Business Information Processing. Springer, 2021, pp. 39–47. DOI: 10.1007/978-3-030-79022-6_4.

[7] *Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. "A General Model for Fair and Explainable Recommendation in the Loan Domain (Short paper)." In: *KaRS/ComplexRec@RecSys*. Vol. 2960. CEUR Workshop Proceedings. CEUR-WS.org, 2021.

[8] *Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. "Improving the User Experience and the Trustworthiness of Financial Services." In: *Human-Computer Interaction – INTERACT 2021*. Ed. by Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen. Cham: Springer International Publishing, 2021, pp. 264–269. ISBN: 978-3-030-85607-6.

[9] *Giandomenico Cornacchia, Giulio Zizzo, Kieran Fraser, Muhammad Zaid Hamed, Ambrish Rawat, and Mark Purcell. "MoJE: Mixture of Jailbreak Experts, Naive Tabular Classifiers as Guard for Prompt Attacks." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, 2024. arXiv: 2409.17699.

[10] *Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, and Tommaso Di Noia. "On Popularity Bias of Multimodal-Aware Recommender Systems: A Modalities-Driven Analysis." In: MMIR '23 (2023), pp. 59–68. DOI: 10.1145/3606040.3617441.

[11] Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, Felice Antonio Merra, Tommaso Di Noia, and Eugenio Di Sciascio. "Formalizing Multimedia Recommendation through Multimodal Deep Learning." In: *ACM Transaction on Recommender System* (Apr. 2024). DOI: 10.1145/3662738.

[12] *Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, and Tommaso Di Noia. "Disentangling the Performance Puzzle of Multimodal-aware Recommender Systems." In: *EvalRS@KDD*. Vol. 3450. CEUR Workshop Proceedings. CEUR-WS.org, 2023.

[13] Francesco Paolo Schena, Vto Walter Anelli, Giandomenico Cornacchia, Tommaso DI Noia, Maria Stangou, Aikaterina Papagianni, and Rosanna Coppo. "FC048: New Tool to Predict the Clinical Course and Renal Failure in Patients with Immunoglobulin a Nephropathy." In: *Nephrology Dialysis Transplantation* 37.Supplement_3 (May 2022), gfac105.004. ISSN: 0931-0509. DOI: 10.1093/ndt/gfac105.004. eprint: https://academic.oup.com/ndt/article-pdf/37/Supplement_3/gfac105.004/43535905/gfac105_004.pdf.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The Cambridge Dictionary defines *discrimination* as the act of "*treating a person or particular group of people differently, especially in a worse way from how you treat other people, because of their skin colour, sex, sexuality, etc.*"[1]. Recently, various regulations have been designed to face the discrimination problem. For instance, Article 21 of the EU Charter of Fundamental Rights defines the non-discrimination requirements: "*any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited*"[2] [40]. In 2015, the United Nations General Assembly set up the Sustainable Development Goals (SDGs) or Global Goals, a collection of 17 interlinked global goals designed to be a "*blueprint for achieving a better and more sustainable future for all*"[3]. Most of the SDGs are somehow related to the discrimination problem, such as *no poverty*, *zero hunger*, *gender equality*, and *reduced inequality*. The discrimination problem is well-known and recognized in the financial domain where, for example, the decision to approve or deny credit has been regulated with precise and detailed regulatory compliance requirements (i.e., Equal Credit Opportunity Act[4] [49], Federal Fair Lending Act[5] [86], and Consumer Credit Directive for EU Community[6] [43]). However, these laws were set to prevent discrimination in human decision-making processes and not in automated ones, such as those exploiting Machine Learning (ML)

[1] https://dictionary.cambridge.org/dictionary/english/discrimination

[2] https://fra.europa.eu/en/eu-charter/article/21-non-discrimination

[3] United Nations (2017) Resolution adopted by the General Assembly on 6 July 2017

[4] https://www.ftc.gov/enforcement/statutes/equal-credit-opportunity-act

[5] https://www.fdic.gov/resources/supervision-and-examinations/consumer-compliance-examination-manual/documents/4/iv-1-1.pdf

[6] https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32008L0048

or, more generally, Artificial Intelligence (AI) systems. The EU Commission, in the wake of the GDPR[7] (i.e., a regulation to safeguard personal data), seeks to regulate the use of AI systems with the "Ethics Guidelines for Trustworthy AI" and more recently with "The Proposal for Harmonized Rule on AI". The regulated characteristics are various (e.g., technical robustness, privacy, data governance, transparency, accountability, societal and environmental well-being), and the European legislature deems adopting non-discriminatory AI models crucial. Therefore, the financial domain is the perfect workbench to test these regulations. Indeed, financial services are considered high-risk AI applications on the European AI risk scale (the levels are minimal, limited, high, and unacceptable risk). As a consequence, a financial AI model must demonstrate fairness concerning sensitive characteristics to protect the social context in which it operates.

Since unfair treatment is strictly related to discriminatory behaviour, fairness can be seen as the antonym of discrimination. Unfortunately, finding a strict and formal definition of fairness is challenging, and the subject is still under debate. Mehrabi et al. [123] proposed a definition that could fit the financial domain and its discrimination-derived risks. They defined *fairness* as "*the absence of any prejudice or favouritism toward an individual or a group based on their inherent or acquired characteristics*." Another relevant aspect of fairness is highlighted by Ekstrand et al. [80] that refers to *unfairness* when a system treats people, or groups of people, in a way that is considered "unfair" by some moral, legal, or ethical standard. The exciting aspect is that, in that case, "fairness" is related to the normative aspects of the system and its effects. For this work, the *counterfactual fairness* as defined by Pitoura et al. [142] is particularly relevant. The intuition, in this case, is that an output is fair towards an entity if it is the same in both the actual world and a counterfactual world where the entity belongs to a different group. Causal inference is used to formalize this notion of fairness. This definition inspired the design of our model. From a geometrical perspective that considers how a decision model works, Dwork et al. [76] says that items that are close in construct space shall also be close in decision space, which is widely known as individual fairness: similar individuals should receive similar outcomes. In contrast to individual fairness, Deldjoo et al. [68] define group fairness as aims to ensure that "similar groups have similar experiences". Typical groups in such a context are a majority or dominant group and a protected group (e.g., an ethnic minority). Following this overview, some critical aspects of this work emerged: the legislation,

[7] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

the counterfactual, and the group. More specifically, the legislation is the primary motivation behind this work, the counterfactual generation is the strategy we exploited for detecting unfairness, and the group is the subject of discrimination we want to catch. Furthermore, the counterfactual generated can also be used for explainability purposes and personalized, actionable recommendations in the loan domain. Although system designers train a model without any discriminatory purpose, several studies demonstrated that using AI systems without considering ethical aspects can promote discrimination [14, 51, 74]. Moreover, while the financial domain regulations strictly prohibit using sensitive characteristics for decision-making, some researchers defend their usage and believe the model should avoid unfair treatments (i.e., active bias detection) [81, 150]. Nevertheless, only avoiding using sensitive features for training AI models does not guarantee the absence of biases in the outcome [2]. Indeed, there could be features in the dataset that can represent an implicit sensitive feature. In this study, we name these independent features as *proxy features* for the sensitive one. For instance, education, smoking and drinking habits, pet ownership, and diet can be proxy variables for the feature "*age*". The relationship between proxy and sensitive features generally depends on multicollinearity, namely a strong linear relationship between more than two variables. Unfortunately, non-linear relationships are more challenging to detect.

1.1 Thesis Statement

The investigation of this thesis finds its backbone and relies on the *"Fairness Under Unawareness"* –or *"blindness"* Pitoura et al. [142]– definition (i.e., *"an algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process"* [32]). The choice of this definition as a fairness setting is a logical operational consequence of current regulations. Indeed, like for other high-risk applications, the law dictates that AI applications in the financial domain cannot use sensitive information.

The second most important concept in this dissertation is *Counterfactual Reasoning*. Counterfactual reasoning stands as a pivotal tool in the realm of machine learning and artificial intelligence, especially in the context of fairness understanding and model explanation. Its significance lies in its capacity to unveil the intricate mechanisms underpinning algorithmic decisions. By generating counterfactual instances that depict alternative scenarios, it offers a profound glimpse into how a model would have acted differently under varying circumstances. This not only enhances transparency and interpretability but also empowers stakeholders, model developers, and end-users to discern the factors contributing to a model's predictions. Furthermore, counterfactual reasoning plays a fundamental role in addressing and rectifying biases, as it enables the identification of discriminatory features or tendencies in machine learning models, ultimately paving the way for more equitable and accountable AI systems. As we navigate the complexities of deploying AI in critical domains like finance, healthcare, and criminal justice, counterfactual reasoning emerges as a critical tool in fostering trust, mitigating bias, and ensuring that AI models operate with fairness and transparency as guiding principles.

More specifically, the dissertation is divided into two main parts. The first part (i.e., Chapter 4 and Chapter 5) investigates a strategy to detect decision biases in a realistic scenario where sensitive features are absent, and there could be an unknown number of proxy features. The second part (i.e., Chapter 6), tries to quantify the same, proposing an explanation methodology for the loan domain. The link between these two macro-sections is the *Counterfactual Reasoning* methodology.

We propose to tackle this challenging task by designing a system composed of three main modules. The first module encapsulates the classifier to analyze, named the **outcome classifier**. This predictor, as regulations demand, is trained without any sensitive features. The second module trains a separate classifier, named **sensitive feature classifier**, on the same features to predict the sensitive characteristics. The third module calculates the minimal counterfactual samples, i.e., variants of the original sample, by modifying the values of non-sensitive features to obtain a different outcome with the outcome classifier. Finally, the sensitive feature predictor classifies the generated samples to check whether the samples still belong to the original sensitive class. If this does not occur, the outcome predictor is biased, and its unfairness can be quantified.

To better explain the idea behind our approach, let us introduce a simple example regarding the loan granting process. Suppose our goal is to assess whether our loan classifier discriminates against women. In this example, the protected class is women, and the sensitive feature is gender. The outcome classifier is a state-of-the-art classification model trained without gender information. The sensitive feature classifier will then distinguish men from women by exploiting the other non-sensitive features in the dataset (e.g., car type, job, education). A customer loan request triggers the system's operation: the classifier rejects her request. Therefore, the counterfactual module perturbs the values of her non-sensitive features until the loan is approved (e.g., increasing income, reducing the loan duration). The sensitive feature classifier then classifies the new approved counterfactual sample. Is she still classified as a woman by the system? What could we say if the features changed and those responsible for the approval were the same responsible for classifying her as a man? The decision model may still be biased and thus unfair, and since it does not use sensitive features, this is due to proxy features. The same methodology, except for the sensitive feature classifier, is part of the second goal of this research, which is explaining a loan decision. Counterfactual sample generation can have multiple purposes. Indeed, they can be used to explain a decision, to propose and recommend a counterfactual actionable change in a loan recommendation platform, and can be user-specific by interacting with the platform by recommending personalized, actionable steps based on the user-specific requirements and possibilities. Notwithstanding, the type of explanation can depend on the stakeholder which is delivered. All these contributions will be exploited in the remaining part of the thesis.

1.2 Research Contributions

In this dissertation, the concepts that will be exploited are fairness under unawareness, counterfactual reasoning, proxy features, and explainability.

To the best of our knowledge, the interconnection between these concepts is still unexplored, which brings novelty to this dissertation. Overall, this study proposes an approach for detecting bias in machine learning models using counterfactual reasoning, even when those models are trained without sensitive features, i.e., in the case of *Fairness Under Unawareness*. In addition, we investigate the presence of bias in an algorithm using counterfactual reasoning as an effective strategy for bias detection and evaluate if different counterfactual strategies have dissimilar efficacy in detecting biases. Furthermore, as a second contribution, we aim to build an explainable framework taking into account counterfactual reasoning. The framework not only can handle discriminative analysis with the investigation and detection of proxy features-outcomes relationships but also the explanation and actionable steps in a loan recommendation domain. In detail, the dissertation intends to answer the following research questions:

- **RQ1:** Is there a principled way to identify if proxy features exist in a dataset?
- **RQ2:** Does Fairness Under Unawareness setting ensure that decision biases are avoided?
- **RQ3:** Is counterfactual reasoning suitable for discovering decision biases?
- **RQ4:** Is it possible to define a strategy for identifying the proxy features?

- **RQ5:** Can counterfactual reasoning be useful to explain a loan decision?
- **RQ6:** What is the most suitable explanation strategy depending on each stakeholder in the loan domain?

1.2.1 Publications

Following, the list of works taken into account for the thesis dissertation are listed in chronological order. Most of them have been published at journals and conferences on algorithmic fairness, information retrieval, recommender system, and artificial intelligence [52, 54–58, 60, 62, 136]. A comprehensive description of the work published but not taken into account for the thesis dissertation is reminded to Section 1.4.

Journal Papers:

- Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio. "Auditing fairness under unawareness through counterfactual reasoning." In: *Information Processing & Management* (2023)
- Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "A General Architecture for a Trustworthy Creditworthiness-Assessment Platform in the Financial Domain." In: *Annals of Emerging Technologies in Computing (AETiC)* 7.2 (2023).

Conference Papers:

- Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Reasoning for Bias Evaluation and Detection in a Fairness under Unawareness setting." In: *ECAI*. Vol. 372. Frontiers in Artificial Intelligence and Applications. IOS Press, 2023, pp. 477–484.
- Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Reasoning for Decision Model Fairness Assessment." In: *Companion Proceedings of the ACM Web Conference 2023*. WWW'23 Companion. Austin, TX, USA: Association for Computing Machinery, 2023, pp. 229–233.
- Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. "Improving the User Experience and the Trustworthiness of Financial Services." In: *Human-Computer Interaction – INTERACT 2021*. Ed. by Carmelo Ardito, Rosa

Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen. Cham: Springer International Publishing, 2021, pp. 264–269.

Workshop Papers:

- Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Reasoning for Responsible AI Assessment." In: *Ital-IA*. Vol. 3486. CEUR Workshop Proceedings. CEUR-WS.org, 2023, pp. 347–352.
- Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. "A General Model for Fair and Explainable Recommendation in the Loan Domain (Short paper)." In: *KaRS/ComplexRec@RecSys*. Vol. 2960. CEUR Workshop Proceedings. CEUR-WS.org, 2021.

The following section provides additional details on the thesis's organization. This is intended to guide the reader on the structure of the thesis and on the main research path that motivated our research goals.

1.3 Organization of the Thesis

The chapters of this thesis are self-contained and present the notions of specific problems, architectures, paradigms, data structures, and metrics related to their content.

In the next two chapters, we propose a brief but comprehensive introduction to the dissertation. To understand the urgency in relying on the concept handle in the dissertation, in Chapter 2 we start by examining the regulation that in the financial domain determines fair practice in human decision-making and the consequent novel law regulating the unethical use of AI. Subsequently, in Chapter 3 we introduce the unfairness problem in the machine learning domain with its multiple definitions, evaluations, and proposed solutions to the problem with their complementary limitation. Following, in Chapter 4 we provide the preliminaries of the work while proposing our methodology to close the gap with the literature. Thus, to provide an answer to the first 4 RQs, in Chapter 5 we performed an extensive experimental evaluation on three state-of-the-art datasets, broadly recognised as datasets containing Social Bias. Then, in Chapter 6 we propose different pipelines that try to give an explanation which is user-friendly and recommends actionable and interactive explanations. Furthermore, we try to differentiate possible useful explanations based on the considered stakeholders. In the conclusive Chapter 7, we have the closing remarks.

1.4 Bibliographical Notes

This section briefly describes the research articles published during the PhD but not discussed in the dissertation. We decided not to include the following article so as not to lose the focus and pertinence with the previously introduced concepts. Indeed, the following works have been conducted as simultaneous topics whose research questions have been raised while studying the literature.

Recommender systems, which provide personalized suggestions based on user preferences, are increasingly influential across digital platforms. Ensuring their trustworthiness—through transparency, fairness, and the mitigation of bias—is crucial for building user trust and satisfaction. One of the crucial issues for a trustworthy Recommender System (RS) is explanations. Research contributions in this direction have become attractive again due to the renewed interest in eXplainable Artificial Intelligence (XAI). In order to improve the effectiveness, efficiency, persuasiveness, and user satisfaction of recommender systems, explainable recommendation refers to the personalized recommendation algorithms that address the problem of why – they not only give the user the suggestions but also make the user aware of why such items are recommended by generating recommendation explanations. Motivated by these results, we proposed a formal approach for generating explanations from Multi-Stakeholder type RSs (MS-RSs). In this context, we considered the point of view of counterfactual explanations. We highlighted the pros and cons of their application in a scenario that explained the recommendation for the consumer and the policy adopted by the system for the considered stakeholder. The proposed model for generating counterfactual explanations in a Multi-Stakeholder context was presented at the Advanced Information Systems Engineering Workshops held in conjunction with the 33rd International Conference on Advanced Information Systems Engineering (CAiSE) 2021 under the title *"Explanation in multi-stakeholder recommendation for enterprise decision support systems"* [59].

While recommendation systems leveraging multimedia content have long been established as successful and efficient approaches in the literature, their application of *multimodal* deep learning strategies remains not clearly defined, formalized, and empirically analyzed. The survey literature, under the title *Formalizing Multimedia Recommendation through Multimodal Deep Learning* [118], has been accepted as a journal at the ACM Transactions on Recommender Systems.

This literature is grounded in an innovative analysis of the performance of state-of-the-art graph-based recommender systems leveraging multimodal information under several evaluation perspectives encompassing, among others, novelty, diversity, bias, and fairness recommendation measures. The analysis is a condensed part of two works: (i) *Disentangling the Performance Puzzle of Multimodal-aware Recommender Systems* presented at the 3rd EvalRS workshop at the Knowledge and Data Mining (KDD) Conference 2023 [119], and (ii) *On Popularity Bias of Multimodal-aware Recommender Systems: A Modalities- driven Analysis* presented at the 1st International Workshop on Deep Multimodal Learning for Information Retrieval at the 31st ACM International Conference on Multimedia (ACM MM) 2023 [117].

Beyond RS, the security of Large Language Models (LLMs) against jailbreak attacks has become increasingly relevant. These attacks compromise LLM safety mechanisms, risking data integrity and privacy. To address this, the MoJE (Mixture of Jailbreak Experts) architecture was proposed, introducing a novel guardrail framework that detects 90% of jailbreak attacks without affecting benign prompts. This research demonstrates a balance between security and computational efficiency, enhancing LLM protection against adversarial threats. The work will be presented at AIES 2024 [63].

Chapter 2

Regulation Compliance in the Credit Domain

Dedicating a chapter to the historical discrimination laws and current legal regulations on AI applications is crucial. It provides essential context by tracing the evolution of societal values and legal frameworks in the fight against discrimination. Additionally, it offers a clear understanding of how AI's integration into society raises complex ethical and legal issues. Examining current AI laws highlights the legal system's response to emerging challenges related to bias, transparency, accountability, and privacy. This chapter equips the reader with a comprehensive grasp of how AI intersects with fairness, equity, and legality, fostering an informed approach to AI development and deployment.

2.1 Introduction

Financial institutions and banks have always been under scrutiny for ethics, safeguarding citizens, and ensuring non-discrimination of sensitive groups. Over the years, several laws have been passed with the primary objective of providing fair access to credit. Despite this, Federal Reserve data showed that discrimination has not entirely disappeared over the years in areas such as home loans and small business credit, which continue to pose challenges for public policymakers [12].

Discrimination and equal credit opportunities have been some of the main objectives of the financial market laws. The discrimination in-laws and economics can be categorised as *statistical* [3, 140], in which an agent discriminates indirectly without any causal relation if any sensitive category is not considered, and *taste-based* [102], in which an agent can discriminate based on personal prejudice (sexism, racism, etc.). Furthermore, both categories focused on utility. Going deeper, with statistical discrimination, the sensitive attribute is used to make inferences. Instead, with discrimination based on preference, decision-makers act as if they have a preference or "taste for prejudice", sacrificing profit to avoid certain transactions [50]. In particular, the equal protection law, established by the U.S. Constitution's Fourteenth Amendment, prohibits government agents from acting with "discriminatory purpose"[1]. Indeed, it prevents preference-based actions (since it means sacrificing utility) but allows the limited use of protected attributes since it represents a form of statistical discrimination.

2.1.1 Credit regulation

One of the first regulations in the financial services industry was the Truth in Lending Act (TILA, 1968), a U.S. federal law [48]. It was designed to promote informed consumer credit use by standardising the disclosure of loan terms and costs [71].

Attribute	**FHA** [47]	**ECOA** [49]	**CCD** [43]
Race	✓	✓	✓
Color	✓	✓	✓
National origin	✓	✓	✓
Religion	✓	✓	✓
Sex	✓	✓	✓
Familial status	✓		✓
Disability	✓		✓
Exercised rights under CCPA		✓	
Marital status		✓	✓
Recipient of public assistance		✓	✓
Age		✓	✓
Language			✓
Opinion (political or any other)			✓
Genetic feature			✓
Sexual orientation			✓

Table 2.1 A list of the protected attributes as specified in the Fair Housing Act, Equal Credit Opportunity Act, and Consumer Credit Directive through the EU Charter of Fundamental Rights (FHA, ECOA, CCD), partially from [32].

Furthermore, the TILA law protects borrowers against inaccurate and unfair credit billing and credit card practices by giving them the right to withdraw. It requires lenders to provide information on the loan cost so that borrowers can decide whether to make or not the loans and, in a negative case, compare with other institutions.

[1] https://supreme.justia.com/cases/federal/us/426/229/

The Fair Housing Act (FHA, 1968) [47], also known as Titles VIII through IX of the Civil Rights Acts of 1968 (follow-up of the precedent Civil Rights Act of 1964), is a federal law enacted in 1968 that prohibits discrimination in the purchase, sale, rental, or financing of housing - private or public - based on race, colour, sex, nationality or religion (see Table 2.1). Thus, the Fair Housing Act succeeded in the main purpose of the previous law, the Civil Rights Act, to prevent various types of discrimination.

The U.S. Department of Housing and Urban Development (HUD) is the primary enforcer of the Fair Housing Act. Some examples of discriminatory practices include different prices for selling or renting a home, delaying or failing to maintain or repair homes for certain tenants, or limiting the privileges, services, or facilities of a home based on a person's gender, nationality, or racial characteristics.

These laws provided the impetus for a regulatory movement that led to new laws forming the basis for fair lending. Indeed, between 1968 and 1974, different laws were legislated. First, the Consumer Credit Protection Act (CCPA, 1968) protects consumers from harm caused by creditors, banks, and credit card companies and requires that the total cost of a loan or credit product be disclosed, including how the interest is calculated and any fees. In addition, it regulates the fair reporting of a client's financial information, as well as prohibits misleading advertising and discrimination by creditors, and makes the terms of loans more transparent and easier to understand for borrowers who may not be financial or banking experts (see Code of Federal Regulations. "12 CFR 1005.4."[2]). Second, in relation to the previous law, the Fair Credit Reporting Act (FCRA, 1970), in addition to the CCPA, regulates the collection of consumers' credit information and access to their credit reports addressing the fairness, accuracy, and privacy of the personal information contained in the files of the credit reporting agencies. The last one is the Equal Credit Opportunity Act (ECOA, 1974), a considerable improvement on the FHA of 1968. The ECOA, as shown in the table, prohibits discrimination in credit transactions based on race or colour, national origin, religion, sex, marital list status, age, whether an applicant receives income from a public assistance program, and the exercise by an applicant, in good faith, of any right under the Consumer Credit Protection Act [67]. Furthermore, Regulation B of the ECOA requires that an adverse action notice (AAN) be served within 30 days if credit is denied, and Appendix C of Reg. B requires that a list of no more than 4 reasons, known as "principle reason explanations", must be provided. For

[2] https://www.ecfr.gov/current/title-12/chapter-X/part-1005/subpart-A/section-1005.4

the U.S., the regulations exploited so far are still operative and continuously updated over time.

While the United States could continue along the line presented in the 1970s as they were among the first to introduce anti-discriminatory laws, in Europe, due to the different nature of its member countries and because of its relatively recent foundation, non-discriminatory laws have only been implemented in the last twenty years.

One of the most important milestones in the EU's commitment against racial and gender discrimination has been the Racial Equality Directive adopted in 2000 [41] and the Gender Equality Directive adopted in 2006 [42]. Considering the credit domain, in 2008, the EU Commission introduced the Consumer Credit Directive, which established lending responsible practice and consumer protection [43]. It harmonized consumer credit laws across member states, ensuring transparent information and non-discriminatory access to credit, thereby promoting fair and equal treatment for all consumers. The non-discrimination directive is based on the EU Charter of Fundamental Rights [40] exploited in Chapter 1 and in Table 2.1.

All these discrimination constraints and explanation requirements, to which financial services firms need to pay attention, can be considered the base and building block for the recent AI regulations.

2.2 AI regulations according to the European Commission

In more recent years, thanks to the revolution brought about by Big Data and the increasing use of artificial intelligence, there has been a speeding up of credit application processes. These systems have proven to be highly accurate in their predictions and emphasize the historical problems they have been trying to eliminate. In fact, in applying risk assessment models, discriminatory intent was never a primary concern because decision-making was delegated to predictive model algorithms. However, powerful predictive models can be hazardous, leading to very often unintended discriminatory decisions. The European Commission, initially with the GDPR, and other countries such as the US, Canada, and the UK have tried to regulate the possible artificial intelligence applications such that they can be considered trusted, safe, fair, and human-centred.

With the "*General Data Protection Regulation*" (EU) n. 2016/679 (GDPR) [44], the European Union has addressed EU law on the protection and privacy of personal data concerning citizens of countries in the European Economic Area (EEA). The GDPR was adopted on 27 April 2016 and officially entered into force on 25 May 2018. The GDPR is not only aimed at protecting the privacy of personal data but also addresses the issue of exporting personal data outside the EU and obliges all data controllers (including those with registered offices outside the EU) who process the data of EU residents to observe and fulfil their obligations by making the data flow compliant. Since it entered into force, the GDPR has replaced the contents of the Data Protection Directive (Directive 95/46/EC) [39].

The GDPR has triggered initiatives to regulate AI by European countries, which may have autonomous laws on the subject, but which converge on a unified European strategy [3]. The main objectives of the European strategy are to make Europe a world leader in this field, to initiate and promote a radical socio-economic change in the countries of the European Economic Area and to ensure the ethical, legal and safe use of the same technology. In drafting the European strategy on artificial intelligence, the European Commission was assisted by a group of high-level experts on AI. Indeed, on 8 August 2019, they presented the "*Ethics Guidelines for Trustworthy Artificial Intelligence* [4]". The Ethics Guidelines introduced the concept of Trustworthy AI, which is the point of convergence of 7 key requirements that AI providers must absolutely comply with. These are as follows: human Agency and oversight, technical robustness and safety, privacy and data governance, transparency, non-discrimination and fairness, societal and environmental well-being, and accountability. These guidelines have been presented to more than 350 different stakeholders through a piloting process to be improved and enhanced. Indeed, on 17 July 2020, the High-Level Expert Group on Artificial Intelligence (AI HLEG) presented its final Assessment List for Trustworthy Artificial Intelligence. This is a revised version of the Ethics Guidelines, in which the AI principles are translated into an accessible and dynamic checklist that guides developers and deployers of AI in implementing such principles in practice[4].

On 2 February 2020, the European Commission released its "*White Paper on Artificial Intelligence - A European approach to excellence and trust*", consisting of two main building blocks [45]. The first analyses and emphasizes AI's benefits and having a European strategy for achieving excellence in the international arena. The second one analyzes the weaknesses and dangers of unregulated, unaccountable, and unsafe use of artificial intelligence to identify its potential risk. Indeed, a risk-based fine-tuning criterion emerges for the first time in the White Paper.

[3] https://digital-strategy.ec.europa.eu/en/policies/strategy-artificial-intelligence

[4] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

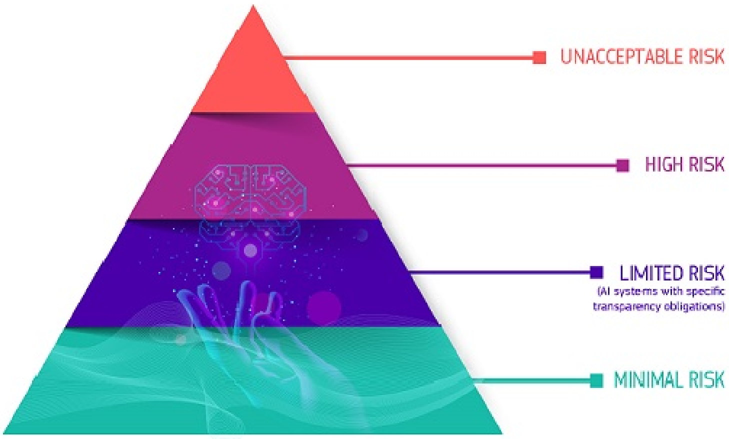


Fig. 2.1 Regulatory framework proposal on Artificial Intelligence, a risk-based approach

In the same direction of the white paper, the European Commission, on 21 April 2021, presented the proposed law "*Proposal for a Regulation laying down harmonized rules on artificial intelligence*" [46], also known as "*EU AI Act*" that has been approved on the 14 June 2023. This proposal remarks on the importance of monitoring the deployed AI system based on a scale of risk (see Figure 2.1). The risk-based approach divides AI system into four different categories depending on the risk of the use case:

- **Unacceptable risk:** all systems that are considered a clear threat to security, livelihoods, and people's rights (e.g., social scores or toys that encourage dangerous behaviour) will be banned;
- **High-risk:** all systems that are used in critical infrastructures (e.g., transport), educational or vocational training (e.g., scoring of exams), safety components of products (e.g., AI application in robot-assisted surgery), employment, workers management, and access to self-employment (e.g., CV-sorting software for recruitment procedures), essential private and public services (e.g., credit scoring denying citizens opportunity to obtain a loan), Law enforcement (e.g., evaluation of the reliability of evidence), migration, asylum, and border control management (e.g., verification of the authenticity of travel documents), and administration of justice and democratic processes (e.g., applying law to a concrete set of facts);
- **Limited risk:** all system that needs some transparency obligations (i.e., chatbots, in which the user must be aware of interacting with a machine so they can take an informed decision to continue or step back);
- **Minimal risk:** all systems that represent a minimal risk or no risk (e.g., AI-enabled in video games or spam filter) allowing the free use of AI;

Therefore, as can be seen from the list, the use of artificial intelligence in credit institutions is considered to be high-risk. Following, it is reported on the two most important recitals concerning systems used to assess creditworthiness (i.e., Recital 37) and the importance of regulators in the financial sector (i.e., Recital 80).

EU AI Act - Recital 37

Creditworthiness of Natural Persons

"Another area in which the use of AI systems deserves special consideration is the access to and enjoyment of certain essential private and public services and benefits necessary for people to participate in society fully or to improve one's standard of living. In particular, AI systems used to evaluate the credit score or creditworthiness of natural persons should be classified as high-risk AI systems, since they determine those persons' access to financial resources or essential services such as housing, electricity, and telecommunication services. AI systems used for this purpose may lead to discrimination of persons or groups and perpetuate historical patterns of discrimination, for example, based on racial or ethnic origins, disabilities, age, and sexual orientation, or create new forms of discriminatory impacts. Considering the very limited scale of the impact and the available alternatives on the market, it is appropriate to exempt AI systems for creditworthiness assessment and credit scoring when put into service by micro or small enterprises, as defined in the Annex of Commission Recommendation 2003/361/EC for their own use. Natural persons applying for or receiving essential public assistance benefits and services from public authorities are typically dependent on those benefits and services and are in a vulnerable position in relation to the responsible authorities. If AI systems are used to determine whether such benefits and services should be denied, reduced, revoked, or reclaimed by authorities, including whether beneficiaries are legitimately entitled to such benefits or services, those systems may have a significant impact on persons' livelihood and may infringe their fundamental rights, such as the right to social protection, non-discrimination, human dignity or an effective remedy. Those systems should therefore be classified as high-risk. Nonetheless, this Regulation should not hamper the development and use of innovative approaches in public administration, which would stand to benefit from a wider use of compliant and safe AI systems, provided that those systems do not entail a high risk to legal and natural persons. Finally, AI systems used to dispatch or establish priority in the dispatching of emergency first response services should also be classified as high-risk since they make decisions in very critical situations for the life and health of persons and their property. AI systems are also increasingly used for risk assessment in relation to natural persons and pricing in the case of life and health insurance which, if not duly designed, developed, and used, can lead to serious consequences for people's lives and health, including financial exclusion and discrimination. To ensure a consistent approach within the financial services sector, the above-mentioned exception for micro or small enterprises for their own use should apply, insofar as they themselves provide and put into service an AI system for the purpose of selling their own insurance products."

EU AI Act - Recital 80

Designation of Financial Services Authorities as Competent Authorities

“Union legislation on financial services includes internal governance and risk management rules and requirements which are applicable to regulated financial institutions in the course of the provision of those services, including when they make use of AI systems. In order to ensure coherent application and enforcement of the obligations under this Regulation and relevant rules and requirements of the Union financial services legislation, the authorities responsible for the supervision and enforcement of the financial services legislation should be designated as competent authorities for the purpose of supervising the implementation of this Regulation, including for market surveillance activities, as regards AI systems provided or used by regulated and supervised financial institutions unless Member States decide to designate another authority to fulfil these market surveillance tasks. Those competent authorities should have all powers under this Regulation and Regulation (EU) 2019/1020 on market surveillance to enforce the requirements and obligations of this Regulation, including powers to carry our ex-post market surveillance activities that can be integrated, as appropriate, into their existing supervisory mechanisms and procedures under the relevant Union financial services legislation. It is appropriate to envisage that when acting as market surveillance authorities under this Regulation, the national authorities responsible for the supervision of credit institutions regulated under Directive 2013/36/EU, which are participating in the Single Supervisory Mechanism (SSM) established by Council Regulation No 1024/2013, should report, without delay, to the European Central Bank any information identified in the course of their market surveillance activities that may be of potential interest for the European Central Bank’s prudential supervisory tasks as specified in that Regulation. To further enhance the consistency between this Regulation and the rules applicable to credit institutions regulated under Directive 2013/36/EU of the European Parliament and of the Council27, it is also appropriate to integrate some of the providers’ procedural obligations in relation to risk management, post-marketing monitoring and documentation into the existing obligations and procedures under Directive 2013/36/EU. In order to avoid overlaps, limited derogations should also be envisaged in relation to the quality management system of providers and the monitoring obligation placed on users of high-risk AI systems to the extent that these apply to credit institutions regulated by Directive 2013/36/EU. The same regime should apply to insurance and re-insurance undertakings and insurance holding companies under Directive 2009/138/EU (Solvency II) and the insurance intermediaries under Directive 2016/97/EU and other types of financial institutions subject to requirements regarding internal governance, arrangements or processes established pursuant to the relevant Union financial services legislation to ensure consistency and equal treatment in the financial sector.”

As can be seen from Figure 2.2, before the model can be placed on the market, it must pass specific verification steps. For example, in the case of AI models in the financial sector and high-risk models, they must pass the assessment to be deemed compliant with the regulations. Subsequently, it must be registered in an EU database, receive the CE mark, and then be placed on the market. Once the AI system is on the

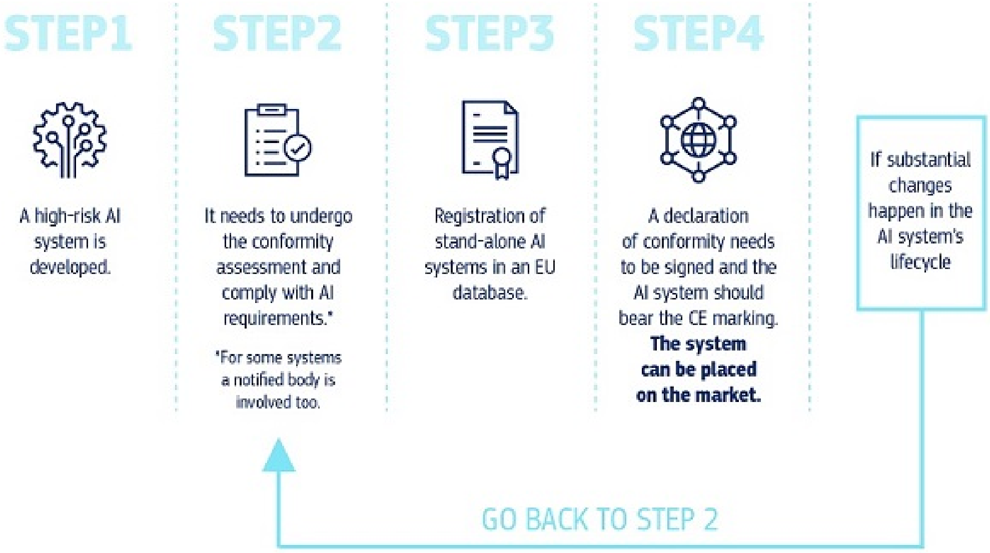

STEP1
A high-risk AI system is developed.
STEP2
It needs to undergo the conformity assessment and comply with AI requirements.*
*For some systems a notified body is involved too.
STEP3
Registration of stand-alone AI systems in an EU database.
STEP4
A declaration of conformity needs to be signed and the AI system should bear the CE marking.
The system can be placed on the market.
If substantial changes happen in the AI system's lifecycle
GO BACK TO STEP 2

Fig. 2.2 How does it all work in practice for providers of high-risk AI systems?

market, the authorities are responsible for market surveillance and control of providers who must also ensure full compliance with the law[5].

In the same direction as the EU Commission, in 2015, the UK presented the "*Digital Economy Strategy 2015-2018*" supporting the application and development of AI in business. In the UK, the Financial Conduct Authority (FCA) requires firms to explain why a more expensive mortgage has been chosen if a cheaper option is available. The G20 has adopted the OECD AI Principles for a trustworthy AI where it is underlined that users should not only understand AI outcomes but also be able to challenge them [78]. In the same way, the US issued in 2016 the "*National Strategic Research and Development Plan for Artificial Intelligence*" [101], and, on 7 January 2019, the White House's Office of Science and Technology Policy presented a draft "*Guidance for Regulation of Artificial Intelligence Applications*", which includes ten principles for United States agencies when deciding whether and how to regulate AI.

All the regulations that have been presented enforce the importance of having a strategy in an international field to tackle undesired problems with AI and regulate unfair and untrustworthy actions, with a particular focus on financial services.

[5] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Chapter 3

Fairness in Machine Learning

The primary goal of this dissertation is to introduce a comprehensive strategy for the identification of bias in machine learning models, employing Counterfactual Reasoning as a fundamental tool. To achieve this, the following chapter is dedicated to establishing the necessary foundation for the reader. It delves into an exploration of the most pertinent contributions within the realms of Fairness and Counterfactual Reasoning research, with the limitation that no regulation could deal with providing essential context for the subsequent discussions.

3.1 Introduction

Artificial Intelligence technologies offer a set of powerful techniques for the financial service domain to handle challenging tasks. In particular, Deep Learning (DL) models have been shown to outperform classical statistical techniques as well as more recent machine learning algorithms. However, none of the current DL-based algorithms can help users recognize where unfair treatment or discrimination will come from. In some cases, the more effective the algorithms are at classification, regression, and time-series detection tasks, the more they are prone to amplify the bias present in the data. These applications directly affect our lives and might harm people if not designed and engineered correctly having fairness considerations in mind. As expressed by Osoba et al. [134]:

> *"While machine learning and AI are technologies often dissociated from human thinking, they are always based on algorithms created by humans. Moreover, like anything created by humans, these algorithms are prone to incorporating the biases of their creators."*

3.1.1 Discrimination in the Financial Domain

Discrimination is not a recent concern in the financial world. Fairness in the financial services domain has had crucial importance since the government tried to address demographic, gender, and racial discrimination as regulatory compliance requirements. The set of laws that represents the foundation for fair acting in financial services firms was legislated between the 1960s and 1970s (e.g., *Fair Housing Act* of 1968 [47], *Truth in Lending Act* of 1968 [48], *Equal Credit Opportunity Act* of 1974 [49]). These laws take into account different aspects, as in the past the discrimination in the Financial decision-making process reflected not only social bias but also the lack of statistical information (see *Statistical Discrimination* in Section 3.2).

Today, we face an overabundance of poor-quality credit lending practices (e.g., high interest rates and fees, abusive debt traps) and there are concerns over the usage of too many data sources that can be used as a proxy for sensitive attributes (e.g., gender, age, country, race) leading to illegal discrimination. Although the law prohibits using gender to determine credit eligibility or pricing, countless proxies for gender exist, ranging from the type of deodorant a person buys to the movies they watch [1]. Credit scoring computation lies very often in the hands of ML algorithms, if we discard ethical considerations, the outcomes can negatively influence decisions, preventing customers from accessing opportunities for which they are instead qualified. Since the previous norms were not set to prevent discrimination in not-human decision-making setting (as in the case of ML algorithms), the EU Commission released the "*Ethics guidelines for a Trustworthy AI*" [4] and "*The White Paper*" [45] to give guidance on the ethical and safe use of AI. Some critical key requirements are "equity, diversity and not-discrimination" enclosed in the concept of fairness. In the same way, the *"Proposal for a Regulation laying down harmonised rules on artificial intelligence"* of the EU Commission remarks the need for model transparency to protect fundamental rights, such as non-discrimination and equality between genders [46].

[1] https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/

Now that we have established the ethical and legal foundation upon which the concept of fairness rests, we can introduce the subsequent sections that are organized according to the following logic. Section 3.2 presents the various types of social biases that can arise and how they can manifest themselves to users. In Section 3.3, the various types of bias are presented and related according to which stage of the data loop and machine learning predictions they originate. Section 3.4 is devoted to the mathematical formulation of the three statistical criteria (i.e. *independence*, *separation*, and *sufficiency*) to which each definition of fairness belongs to. Furthermore, the various metrics derived from the respective definitions of fairness are presented. Finally, Section 3.5 is devoted to listing the various bias-mitigation methodologies divided into pre-processing, in-processing, and post-processing.

3.2 Social harmful discrimination

Philosophers and psychologists have long studied fairness. However, no universal law acceptable in different dimensions has been drawn [166]. Before exploring the various definitions of fairness and the various statistical criteria on which the evaluation metrics are based, it is helpful to understand how a user can be perceived to have been hit by a discriminating judgment in the financial field. The various concepts of fairness are closely linked to the different types of discrimination and vice versa. Fairness can be achieved in different ways but it can be nullified by the occurrence of a parallel discriminating event affecting one or more different fairness definitions. This is why it is essential to determine every single type of discriminating phenomenon to have an overview of how it can be perceived and how it can harm a user [153]. Three principal kinds of discrimination can occur: *Direct, Indirect, Statistical*.

1. *Direct Discrimination* is when someone is treated unfairly because of a protected characteristic, such as gender, race, disability, or age (see Table 2.1). This type of discrimination arises when people belonging to one or more protected characteristics turn out to be disadvantaged in the outcome [176]. Taking into account racial discrimination, it appears any time an individual is discriminated against based on their *skin colour* or *racial* or *ethnic origin*. In financial Services, these traits have been regulated for a long time as specified in the laws *Fair Housing Act* [47] and *Equal Credit Opportunity Act* [49]. Although laws regulate the behaviour of financial institutions, several studies have demonstrated that racial discrimination still persists today. Indeed, the study of Cohen-Cole [38] found a significant difference in the amount of credit offered to similarly qualified

applicants living in black areas w.r.t. white areas. Subsequently, the *Census of Federal Deposit Insurance Corporation* [64] shows black and Hispanic Americans are more likely to go underbanked or deprived of conventional banking services than white or Asian Americans. Bartlett et al. [9] compare the discrimination of face-to-face decision w.r.t. algorithm scoring. He found that lenders charge Latin/African-American borrowers 7.9 and 3.6 basis points more for purchase and refinance mortgages, respectively, costing them $765 million in aggregate per year in extra interest. FinTech algorithms also discriminate but 40% less than face-to-face lenders [177]. Another longstanding form of direct discrimination is related to *gender*. Fay et al. [85] state that women can experience gender discrimination when seeking start-up capital. In that direction, Baden et al. [5] claim that financial markets experience *gender-based distortions*, disadvantaging female borrowers and savers, in addition to the lack of collateral requirements that limit women's access to finance. The study conducted by Ongena et al. [133] shows that women who own firms face more difficulties obtaining credit compared to similar firms owned by men, even when female-owned firms perform better than male-owned firms. All type of discrimination leads to social deterioration, to a loss of earnings both by women and financiers. As proof of what was previously stated, Sahay et al. [151] found that greater inclusion of women as users, providers, and regulators of financial services would have benefits beyond addressing gender inequality. Narrowing the gender gap would foster more excellent stability in the banking system and enhance economic growth. It could also contribute to more effective monetary and fiscal policy [177]. The study of Cohen-Cole also found *age discrimination*. Indeed, age discrimination is not new in financial firms. In fact, in 2002 the Federal Reserve Board [22] presented a study that shows people in underserved populations may be unfamiliar with tools of the financial system (e.g. credit cards) and may feel distrust in the use of the same. A combination of growing complexity increases in consumer responsibility, as well as the noted changes in the structure of personal nuance to include more individual credit have contributed to differences in financial literacy [177].

2. *Indirect Discrimination* happens when there are policy and rules that apply in the same way for everybody, but disadvantages a group of people who share a protected characteristic. However, protected groups or individuals still get to be treated unjustly due to implicit effects from their protected attributes. As an example, the residential zip code, or more generic geographical information of a person, can be used in decision-making processes such as loan applications.

However, this can still lead to racial discrimination. Even though zip code is not a non-sensitive attribute in principle, it might be correlated with race because of the population of residential areas [176]. If this happens, the person or organisation applying the policy must show a good reason for it, but if there is a direct correlation and causality between a sensitive characteristic and a non-sensitive one, it must be addressed carefully.

3. *Statistical Discrimination* is a theorized statistical behaviour in which statistical characteristics are used as a proxy for either hidden or more difficult-to-determine characteristics relevant to the outcome. For instance, if a financial agent has to decide whether or not to grant a customer the requested credit but, having imperfect information about the customer, decides to use statistical information (e.g., correlation or mutual information) according to the group to which he/she belongs as a proxy for the decision, then we are in the presence of statistical discrimination. According to this theory, inequality may exist and persist between demographic groups even when economic agents are rational and non-prejudiced [122, 140]. It stands in contrast with taste-based discrimination, which uses racism, sexism, and the like to explain different labour market outcomes of groups.

3.3 Where does bias in data come from?

As said in the previous sections, fairness has been widely regulated in financial firms. However, these kinds of regulations are aimed at human beings involved in decision-making processes in financial activities. In the era of Big Data and with the advent of automated decision-making systems, responsibility for unethical actions is hard to determine: who is accountable for the decision taken by these algorithms? Olteanu et al. [131], Suresh et al. [159] and Mehrabi et al. [122] have provided an extensive list of data biases. In Mehrabi et al. [122], the different types of data biases can occur in different steps of the machine learning operation loop: data collection and manipulation (Data), prediction (Algorithms), and data ingestion (User-Interactions).

Going deeper, we have the feedback loop phenomenon when the prediction outcome is exploited as a new ground-truth to update the model. These decisions will affect future data that will be collected for subsequent training rounds or new models [35]. In addition, this infinite loop also involves the interaction with the user who can provide feedback that influences the model [30].

Table 3.1 Definition and type of bias, partially from [122]

Bias	Definition
Historical bias	*Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even if data is perfectly measured and sampled [158].*
Social bias	*Social bias happens when other people's actions or content coming from them affect our judgment [6].*
Temporal bias	*Temporal bias arises from differences in populations and behaviors over time [131].*
Representation bias	*Representation bias occurs when the development sample under-represents some part of the population, and subsequently fails to generalize well for a subset of the use population [158].*
Population bias	*Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset or platform from the original target population [131].*
Aggregation bias	*Aggregation bias happens when false conclusions are drawn for a subgroup based on observing other different subgroups or generally when false assumptions about a population affect the model's outcome and definition [158].*
Sampling bias	*Sampling bias arises when there is a non-random sampling of subgroups and, as a result, the development sample will represent a skewed subset of the target population [158].*
Popularity bias	*Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation—for example, by fake reviews or social bots [37].*
Evaluation bias	*Evaluation bias occurs when the benchmark data used for a particular task does not represent the use population [158].*
Algorithmic bias	*Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm [6].*
User-Interaction bias	*User Interaction bias is a type of bias that can not only be observed on the Web but also get triggered from two sources—the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction [6].*
Emergent bias	*Emergent bias happens as a result of use and interaction with real users. This bias arises as a result of change in population, cultural values, or societal knowledge usually some time after the completion of design [91].*
Statistical bias (SB)	*Statistical bias is a term that refers to any type of error or distortion that is found with the use of statistical analyses*[4].
SP-Simspons's Paradox	*Simpson's Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations*[3].
SB-Omitted Variable bias	*Omitted variable bias occurs when one or more important variables are left out of the model*[4].
SB-Cause-Effect bias	*Cause-effect bias can happen as a result of the fallacy that correlation implies causation*[4].

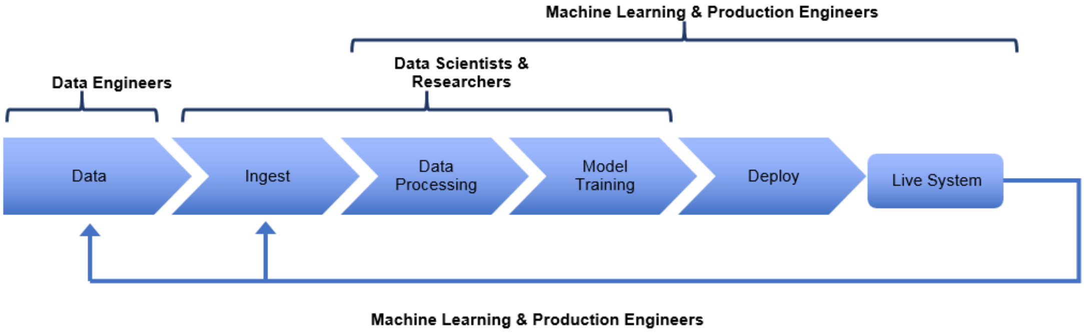

Machine Learning & Production Engineers
Data Scientists & Researchers
Data Engineers
Data
Ingest
Data Processing
Model Training
Deploy
Live System
Machine Learning & Production Engineers

Fig. 3.1 Example of Machine Learning feedback loop

Considering a financial institution that wants to automate the decision-making process of a loan application, the data biases that might affect the fairness and present discriminatory biases in the main ML loop steps are those listed in Table 3.1.

Historical and Social biases are the main biases that can occur in the model. Suppose that the data is characterized by historical choices influenced by discriminatory Social biases on sensitive characteristics. Accordingly, these Social biases may be part of the Historical ones. In that case, the algorithm will be influenced by such biased data leading to equally discriminatory and biased decisions based on historical bias.

Although not related to finance, the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) case[2] is a prime example of how an algorithm affected from Historical and Social biases can discriminate and, consequently, be considered unreliable. The COMPAS risk scores, considered by judges during sentencing, predict black people to commit a future crime of any kind with higher probability than white people. However, these scores do not reflect the real risks of black and white people to re-offend. Hence, we can state that the algorithm is affected by historical and social biases related to discrimination of black people.

Historical bias is closely related to the temporality of the phenomenon and can also be seen as Temporal bias from a certain point of view. Let us suppose that in a given period of time, the dataset of financial loans' application contains only users belonging to the well-off category of a population who got a credit. In the same dataset, there have been no loans for less well-off categories. This dataset characteristic will automatically lead the algorithm to avoid approving a credit towards the lower-middle class of the population. However, even though a change in characteristics of the population that apply for a credit occurs, the ML algorithm will remain biased by the *old* credit-applicant profile.

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Another type of bias is Representation bias. This type of bias is closely linked to data generation and sampling used for the model training. Furthermore, it can be considered a macro category of other biases such as Population bias, Aggregation bias, and Sampling bias. Suppose a data engineer needs to create a dataset for the credit application. If the available data represents only a specific portion of the population (e.g., people over 50), we can speak of Population bias. The same applies if there is an aggregation of two or more subgroups with specific peculiarities. In that case, the new merged group will probably lose these peculiarities in favour of the most shared characteristics. Hence, Aggregation bias can lead to a model that is not optimal for any group or a model that is fit to the dominant population [158]. In the case where the data provides a broad spectrum of the population but, during the sampling phase before model training, there is an imbalance or lack of representation of some subgroups, this can be called Sampling bias. In the case the dataset used in the evaluation stage is affected by one or more of the above biases, we fall in the case of the Evaluation bias.

A further bias closely related to the Representation bias category is Popularity bias. Popularity bias is a type of bias of particular importance in Information Retrieval and Recommender Systems. Recommender Systems (RS) can have different applications in the financial world, such as portfolio optimization, personalized stock prediction, or peer-to-peer lending [178]. When a financial product is more present in the data (blockbusters) than another, the RS will recommend that product more frequently than niche products (long tails). This issue is widely known as the Accuracy-diversity trade-off, or simply Popularity bias [1].

A statistical and evaluation problem on unfairness criteria is the *Simpson's Paradox*[3]. The Simpson's Paradox is related to the subgroup representation of a dataset as well. It could manifest itself when a population characteristic (e.g., gender) appears to be disadvantaged when the entire population is analyzed. However, the same characteristic is advantaged in some subgroups when the analysis is performed at a finer-grained level. For example, credit card holders are in general unbalanced towards employers with a permanent contract. However, if the analysis is focused on pre-paid credit cards, people with fixed-term contracts represent the predominant part of customers. This type of paradox is widespread in statistics. However, this is not the unique type of Statistical bias[4] that can arise in data and problem formulation. Suppose that in a loan application, a sensitive variable is omitted (e.g., the race). The algorithm can discriminate sensitive groups without basing their decision on the same sensitive

[3] https://plato.stanford.edu/entries/paradox-simpson/

[4] https://data36.com/statistical-bias-types-explained/

variable but through other variables that implicitly contain the information of the sensitive variables. This situation is known as Omitted Variable bias. Similarly, correlation-causality bias, known as Cause-Effect bias, is a statistical error that arises when the correlation between two events is mistaken for causation. For example, if there is a correlation between the race and the defaulting characteristic, there may be a risk that the algorithm will erroneously learn this correlation.

Biases can also emerge in data acquired in the third phase of the feedback loop. Emergent bias is one of them. In that case, the bias emerges only at the production stage. Accordingly, the model did not show biases in the training phase, but when the characteristics of real users become different from the training dataset, it generates biased outcomes. The biases presented so far referred to biases in the data or during the generation of the data. Naturally, as emerged during our discussion, there is not a sharp separation between one bias and another. Indeed, some biases share similar characteristics, such as the case of Temporal bias and Emergent bias, which are both influenced by the temporal dimension.

This type of bias can also be influenced by poor User Interaction design, also known as User-Interaction bias. Algorithm bias, on the other hand, is a type of bias directly added by the algorithm. It is caused either by premeditation or by a design error on the part of the data scientist. Unquestionably there are other types of biases, but to the best of our knowledge, these are the main biases that can occur when using ML techniques in financial tasks.

3.4 Non-discrimination statistical criteria and fairness definitions

Fairness has been studied for a long time by philosophers and psychologists as one of the principles on which to base justice and the rights of all citizens. Despite the work done and the laws enacted in all these years, a clear and precise vision of how this principle can be applied to everyday life is still hard to achieve. Overall, in the context of decision-making, the concept of fairness can be briefly and ideally summarized as *"the absence of any prejudice or favouritism toward an individual or a group based on their inherent or acquired characteristics"*[122]. However, although it can give an idea of fairness, it cannot be helpful in a practical way since unfairness and discrimination can appear differently. According to the analysis provided in the previous section, from our point of view, fairness can be seen as the total absence of any kind of bias.

Following, the problem of fairness and the interconnection between data, sensitive variables, and outcomes is outlined from a statistical point of view.

Next, a mathematical formulation of a classification model will be given, and the statistical criteria underpinning the various definitions and metrics of fairness will also be presented.

3.4.1 Problem formulation

Considering the case of financial firms' activities and precisely the case of the loan application, Machine Learning models are trained to find generalizable predictive patterns from previous customers' data. A loan application is a classification task in which we might try to predict whether a loan applicant will pay back her loan by looking at various characteristics such as credit history, income, and net worth.

Let $X \in \mathbb{R}^n$ the space of possible n non-sensitive feature values of a loan applicant, $Y \in \{0, 1\}$ the target variable denoting if the loan applicant repays the loan (i.e., $Y = 1$) or defaults (i.e., $Y = 0$), and $S \in \mathbb{R}^l$ be the space of l protected sensitive attributes. financial firms make use of ML models to predict the risk score $f(X) = \mathbb{P}(Y = 1 \mid X)$ equal to the probability that a new loan applicant will repay based on his characteristics. This probability can be exploited for classification purposes by assigning a positive label (i.e., will repay) to customers above a given cutoff τ which is generally defined as equal to 0.5. We can now introduce the output prediction space as:

$$\hat{Y} = \begin{cases} 1 & \text{if } f(X) > \tau \\ 0 & \text{if } f(X) < \tau \end{cases}$$

The n features of any X customer applying for a loan can implicitly contain or encode sensitive individual characteristics. Let us suppose we divide the features into "*neutral*" (i.e., X) and "*protected*" (i.e., S), removing the protected from the training data or for monitoring tasks cannot guarantee the model to be discrimination-free for multiple reasons (see *Indirect Discrimination* and *Statistical Discrimination* in Section 3.2). We will discuss this deeply in Section 3.4.5 under the concept of *fairness under unawareness*.

Many fairness criteria have been proposed over the years, each aiming to formalize different desiderata. Without these considerations and following the Kozodoi et al. [110] differentiation, in what follows, we will introduce three non-discrimination statistical criteria (i.e., Independence, Separation, and Sufficiency) and the correspondent Fairness Definitions (FDs) for reaching them.

3.4.2 Independence

Independence is a fundamental notion of *Probability Theory.* In statistic, independence states that "*given n variables $A_1, A_2, ..., A_n$, they are independent if and only if their joint probability can be factorized into their marginal probabilities as* $\mathbb{P}(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} \mathbb{P}(A_i)$". Furthermore, independence is the backbone of the *Statistical* notion of discrimination (see Section 3.2).

Taking up the case of a loan classification task in which S_i corresponds to a sensitive feature, for simplicity we consider a binary case, the independence criteria can be formulated as:

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

where the score function $f(X)$ satisfies the independence criteria if the classifier prediction is statistically independent from the sensitive attributes, so the model function can be also written as $\hat{Y} \perp S_i$. Let us suppose to have gender as a sensitive feature where the disadvantaged, or unprivileged, group is the *female* gender. Then, the algorithm used by the financial institution has achieved statistical independence from that particular sensitive variable if it proves to have the same statistics in predicting the good creditor both with "*female*" and the advantaged "male" gender. Then the probabilistic formulation is as follows:

$$\mathbb{P}[\hat{Y} \mid S_i = \text{“female”}] = \mathbb{P}[\hat{Y} \mid S_i = \text{“male”}]$$

Thus, the algorithm can be considered independent of the sensitive variable *gender*. An example of a fairness definition (FD) based on independence criteria is the Statistical or Demographic Parity.

FD 3.4.1 (*Demographic Parity or Statistical Parity*)**.** *A predictor satisfies demographic parity if the probability of obtaining a positive outcome does not depend on the sensitive characteristics* [77, 112].

Demographic Parity can be seen both as an ex-ante and ex-post metric. The ex-ante metric refers to the Difference in Positive proportion Labels (DPL), which measures the difference between the ratio of positive labels for the sensitive or disadvantaged groups and the ratio of positive labels for advantaged groups in the dataset. This metric is also known as *ex-ante Difference in Statistical Parity* (ex-ante DSP). DPL can be formulated as:

$$DPL = q_{S_i=1} - q_{S_i=0}, \tag{3.2}$$

where $q_{S_i=1} = \frac{\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)}=1, Y^{(j)}=1\right)}{\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)}=1\right)}$ [5] is the ratio between positive data (or good creditor) of an advantaged group $\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)} = 1, Y^{(j)} = 1\right)$ and all the sample belonging to the same advantage group $\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)} = 1\right)$, while $q_{S_i=0} = \frac{\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)}=0, Y^{(j)}=1\right)}{\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)}=0\right)}$ is the ratio between positive data (e.g. good creditor) of a disadvantaged group $\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)} = 0, Y^{(j)} = 1\right)$ and all the sample belonging to the same disadvantage group $\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)} = 0\right)$. The closer DPL is to 0, the fairer the dataset is. The ex-post metrics are the Difference in the Positive proportion of Predicted Labels (DPPL) and the Disparate Impact (DI). The first is similar to the DPL but refers to the predicted labels and is formulated as:

$$DPPL = \hat{q}_{S_i=1} - \hat{q}_{S_i=0}, \tag{3.3}$$

where $q_{\hat{S_i}=1} = \frac{\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)}=1, \hat{Y}^{(j)}=1\right)}{\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)}=1\right)}$ is the ratio between the positive predicted outcome of an advantage group $\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)} = 1, \hat{Y}^{(j)} = 1\right)$ and all the outcome belonging to the same advantage group $\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)} = 1\right)$, while $q_{\hat{S_i}=0} = \frac{\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)}=0, \hat{Y}^{(j)}=1\right)}{\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)}=0\right)}$ is the ratio between positive predicted outcome of a disadvantaged group $\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)} = 0, \hat{Y}^{(j)} = 1\right)$ and all the sample belonging to the same disadvantage group $\sum_{j=1}^{N} \mathbb{I}\left(S_i^{(j)} = 0\right)$. The interpretation of this metric is very similar to DPL and is known as *ex-post Difference in Statistical Parity* (ex-post DSP) or simply *Difference in Statistical Parity* (DSP). The Disparate Impact is the ratio of the proportion in the predicted positive outcome of disadvantaged groups $q_{\hat{S_i}=0}$ and advantaged groups $q_{\hat{S_i}=1}$ [87, 107], as:

$$DI = \frac{\hat{q}_{S_i=0}}{\hat{q}_{S_i=1}}. \tag{3.4}$$

The Disparate Impact metric should be within $(0.80 - 1.20)$ range for the "*rule of thumb*" [6].

[5] An indicator function $\mathbb{I}(\cdot)$ is used to denote whether a condition is true (1) or false (0).

[6] In an employment context in the US, the regulation of The Equal Opportunity Act is known as "80% rule" or as a "rule of thumb" for measuring disparate impact [67]. In fact, the DI value should be between 0.8 and 1.2

Another type of fairness definition that stands under the Independence Criteria is the Conditional Statistical Parity.

FD 3.4.2 (*Conditional Statistical Parity*)**.** *Conditional statistical parity means that an equal proportion of defendants are detained within each sensitive group, controlling for a limited set of "legitimate" risk factors* [51].

Formally, for a set of legitimate factors L which can correspond also to a specific value of features in the X space, the predictor $f(X)$ satisfies conditional statistical parity if $\mathbb{P}[\hat{Y} \mid L=1, S_i=0] = \mathbb{P}[\hat{Y} \mid L=1, S_i=1]$. In essence, Conditional statistical parity states that people in both protected and unprotected groups should have an equal probability of being assigned to a positive outcome given a set of legitimate factors L (e.g., credit history, employment) [164].

Independence has convenient technical properties [8]. However, decisions based on a classifier that satisfies independence can have undesirable properties (and similar arguments apply to other statistical criteria). For example, if the positive and negative outcomes are differently distributed between groups, using Independence criteria can worsen the model's performance. It all depends on how we want to consider homogeneous or heterogeneous the two groups concerning the target variable.

3.4.3 Separation

In statistics, Separation occurs if *"the predictor is associated with only one outcome value when the predictor range is split at a certain value"*, and, in this case, if the sensitive characteristic may be correlated with the target variable. The separation criterion allows the correlation between the outcome and the sensitive attribute to the extent that the target variable justifies it. For instance, in the case of a loan classification task in which S_i corresponds to a sensitive feature, the separation criteria can be formulated as:

$$\begin{cases} \mathbb{P}[\hat{Y} = 1 \mid Y = 1, S_i = 0] = \mathbb{P}[\hat{Y} = 1 \mid Y = 1, S_i = 1] \\ \mathbb{P}[\hat{Y} = 0 \mid Y = 0, S_i = 0] = \mathbb{P}[\hat{Y} = 0 \mid Y = 0, S_i = 1] \end{cases} \tag{3.5}$$

The score function $f(X)$ satisfies the separation criteria if both the positive outcomes ($\hat{Y}=1$) and the negative outcomes ($\hat{Y}=0$) are independent of the sensitive attributes given the ground truth Y, so the model function can be also written as $\hat{Y} \perp S_i \mid Y$.

Let us suppose to have the gender as a sensitive feature (S_i) where the disadvantaged group is the *female* values. Then, the algorithm used by the financial institution has achieved statistical separation from that particular sensitive variable if the probability of obtaining both a positive outcome (good creditor) and a negative outcome (bad creditor) for the "female" gender is the same as the advantaged group of the "male" gender. Then the probabilistic formulation is as follows:

$$\begin{cases} \mathbb{P}[\hat{Y}=1 \mid Y=1, S_i=\text{“female”}] = \mathbb{P}[\hat{Y}=1 \mid Y=1, S_i=\text{“male”}] \\ \mathbb{P}[\hat{Y}=0 \mid Y=0, S_i=\text{“female”}] = \mathbb{P}[\hat{Y}=0 \mid Y=0, S_i=\text{“male”}] \end{cases}$$

In line with the above, several definitions of fairness that fall within the separation criterion will be presented below.

FD 3.4.3 (Treatment Equality)**.** *Treatment Equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories*[13].

Treatment Equality can be applied only to the outcome as an ex-post metric. Berk et al. [13] defined this as the ratio of false negatives to false positives (or viceversa):

$$TE = \tau_{S_i=0} - \tau_{S_i=1} \tag{3.6}$$

where both classes are $\tau_{S_i=1} = FN_{S_i=1}/FP_{S_i=1}$ and $\tau_{S_i=0} = FN_{S_i=0}/FP_{S_i=0}$.

Another important fairness definition is Equal Opportunity. This is also considered a relaxation of the *Equalized Odds* definition that will be exploited in this section.

FD 3.4.4 (Equal Opportunity)**.** *A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to a sensitive feature S_i and Y if $\hat{Y}=1$ and S_i are independent conditional on $Y=1$*[97].

The Equal Opportunity definition means that the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected (female and male) group members, so $\mathbb{P}[\hat{Y}=1|S_i=0,Y=1] = \mathbb{P}[\hat{Y}=1|S_i=1,Y=1]$ [164]. In other words, the equal opportunity definition states that the protected and unprotected groups should have equal true positive rates or Recall:

$$DEO = Recall_{S_i=1} - Recall_{S_i=0} \approx 0 \tag{3.7}$$

FD 3.4.5 (Equalized Odds)**.** *A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute S_i and outcome Y, if $\hat{Y}$ and S_i are independent conditional on Y.*

The equalized odds definition, provided by Hardt et al. [97], states that the protected and unprotected groups should have equal rates for true positives and false positives known as *Difference in Average Odds* (DAO) and formalized as:

$$DAO = \frac{1}{2}\left(\left|FPrate_{S_i=1} - FPrate_{S_i=0}\right| + \left|TPrate_{S_i=1} - TPrate_{S_i=0}\right|\right) \tag{3.8}$$

The *Difference in Rejection Rate* is another metric that measures whether qualified applicants from the advantages and disadvantages class are rejected at the same rates and is formulated as:

$$DRR = TNrate_{S_i=0} - TNrate_{S_i=1} n_{\widehat{S_i=1}}^{(0)}.$$

3.4.4 Calibration and Sufficiency

In statistics, a statistic is sufficient with respect to a statistical model and its associated unknown parameter if "*no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter*" [90].

In the case of a loan classification task in which S_i corresponds to a sensitive feature, the sufficiency criteria can be formulated as:

$$\mathbb{P}[Y=1 \mid f(X) > \tau, S_i=0] = \mathbb{P}[Y=1 \mid f(X) > \tau, S_i=1] \tag{3.9}$$

where the score function $f(X)$ satisfies the separation criteria if the likelihood of positive outcomes (i.e., $f(X) > \tau$) is independent of the sensitive attributes, so the model function can also be written as $Y \perp S_i \mid \hat{Y}$.

To be more specific, let us suppose to have the gender as a binary sensitive feature (S_i) where the disadvantaged group is the *female* gender (i.e., $S_i = 0$) and *male* the advantage one (i.e., $S_i = 1$). Then, the algorithm used by the financial institution is statistically sufficient for the sensitive variable if the probability of being judged a good creditor does not change, for instance, between male and female. Then, the probabilistic formulation is as follows:

$$\mathbb{P}[Y=1 \mid f(X) > \tau, S_i=\text{“female”}] = \mathbb{P}[Y=1 \mid f(X) > \tau, S_i = \text{“male”}]$$

Thus, the sensitive variable does not contribute to the strictly positive prediction of the outcome.

FD 3.4.6 (Conditional use accuracy equality)**.** *A classifier satisfies "conditional use accuracy equality" if the subjects in the protected and unprotected groups have equal Positive Predicted Value (PPV) and equal Negative Predicted Value (NPV).*

Furthermore, Chouldechova [34] defines the sufficiency metric SF^+ as the absolute difference between the group-wise PPV:

$$SF^+ = \mid PPV_{S_i=1} - PPV_{S_i=0} \mid$$

where $PPV_{S_i=1} = \frac{TP_{S_i=1}}{TP_{S_i=1}+FP_{S_i=1}}$ and $PPV_{S_i=0} = \frac{TP_{S_i=0}}{TP_{S_i=0}+FP_{S_i=0}}$[7].

FD 3.4.7 (Test Fairness or Calibration)**.** *A classifier satisfies test-fairness if individuals with the same predicted probability score R have the same probability of being classified in the positive class when they belong to either the protected or the unprotected group* [34]*.*

The definition of "Test fairness" can be written as $\mathbb{P}(Y=1|f(X)=r,S_i=0) = \mathbb{P}(Y=1|f(X)=r,S_i=1)$. The following is the definition of "Well-Calibration", which completes the definition of test fairness and extends it.

FD 3.4.8 (Well-Calibration)**.** *A classifier is considered well-calibrated between groups if, given the same predicted probability score R to individuals inside or outside the protected group, they must have the same probability of being classified in the positive class, and this probability must be equal to r.*

So, following the "Test Fairness" formulation, a classifier is considered "well-calibrated" between two groups if $\mathbb{P}(Y=1|f(X)=r,S_i=0) = \mathbb{P}(Y=1|f(X)=r,S_i=1) = r$.

Sufficiency often could come for free (at least approximately) due to standard machine learning practices. The flip side is that imposing sufficiency as a constraint on a classification system may not be much of an intervention. In particular, it would not affect a substantial change in current practices [8].

3.4.5 Other Fairness Criteria

The three principal fairness criteria have been presented in the previous Sections. These criteria introduced above comprise several other fairness concepts, which have been proposed in prior work. These criteria are slightly varying statistical formulations

[7] In an equivalent manner, and if the task is to get equivalent negative prediction between groups, can be defined the adjacent metric SF^- as the absolute difference between the group-wise NPV:

$$SF^- = \mid NPV_{S_i=1} - NPV_{S_i=0} \mid$$

where $NPV_{S_i=1} = \frac{TN_{S_i=1}}{TN_{S_i=1}+FN_{S_i=1}}$ and $NPV_{S_i=0} = \frac{TN_{S_i=0}}{TN_{S_i=0}+FN_{S_i=0}}$.

of the same fairness criteria of Independence, Separation, and Sufficiency [8, 127]. However, these fairness criteria belong to a specific type of differentiation of fairness criteria. Indeed, one way to evaluate fairness in the classification or regression task is based on identifying *groups*, *subgroups*, *individuals*, and finally *counterfactuals*. Besides these concepts, there is another one that has emerged due to privacy compliance with the not use of sensitive information and is known as *fairness under unawareness*.

Group-based fairness metrics essentially compare the outcome performance of the classification algorithm between two or more predefined groups. These groups can be based on attributes like race, gender, age, or any other protected characteristic.

Subgroup fairness is an extension of group fairness, but it considers fairness within subgroups or smaller categories within the protected groups. It looks at fairness at a more granular level.

Individual-based fairness metrics do not compare different groups as defined by a sensitive variable but consider the outcome for each participating individual regardless of their group or subgroup. It emphasises that similar individuals should receive similar treatment [77].

Counterfactual fairness notion derives from Pearl's causal model [154], which considers a model is fair if for a particular individual its prediction in the real world is the same as that in the counterfactual world where the individual had belonged to a different demographic group [169]. Kusner et al. [112] propose the concept of counterfactual fairness, which builds on causal fairness models and can be considered as the intersection of both *individual*- and *group-based* fairness concepts.

Fairness under Unawareness is a concept in the field of machine learning and algorithmic fairness that addresses fairness in situations where an algorithm or decision-making process is designed to be fair without having direct access to or awareness of certain sensitive attributes or characteristics of individuals. In other words, it aims to mitigate bias and ensure fairness without using or considering protected attributes like race, gender, or religion.

Following, we will define and exploit *fairness under unawareness* and *counterfactual fairness* related works as the backbone of our study.

Fairness under Unawareness

The first domains to take an interest in the theme were financial Services, Banking, and Health. In fact, due to the critical impact of decision-making in these domains on people's well-being, today, the use of sensitive characteristics is strictly prohibited. The decisional tasks, i.e., regression and classification tasks with models deprived of sensitive features, took the name of *fairness under unawareness* assessment. However, companies and institutions must demonstrate the fairness and impartiality of their systems despite the absence of such sensitive characteristics [31].

While designing the decision-making algorithm not to leverage sensitive information is simple, assuring the same accuracy as before and demonstrating that the predictor is unbiased is another matter. Thus, even if the regulation requires the use of the unawareness setting for model training, the assumption is still too strong to guarantee a fair model behaviour.

FD 3.4.9 (Fairness under Unawareness)**.** *An algorithm is fair as long as any protected attributes S are not explicitly used in the decision-making process* [94, 112]*.*

For tasks like granting credit cards or approving loans and mortgages, financial companies should collect and use sensitive features to ensure their tools are non-discriminatory. On this point, the EU Commission proposes a conformity assessment before AI systems are put into service or placed on the market [8]. In fact, their tools are subject to fair and trustworthy audit assessments to check their conformity. However, is a shallow check of the input characteristics sufficient to determine that a predictor will not suggest unfair treatment? Even though the user does not provide protected characteristics, the system could predict sensitive features from variables, i.e., proxy variables, that still represent protected characteristics. The models that infer sensitive features from proxy variables are known as "probabilistic proxy models" [24, 32].

Most of the approaches proposed in the literature for identifying proxy features rely on techniques capable of discovering multicollinearity between variables. If the correlation between two independent variables is 1 or -1, we have perfect multicollinearity between them [2]. Methods for discovering multicollinearity are based on Linear Regression, Variance Inflation Factor, and Pearson correlation coefficient [171]. However, the relationships may not be linear. In that case, cosine similarity and mutual information are the most used approaches [2].

Elliott et al. [82] investigated, in their work, whether from customer characteristics such as name and geolocation information (e.g., residence address) the information about the race can be inferred. Using a Bayesian classifier model, they demonstrated that first-name listings might improve prediction estimates. In particular, they showed that in some Asian and black subgroups, first names tend to have low sensitivity. Conversely, imputing native American and multiracial identities from surname and residence remains challenging. Chen et al. [32] studied the relationship between proxy

[8] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

features and sensitive variables (i.e., geolocation and race). In their work, bias seems to depend on the chosen threshold, suggesting an ad-hoc threshold estimation to produce fair thresholded classifiers and probabilistic proxy models.

Fabris et al. [83] use a quantification approach to measure group fairness when sensitive features are unknown. The advantage is that quantification-based estimates are robust to distribution shifts and do not allow the inference of sensitive attributes at the individual-class level. Biswas et al. [16] likewise employ quantification techniques. In detail, they propose a mitigation model in which training and test population subgroups structurally differ. The proposed model, CAPE (i.e., Combinatorial Algorithm for Proportional Equality), aims to minimize a peculiar loss to obtain a Proportional-Equality-fair model.

The exposure of some groups on a geographic and demographic basis is also a problem that impacts the Recommender Systems community. In this direction, there are some attempts to analyze and mitigate this type of issue. One possible solution is the re-ranking strategy [93], to balance the items produced in a continent and the ranking of the items. Another recent proposal is FairLens [135], a framework to discover the bias of a generic Decision Support System model. The authors tested the approach in the medical domain. Interestingly, this strategy involves human experts in analyzing misclassifications. Specifically, the expert describes which aspects of the impacted patients' clinical history are responsible for the model error in the considered groups. It is essential to underline that the human expert, who thoroughly analyzes potential fairness issues, plays a crucial role in the operational loop.

Counterfactual Reasoning as Fairness Perception

Counterfactual Reasoning is an active and flourishing field in artificial intelligence research [92, 125]. This research was initially born to investigate causal links [137], and today it can count on several contributions [89]. Most of them define and employ counterfactuals as helpful tools to explain the decisions taken by modern decision support systems. The underlying rationale is that some aspects of past events could predict future events. In detail, some studies focus on identifying causality-related aspects to discover the link between the counterfactuals and the analyzed phenomenon [69].

Counterfactual Reasoning finds application in various fields. To summarize what we have briefly detailed before, machine learning research has positively valued these contributions ranging from Explainable AI [128] to the most recent counterfactual fairness measures [103, 112]. Beyond the theoretical aspects, Counterfactual Reasoning is extensively applied to interactive systems [19, 61, 160, 162]. Unfortunately, this important application showed some limitations. These systems employ machine learning models that reflect the data they use for learning. Consequently, the same information influences the reasoning, and the contribution of Counterfactual Reasoning could be limited or somehow biased. The explaining policy, coming from Counterfactual Reasoning, exhibits a bias toward the implemented learning model. Researchers devoted considerable effort to tackle this issue and proposed new models such as doubly robust estimators [75].

Overall, even though limitations that need a solution, Counterfactual Reasoning is taking over Explainable AI, and it is becoming the de facto standard for explaining decisions taken by autonomous systems. In this respect, the European Union's "right to explanation" played a crucial role in arousing a further interest in this methodologies [109]. Indeed, they are compliant with the regulation and easily interpreted by either a domain expert or a layperson [156].

Decision support systems particularly benefited from these models. However, the more the application domain is vital, the more the fairness problem emerges. For instance, the issue cannot be overlooked in sensitive domains such as justice, risk assessment, or clinical risk prediction. This need promoted the most promising research in the Counterfactual Reasoning field to analyze and mitigate this issue. Kusner et al. [112] proposed a metric exploiting casual inference to assess fairness at an individual level (i.e., *Counterfactual Fairness*) by requiring that a sensitive attribute not be the cause of a change in a prediction.

In order to define Counterfactual Fairness, let us assume that we are given a causal model (U,V,F), where $V \equiv S \cup X$ are the observable variables, U is a set of latent background variables, and F is a set of functions $\{f_1,\ldots,f_n\}$, one for each $V_i \in V$ known as structural equations [17].

FD 3.4.10 (Counterfactual Fairness)**.** *Given a set of attributes n in a random space $X \in \mathrm{R}^n$, a classifier $\hat{Y}: X \to Y$ is counterfactually fair if, under any observational condition, we have:*

$$\mathbb{P}[\hat{Y}_{S_i \leftarrow 1} \mid S_i] = \mathbb{P}[\hat{Y}_{S_i \leftarrow 0} \mid S_i, X],$$

where $\hat{Y}_{S_i \leftarrow 1}$ and $\hat{Y}_{S_i \leftarrow 0}$ is the causal effect of the sensitive variable set to 1 or 0 on the outcome. The definition ensures that the prediction for an individual coincides with the decision if the sensitive variable would have been its counterfactual value [28]. Thus, the sensitive variable S should not be a cause of $\hat{Y}$ in any individual instance and will not change its distribution.

Pfohl et al. [139] further extended the approach for clinical risk assessment. They aim to mitigate the exposure of medical care disparities due to bias implicitly embedded in data for historically underrepresented and mistreated groups. For what concerns the risk assessment domain, Mishler et al. [126] put forward a similar working hypothesis. They propose a counterfactual equalized odds ratio criterion to train predictors operating in the post-processing phase. They extend and adapt previous post-processing approaches [97] to the counterfactual setting and employ doubly robust estimators.

In contrast to the majority of the mentioned studies, our investigation aims to leverage a counterfactual generation tool to reveal the presence of implicit biases in a decision support system. Interestingly, this motivation is similar to Bottou et al. [20]. In fact, both aim to answer the question: "How would the system have decided if we had replaced some user characteristics?". Beyond this commonality, the two studies differ significantly. Indeed, they focus on measuring the fidelity level of the system and robustifying the model. Instead, our study is in line with the goal of other investigations [70, 124] that aim to use the counterfactual approach to uncover the bias present in the dataset that plagues the predictive model itself.

3.5 Bias Mitigation Methodologies

The previous section has presented different statistical criteria based on which all fairness definitions refer from. These different definitions refer specifically to different ways or perspectives based on which to measure fairness. Most of the metrics are post-training metrics and evaluate the impact of the model prediction on the different groups. However, the bias mitigation is not exclusively a post-processing step but can also be a regularization step during the training of the model or a pre-training cleaning and re-balance. Indeed, bias mitigation methodologies can be divided into pre-training or pre-processing, in-training or in-processing (through algorithms modification or regularization), post-training, or post-processing. In the following, the various methodologies will be presented for each type of mitigation.

3.5.1 Pre-processing

Pre-processing methodologies refer to tools and strategies that Data scientists can use to give the model "as fair as possible" input data. [65] Examples of fair Pre-processing methodologies we can have are *Reweighting*, *fair representation*, *Class rebalance* and *Removing sensitive feature*. We want to highlight that "*Removing sensitive features*" is not a good strategy since they can be hidden in other proxy features. Some methodologies will be synthetically presented below.

Reweighting: Each instance (or tuple) in the training dataset is assigned a weight based on its sensitive attribute, allowing the dataset to be balanced with respect to that attribute without altering the original labels [26]. The goal of this method is to minimize the dependency between the sensitive attribute and the predicted outcome, effectively reducing bias according to the *Independence Criterion*. By reweighting the instances, the method ensures that the overall proportion of positive class labels is preserved, leading to the development of a classifier that is independent of the sensitive attribute. This process replaces underrepresented or overrepresented instances with others, ensuring the model learns in a balanced and unbiased manner.

Rebalance: Rebalancing methods address imbalances in the rates of positive and negative outcomes between different groups defined by sensitive attributes. For example, in a loan application scenario, if the rate of loan approvals (positive outcomes) is higher for the advantaged group compared to the disadvantaged group, the rebalance strategy seeks to equalise these rates. This can be achieved through oversampling techniques, where additional instances are generated for the disadvantaged group to balance their representation. Similarly, if the rate of loan rejections (negative outcomes) is higher for one group, rebalancing can adjust these instances accordingly. However, this approach may reduce accuracy when the outcome is inherently correlated with the sensitive attribute. Alternatively, random perturbation of class labels can be used, although it may lead to variability in results across different iterations. Another approach involves transforming the features to maintain their distribution while minimising their correlation with the class label, aiming to reduce any bias linked to sensitive attributes as much as possible [67].

Disparate impact remover: Feldman et al. [87] introduce this method to achieve fairness by ensuring independence between the sensitive attribute S_i, the features X, and the outcome Y.

This method refers to the *Disparate Impact* fairness metric previously examined. Once the Disparate Impact presence has been certified, it requires the dataset $D = (X, Y, S)$ to be changed to $\hat{D} = (\hat{X}, Y, S)$ so that would be certified as fair. The repaid term is the label X with the $\hat{X}$ so that the cumulative probability of $F_{S_i=1}(X)$ is equal to $F_{S_i=1}(\hat{x})$, thus preserving the ability to predict the class Y.

Learning fair representations: This algorithm proposed by Zemel et al. [173] formulate fairness as an optimisation problem of finding a good representation of the data with two competing goals: to map the data $D = (X, Y, S)$ as well as possible to a prototype set $\hat{D} = (Z, Y)$ while simultaneously obfuscating any information about membership in the protected group by nulling out the difference in statistical parity.

According to Zemel et al. [173], this algorithm seems to achieve both group fairness and individual fairness. On the same line of this work, different versions of algorithm accounting variational fair autoencoder [114] or fair normalizing flows [7] have been proposed with the same goal.

Optimized preprocessing: Pin Calmon et al. [141] proposed a convex optimization for learning a data transformation with three goals: controlling discrimination, limiting distortion in individual data samples, and preserving utility. It transforms the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives. This method also enables an explicit control of individual fairness and the possibility of multivariate, non-binary protected variables [177].

3.5.2 In-processing

In-processing methodologies refer to techniques that try to modify and change classic learning algorithms to prevent unfair and discriminatory outcomes during the model training process [65]. Suppose it is allowed to change the learning procedure for a machine learning model. In that case, in-processing can be used during the training of a model, either by incorporating a regularisation term, changes into the objective function, or imposing a constraint. The main measures used in this approach are false positive rate, false negative rate, or any misclassification rate. One constraint or more can be added, and the equality of false negative rates implies the equality of true positive rates, which means equal opportunity (Separation Criteria). After adding the restrictions to the problem, it may turn intractable, so a relaxation on them may be needed. This technique obtains good results in improving fairness while keeping high accuracy and lets the programmer choose the fairness measures to improve. However, each machine learning task may need a different method to be applied, and the code in the classifier needs to be modified, which is not always possible [8]. Some methodologies will be synthetically presented below.

Prejudice remover: Kamishima et al. [106] proposed prejudice remover, a fairness-driven regularized classification model. This is obtained by adding a regularization term to the loss function and analyzing the model fairness-based independence criteria named as prejudice index (PI):

$$\arg\min_{f} L[f(X), Y] + \eta \mathrm{PI}$$

where $L(\cdot)$ is the loss function of the model $f(X)$, and η hyper-parameter that regulates the influence of the regularization terms PI. PI measures the amount of mutual information between Y and S_i, and is equal to:

$$PI = \sum_{\hat{Y},S_i} \mathbb{P}(\hat{Y},S_i) \ln \frac{\mathbb{P}(\hat{Y},S_i)}{\mathbb{P}(S_i)\mathbb{P}(\hat{Y})}.$$

This regularization approach can be applied to general prediction algorithms within a training rutine ensuring that the sensitive attributes become less influential in the outcome.

Adversarial debiasing: Zhang et al. [174] proposed a framework for mitigating such biases. In the proposed architecture, they try to maximise the accuracy of outcome prediction on Y and minimise the accuracy of adversary outcome's predictor on sensitive attribute S_i. The model is thus composed of two neural networks (the *predictor* and the *adversary*), one with opposite objectives to the other. The *predictor* network tries to accomplish the task of predicting the target variable Y, given X, modifying the model weights parameter W to minimise the loss function $L_P(f_p(X),Y)$. The *adversary* network tries to accomplish the task of predicting the sensitive class S_i, given $\hat{Y} = f_p(X)$, also modifying, in this case, the weights U to minimising the loss function $L_{S_i=1}(f_{S_i=1}(\hat{Y}),S_i)$. The weights U of the *adversary* are updated in order to minimize $L_{S_i=1}(\hat{S}_i,S_i)$ at each training step according to the gradient $\nabla_U L_{S_i=1}$. The weights W of the *predictor* are propagated to the gradient in order to minimise its loss function $L_p(\hat{Y},Y)$, and simultaneously maximising the loss function of the adversary, formulated as:

$$\nabla_W L_P(\hat{Y},Y) - proj_{\nabla_W L_{S_i=1}(\hat{S}_i,S_i)} \nabla_W L_P(\hat{Y},Y) - \alpha \nabla_W L_{S_i=1}(\hat{S}_i,S_i)$$

where the $proj_{\nabla_W L_{S_i=1}(\hat{S}_i,S_i)} \nabla_W L_P(\hat{Y},Y)$ prevents the *predictor* from moving in a direction that helps the adversary decrease its loss while the last term, $\alpha \nabla_W L_P$, attempts to increase the adversary loss.

Xu et al. [170] introduced FairGAN, which generates synthetic data free from discrimination and is similar to real data. FairGAN consists of two components: a *generator* which generates the fake data conditioned on the protected attribute, and two *discriminator* that are trained to identify the fake sample from the real ones. For achieving fairness constrained, one discriminator is trained in order to identify if the outcome is of a sensitive group or not [170].

Fair constraints: Recent works use constraints on the classifier, formulating it as a constrained optimization problem to satisfy specific group fairness and simultaneously maximize accuracy. One of the most interesting is the work of Celis et al. [29] that proposes a meta-fair classification algorithm designed to achieve fairness according to several different fairness criteria. This meta-algorithm for classification takes as input a large class of fairness constraints regarding multiple non-disjoint sensitive attributes, which come with provable guarantees. This is achieved by first developing a meta-algorithm for a large family of classification problems with convex constraints and then showing that classification problems with general types of fairness constraints can be reduced to those in this family. Each criterion has a fairness metric, which measures the equality (or discrimination) between groups. So the main idea is that if the metric is similar across groups, the level of fairness is high (group fairness).

Donini et al. [72] propose a fairness risk measure during model learning. In the same way, Williamson et al. [168] start from the notion of perfect fairness in terms of *Demographic Parity* and *subgroup losses* parity (as the average of subgroup losses deviation) and build a convex fairness-aware objective based on minimizing the *Conditional Value at Risk* (CVaR) [148] and demonstrate the relation between fairness risk measures and risk measure of mathematical finance. Martinez et al. [120] propose a Pareto constraints method known as Blind Pareto Fairness (BPF) leveraging recent methods in no-regret dynamics [33] in order to address the worst-case of subgroup robustness. Similar to Martinez et al. [120], Chzhen et al. [36] propose a framework based on Pareto frontier optimization with Demographic Parity constraint.

A significant concern of fair classification with constraints is the trade-off between accuracy and fairness. As demonstrated by the previous works [29, 120], and specifically in loan application by Zhang et al. [177], fairness constraints can have on one hand a big impact on the accuracy and, being the accuracy strictly related to the profit of the financial institution and the long term profit of an applicant, from the other hand a loss of revenue or impoverishment for all the stakeholders. For these reasons, Liu et al. [113] propose a *Relaxation Constraints Fairness* and do not recommend base constraints on classical fairness criteria but through firstly understanding the causality between variable and outcome.

3.5.3 Post-processing

Post-processing methodologies refer to techniques that try to modify and change the outcome of classic learning algorithms to improve prediction fairness. In the loan classification task, the classifier will return a score that reflects the posterior probability that a candidate could be or not be a defaulter. High scores are likely to get a positive outcome, while low scores are likely to get a negative one, but we can adjust the cutoff to determine when to answer yes as desired, affecting the trade-off between the rates for true positives and true negatives.

The advantages of post-processing include that the technique can be applied after any classifiers, without modifying it, and has a good performance in fairness measures. The cons are the need to access to the protected attribute in test time and the lack of choice in the balance between accuracy and fairness [8].

Cutoff post modelling: The cutoff probability is usually set at $\tau = \frac{1}{2}$. The score function will be fair if the model is independent of the protected attribute. Then any choice of the cutoff will also be fair, but classifiers of this type tend to be biased, so a different cutoff may be required for each protected group to achieve fairness. If the classifier is biased, then the cutoff for the advantaged class can be adjusted to $\tau + \delta$, and the cutoff for the disadvantaged class will be reduced to $\tau - \delta$ or vice versa, and the hyperparameter δ can be tuned appropriately until the desired level of fairness and accuracy is achieved [97]. A way to do this is plotting the true positive rate against the false negative rate at various cutoff settings and find a cutoff where the rates for the protected group and other individuals are equal and also trying to maintain as high as possible the accuracy [97].

Reject Option Based Classification: Reject option classification, proposed by Kamiran et al. [105], defines a critical region of high uncertainty and reassigns labels for customers that have predicted scores within this region, such that members of the unprivileged group receive a positive label ($Y = 1$) and vice versa. Formally, the critical region is defined as:

$$\max(\mathbb{P}[f(X) = 1 \mid X], 1 - \mathbb{P}[f(X) = 1 \mid X]) \leq \theta$$

where $0.5 < \theta < 1$. Given a set of predicted scores and the true outcomes, a suitable value of θ and the number of posterior classifications can be optimized for a fairness criterion (e.g., independence) based on the allowed fairness bound $\sigma = [\sigma_1, \sigma_2]$ for the corresponding constraint.

Discrimination-Aware Ensemble: Kamiran et al. [104] proposed a second solution. It makes an ensemble of (probabilistic, non-probabilistic, or mixed) classifiers discrimination-aware by exploiting the disagreement region among the classifiers. A standard ensemble classifier predicts and classifies new instances by assigning the majority class label. The solution deviates from this standard procedure to neutralize the effect of discrimination. Specifically, if all member classifiers predict the same label, the agreed class label is assigned; otherwise, we compensate the instances belonging to the deprived group by assigning them the positive label and penalize the instances belonging to the favored group by giving the negative label [104]. Then having more classifiers in the ensemble may neutralize the discriminatory effect of the ensemble due to the fair classifiers. Thus, using ensembles is very useful by nature towards the solution of the discrimination-aware classification problem.

Equalized odds post-processing: Pleiss et al. [144] proposed "*Equalized odds post-processing*" as a post-processing technique that solves a linear program to find probabilities with which to change output labels to optimize equalized odds [97, 144]. Equalized odds processor uses a different logic to post-process classifier predictions. It finds a cutoff value τ that optimizes the predictive performance while satisfying the separation criterion, i.e., ensuring the same false negative and false positive rate per group.

Platt scaling: Platt [143] proposed a post-processing method known as *Platt scaling* based on the calibration criteria for support vector machine output. Calibration addresses the problem that some classification algorithms cannot make a statement about the certainty of their prediction, i.e., the probability with which an instance belongs to a certain class. In credit scoring, the predicted score could be an indicator of default risk but not the actual probability of default. A score $f(X)$ is calibrated if:

$$\mathbb{P}(Y = 1 \mid f(X) = \tau) = \tau$$

When extending the calibration condition to the group level, it becomes apparent that it implements the sufficiency criterion [8]:

$$\mathbb{P}(Y = 1 \mid f(X) = \tau, S_i = 1) = \mathbb{P}(Y = 1 \mid f(X) = \tau, S_i = 0) = \tau.$$

To achieve calibration for each group, Platt scaling is applied separately to each sensitive group. The method uses the output of a possibly uncalibrated score $f(X)$ as input for logistic regression fitted against the target variable Y. Based on the loss function of the logistic regression, the result is a new calibrated score that represents the probability that an instance belongs to the positive class. Formally, Platt scaling minimizes the log-loss equal to $\mathbb{E}(Y\ln(\sigma)+(1-Y)\ln(\sigma))$ by finding the optimal parameters A and B of the sigmoid function $\sigma = \frac{1}{1+exp(Af(X)+B)}$ [143].

3.5.4 Fairness tools

In the scientific community, fairness has long been the subject of attention and study in searching for techniques that can solve this problem in the *pre*, *in*, and *post-processing* phases, as seen above. Several companies and developers have contributed to using these techniques through the development of open-source libraries or paid tools to stimulate a conscious, responsible, and trustworthy use of machine learning techniques. Some of the best-known libraries are *What-if tools*, *Audit AI*, *AIF360*, *fairlearn*, *Aequitas*, and *Amazon SageMaker Clarify*.

The What-if tools (WIT) [167] developed by Google is an open-source, model-agnostic interactive visual tool for model understanding and fairness measure, as part of TensorBoard. AIF360 is an open-source library that can help detect and remove bias in machine learning models [10]. It has been developed by IBM and contains metrics and techniques developed by the researcher to mitigate bias and prevent undesired bias in the classification task. Audit AI[9] is another open-source library developed by the Data Scientist at Pymetrics[10] and has almost the same utilities of AIF360, as the Microsoft open-source library, fairlearn [15]. Aequitas is an open-source library developed by the University of Chicago in 2018 to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive tools [152]. Amazon SageMaker Clarify [96] developed by Amazon Web Service is a feature for bias detection and model explanation integrated into Amazon SageMaker, a fully managed service for build, train, and deploy ML models at any scale.

[9] https://github.com/pymetrics/audit-ai

[10] https://www.pymetrics.com/

Chapter 4

Fostering Counterfactual Reasoning for Auditing Fairness: the Showcase

To the best of our knowledge, and quite unexpectedly, the idea of adopting counterfactual reasoning along with learning a classifier on sensitive features for discovering biases is unexplored in the financial domain literature. Furthermore, given the regulator's intervention, the concept of *fairness under unawareness* has assumed a crucial role in financial decision-making systems. However, the research on detecting bias for models trained in a fairness-under-unawareness setting is still in a very early stage. The experimental setup adopted in this investigation rigorously follows the best practices proposed in the recent literature and complies with the regulations. Nevertheless, the study shows that removing sensitive features from a decision support system does not guarantee a fair outcome. Concerning existing state-of-the-art approaches, the analysis tackles the fairness theme in the financial domain and proposes a general approach to identify implicit bias in a decision support system. Finally, instead of leveraging Counterfactual Reasoning to explain outcomes, the approach exploits the causal link between the counter-facts and the prediction to reveal the otherwise unnoticed bias.

4.1 Social, Theoretical, and Practical Implications on Information Access Systems

The UN Agenda 2030 for Sustainable Development sets out 17 Sustainable Development Goals, which are part of a broader program of actions consisting of 169 associated targets to be achieved in the environmental, economic, social, and institutional domains by 2030. Among them, there are "*gender equality*", "*reducing inequalities*", and "*responsible* *consumption and production*" –i.e., goals 5, 10, and 12, respectively. As a consequence, current and impending regulations affecting high social impact tasks will comply with the UN Agenda 2030. Among the others, the financial sector is a high-risk domain, as unethical use of AI can have significant repercussions from a social point of view, such as, for instance, discriminatory access to credit.

Several works attempted to tackle the fairness problem or provide model explainability for tasks ranging from classification to loan recommendation [31, 61, 62, 66]. The "*Fairness Under Unawareness*" setting mitigated the discrimination. However, the evaluation and the quantification of bias in a situation of "*Fairness Under Unawareness*" are of worryingly little interest to researchers.

The investigation at hand proposes a theoretical approach to identify the existence of bias even when sensitive information is not exploited in the training of the machine learning model. The proposed approach is general enough to neglect what kind of classifier is adopted under the hood and could be used in any classification task. The whole approach could be practically very useful for any practitioner since it could be used as a black box that measures and returns several pieces of information regarding the potential bias. Finally, the approach is designed to be a support tool for several kinds of Information Access Systems. The prominent potential application of the proposed approach is in Conversational-Agent systems that rely on lending recommendations (e.g., peer-to-peer lending) in which social bias may imply different access to credit. In that setting, the proposed system sheds light not only on the features that are necessary to reverse the decision but also on the potential biases of the decision maker. More generally, every Information Access System exploiting machine learning models that imply life-changing decisions can use our methodology to assess the bias in the models.

4.2 Preliminaries

This section introduces some useful notation that is extensively used in the rest of the dissertation. To ease the reading and for a rapid understanding, the definition of protected groups has some commonalities with Chen et al. [32], while some other aspects necessarily diverge from it due to the different nature of the study. It is important to note that the following notation diverges from that used in Section 3.4.1. While Section 3.4.1 employs population-level statistical criteria to assess fairness across entire datasets, the following notation transitions to a sample-wise notation. Despite the transition, the mathematical background remains consistent between the two sections.

This shift enables the application of counterfactual reasoning and fairness evaluation at an individual level, providing a more detailed understanding of model behaviour and fairness for specific instances rather than the broader population. Throughout the remainder of this dissertation, the notation from Section3.4.1 will be used when presenting aggregation or statistical-based results, while the sample-wise notation introduced here will be used to express results based on individual data points.

Table 4.1 List of the main notational conventions used in this document.

Notation	Description
$\mathbf{x}$	a vector of values for non-sensitive features $\mathbf{x} =< x_1, x_2, ..., x_n >$.
$\mathbf{s}$	a vector of binary values for sensitive features $\mathbf{s} =< s_1, s_2, ..., s_l >$. When no confusion arises, s is reported instead of s_i
y	a binary class value from the target domain for a single data point, with $y \in \{0,1\}$
$\mathbf{p}$	a vector of values for proxy features, i.e., a subvector of $\mathbf{x}$, with $h(\cdot)$ being an unknown function s.t. $h(\mathbf{p}) = s_i$
$\hat{y}$	a binary class prediction value from the target domain for a single data point, with $\hat{y} \in \{0,1\}$
$\hat{s}_i$	a binary prediction value of the i-th sensitive feature, with $\hat{s}_i \in \{0,1\}$
$f(\mathbf{x}) = \hat{y}$	a binary classification function of the target variable y
$f_s(\mathbf{x}) = \hat{s}_i$	a binary classification function of the sensitive variable s_i
$g(\mathbf{x}) = \mathcal{C}_{\mathbf{x}}$	a function that, given a data point $\mathbf{x}$, returns k counterfacts.
$\mathcal{X}^-$	set of samples negatively predicted by the decision maker and correctly predicted by the sensitive features classifier (i.e., $\mathbf{x} \vert f(\mathbf{x}) = 0 \wedge f_{s_i} = s_i$).
$\mathbf{c_x} \in \mathcal{C}_{\mathbf{x}}$	a counterfact of $\mathbf{x}$. $\mathbf{c_x}$ is a vector $\mathbf{c_x} =< c_{x_1}, c_{x_2}, ..., c_{x_n} >= \mathbf{x} \pm \epsilon$, with ϵ being a perturbation such that $f(\mathbf{c_x}) = 1 - f(\mathbf{x}) = 1 - \hat{y}$.

In the following, we assume the dataset $\mathcal{D}$ is an m-dimensional space containing n non-sensitive features, l sensitive features, and a target attribute. In other words, we have $\mathcal{D} \subseteq \mathbb{R}^m$, with $m = n + l + 1$.[1] A data point $d \in \mathcal{D}$ is then represented as $d = \langle \mathbf{x}, \mathbf{s}, y \rangle$, the concatenation of a vector $\mathbf{x}$ containing values of non-sensitive features and a vector $\mathbf{s}$ containing values for sensitive features.

Non-sensitive Features: We use $\mathbf{x} =< x_1, x_2, ..., x_n >$ to represent a vector of values for non-sensitive features defined as $\mathbf{x} \in X = \mathbb{R}^n \subset \mathbb{R}^m$. The value of x_i, with $1 \leq i \leq n$, can be categorical (set of discrete values) or numerical (set of continuous values).

[1] Without loss of generality, we assume that categorical features can always be transformed into features in $\mathbb{R}$ via one-hot-encoding.

Sensitive Features: We use $\mathbf{s} =< s_1, s_2, ..., s_l >$ to represent a vector of values for sensitive features in $dom(\mathcal{D})$. When no confusion arises, s is reported instead of s_i. Without loss of generality, we assume the value of s_i, with $1 \leq i \leq l$, as binary, i.e., $s_i \in \{0,1\}$. Based on the value of s_i, the advantaged group is referred to as *privileged* and associated with $s_i = 1$; the disadvantaged group is referred to as *unprivileged* and associated with $s_i = 0$.

Target Labels: Given a target feature $y \in \{0,1\}$, we use y^* to represent the positive outcome $y = 1$ (the negative outcome is associated to $y = 0$).

Proxy Features: Let $\mathbf{p} \subseteq \mathbf{x}$ be a subset of $\mathbf{x}$, and $h(\cdot)$ be a function that can maps the relation $h : \mathbf{p} \mapsto s_i$ such that $h(\mathbf{p}) = s_i$, i.e., the value returned by h applied to the values associated to the features in $\mathbf{p}$ is equal to the values associated to s_i. We say that $\mathbf{p}$ is a set of proxy features for the sensitive feature s_i.
In practical terms, if we knew $h(\cdot)$, a set of proxy features could be used to predict a certain sensitive feature.

Outcome prediction: Let $\hat{y} \in \{0,1\}$ be the prediction for a given data point. The notation $\hat{y} = 1$ denotes a *favorable* prediction (e.g., loan application approved), while $\hat{y} = 0$ an *unfavorable* one (e.g., loan application rejected). Let $f(\cdot)$ be a function such that $f(\mathbf{x}) = \hat{y}$.

Sensitive Feature Prediction: Let $\hat{s}_i \in \{0,1\}$ be the prediction of the i-th sensitive feature. The notation $\hat{s}_i = 1$ denotes the prediction to belong to a *privileged* group, while $\hat{s}_i = 0$ denotes the prediction to belong to an *unprivileged* group.

Let $f_s(\cdot)$ be a function able to predict the value of a sensitive feature given the value of non-sensitive ones, i.e., $f_s(\mathbf{x}) = \hat{s}_i$. Since the set of proxy features $\mathbf{p}$ is unknown, we can use $f_s(\cdot)$ to predict the value of s_i.

Negatively-predicted samples: Our work is focused on samples negatively predicted by the *Decision Maker* (i.e., $\forall d \in \mathcal{D}$ s.t. $f(\mathbf{x}) = 0$) and correctly predicted by the *Sensitive-Feature Classifier* (i.e.,$\forall d \in \mathcal{D}$ s.t. $f_s(\mathbf{x}) = s$). For simplicity, we denote the set of such samples with $\mathcal{X}^-$, with $\mathcal{X}^- \subseteq \mathcal{D}$. For clarity, this set depends on the $f(\cdot)$ used to predict the sample and varies for each Decision Maker taken into account.

Counterfactual samples: Given a vector $\mathbf{x}$ and a perturbation ϵ, we say that a vector $\mathbf{c_x} =< c_{x_1}, c_{x_2}, ..., c_{x_n} >= \mathbf{x} \pm \epsilon$ is a counterfactual of $\mathbf{x}$ if $f(\mathbf{c_x}) = 1 - f(\mathbf{x}) = 1 - \hat{y}$. We use the set $\mathcal{C}_\mathbf{x} \in (\mathbb{R}^m)^k$, with $|\mathcal{C}_\mathbf{x}| = k$, to denote the set of possible

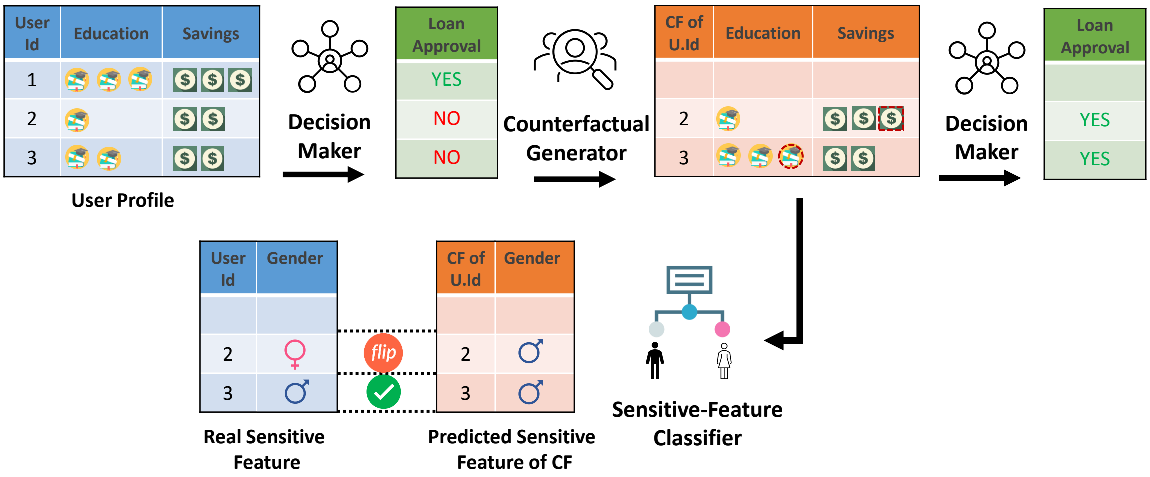


Fig. 4.1 An example of a loan-approval decision process analysed through our model. From left to right, we have the decision made on the original user profiles, the counterfactual generation for users with loan denied, the sensitive-feature classification of the original profile, and counterfactual profiles with the decision changed. For user 2, both counterfactual profiles change the sensitive feature category.

counterfactuals for $\mathbf{x}$. A function $g(\mathbf{x})$ that is used to compute k counterfactuals for $\mathbf{x}$ such that $g(\mathbf{x}) = \mathcal{C}_{\mathbf{x}}$.

Our investigation focuses on unfavourable outcome predictions. Consequently, all the generated counterfactuals are associated with a favourable $f(\mathbf{c}_{\mathbf{x}}) = 1$. When no confusion arises, $\mathbf{c}$ and $\mathcal{C}$ are reported instead of $\mathbf{c}_{\mathbf{x}}$ and $\mathcal{C}_{\mathbf{x}}$, respectively. For simplicity, we denote $f(\cdot)$, $f_{s_i}(\cdot)$, and $g(\mathbf{x})$ as the **Decision Maker**, the **Sensitive-Feature Classifier**, and the **Counterfactual Generator** respectively.

4.3 Methodology

The *fairness under unawareness* setting (see Section 3.4.5) poses several challenges to the identification of discriminatory behaviours performed by intelligent systems. Proxy traits can be non-linearly associated with sensitive ones, making typical statistical procedures ineffective. Figure 4.1 depicts the principal components of our model, namely the *Decision-Maker*, the *Counterfactual Generator*, and the *Sensitive-Feature Classifier*, as well as the flow of our pipeline.

Following, we are introducing the key component of our methodology.

4.3.1 Decision-Maker

The *decision-maker* is the key component of the decision support system. Even though the nature of the decisions can be heterogeneous, the decision-maker implements a machine-learning algorithm trained using past human decisions. Although it does not use sensitive features in the learning phase, we assume the predictive model is not necessarily bias-free, thanks to current regulations. This phenomenon could be due to proxy features.

To keep the approach as general as possible, to implement the *decision-maker*, we have chosen four largely adopted approaches to tackle the classification task. As far as possible, we avoided domain-specific models, preprocessing steps, and operations, and we relied on the general best practices that apply to a broader set of machine learning domains.

Our choice was to sacrifice a small quantity of accuracy (even though the performance remains highly competitive) to gain the generality of the approach. In detail, we opted for Logistic Regression (LR), Decision Tree (DT), Support-Vector Machines (SVM), XGBOOST[2] (XGB), LightGBM[3] (LGBM), Random Forest (RF), and Multi-Layer Perceptron[4] (MLP). LR is a linear statistical model that predicts the probability of one event taking place through a linear combination of independent variables. SVM is a pattern classification technique aiming to minimise an upper bound of the generalisation error by maximising the margin between the separating hyperplane and data instances [18]. DT is a tree-like structure used for classification and regression tasks. It splits the data into subsets based on the values of different features, making decisions at each node until it reaches leaf nodes that provide the final predictions. RF is an ensemble learning method that combines multiple decision trees to make predictions. It builds a forest of decision trees, each trained on a different subset of the data and with a random selection of features. The final prediction is made by aggregating the predictions of individual trees, often using a majority vote or averaging. The MLP is a type of artificial neural network with multiple layers of interconnected neurons (perceptrons). It's used for a wide range of tasks, including image recognition, natural language processing, and regression. An MLP typically consists of an input layer, one or more hidden layers, and an output layer. It's capable of modelling complex relationships in data, and training involves backpropagation and

[2] XGB: https://github.com/dmlc/xgboost

[3] LGBM: https://github.com/microsoft/LightGBM

[4] LR, DT, SVM, RF, MLP: https://scikit-learn.org/

gradient descent. We exploited LR, SVM, DT, RF, and MLP's Scikit-learn[5] implementation. XGB stands for eXtreme Gradient Boosting, and it implements gradient boosting machines guaranteeing high computational speed and performance. XGB learns both classification and regression models employing gradient-boosted decision trees. LGBM stands for Light Gradient Boosting Machine and uses an approach similar to XGB, thus favouring speed to robustness. Since the two approaches are state-of-the-art solutions yielding the best results in many competitions, we considered them despite their similarity.

Debiased Decision-Makers

To evaluate whether debiasing algorithms can reduce discriminatory behaviour even in a "*fairness under unawareness*" setting, we also considered *decision-makers* that exploit debiasing approaches. The overall system is the same as the one depicted in Figure 4.1. This variation aims to assess whether debiasing models guarantee fair behavior, and counterfactual reasoning can help discover discrimination even when these models are chosen as decision-makers. The debiasing algorithms we chose to investigate are Linear Fair Empirical Risk Minimization (LFERM) [6] [72], Adversarial Debiasing (Adv) [7] [175], and Fair Classification (FairC)[8] [172].

- **Adversarial Debiasing:** Zhang et al. [175] propose an adversary framework for debiasing algorithms (AdvDeb). The model comprises two elements: a target predictor and an adversary. The target label predictor consists of a Deep Neural Network that, given a general input $\mathbf{x}$, tries to predict the target label y. The adversary is a simple Neural Network that, fed by the predicted output of the DNN $\hat{y}$, tries to predict the sensitive label s. The DNN and the Adversary Network (AN) are trained to optimise both their model weights, W (for DNN) and U (for AN), by minimising the losses $L_P(\hat{y}, y)$ and $L_A(\hat{s}, s)$, respectively. $L_P(\hat{y}, y)$ is the target discrimination loss of the classification task, typically a cross-entropy loss. $L_A(\hat{s}, s)$ is the loss the adversary aims to maximise to predict the sensitive label. To ease the understanding of the adversarial learning process, $L_A(\hat{s}, s)$ is herein used with an opposite sign concerning the original paper, in

[5] https://scikit-learn.org/

[6] LFERM: https://github.com/jmikko/fair_ERM

[7] Adv: https://github.com/Trusted-AI/AIF360

[8] FairC: https://github.com/mbilalzafar/fair-classification

which the adversary aims to minimize $L_A(\hat{s}, s)$.

$$\underbrace{\arg\min_{W} L_P(\hat{y}, y) \; - \; \underbrace{[\; \underset{U}{\mathrm{argmax}}\; proj_{\nabla_W L_A(\hat{s},s)} L_P(\hat{y}, y) + \alpha L_A(\hat{s}, s) \;]}_{\text{best-case loss } L_A \,=\, \textbf{optimal prediction of the sensitive feature}}}_{\textbf{robust classification} \text{ against the prediction of the sensitive feature}} \tag{4.1}$$

The overall learning process resembles a min-max game in which the discriminator tries to minimise the loss of the predictor while the adversary tries to maximise its utility (see Equation 4.1). The middle term (i.e., $proj_{\nabla_W L_A(\hat{s},s)}$) limits the predictor from moving in a direction that promotes the adversary's loss reduction. For reproducibility, we adopt the IBM implementation available in the AIF360[9] framework.

- **Linear Fair Empirical Risk Minimization:** Donini et al. [72] propose a method that applies a fairness constraint to the loss function of an SVM classifier. In detail, they constrain the Hinge-loss to respect the "Equality of Opportunity" condition. The underlying goal is to remove the discrepancy between the false-negative rates of the privileged and unprivileged groups. The fairness condition is implemented by imposing an orthogonality constraint directly on the sample. Specifically, the sample vector is required to be orthogonal to the vector formed by the difference between the barycenters of the positive input samples in the two groups. Let $\mathbf{u} = u_{priv} - u_{unpriv}$ be the difference between the two barycenter vectors of the privileged and unprivileged groups, respectively, and let $|u_i|$ be the maximum valued feature in the vector, and x be a sample in the original space. The fairness-constrained representation $\tilde{x}$ is then calculated as follows:

$$\tilde{x}_j = x_j - x_i \frac{u_j}{u_i}, \quad j \in \{1, \ldots, i-1, i+1, \ldots, d\} \tag{4.2}$$

with d being the number of features. In this study, to ensure the reproducibility of the results, the implementation provided by the authors[10] is used. Specifically, the reader can refer to the linear implementation of Fair SVM, named *linear fair empirical risk minimisation* (LFERM) therein.

- **Fairness Classification:** Zafar et al. [172] propose incorporating fairness constraints into the model optimization process. These constraints are mathematical

[9] https://github.com/Trusted-AI/AIF360

[10] https://github.com/jmikko/fair_ERM

expressions that specify fairness requirements that the model should satisfy. For example, fairness constraints might ensure that the false positive and false negative rates for different groups (e.g., different demographic groups) are approximately equal. Specifically, the authors introduce a flexible mechanism quantifying the relation between the classifier's decision boundary with the sensitive attributes. Indeed, their (un)fairness measures is defined as the covariance between the users' sensitive attributes, $\{\mathbf{s}_i\}_{i=1}^{N}$, and the signed distance from the users' feature vectors to the decision boundary, $\{d_\theta(\mathbf{x}_i)\}_{i=1}^{N}$, i.e.:

$$\begin{aligned} \operatorname{Cov}(\mathbf{s}, d_{\boldsymbol{\theta}}(\mathbf{x})) &= \mathbb{E}\left[(\mathbf{s}-\overline{\mathbf{s}})d_{\boldsymbol{\theta}}(\mathbf{x})\right] - \mathbb{E}[(\mathbf{s}-\overline{\mathbf{s}})]\bar{d}_{\boldsymbol{\theta}}(\mathbf{x}) \\ &\approx \frac{1}{N}\sum_{i=1}^{N}(\mathbf{s}_i-\overline{\mathbf{s}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i), \end{aligned} \tag{4.3}$$

where $\mathbb{E}[(\mathbf{s}-\overline{\mathbf{s}})]\bar{d}_\theta(\mathbf{x})$ cancels out since $\mathbb{E}[(\mathbf{s}-\overline{\mathbf{s}})] = 0$. Since in linear models for classification, such as logistic regression or linear SVMs, the decision boundary is simply the hyperplane defined by $\boldsymbol{\theta}^T\mathbf{x} = 0$, Equation 4.3 reduces to $\frac{1}{N}\sum_{i=1}^{N}(\mathbf{s}_i-\overline{\mathbf{s}})\,\boldsymbol{\theta}^T\mathbf{x}_i$. To this end, the authors find the optimal θ parameters by minimising the corresponding loss function over the training set under the previous fairness constraints, i.e.:

$$\begin{array}{rl} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N}(\mathbf{s}_i-\overline{\mathbf{s}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i) \leq \mathbf{c} \\ & \frac{1}{N}\sum_{i=1}^{N}(\mathbf{s}_i-\overline{\mathbf{s}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq -\mathbf{c} \end{array} \tag{4.4}$$

where $\mathbf{c}$ is the covariance threshold, which specifies an upper bound on the covariance between each sensitive attribute and the signed distance from the feature vectors to the decision boundary, trading off fairness and accuracy.

4.3.2 Counterfactual Generator

This study leverages the counterfactual reasoning approach to explore the decision-maker boundary in the feature space. Thanks to the sample generation process, this strategy can ease the analysis of the decision boundary even though the decision-maker is a black-box model. Moreover, the proposed model is utterly agnostic about the algorithm chosen as the decision-maker.

The input of the counterfactual generator is the same sample previously evaluated by the decision-maker. When the system takes a decision adverse to the user (e.g., loan request rejected, income under a given threshold), the counterfactual generator is called in, and it produces new samples that would lead to a favourable outcome, as we discussed in Section 4.2. Under the hood, it modifies user characteristics following various strategies (e.g., increasing savings or changing education level). Each generated counterfactual feeds the decision-maker, and all the counterfactuals that switch the decision outcome, e.g., granting the loan, constitute the input of the next module of the system. For the sake of reproducibility and reliability, the counterfactuals are generated with an external counterfactual framework. We opted for DiCE [129], an open-source framework developed by Microsoft [11]. Mothilal et al. [128] built their framework to satisfy two fundamental requirements. The generated counterfactuals should be (1) plausible and associated with actions that could be actionable by users and (2) diverse from each other. Both requirements fit the goals of our work. The first ensures that generated counterfactuals are close to the original sample and thus realistic. The second guarantees that they are all different, thus suggesting various strategies to solve the problem. The diversity requirement is fulfilled thanks to determinantal point processes (DPP), commonly used in selection problems with diversity constraints [111].

For the sake of completeness, we briefly introduce the DiCE counterfactual generation process using the notation adopted in this study. Let $\mathbf{x}$ be a candidate sample, $\mathcal{C}_{\mathbf{x}} = \{\mathbf{c}_{\mathbf{x}}^1, \mathbf{c}_{\mathbf{x}}^2, \ldots, \mathbf{c}_{\mathbf{x}}^k\}$ be a set of k candidate counterfactual samples, with k being the desired number of counterfactuals, and $f(\cdot)$ being a predictor function, i.e., a machine learning model. The optimisation function of the module that generates counterfactual samples is then the following:

$$g(\mathbf{x}) = \underset{\mathbf{c}_{\mathbf{x}}^1,\ldots,\mathbf{c}_{\mathbf{x}}^k}{\arg\min} \frac{1}{k}\sum_{i=1}^{k} yloss(f(\mathbf{c}_{\mathbf{x}}^i), y^*) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(\mathbf{c}_{\mathbf{x}}^i, \mathbf{x}) - \lambda_2 dppd(\mathbf{c}_{\mathbf{x}}^1, \ldots, \mathbf{c}_{\mathbf{x}}^k) \tag{4.5}$$

where $yloss(.)$ is a metric (e.g., ℓ_1-loss, ℓ_2-loss, or hinge-loss) minimising the distance between the predicted output of $\mathbf{c}_{\mathbf{x}}^i$ and the desired y^*; *dist* is a proximity function that quantifies the distance between $\mathbf{c}_{\mathbf{x}}^i$ and $\mathbf{x}$; $dppd(\cdot)$ is the *determinantal point processes diversity*, i.e., the determinant of the kernel matrix of the inverse distance between counterfactuals. More formally:

$$dppd = det(\mathbf{K}), \quad \text{with } \mathbf{K_{i,j}} = \frac{1}{1 + dist(\mathbf{c}_{\mathbf{x}}^i, \mathbf{c}_{\mathbf{x}}^j)} \tag{4.6}$$

[11] https://github.com/interpretml/DiCE

where *dist* in the previous equation denotes a generic distance metric between counterfactuals. Finally, λ_1 and λ_2 are hyperparameters that balance the contribution of the distance and the diversity part, respectively.

DiCE offers several strategies for generating candidate counterfactual samples. We decided to use two different approaches, i.e., Genetic and KDtree generation. The choice of these strategies allows (i) to assess whether it is possible to generate a large enough number of counterfactuals from a sample, (ii) to investigate which strategy is most effective for our purposes, and (iii) to find the most robust and valid method in generating plausible counterfactuals. The DiCE framework also offers a third strategy, i.e. Random, that we decided not to use at this time due to not effectively generating counterfactual samples [53]. The Random strategy randomly selects a set of features to perturb and replace the original sample. The perturbation goes ahead until the counterfactual satisfies the requirement $f(\mathbf{c_x}) = y^*$. The KDtree strategy computes a tree-based distance between all the dataset samples; it chooses the samples that are close to the original one and switches the outcome prediction to y^*. The Genetic strategy can start with a Random initialisation or a KDtree initialisation and then iterates by generating new samples close to the original one that switches the outcome prediction to y^*.

4.3.3 Sensitive-Feature Classifier

The *sensitive-feature classifier* performs a classification of the sample generated by the *counterfactual generator* (that caused a decision flip) into one of the sensitive categories. This component plays a crucial role in our methodology since it allows the system to discover hidden discriminatory models. For each sensitive feature (e.g., *gender*, *race*, *age*, etc.), a classifier is thus learned. In Figure 4.1, the counterfactual sample that caused the flip becomes the input of the sensitive-feature classifier. If the sensitive-feature classification predicts a category different from the one initially (i.e., before generating counterfactuals) associated with the sample (e.g., from female to male), a bias in the decision-making process could occur. In fact, a change in the sensitive-feature classification means that there are some non-sensitive features (whose values have been changed by the counterfactual generator) that allow the system to recognize the counterfactual sample as belonging to the *privileged* class (i.e., male). Hence, the *sensitive-feature classifier* gives us an indication of the existence of a function that links non-sensitive features to sensitive ones, namely a proxy feature. We exploited RF, MLP, and XGB for implementing this component due to their capability to learn non-linear dependencies.

4.4 The Model at Work

As a relevant case study, we refer to the financial domain, considering the tasks of predicting loan-repayment default. However, focusing on this specific domain does not compromise the model generality.

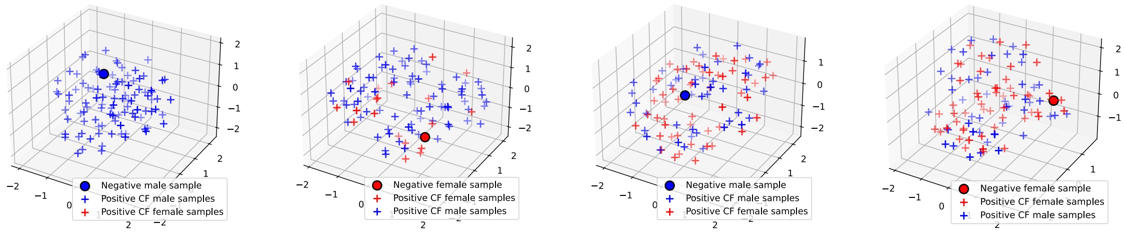
Loan Request Example

Let us imagine that some users with certain characteristics apply for a loan (see Figure 4.1). The Decision Maker analyzes their requests and computes a positive or negative decision. If a request is denied (e.g., for users 2 and 3), the counterfactual generator starts to produce a series of counterfactuals until it gets a positive decision ($\hat{y} = 1$, loan accepted). In the example, only two counterfactuals for users 2 and 3 are generated. Once the decision has been changed, the Sensitive-Feature Classifier analyzes the original characteristics of users' 2 and 3 profiles and the newly generated counterfactuals to assess how many counterfactuals changed the sensitive-feature gender. User 2 was originally classified as female and, then, as male for both counterfactuals (profile with counterfactual changes for getting loan approval). For user 3, this does not happen, the gender classification is the same before and after the counterfactual changes (loan approved).

Excluding sensitive features makes verifying that all users are treated equally incredibly challenging. However, counterfactual reasoning can be an effective tool to propose actionable steps for reaching a positive outcome.

In a nutshell, our process pipeline, described in Section 4.4, is as follows: the **Decision Maker** makes decisions without exploiting sensitive features, then if the outcome is negative (e.g. loan rejected), the **Counterfactual Generator** is exploited to propose modifications to user characteristics and request for reaching a positive outcome (e.g. loan approved). For each data point d with a negative prediction $f(\mathbf{x}) = 0$, we generate a set of counterfactual samples $\mathcal{C}_{\mathbf{x}}$ that reach a positive outcome (i.e., $\forall \mathbf{c}_{\mathbf{x}} \in \mathcal{C}_{\mathbf{x}}$ s.t. $f(\mathbf{c}_{\mathbf{x}}) = 1$). Afterward, each counterfactual (CF) sample is evaluated by the **Sensitive-Feature Classifier** that predicts the value of the (omitted) sensitive feature for the given CF sample. If the CF sample is classified as e.g. male (privileged group), while the original sample was e.g. female (unprivileged group), the decision model could be biased and its unfairness can be quantified (Equations 4.15 and 4.17).

Indeed, each CF sample derives from the original sample $\mathbf{x}$ plus a perturbation ϵ, where $|\epsilon|$ is the *distance* from the original sample for getting a positive outcome, and it should be independent from the user-sensitive characteristics.



(a) male on Classic ML model (b) female on Classic ML model (c) male on Debiasing model (d) female on Debiasing model

Fig. 4.2 Adult t-SNE visualizations of a random male (a-c) and female (b-d) sample with a negative outcome and their CF samples with a positive outcome, respectively, for a Classic ML model (i.e. XGB) and a Debiasing model (i.e. Adversarial Debiasing).

Figure 5.1 depicts a scenario in which *male* (blu color) is the privileged category, and *female* (red color) is the unprivileged one. For each subfigure, a sample with an unfavorable decision and its corresponding CFs are depicted. A classic ML model (i.e., XGB) is compared with a debiasing ML model (i.e., AdvDeb). We can observe that for the male sample and classic ML model (Figure 5.1(a)), the CF samples belong to the same sensitive category (i.e., male). For the female sample (Figure 5.1 (b)), this is not true, revealing a bias of the model. Conversely, the debiasing model (Figure 5.1 (c) and (d)) shows no predominance in the generated counterfactuals of one value of the sensitive class. However, a change of the outcome, e.g. from negative to positive, should not be determined by a flip of the value(s) of the sensitive feature(s).

4.5 Counterfactual Fair Flip and Opportunity

To define a discrimination score of a given decision model, we propose a metric that we call *Counterfactual Flips* providing a snapshot of the discriminatory behavior the model might put in place.

Definition 1 (*Counterfactual Flips*)**.** *Given a sample* $\mathbf{x}$ *belonging to a demographic group* s *whose model output is denoted as* $f(\mathbf{x})$*, generated a set* $\mathcal{C}_{\mathbf{x}}$ *of* k *counterfactuals with a desired* y^* *outcome* $f(\mathbf{c}^i_{\mathbf{x}}) = y^* \quad \forall \mathbf{c}^i_{\mathbf{x}} \in \mathcal{C}_{\mathbf{x}}$*, the Counterfactual Flip indicates the percentage of counterfactual samples belonging to another demographic group (i.e.,*

$f_{S_i}(\mathbf{c}_\mathbf{x}^i) \neq f_{S_i}(\mathbf{x})$, with $f_{S_i}(\mathbf{x}) = s$).

$$\text{CFlips}(\mathbf{x}, \mathcal{C}_\mathbf{x}, f_{S_i} = \frac{\sum_{i=1}^{k}(\mathbb{1}(\mathbf{c}_\mathbf{x}^i))}{k} \tag{4.7}$$

where the function $\mathbb{1}(\mathbf{c}_\mathbf{x}^i)$ correspond to:

$$\mathbb{1}(\mathbf{c}_\mathbf{x}^i) = \begin{cases} 1 & \textit{if } f_{S_i}(\mathbf{c}_\mathbf{x}^i) \neq f_{S_i}(\mathbf{x}) \\ 0 & \textit{if } f_{S_i}(\mathbf{c}_\mathbf{x}^i) = f_{S_i}(\mathbf{x}) \end{cases} \tag{4.8}$$

and it measures if a counterfactual sample is linked with a Flip, i.e. a change, for the sensitive characteristic [12].

The bigger the CFlips value is, the stronger the biases and the discrimination the model suffers from. In Example 4.4, $CFlips_i = 1$ for user 2, and $CFlips_i = 0$ for user 3, thus the sensitive classification changed 2 out of 2 CF samples for user 2 and 0 out of 2 CF samples for user 3, unveiling implicit bias in the *Decision Maker* in favour of male characteristics. All the following metrics assume that the sensitive feature classifier make perfect predictions. Therefore, we can introduce the proposed *Counterfactual Fair Flips* desiderata.

Definition 2 (Counterfactual Fair Flips). *-A binary classifier shows Counterfactual Fair Flips if the probability of generating Counterfactual samples belonging to a different demographic group (privileged vs unprivileged) is the same:*

$$\mathbb{P}(f_{S_i}(\mathcal{C}_{\mathcal{D}|_{S_i=0}}) \neq S_i \mid f(\mathcal{C}_{\mathcal{D}|_{S_i=0}}), S_i = 0) = \mathbb{P}(f_{S_i}(\mathcal{C}_{\mathcal{D}|_{S_i=1}}) \neq S_i \mid f(\mathcal{C}_{\mathcal{D}|_{S_i=1}}), S_i = 1) \tag{4.9}$$

which implies the complement*:*

$$\mathbb{P}(f_{S_i}(\mathcal{C}_{\mathcal{D}|_{S_i=0}}) = S_i \mid f(\mathcal{C}_{\mathcal{D}|_{S_i=0}}), S_i = 0) = \mathbb{P}(f_{S_i}(\mathcal{C}_{\mathcal{D}|_{S_i=1}}) = S_i \mid f(\mathcal{C}_{\mathcal{D}|_{S_i=1}}), S_i = 1) \tag{4.10}$$

where $\mathcal{C}_{\mathcal{D}|_{S_i}}$ refers to all counterfactual samples generated for all dataset sample belonging to specific S_i.

In our work, we only take into account samples negatively predicted by the *Decision Maker* (i.e., $f(\mathbf{x}) = 0$), that we denote as $\mathcal{X}^- \subseteq \mathcal{D}$, as we are interested in quantifying the discrimination of the minority group (e.g. women) in the process to achieving a

[12] Without loss of generality $\mathbb{1}(\mathbf{c}_\mathbf{x}^i)$ correspond to $\mathbb{1}(f_{S_i}(\mathbf{c}_\mathbf{x}^i) = f_{S_i}(\mathbf{x}))$ as more compact notation.

positive counterfactual result (i.e., $f(\mathbf{c_x}) = 1 \wedge f_{S_i}(\mathbf{c_x}) \neq s$). Thus, we want to restrict the Definition 2 to only the positively predicted counterfactual samples that belong to the set $\mathcal{X}^-$. We introduce now the "*Counterfactual Fair Opportunity*" notion.

Definition 3 (Counterfactual Fair Opportunity)**.** *"A decision model is fair if the counterfactual samples of individuals with unfavourable decisions (i.e., $\mathcal{X}^-$), to reach a positive outcome (i.e., $f(\mathcal{C}_{\mathcal{X}^-}) = 1$), maintain the same sensitive behaviour. This behavior must be guaranteed for the privileged and for the unprivileged group."*

$$\mathbb{P}(f_{S_i}(\mathcal{C}_{\mathcal{X}|^-_{S_i=0}}) \neq S_i \mid f(\mathcal{C}_{\mathcal{X}|^-_{S_i=0}}) = 1, \mathcal{X}|^-_{S_i=0}) = \mathbb{P}(f_{S_i}(\mathcal{C}_{\mathcal{X}|^-_{S_i=1}}) \neq S_i \mid f(\mathcal{C}_{\mathcal{X}|^-_{S_i=1}}) = 1, \mathcal{X}|^-_{S_i=1}) \tag{4.11}$$

which implies the complement:

$$\mathbb{P}(f_{S_i}(\mathcal{C}_{\mathcal{X}|^-_{S_i=0}}) = S_i \mid f(\mathcal{C}_{\mathcal{X}|^-_{S_i=0}}) = 1, \mathcal{X}|^-_{S_i=0}) = \mathbb{P}(f_{S_i}(\mathcal{C}_{\mathcal{X}|^-_{S_i=1}}) = S_i \mid f(\mathcal{C}_{\mathcal{X}|^-_{S_i=1}}) = 1, \mathcal{X}|^-_{S_i=1}) \tag{4.12}$$

Definition 3 works in a context where counterfactual samples are used for suggesting actionable steps to achieve the desired positive outcome that corresponds to an opportunity (e.g., to achieve the loan). However, the outcome behaviour should not depend on a specific sensitive group. Thus, the degree of sensitive (un)fidelity of counterfactual samples must be equal between the two sensitive groups.

The bigger the CFlips value is, the stronger the discriminatory bias the model suffers from. From a probabilistic perspective, the CFlips can be considered as the probability of Counterfactual, generated to reach an opposite outcome from the original sample, to be predicted by the sensitive feature classifier as opposite to the original sample (see Equation 4.13).

$$\mathbb{P}(f_{S_i}(\mathcal{C}_\mathbf{x}) \neq S_i \mid f(\mathcal{C}_\mathbf{x}) = 1 - f(\mathbf{x}), f(\mathbf{x}), s) \quad \text{with } S_i \in \{0,1\} \tag{4.13}$$

For the set of samples $\mathcal{X}^-$, the metric in Equation 4.7 can be generalized to the *privileged* and *unprivileged* group (Equation 4.14 is restricted to the *privileged* samples negatively predicted, while Equation 4.15 is restricted to the *unprivileged* samples negatively predicted).

$$\boldsymbol{\mathit{Privileged}} \begin{cases} \text{CFlips}_{S_i=1} = \frac{\sum_i \text{CFlips}(\mathbf{x_i}, \mathcal{C}_{\mathbf{x_i}}, f_{S_i}}{|\mathcal{X}|^-_{S_i=1}|} \quad \text{with } \mathbf{x} \in \mathcal{X}|^-_{S_i=1} \end{cases} \tag{4.14}$$

$$\boldsymbol{\mathit{Unprivileged}} \begin{cases} \text{CFlips}_{S_i=0} = \frac{\sum_i \text{CFlips}(\mathbf{x_i}, \mathcal{C}_{\mathbf{x_i}}, f_{S_i}}{|\mathcal{X}|^-_{S_i=0}|} \quad \text{with } \mathbf{x} \in \mathcal{X}|^-_{S_i=0} \end{cases} \tag{4.15}$$

Definition 2 requires that the quantities of counterfactuals flipped in Equation (4.14) and Equation (4.15) must be equal. Thus, we are interested in the difference between the result of the two equations, i.e. ΔCFlips, being close to zero (see Equation 4.16).

$$\Delta\text{CFlips}_s = |\text{CFlips}_{S_i=1} - \text{CFlips}_{S_i=0}| \tag{4.16}$$

A limitation of the CFlips metric is that it does not measure the distance of each CF sample from the original data point. However, from an individual-fairness wise, a debated issue is the definition of a metric that considers that distance [77]. Accordingly, we propose a new metric that considers CFs ranked based on the Mean Absolute Deviation from the original sample and other criteria [128]. The insight behind this metric is that the more the CF is ranked high, the more its impact on the metric value. Thus, the metric penalises CFs ranked in the top positions for which the value of the sensitive feature is flipped. More formally:

Definition 4 (Discounted Cumulative Counterfactual Fairness)**.** *Given a set of Counterfactuals $\mathcal{C}_{\mathbf{x}}$ for a sample $\mathbf{x}_i$, the Discounted Cumulative Counterfactual Fairness* $\text{DCCF}_{\mathbf{x}_i}$ *measures the cumulative gain of the ranking of counterfactuals with respect to the sensitive group of the original sample:*

$$\text{DCCF}_{\mathbf{x}_i} \triangleq \sum_{p_j, \mathbf{c}^j_{\mathbf{x}_i} \in \mathcal{C}_{\mathbf{x}_i}} \frac{2^{(1-\mathbb{1}(c^j_{\mathbf{x}_i}))} - 1}{\log_2(p_j+1)} \tag{4.17}$$

where p_j is the rank of $\mathbf{c}^j_{\mathbf{x}_i}$ in $\mathcal{C}_{\mathbf{x}_i}$ and $\mathbb{1}(c^j_{\mathbf{x}_i})$ from Equation 4.7.

DCCF rewards the CF samples in the ranking that did not flip. If more CF samples belonging to the same sensitive group as the original data point are in a higher ranking position, the result will be a higher DCCF. Thereby, we can formulate the *Ideal Discounted Cumulative Counterfactual Fairness* (IDCCF) correspond to an ideal ranking in which each CF sample $\mathbf{c}_{\mathbf{x}_i}$ belongs to the same sensitive group as the starting sample $\mathbf{x}_i$ (Definition 4.5), and the *normalized* DCCF (nDCCF) (Definition 4.5).

Definition 5 (Ideal Discounted Cumulative Counterfactual Fairness)**.** *The Ideal Discounted Cumulative Counterfactual Fairness is the ideal ranking of the estimated sensitive information of counterfactuals with respect to the sensitive information of the original sample. In an ideal ranking, each counterfactual belongs to the same sensitive group of the original sample.*

$$\mathrm{IDCCF}_{x_i} \triangleq \sum_{p_j, \mathbf{c}^j_{\mathbf{x}_i} \in \mathcal{C}_{\mathbf{x}_i}} \frac{1}{\log_2(p_j+1)} \tag{4.18}$$

Definition 6 (normalized Discounted Cumulative Counterfactual Fairness)**.** *The normalized Discounted Cumulative Counterfactual Fairness (nDCCF) is the normalization of the current counterfactual rank DCCF with respect to the ideal rank IDCCF.*

$$\mathrm{nDCCF}_{x_i} \triangleq \frac{\mathrm{DCCF}_{x_i}}{\mathrm{IDCCF}_{x_i}} \tag{4.19}$$

In the same way as CFlips, given a set of samples $\mathcal{X}^- \subseteq \mathcal{D}$ predicted by the decision maker as negative, the metric in Equation 4.19 can be generalized to the *unprivileged* and *privileged* group (Equation 4.20 and 4.21).

$$\mathrm{nDCCF}_{S_i=0} \triangleq \frac{1}{|\mathcal{D}|_{S_i=0}|} \sum_{\mathbf{x}_i} \mathrm{nDCCF}_{\mathbf{x}_i} \quad \text{with } \mathbf{x}_i \in \mathcal{X}|^-_{S_i=0} \tag{4.20}$$

$$\mathrm{nDCCF}_{S_i=1} \triangleq \frac{1}{|\mathcal{D}|_{S_i=1}|} \sum_{\mathbf{x}_i} \mathrm{nDCCF}_{\mathbf{x}_i} \quad \text{with } \mathbf{x}_i \in \mathcal{X}|^-_{S_i=1} \tag{4.21}$$

For both CFlips and nDCCF, we are interested in the difference (i.e., ΔCFlips and ΔnDCCF), between the *privileged* and *unprivileged* groups, being close to zero. In fact, even though those metrics are individual-based computed they can serve as group-based fairness measures thanks to the Δ (see Equation 4.22).

$$\Delta\mathrm{nDCCF}_s = |\mathrm{nDCCF}_{S_i=1} - \mathrm{nDCCF}_{S_i=0}| \tag{4.22}$$

4.6 Algorithm

For reproducibility reasons, the framework adopted for the generation of the Counterfactual samples is DiCE (see Section 4.3.2) with two different strategies: Genetic and KDtree generation. We first train the models for the target label binary classification task $f(\cdot)$, i.e., the decision-maker. Analogously, we train the models for the sensitive classification task $f_s(\cdot)$, i.e., the sensitive-feature classifiers. The counterfactual module generates k counterfactuals for each original sample. Whether the sample is associated with a negative outcome (i.e., $f(\mathbf{x}) = 0$), it belongs to a privileged group (i.e., $s = 1$), and it is correctly predicted to belong to the same sensitive class (i.e., $f_s(\mathbf{x}) = 1$), then the sample and its counterfactuals are added to the set $\mathcal{A}$. Alternatively, if the sample is associated with a negative outcome (i.e., $f(\mathbf{x}) = 0$), it belongs to an unprivileged group (i.e., $s = 0$), and it is correctly predicted to belong to the same sensitive class (i.e., $f_s(\mathbf{x}) = 0$), then the sample and its counterfactuals are added to the set $\mathcal{B}$. The union of the sets $\mathcal{A}$ and $\mathcal{B}$ correspond to the previously mentioned negatively predicted set (i.e., $\mathcal{X}^-$). In detail, for each sample, a tuple of objects is stored, including (i) the original sample $\mathbf{x}$, (ii) the predicted target label $f(\mathbf{x})$, (iii) the sensitive feature of the sample as it is predicted by the dedicated classifier $f_s(\mathbf{x})$, (iv) the set of counterfactual samples $\mathcal{C}_\mathbf{x}$, (v) and the predictions of the sensitive labels performed on the counterfactuals $f_s(c_\mathbf{x})\ \forall\ c_\mathbf{x} \in \mathcal{C}_\mathbf{x}$. The process is summarized by Algorithm 1.

The sets set $\mathcal{A}$ and set $\mathcal{B}$ are evaluated using the counterfactual metric CFlips (see Equation 4.7). Specifically, the metric CFlips applies for each tuple in $\mathcal{A}$ and $\mathcal{B}$ to all the counterfactuals therein. The CFlips values are then averaged to obtain an overall value for $\mathcal{A}$ and $\mathcal{B}$, respectively. The evaluation pipeline is graphically depicted in Figure 4.1. The procedure can be repeated for different values of k and the different counterfactual generation strategies. To efficiently compute the metric CFlips for several values of k, two vectors (i.e., for $\mathcal{A}$ and $\mathcal{B}$) of size k can be created to accumulate the CFlips values before averaging them. These vectors can be used to plot how CFlips vary over the number of considered counterfactuals (see plots in Section 5.4.1). The optimized procedure is condensed into Algorithm 2. The same procedure can be used for the nDCCF evaluation and is condensed into Algorithm 3.

Algorithm 1: Algorithm for model training and counterfactual generation

Input:

- the Train and Test datasets D_{train} and D_{test}, where $D_{train} = \{\mathcal{X}_{train}, \mathcal{Y}_{train}, \mathcal{S}_{train}\}$, and $D_{test} = \{\mathcal{X}_{test}, \mathcal{Y}_{test}, \mathcal{S}_{test}\}$,
- the target label Classifier = $f(\cdot)$,
- the sensitive label Classifier = $f_{S_i}(\cdot)$,
- the classification loss Loss$(\cdot)$
- the number of train epochs *Epochs*,
- the number of counterfactuals to be generated for each sample N_{CF},
- the counterfactual generator $g(\cdot)$.

Result:

- the set $\mathcal{A}$ composed of tuples of objects related to the samples associated with a **negative outcome** (i.e., $f(\mathbf{x}) = 0$), belonging to an **privileged group** (i.e., $s_i = 1$), and correctly predicted to belong to the same sensitive class (i.e., $f_{S_i}(\mathbf{x}) = 1$),
- the set $\mathcal{B}$ composed of tuples of objects related to the samples associated with a **negative outcome** (i.e., $f(\mathbf{x}) = 0$), belonging to an **unprivileged group** (i.e., $s_i = 0$), and correctly predicted to belong to the same sensitive class (i.e., $f_{S_i}(\mathbf{x}) = 0$).

```
Randomly initialize $\theta_1$ for target output classifier $f(\cdot)$, and $\theta_2$ for sensitive label
  classifier $f_{S_i}(\cdot)$;
for $epoch \leftarrow 1$ to $Epochs$ do
    $\mathcal{X}_{train}, \mathcal{Y}_{train}, \mathcal{S}_{train} \leftarrow D_{train}$;
    $\hat{\mathcal{Y}}_{train} \leftarrow f(\mathcal{X}_{train})$;
    $\hat{\mathcal{S}}_{train} \leftarrow f_{S_i}(\mathcal{X}_{train})$;
    $\theta_1^* \leftarrow \arg\min_{\theta_1} Loss(\hat{\mathcal{Y}}_{train}, \mathcal{Y}_{train})$;
    $\theta_2^* \leftarrow \arg\min_{\theta_2} Loss(\hat{\mathcal{S}}_{train}, \mathcal{S}_{train})$;
endfor
for $\mathbf{x}^{(l)}, y^{(l)}, s_i^{(l)} \in D_{test}$ do
    $\hat{y}^{(l)} \leftarrow f(\mathbf{x}^{(l)})$;
    $\hat{s}^{(l)} \leftarrow f_s(\mathbf{x}^{(l)})$;
    $\mathcal{C}_{\mathbf{x}^{(l)}} = \{g(\mathbf{x}^{(l)}) \ : f(\mathbf{c}_{\mathbf{x}}) = y^* \}$;
    $\hat{\mathbf{s}}_{CF} \leftarrow f_s(\mathbf{c}_{\mathbf{x}}^{(l)})$ for $\mathbf{c}_{\mathbf{x}}^{(l)} \in \mathcal{C}_{\mathbf{x}^{(l)}}$;
    if $\hat{y}^{(l)} = 0$ then
        if $\hat{s}^{(l)}=1 \wedge s_i^{(l)}=1$ then
            $\mathcal{A} \leftarrow \mathcal{A} \cup \{\langle \mathbf{x}^{(l)}, \mathcal{C}_{\mathbf{x}^{(l)}}, \hat{y}^{(l)}, \hat{s}^{(l)}, \hat{\mathbf{s}}_{CF} \rangle\}$;
        end
        if $\hat{s}^{(l)}=0 \wedge s_i^{(l)}=0$ then
            $\mathcal{B} \leftarrow \mathcal{B} \cup \{\langle \mathbf{x}^{(l)}, \mathcal{C}_{\mathbf{x}^{(l)}}, \hat{y}^{(l)}, \hat{s}^{(l)}, \hat{\mathbf{s}}_{CF} \rangle\}$;
        end
    end
endfor
```

Algorithm 2: Counterfactual Flips (CFlips) Evaluation

Input:

- the number of counterfactuals to be generated for each sample N_{CF},
- the set $\mathcal{A}$ composed of tuples of objects related to the samples associated with a **negative outcome** (i.e., $f(\mathbf{x}) = 0$), belonging to an **privileged group** (i.e., $s_i = 1$), and correctly predicted to belong to the same sensitive class (i.e., $f_{S_i}(\mathbf{x}) = 1$),
- the set $\mathcal{B}$ composed of tuples of objects related to the samples associated with a **negative outcome** (i.e., $f(\mathbf{x}) = 0$), belonging to an **unprivileged group** (i.e., $s_i = 0$), and correctly predicted to belong to the same sensitive class (i.e., $f_{S_i}(\mathbf{x}) = 0$).

Result:

- the vector $\textbf{CFlips}_{priv}$ of size N_{CF} that contains averaged CFlips values, across all samples, of counterfactuals in $\mathcal{A}$ sorted in descending order of similarity, as returned by the counterfactual generator. The ith element of $\textbf{CFlips}_{priv}$ is the average of CFlips values considering i counterfactuals for all the samples.
- the vector $\textbf{CFlips}_{unpriv}$ of size N_{CF} that contains averaged CFlips values, across all samples, of counterfactuals in $\mathcal{B}$ sorted in descending order of similarity, as returned by the counterfactual generator. The ith element of $\textbf{CFlips}_{unpriv}$ is the average of CFlips values considering i counterfactuals for all the samples.

```
Initialize $\textbf{CFlips}_{priv} = [0,0,\dots,0]$, and $\textbf{CFlips}_{unpriv} = [0,0,\dots,0]$;
for $k \leftarrow 1$ to $N_{CF}$ do
    $n_p \leftarrow 0$;
    for $l_p^i \in \mathcal{A}$ do
        $\mathbf{x}^{(i)}, \mathcal{C}_{\mathbf{x}^{(i)}}, \hat{y}^{(i)}, \hat{s}_i^{(i)}, \hat{\mathbf{s}}_{i-CF} \leftarrow l_p^i$;
        $n_p \leftarrow n_p + 1$;
        $\textbf{CFlips}_{priv}[k] \leftarrow \textbf{CFlips}_{priv}[k] + \mathrm{CFlips}(\mathbf{x}^{(i)}, \mathrm{sorted}(\mathcal{C}_{\mathbf{x}^{(i)}})[:k], \hat{\mathbf{s}}_{i-CF}[:k])$;
    end
    $\textbf{CFlips}_{priv}[k] \leftarrow \textbf{CFlips}_{priv}[k]/n_p$;
    $n_{unp} \leftarrow 0$;
    for $l_{unp}^i \in \mathcal{B}$ do
        $\mathbf{x}^{(i)}, \mathcal{C}_{\mathbf{x}^{(i)}}, \hat{y}^{(i)}, \hat{s}_i^{(i)}, \hat{\mathbf{s}}_{i-CF} \leftarrow l_{unp}^i$;
        $n_{unp} \leftarrow n_{unp} + 1$;
        $\textbf{CFlips}_{unpriv}[k] \leftarrow \textbf{CFlips}_{unpriv}[k] + \mathrm{CFlips}(\mathbf{x}^{(i)}, \mathrm{sorted}(\mathcal{C}_{\mathbf{x}^{(i)}})[:k], \hat{\mathbf{s}}_{i-CF}[:k])$;
    end
    $\textbf{CFlips}_{unpriv}[k] \leftarrow \textbf{CFlips}_{unpriv}[k]/n_{unp}$;
end
```

Algorithm 3: normilized Cumulative Counterfactual Fairness (nDCCF) Evaluation

```
Input:
  • the number of counterfactuals to be generated for each sample $N_{CF}$,
  • the set $\mathcal{A}$ (see Algorithm 2),
  • the set $\mathcal{B}$ (see Algorithm 2).
Result:
  • the vector $\mathbf{nDCCF}_{priv}$ of size $N_{CF}$ that contains averaged nDCCF values, across all
    samples, of counterfactuals in $\mathcal{A}$ sorted in descending order of similarity, as returned
    by the counterfactual generator. The $i$th element of $\mathbf{nDCCF}_{priv}$ is the average of
    nDCCF values considering $i$ counterfactuals for all the samples.
  • the vector $\mathbf{nDCCF}_{unpriv}$ of size $N_{CF}$ that contains averaged nDCCF values, across
    all samples, of counterfactuals in $\mathcal{B}$ sorted in descending order of similarity, as
    returned by the counterfactual generator. The $i$th element of $\mathbf{nDCCF}_{unpriv}$ is the
    average of nDCCF values considering $i$ counterfactuals for all the samples.
Initialize $\mathbf{nDCCF}_{priv} = [0, 0, \dots, 0]$, and $\mathbf{nDCCF}_{unpriv} = [0, 0, \dots, 0]$;
for $k \leftarrow 1$ to $N_{CF}$ do
    $n_p \leftarrow 0$;
    for $l_p^i \in \mathcal{A}$ do
        $\mathbf{x}^{(i)}, \mathcal{C}_{\mathbf{x}^{(i)}}, \hat{y}^{(i)}, \hat{s}^{(i)}, \hat{\mathbf{s}}_{CF} \leftarrow l_p^i$;
        $n_p \leftarrow n_p + 1$;
        $\mathbf{nDCCF}_{priv}[k] \leftarrow \mathbf{nDCCF}_{priv}[k] + \frac{\mathrm{DCCF}(\mathbf{x}^{(i)}, \mathrm{sorted}(\mathcal{C}_{\mathbf{x}^{(i)}})[:k], \hat{\mathbf{s}}_{CF}[:k])}{IDCCF[:k]}$;
    end
    $\mathbf{nDCCF}_{priv}[k] \leftarrow \mathbf{nDCCF}_{priv}[k] / n_p$;
    $n_{unp} \leftarrow 0$;
    for $l_{unp}^i \in \mathcal{B}$ do
        $\mathbf{x}^{(i)}, \mathcal{C}_{\mathbf{x}^{(i)}}, \hat{y}^{(i)}, \hat{s}^{(i)}, \hat{\mathbf{s}}_{CF} \leftarrow l_{unp}^i$;
        $n_{unp} \leftarrow n_{unp} + 1$;
        $\mathbf{nDCCF}_{unpriv}[k] \leftarrow \mathbf{nDCCF}_{unpriv}[k] + \frac{\mathrm{DCCF}(\mathbf{x}^{(i)}, \mathrm{sorted}(\mathcal{C}_{\mathbf{x}^{(i)}})[:k], \hat{\mathbf{s}}_{CF}[:k])}{IDCCF[:k]}$;
    end
    $\mathbf{nDCCF}_{unpriv}[k] \leftarrow \mathbf{nDCCF}_{unpriv}[k] / n_{unp}$;
end
```

Chapter 5

Fairness under Unawareness is not a reliable fairness setting

This section details our experimental settings, designed to answer the research questions defined in Section 1.2. Two different models are trained: on the one hand, we train a model for making decisions for a specific task (i.e., income prediction or loan prediction), and on the other hand, we train the sensitive-feature classifiers to predict the sensitive group the samples belong to.

Specifically, we focus on the samples predicted as negative by the main task classifier. Next, we exploit counterfactual reasoning: starting from these samples classified as negative, we aim to modify features to cause a *flip* concerning the final prediction class (i.e., the prediction class goes from 0 to 1 by modifying one or more features). Subsequently, these new counterfactual samples feed the classifier for the sensitive features to predict the demographic group they belong to. In this way, we check if the counterfactual modifications have caused a flip concerning the sensitive group to which the sample belongs. The intuition here is that counterfactual-generated data are more explanatory in showing the model unfairness resulting from proxy features. The system's fairness can be evaluated by analyzing, for each test sample, any existing correlations between the target classification task and the protected classes inferred from counterfactuals.

5.1 Experimental Evaluation

Before addressing the initial three research questions, we provide a concise exposition of the comprehensive experimental setup for our study. This includes an elucidation of the utilized datasets, preprocessing procedures, metrics, evaluation protocols, and a

Table 5.1 Adult, Adult-debiased, Crime, and German datasets' characteristics. $|X|$ to the number of feature w/o the sensitive one (i.e., *gender* and *crime*).

Dataset	Split	$\|\mathcal{D}\|$	$\|X\|$	Target (Y)	$Y = 1$
Adult	**Train**	40699	13	income	$\geq \$50,000$
	Test	4523	13	income	$\geq \$50,000$
Adult-debiased	**Train**	40699	6	income	$\geq \$50,000$
	Test	4523	6	income	$\geq \$50,000$
Crime	**Train**	1794	98	Violent State	$<$*median*
	Test	200	98	Violent State	$<$*median*
German	**Train**	900	17	credit score	Good
	Test	100	17	credit score	Good

detailed exploration of hyperparameters for each component of our framework. This meticulous presentation ensures the provision of all necessary information for a rigorous reproduction of the thesis's findings.

5.1.1 Datasets and Preprocessing

Experiments are conducted on three popular datasets, used as benchmarks in several works [7, 66, 72, 84, 138]. Despite their small dimension, as stated by Rossini et al. [149], these datasets are useful to evaluate fairness approaches because they represent real-world problems and provide a wide range of attributes that can be used to develop ethical standards. These are: Adult [108], a real-world dataset used for income prediction[1], German [99], a real-world dataset for default prediction[2], and Crime [145], a real-world Census dataset for violent state prediction[3] (i.e., a state is violent if the number of crimes in a state is higher with respect to the median ($|\mathcal{C}_{\mathbf{x}}|$) of all the states). For Adult and German, the sensitive attribute we considered is *gender*, with *male* and *female* corresponding to the *privileged* and *unprivileged* group respectively. For the Crime dataset, the sensitive attribute we considered is *race* that indicates the race with the largest number of crimes committed in a specific state. As a second sensitive feature, for Adult, we chose *maritalStatus*, with *married* and *not married* as *privileged* and *unprivileged*; for German, we chose age as > 25 *years* and $<= 25$ *years* as *privileged* and *unprivileged.* In this dataset, each sample consists of the name of the *state* and the number of crimes associated with each race, i.e., *white*, *black*, *asian*, and *hispanic.*

[1] Adult: https://archive.ics.uci.edu/ml/datasets/adult

[2] German: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[3] Crime: https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)

Table 5.2 Overview of relevant dataset information, including sensitive feature distribution, target distribution, name of privileged group, and ex-ante Statistical Parity respectively for the Adult, Adult-debiased, German, and Crime datasets.

Dataset	S_i	privileged ($S_i = 1$)	$\Phi(S)^{\dagger}$	$\Phi(Y)^{\dagger\dagger}$	ex-ante SP*
Adult	*gender*	*male*	0.68/0.32	0.25/0.75	0.199
	maritalStatus	*married*	0.48/0.52	0.25/0.75	0.378
Adult-deb.	*gender*	*male*	0.68/0.32	0.25/0.75	0.199
	maritalStatus	*married*	0.48/0.52	0.25/0.75	0.378
Crime	*race*	*white*	0.58/0.42	0.50/0.50	0.554
German	*gender*	*male*	0.69/0.31	0.70/0.30	0.075
	age	> 25 *year*	0.81/0.19	0.70/0.30	0.149

† Probability distribution of the *privileged* and *unprivileged* group:

$$\mathbb{P}(S_i = 1)/\mathbb{P}(S_i = 0)$$

†† Probability distribution of the target variable:

$$\mathbb{P}(Y = 1)/\mathbb{P}(Y = 0)$$

* A priori Statistical Parity, based on Independence criteria:

$$\mathbb{P}(Y = 1 \mid S_i = 1) - \mathbb{P}(Y = 1 \mid S_i = 0)$$

For our task, we split races into two groups *White* and *Others* where *Others* groups the crimes of *Black*, *Asian*, and *Hispanic* races. This reproduces the setting of [7]. The *privileged* group is the *White* one, and the *unprivileged* is *Others* (i.e., Blacks, Asians, and Hispanics). More details for each dataset setting can be found following.

Adult Dataset

Adult[4] is a popular UCI Machine Learning dataset extracted from the 1994 US Census database. The prediction task is to determine whether a person earns more than 50K a year. The sensitive attributes consider for this dataset are *gender* which indicates the sex of an individual, and *marital status*, whether an individual is married or not.

In the Adult dataset, there are other sensitive characteristics (i.e., *age*, *relationship*, and *race*). Since Fairness Under Unawareness, the setting most coherent with current AI regulations, requires bereaving the dataset of sensitive information during training, we decided not to use these features to learn the model. For that reason and due to the dimension of the dataset, we decided to create two different settings:
(a) - **Adult**: the original dataset where we only discarded the sensitive features *gender* and *marital-status*; (b) - **Adult-debiased**: where we remove all the sensitive features

[4] https://archive.ics.uci.edu/ml/datasets/adult

(i.e., gender, age, marital status, and race), and all the features highly correlated with at least one of the sensitive features. From the whole set of sensitive features we chose to investigate but not to use in the training phase, only *gender* and *marital-status* as classic sensitive information for benchmarking debiasing models [72, 95, 132]. As regards the non-sensitive features used for training the models, 6 out of 15 were used: *education num*, *occupation*, *work class*, *capital gain*, *capital loss*, *hours per week*. The remaining non-sensitive features are filtered out because they show a high correlation with the sensitive features (Pearson's correlation coefficient greater than 0.35). Furthermore, the feature *work class* is condensed into three classes: *Private*, *Public*, and *Unemployed*. We replace the categories in *work class Private*, *SelfEmpNotInc*, *SelfEmpInc*, with *Private*, the categories *FederalGov*, *LocalGov*, *StateGov*, with *Private*, and the category *WithoutPay* with *Unemployed*. The Adult dataset is imbalanced, as shown in Table 5.2. This can emphasize some biases [72, 173, 175]. The target label *income >= 50K* is strongly unbalanced towards the *privileged* class (male, married). More detailed statistics, including the number of samples, the sensitive feature distribution, and the ex-ante statistical parity, are summarized in Table 5.1 and Table 5.2.

German Dataset

German[5] is another popular UCI Machine Learning dataset extracted from a German bank loan approval history. Demographic and financial characteristics of individuals who applied for a loan are collected in this dataset, along with the decision to grant them a loan or not. The prediction task is the binary decision of approving a loan based on the probability of repaying it. The sensitive characteristics taken into account are *gender* and *age*. As for the Adult dataset, German contains other sensitive characteristics (e.g., *race*) beyond those exploited in this study. Also, in this case, we do not include these features for learning the model for guaranteeing the *fairness under awareness* setting. We exploit 17 non-sensitive features to train the predictive models (i.e., *existingchecking*, *duration*, *credithistory*, *purpose*, *creditamount*, *savings*, *employmentsince*, *installmentrate*, *otherdebts*, *residencesince*, *property*, *otherinstallmentplans*, *housing*, *existingcredits*, *job*, *peopleliable*, *telephone*). As for the Adult dataset, German is imbalanced [72, 173, 175]. Table 5.2 shows that the privileged group is overrepresented for both the sensitive features. Moreover, the ex-ante statistical parity metric indicates that the advantaged target label ($Y = 1$) is strongly associated with the privileged group ($S_i = 1$) compared to the unprivileged group ($S_i = 0$), which

[5] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

confirms that the data is imbalanced and strongly biased. Useful statistical details are reported in Table 5.1 and Table 5.2.

Crime dataset

The Communities and Crime dataset combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. It was created by Michael Redmond and is provided by the UCI machine learning repository (Dua & Graff, 2017). The dataset contains 128 attributes such as county, population, per capita income, and number of immigrants.

The task consists of predicting whether the number of violent crimes per population for a given community is above or below the median. For our task, we reproduced the setting of [7]. Thus, we condense and then split races into two groups *White* and *Others* where *Others* groups the crimes of *Black*, *Asian*, and *Hispanic* races. This reveals who among the two groups (i.e., white, others) committed more crimes in the state. The *privileged* group is the *White* one, and the *unprivileged* is *Others* (i.e., Blacks, Asians, and Hispanics). More details for each dataset setting can be found following. Useful statistical details are reported in Table 5.1 and Table 5.2.

5.1.2 Evaluation Metrics

The evaluation includes two different groups of metrics: accuracy-based and bias-based metrics. The accuracy-based metrics are mainly based on the confusion matrix, which quantifies how many samples are correctly classified or misclassified for both the negative and positive classes. For self-consistency, this section details all the considered metrics. Some are just recalled, reporting the formulas. The others, used in cutting-edge fairness research, are described. The first metric is the Accuracy, which quantifies the overall number of correct classifications over the predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{5.1}$$

The Recall metric measures the number of positive correctly classified samples with respect to all the real positive ones:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.2}$$

Precision measures the ratio of samples correctly classified as positive over the ones classified as positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{5.3}$$

The F1 score is the harmonic mean between recall and accuracy:

$$\text{F1} = \frac{2}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}} = 2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2\cdot TP}{2\cdot TP+FP+FN} \tag{5.4}$$

The primary goal of the F1 score is to combine the precision and recall metrics into a single metric. Indeed, this metric is useful for evaluating classification methods when dealing with imbalanced data. The Area Under the Receiver Operating Characteristic Curve (AUC) is a metric that measures the capability of a classifier to separate the positive class from the negative class correctly. It can be formulated as follows:

$$\text{AUC} = \frac{\sum_{x^-\in X^-}\sum_{x^+\in X^+}(\mathbb{1}(f(x^-)<f(x^+)))}{|X|^- + |X|^+} \tag{5.5}$$

where X^+ is the set of positive sample, X^- is the set of negative sample, $f(\cdot)$ is the result of model prediction, and $\mathbb{1}(\cdot)$ an indicator function [25].

To quantify the presence of bias in the decision of the two classifiers several fairness metrics were used that consider the *Independence* and *Separation* statistical criteria. For the Independence statistical criteria, we used *Difference in Statistical Parity* (DSP) and *Disparate Impact* (DI). DSP measures the difference between the probability that samples belonging to the *privileged* group and to the *unprivileged* group are classified in a positive outcome class [97]. It is the equivalent of the difference between the sum of the TP rate and FP rate of the *privileged* and *unprivileged* group (see Equation 5.6 as reminder of Equation 3.3). A model is considered Fair w.r.t. DSP if the measure is equal or, at least, very close to zero.

$$\begin{aligned} \text{DSP} &= \left|\mathbb{P}(\hat{Y}=1|S_i=1)-\mathbb{P}(\hat{Y}=1|S_i=0)\right| \\ &= \left|(\text{TPrate}_{priv}+\text{FPrate}_{priv})-(\text{TPrate}_{unpriv}+\text{FPrate}_{unpriv})\right| \end{aligned} \tag{5.6}$$

For the *Separation* statistical criteria, we used *Difference in Equal Opportunity* (DEO) and *Difference in Average Odds* (DAO). The former, i.e., DEO, measures the difference between the probability of instances in a *privileged* group and the probability of instances in an *unprivileged* group being correctly classified in a positive outcome class [97]. The formulation of the DEO metric is shown in Equation 5.7, as reminder of Equation 3.7.

$$\begin{aligned}\text{DEO} &= \left|\mathbb{P}(\hat{Y}=1|Y=1,S_i=1)-\mathbb{P}(\hat{Y}=1|Y=1,S_i=0)\right| \\ &= \left|\text{TPrate}_{priv}-\text{TPrate}_{unpriv}\right|\end{aligned} \tag{5.7}$$

The latter, i.e., DAO, measures the difference between the probability of instances in a *privileged* group and the probability of instances in an *unprivileged* group being correctly classified in a positive outcome class, as DEO does. Furthermore, DAO also considers the difference between the probability of instances in a *privileged* group and the probability of instances in a *privileged* group being incorrectly classified in a positive outcome class. DAO gives a broader intuition of how imbalanced the classifier accuracy is between the two groups [97]. The formulation of the DAO metric is shown in Equation 5.8, as reminder of Equation 3.8.

$$\begin{aligned}\text{DAO} = \frac{1}{2}\Big(&\left|\mathbb{P}(\hat{Y}=1|Y=0,S_i=1)-\mathbb{P}(\hat{Y}=1|Y=0,S_i=0)\right| \\ &+\left|\mathbb{P}(\hat{Y}=1|Y=1,S_i=1)-\mathbb{P}(\hat{Y}=1|Y=1,S_i=0)\right|\Big) \\ = \frac{1}{2}\Big(&\left|\text{FPrate}_{priv}-\text{FPrate}_{unpriv}\right|+\left|\text{TPrate}_{priv}-\text{TPrate}_{unpriv}\right|\Big)\end{aligned} \tag{5.8}$$

In either case, for DEO and DAO, a model is considered fair if the measure is equal or, at least, very close to zero.

5.1.3 Evaluation Protocol and Reproducibility

Following, we will give all the details of our experimental pipeline to reproduce our experiments starting in each setting.

Dataset Splitting

The datasets were split with a random 90/10 hold-out method to partition train and test sets, with stratification based on the target variable $\mathcal{Y}$ and the sensitive features $\mathcal{S}$. For the Adult dataset, we have 40699 train samples and 4523 test samples, for the Crime dataset, 1794 train samples and 200 test samples, and for the German dataset, 900 train samples and 100 test samples (see Table 5.1). For reproducibility, we used the Scikit-learn implementation for splitting with a random seed set to 42[6].

[6] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Decision-Maker Hyperparameter Tuning and optimization

The target label classifiers, i.e., LR, SVM, DT, RF, MLP, XGB, and LGB (see Section 4.3.1), have been tuned using a grid search strategy[7]. For hyperparameter tuning and validation, the train data was further split using a k-fold cross-validation strategy, with the number of folds set to five. The best models hyperparameters have been chosen to optimize the Area under the ROC curve metric (AUC) since AUC indicates how well the classifier can separate the positive from the negative class (see Equation 5.5). For reproducibility, the list of explored hyperparameter values is reported in Table 5.3.

Debiased Decision-Makers Hyperparameter Tuning and Optimization

The Debiasing classifiers, i.e., AdvDeb, LFERM, and FairC (see Section 4.3.1), have been tuned using the same evaluation protocol, with a grid search for the hyperparameter values and a 5-fold cross-validation strategy. Conversely, in this evaluation, the best models have been chosen to optimize AUC and Fairness with an overall metric that considers both:

$$\mathrm{AUC_{FAIR}} = \mathrm{AUC} \cdot (1 - \mathrm{DAO}) \tag{5.9}$$

It is straightforward notice that any other Fairness metric could replace DAO. In this work, DAO is chosen to balance fairness in terms of correct predictions for negative and positive samples. The list of explored hyperparameter values is reported in Table 5.3.

Sensitive Feature Classifier Hyperparameter Tuning and optimization

The sensitive label classifiers, i.e., XGB (see Section 4.3.3), are tuned using the same approach, exploiting a grid search exploration[7] for hyperparameter values and a 5-fold cross-validation strategy. Due to the imbalanced nature of the datasets concerning the sensitive classes, the models optimizing the F1 score are chosen (see Equation 5.4). Explored hyperparameter values are shown in Table 5.3.

Counterfactual generation

For the sake of reproducibility, the generation of counterfactual samples makes use of DiCE, as discussed in Section 4.3.2. To avoid the results depending on a single counterfactual generation strategy, we considered three different strategies, i.e., Random,

[7] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Table 5.3 Hyperparameter list, values and type for the classification models reported in this work.

Algorithm	Hyperparameter	Values	Type
All Model	seed	{42}	Integer
Logistic Regression	penalty	{l1,l2}	String
	tol	{0.0001,0.00001}	Float
	C	$\{10^{-4+(\frac{8}{20}i)}$ for i in $range(1,21)\}$	Float
	fit_intercept	{True, False}	Boolean
	class_weight	{ balanced, None}	String
	solver	{newton-cg, lbfgs, liblinear, sag, saga}	String
	warm_start	{True, False}	Boolean
Support Vector Machines	C	{0.1, 1, 10}	Float
	class_weight	{balanced, None}	String
	gamma	{scale, auto}	String
	kernel	{linear, rbf, sigmoid}	String
Decision Tree	ccp_α	{0.1, 0.01, 0.001}	Float
	max_features	{'sqrt', 'log2'}	String
	max_depth	{i for i in range(1,10)}	Integer
	max_samples_split	{i for i in range(1,10)}	Integer
	max_samples_leaf	{i for i in range(1,5)}	Integer
	criterion	{*gini*, *entropy*}	String
Random Forest	bootstrap	{True, False}	Boolean
	max_depth	{None, 10, 20, 40, 60, 80, 100}	Integer
	max_features	{auto, sqrt}	String
	min_samples_split	{2, 5, 10}	Integer
	min_samples_leaf	{1, 2, 4}	Integer
	n_estimators	{50, 100, 200, 400}	Integer
Multi-Layer Perceptron	hidden_layer_sizes	{(32, 64, 128), (32, 64), (64,)}	Tuple[Int]
	activation	{tanh, relu}	String
	solver	{sgd, adam}	String
	alpha	{0.0001, 0.05}	Float
	learning_rate	{constant, adaptive}	String
eXtreme Gradient Boosting	min_child_weight	{1, 5, 10}	Integer
	gamma	{0.01, 0.1, 0.5}	Float
	learning_rate	{0.1, 0.01, 0.001}	Float
	max_depth	{3, 5, 6}	Integer
	subsample	{0.4,0.6,0.8,1.0}	Float
	colsample_bytree	{0.6, 0.8, 1}	Float
	n_estimators	{50, 100, 300,500}	Integer
	reg_alpha	{0.1, 0.01, 0.02}	Float
Light Gradient Boosting	learning_rate	{0.1, 0.05}	Float
	num_leaves	{3, 10, 30, 50, 100, 200}	Integer
	reg_alpha	{None, 0.01, 0.05, 0.1}	Float
	colsample_bytree	{0.6, 0.8,1}	Float
	max_depth	{-1, 3, 5, 8, 10}	Integer
	reg_lambda	{None, 0.01, 0.02, 0.03}	Float
	n_estimators	{50, 100, 300}	Integer
Adversarial Debiasing	adversary_loss_weight	{0.01, 0.05, 0.1}	Float
	num_epochs	{50, 70, 150, 250, 500}	Integer
	batch_size	{64, 128, 256, 512}	Integer
	hidden_units	{64, 128, 256}	Integer
	number_of_layers	{1}*	Integer
Linear Fair Empirical Risk Minimization	C	{0.01, 0.1, 1}	Float
	kernel	{linear}	String
Fair Classification	C	{0.001, 0.01, 0.1, 1}	Float

* AIF360 implementation of Adversarial Debiasing does not allow to change the number of layers.

Genetic, and KDtree. For the Random strategy, the *seed* has been set to 42, the *posthoc sparsity parameter* to 0.1, and the *posthoc sparsity algorithm* to *linear search.* For the Genetic strategy, we set the *initialization* to *kdtree*, the *proximity weight* to 0.2, the *sparsity weight* to 0.2, the *diversity weight* to 5, the *categorical penalty* to 0.1, the counterfactual generation loss to *hinge-loss*, the *feature weights* to *inverse Mean Absolute Deviation* (MAD), the *posthoc sparsity parameter* to 0.1, the *posthoc sparsity algorithm* to *binary search*, and the *max iterations* to 500. For the KDtree strategy, we set the *sparsity weight* to 1, the *feature weights* to *inverse Mean Absolute Deviation* (MAD), the *posthoc sparsity parameter* to 0.1, and the *posthoc sparsity algorithm* to *linear search.* For each sample in the test set, an overall number of 100 counterfactuals was requested (see Algorithm 1). For reproducibility reasons, we use all the previously listed default parameter values of the DiCE tool, except for the *posthoc sparsity algorithm* set to *binary search* in the Genetic strategy for speeding up the search due to the expensive experimental time.

5.2 Unawareness doesn't mean privatization of sensitive information (RQ1)

Predicting sensitive characteristics of users, such as their personal or private information, is a complex and ethically sensitive task that must be handled with great care and adherence to privacy laws and ethical guidelines. Predicting or attempting to infer sensitive information about individuals without their consent or in violation of privacy regulations is unethical and potentially illegal. The *Fairness under Unawareness* setting tries to guarantee fair model behaviour by discarding sensitive information. Discarding sensitive features can be considered a sort of high-level user privatization of sensitive information. However, for fairness purposes, it cannot guarantee that the model predicts without considering characteristics (i.e., non-sensitive features) that can be correlated or can have non-linear dependencies with the sensitive features. These types of characteristics are known in the literature as proxy features. Identifying them is still a hard task. So our first research question is:

RQ1

Is there a principled way to identify if proxy features exist in a dataset?

Thus, the first stage of our experiments aims to assess how well the sensitive-feature classifier can identify if a subject belongs to the *privileged* or *unprivileged* group,

Table 5.4 Complete results on the Adult, Adult-debiased, Crime, and German test set of the Sensitive Feature Classifiers.

			Sensitive feature Classifier		
Dataset	S_i	metric	**RF**	**MLP**	**XGB**
Adult	gender	AUC↑	0.9402	0.9363	**0.9413**
		ACC ↑	0.8539	**0.8559**	0.8463
		Precision ↑	0.9043	0.9065	**0.9549**
		Recall ↑	0.8762	**0.8768**	0.8107
		F1 ↑	**0.8900**	0.8914	0.8769
	maritalStatus	AUC↑	0.9883	0.9882	**0.9907**
		ACC ↑	**0.9830**	0.9825	0.9828
		Precision ↑	**1.0000**	0.9986	**1.0000**
		Recall ↑	0.9644	**0.9649**	0.9640
		F1 ↑	**0.9819**	0.9814	0.9816
Adult-deb	gender	AUC↑	**0.8028**	0.8010	0.7896
		ACC ↑	**0.7482**	0.7480	0.7444
		Precision ↑	0.7699	0.7832	**0.8111**
		Recall ↑	**0.8942**	0.8664	0.8100
		F1 ↑	**0.8274**	0.8227	0.8106
	maritalStatus	AUC↑	0.7286	0.7103	**0.7708**
		ACC ↑	0.6655	0.6611	**0.6918**
		Precision ↑	0.6598	0.6547	**0.6677**
		Recall ↑	0.6211	0.6169	**0.7098**
		F1 ↑	0.6398	0.6353	**0.6879**
Crime	race	AUC↑	0.9893	0.9885	**0.9910**
		ACC ↑	0.9450	**0.9500**	0.9450
		Precision ↑	0.9412	**0.9417**	0.9412
		Recall ↑	0.9655	**0.9741**	0.9655
		F1 ↑	0.9532	**0.9576**	0.9532
German	gender	AUC↑	0.7106	0.5091	**0.7139**
		ACC ↑	**0.7300**	0.6900	0.6900
		Precision ↑	0.7234	0.6900	**0.7879**
		Recall ↑	0.9855	**1.0000**	0.7536
		F1 ↑	**0.8344**	0.8166	0.7704
	age	AUC↑	**0.8876**	0.4756	0.8363
		ACC ↑	**0.8600**	0.8100	0.8100
		Precision ↑	0.8526	0.8100	**0.8605**
		Recall ↑	**1.0000**	**1.0000**	0.9136
		F1 ↑	**0.9205**	0.8950	0.8862

without exploiting sensitive features in the training phase. We trained a sensitive feature classifier for each dataset. The investigated sensitive features are *gender* or *race* and results are shown in Table 5.4.

- The first observation is that every *Sensitive-Feature Classifier* shows to be accurate for all the datasets. Among them, XGB is the algorithm that exhibits the best performance in terms of AUC while RF shows the most promising ACC and F1 performance. For the Adult, Adult-debiased, and Crime datasets, MLP performance is comparable to XGB and RF, while showing a low AUC on the German dataset indicating it is not reliable and useful for our analysis. We recall we seek models with high F1 and AUC since it indicates that the classifiers provide accurate and balanced predictions.

- The careful reader may have noticed that, on the Adult-debiased, the sensitive-feature classifiers exhibit the worst performance in comparison with the original Adult dataset, and for each sensitive feature setting (i.e., *gender* and *marital Status*). This behaviour is due to the debiasing process, as the Adult-debiased dataset has been deprived of features highly correlated with the sensitive ones. Noteworthy, the prediction capability remains high despite the absence of sensitive and sensitive-correlated features. This simple analysis demonstrates how each classifier can deeply model non-linear relations with the remaining features "*deprivatizing*" the sensitive information.

Final comments. *Results show that, due to proxy features, it is possible to train a classifier capable of predicting sensitive characteristics. Moreover, it is still possible to predict sensitive information even when only low correlated features with the sensitive information are available (i.e., Adult-debiased).*

5.3 Unawareness doesn't guarantee a model-agnostic fair behaviour (RQ2)

The results of the first research question (i.e., RQ1) can be considered trivial or unuseful since the existence of proxy features is something already known in the literature [32, 83, 94]. However, this first experimental session laid the foundation for going ahead with our investigation.

The *Fairness Under Unawareness* setting aims to ensure fair treatment by removing sensitive features from training data. However, as demonstrated previously, it is possible to predict sensitive information due to the existence of proxy features. The purpose of this second analysis can be summarized through the following research question:

Table 5.5 Complete accuracy and fairness results on the Adult, Adult-debiased, Crime, and German test set of the Decision Maker Classifiers.

		Decision-Maker $f(\cdot)$									
Dataset	metric	**LR**	**DT**	**SVM**	**LGBM**	**XGB**	**RF**	**MLP**	**LFERM**	**ADV**	**FairC**
Adult	AUC ↑	0.9078	0.8484	0.9073	0.9304	**0.9314**	0.9118	0.9119	0.9031	0.9123	0.8770
	ACC ↑	0.8099	0.8161	0.8541	0.8658	**0.8698**	0.8534	0.8494	0.8428	0.8512	0.8395
	Precision ↑	0.5782	0.6879	0.7570	0.7655	**0.7737**	0.7371	0.7222	0.7324	0.7500	0.7382
	Recall ↑	**0.8608**	0.4719	0.6057	0.6610	0.6708	0.6351	0.6378	0.5763	0.5995	0.5459
	F1 ↑	0.6918	0.5598	0.6729	0.7094	**0.7186**	0.6823	0.6774	0.6450	0.6663	0.6277
	DSP ↓ (*gender*)	0.2947	0.1461	0.1769	0.1850	0.1884	0.1854	0.1902	0.1448	0.1151	**0.0528**
	DEO ↓ (*gender*)	0.0546	0.0760	0.0644	0.0379	0.0635	0.0216	0.0529	**0.0194**	0.1399	0.2451
	DAO ↓ (*gender*)	0.1241	0.0722	0.0692	0.0569	0.0680	0.0545	0.0708	**0.0386**	0.0879	0.1274
	DSP ↓ (*MS*)	0.6290	0.2833	0.3562	0.3648	0.3601	0.3750	0.3430	0.2713	0.3779	**0.0162**
	DEO ↓ (*MS*)	0.4656	0.3461	0.3464	0.2893	0.2874	0.3014	0.3168	**0.1875**	0.3851	0.3710
	DAO ↓ (*MS*)	0.4736	0.2343	0.2554	0.2148	0.2122	0.2388	0.2300	**0.1467**	0.2859	0.2112
Adult-deb	AUC ↑	0.8233	0.7895	0.7944	**0.8596**	0.8578	0.8336	0.8271	0.8017	0.8309	0.7981
	ACC ↑	0.7367	0.8017	0.8061	0.8371	**0.8375**	0.8267	0.8156	0.7953	0.8196	0.8054
	Precision ↑	0.4790	0.8294	**0.8389**	0.8038	0.8063	0.7621	0.7540	0.7079	0.7529	0.7526
	Recall ↑	**0.7119**	0.2516	0.2694	0.4532	0.4532	0.4371	0.3800	0.2962	0.4050	0.3202
	F1 ↑	0.5727	0.3860	0.4078	0.5796	**0.5802**	0.5556	0.5053	0.4176	0.5267	0.4493
	DSP ↓ (*gender*)	0.1567	**0.0438**	0.0534	0.1093	0.1056	0.1058	0.0863	0.0639	0.0957	0.0575
	DEO ↓ (*gender*)	0.0695	0.0492	0.0353	0.0470	0.0400	0.0703	**0.0173**	0.0179	0.0326	0.0529
	DAO ↓ (*gender*)	0.0693	0.0272	0.0227	0.0356	0.0304	0.0461	0.0188	**0.0186**	0.0282	0.0315
	DSP ↓ (*MS*)	0.1793	0.0945	0.0948	0.1702	0.1663	0.1501	0.1249	**0.0336**	0.1316	0.1241
	DEO ↓ (*MS*)	0.1450	0.0468	0.0720	0.0676	0.0645	**0.0460**	0.1266	0.0489	0.1108	0.1128
	DAO ↓ (*MS*)	0.0880	0.0240	0.0362	0.0409	0.0355	**0.0232**	0.0633	0.0262	0.0591	0.0570
Crime	AUC ↑	0.9248	0.8991	**0.9288**	0.9168	0.9099	0.9096	0.9203	0.9100	0.9008	0.8024
	ACC ↑	**0.8700**	0.8200	**0.8700**	0.8400	0.8500	0.8400	0.8650	0.8400	0.8100	0.7500
	Precision ↑	0.8627	0.8265	**0.8776**	0.8400	0.8500	0.8400	0.8544	0.8333	0.8444	0.7500
	Recall ↑	**0.8800**	0.8100	0.8600	0.8400	0.8500	0.8400	**0.8800**	0.8500	0.7600	0.7500
	F1 ↑	**0.8713**	0.8182	0.8687	0.8400	0.8500	0.8400	0.8670	0.8416	0.8000	0.7500
	DSP ↓ (*race*)	0.6535	0.6190	0.6396	0.6363	0.6568	0.6363	0.6622	0.6125	0.5501	**0.2258**
	DEO ↓ (*race*)	0.3294	0.4039	0.3843	0.2824	0.2941	0.2824	0.3294	0.2941	0.1882	**0.1373**
	DAO ↓ (*race*)	0.3438	0.3827	0.3390	0.3525	0.3656	0.3525	0.3599	0.3278	0.2732	**0.0862**
German	AUC ↑	**0.8186**	0.7219	0.8110	0.7614	0.7871	0.7936	0.8162	0.7605	0.7371	0.8152
	ACC ↑	0.7600	0.7600	0.7600	0.7500	**0.7900**	0.7600	0.7600	0.7200	0.7300	0.7400
	Precision ↑	**0.8485**	0.7805	0.7738	0.7848	0.8025	0.7674	0.7738	0.7188	0.7792	0.7619
	Recall ↑	0.8000	0.9143	0.9286	0.8857	0.9286	0.9429	0.9286	**0.9857**	0.8571	0.9143
	F1 ↑	0.8235	0.8421	0.8442	0.8322	**0.8609**	0.8462	0.8442	0.8313	0.8163	0.8312
	DSP ↓ (*gender*)	0.1187	**0.0271**	0.0449	0.1632	0.0519	0.0626	0.0449	0.0355	0.1809	0.0449
	DEO ↓ (*gender*)	0.1400	0.0500	0.0300	0.1900	0.0400	0.0800	0.0300	**0.0200**	0.2200	0.0500
	DAO ↓ (*gender*)	0.1657	0.0537	0.0892	0.1117	**0.0296**	0.0878	0.0892	0.0746	0.1267	0.0728
	DSP ↓ (*age*)	0.2827	0.0845	0.0344	0.1676	0.0006	0.1397	0.2112	**0.0000**	0.0273	0.1956
	DEO ↓ (*age*)	0.3020	0.0693	0.0570	0.0740	0.0847	0.0570	0.2049	**0.0000**	0.0108	0.2049
	DAO ↓ (*age*)	0.1851	0.0347	0.0853	0.0995	0.0651	0.0967	0.1195	**0.0000**	0.0849	0.1252

RQ2

Does the Fairness Under Unawareness setting ensure that decision biases are avoided?

Therefore, our second analysis is structured as follows. Before evaluating fairness metrics, we evaluated the accuracy performance of the classifiers exploited to implement the *Decision Maker*. Subsequently, we evaluated each classifier's performance on classic fairness metrics (i.e., DSP, DEO, and DAO) even if in an *unawareness* setting to check whether it is useful for fairness purposes.

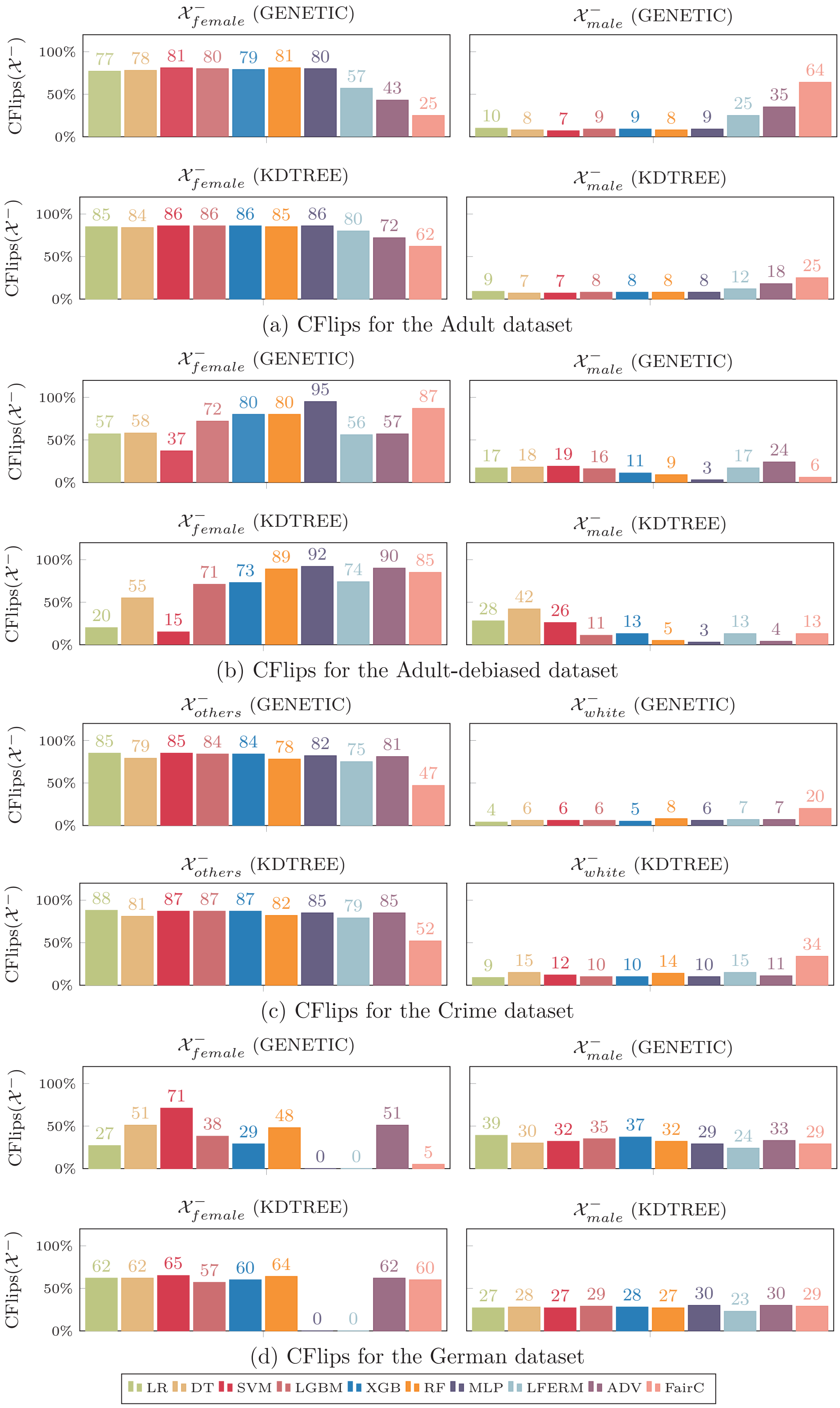
- Table 5.5 indicates that all classifiers work well in terms of accuracy metrics. However, as expected, adopting *Fairness Under Unawareness* – i.e., removing all the sensitive information and, for Adult-debiased, also removing highly correlated features – has caused a worsening of the performance for all the classifiers (see the comparison between Adult and Adult-debiased results). This observation suggests that sensitive and sensitive-correlated information may be “necessary" to predict the target label correctly.

- Table 5.5 also reports the fairness evaluation computing the Difference in Statistical Parity (DSP), Difference in Equal Opportunity (DEO), and Difference in Average Odds (DAO). It is worth noticing that removing the considered sensitive information (i.e., gender, marital status (MS), race, and age) has not improved model equity. This result shows that not removing proxy features makes the *Fairness Under Unawareness* setting useless since the model can implicitly learn them. A clear example is the Adult-debiased dataset, where DSP, DEO, and DAO values are generally better than the Adult dataset. This is probably caused by the removal of highly correlated features with the sensitive ones. However, some degree of discrimination is still present due to non-linear dependencies with proxy features. Furthermore, despite adopting the *in-processing* debiasing constrained optimization and fair hyperparameters tuning strategy, the debiased *Decision Maker* does not seem to improve fairness performance consistently (e.g., ADV or FairC in Adult). As a last remark, we can notice how not always debiasing the dataset corresponds to an improvement of fairness metrics. Thus, in some cases, in order to find a setting that enforces accuracy, it can worsen their performance by disclosing implicit non-linear discriminative bias.

Final comments. *The classifiers seem to be affected by discrimination even when the sensitive information is omitted. Accordingly, imposing Fairness Under Unawareness setting is not sufficient to avoid decision biases and discrimination.*

5.4 Counterfactual Reasoning for (un)fairness assessment (RQ3)

Throughout the first two research questions, we have seen how sensitive information can be recovered with latent nonlinear relationships known as proxy features making



(a) CFlips for the Adult dataset

(b) CFlips for the Adult-debiased dataset

(c) CFlips for the Crime dataset

(d) CFlips for the German dataset

Fig. 5.1 CFlips (i.e., Counterfactual Flips, see Definition 4.7) for samples of the *unprivileged* and *privileged* sensitive group. The groups are *others* ($S_i = 0$) and *white* ($S_i = 1$) for the Crime dataset, and *female* ($S_i = 0$) and *male* ($S_i = 1$) for the Adult, Adult-deb, and German dataset.

unuseful privatizing the dataset with *fairness under unawareness* setting. Furthermore, this setting does not guarantee the model to be absent from discriminative behaviours even when we omit highly correlated features. Another challenge is that some fairness metrics may conflict with each other without giving a deeper quantification of the (un)fairness of the model, making developers confused and disoriented. With our analysis, we want to give a framework that can enhance the clarity for AI practitioners. This experiment aims to unveil potential decision biases by counterfactual reasoning and wants to answer the following research question:

RQ3

Is counterfactual reasoning suitable and effective for discovering hidden decision biases?

Figure 5.1 reports the CFlips values (see Definition 4.7) for each classifier and category – i.e., *privileged*, see Eq. 4.14, and *unprivileged*, see Eq. 4.15, with XGB as $f_{S_i}(\cdot)$ and other sensitive features classifiers are postponed to Section 5.4.1. Following, in Table 5.6 and Table 5.7, we exploit the CFlips, nDCCF, and their Δ between the *privileged* and *unprivileged* group for both the Genetic and KDtree Counterfactual strategy. Before starting to comment on the analysis, we want to remind the reader that the CFlips metric tells us how frequently a change in the decision (from negative to positive) for a sample is followed by a change in the sensitive-feature classification (e.g., from *female* to *male* and vice versa) while the nDCCF metric rewards CF samples in the ranking that did not flip based a higher rank.

- The proposed metric seems to operate as expected since some hidden discriminatory behaviours emerge. For instance, the counterfactuals belonging to the *unprivileged* category, i.e., *female* or *others*, have a much higher CFlips than counterfactuals of *privileged* samples, i.e., *male* or *white*. This high percentage of flips for the unprivileged category means that the counterfactuals for the female (and "others" race) group must show male (and white) characteristics to get a positive decision. In this respect, Adult and Crime are characterized by the highest CFlips values and the largest difference between the *privileged* and *unprivileged* groups (see Table 5.6). We underline that CFlips and ΔCFlips results complement, they explain (e.g., highlighting if privileged group characteristics lead to a positive outcome) but also overturn the DEO metric in Table 5.5, shedding light on how the discriminatory classifiers work.

- The debiasing models perform particularly well (see the DEO metric in Table 5.5) on datasets where sensitive features can be easily identified, i.e., the datasets characterized by accurate sensitive-feature classifiers. Notwithstanding, the ΔCFlips in Table 5.6 highlights that a certain degree of discrimination persists for the datasets i) with features with a low correlation with the sensitive features or ii) composed by just a few samples. For instance, for the Crime dataset, even though all the sensitive-feature classifiers exhibit high accuracy, only FairC succeeds in decreasing discrimination at the expense of a significant loss in accuracy.

- The Adult-debiased dataset shows a smaller number of CFlips than Adult, especially for LR and SVM. However, even the most accurate XGB and LGBM show an evident discrepancy between the *privileged* and *unprivileged* groups. This might indicate that both classifiers learned correlations between the proxy features and the target. The German dataset has a similar trend with MLP and SVM as the most affected by unfairness. Finally, German's small test set size and the low accuracy of XGB as $f_{S_i}(\cdot)$ drive MLP, and LFERM to have 0 flips in the *female* category.

- Figure 5.2 analyzes the impact of the number of generated counterfactuals and the validity of the metric ΔCFlips when different sensitive features are present in the same dataset. The figure shows that the values of ΔCFlips are stable and reliable if the metric is computed on at least 20 counterfactuals for each sample. A different behaviour can be observed for LFERM in Adult-gender and FairC in Crime-race that start with an optimal performance, and then their ΔCFlips increases linearly. This trend needs further investigation, but it could be related to the higher number of counterfactuals. Indeed, with several counterfactuals, some could be farther from the original sample, and this distance probably entails a higher ΔCFlips. However, in fairness research, measuring the distance of a sample from the decision boundary of the classifier is a timely challenge [77]. It is worth mentioning that the analysis can be conducted for different features on the same dataset, even when it contains more than one sensitive feature (see *gender* and *maritalStatus* plots for the Adult dataset).

- The counterfactual generation strategies reveal some important findings: there exist similar real samples (i.e., similar people) for which the switch to the privileged group led to a positive outcome. Indeed, KDtree searches among dataset samples – i.e., $\mathbf{c_x} \in \mathcal{D}$ – thus CFlips ≥ 0 is critical. Genetic strategy, instead, analyzes the unexplored space and confirms the discriminatory behaviour.

Table 5.6 (GENETIC) CFlip and nDCCF results at different $|k|$ number of Counterfactuals for each negatively predicted Test set sample (0^* there are no negative predicted *unprivileged* samples which result in no CF samples for the unprivileged group). We mark the best-performing model for each fairness metric in bold font.

			CFlips@$\|k\|$ (%)									nDCCF@$\|k\|$								
			Privileged			Unprivileged			ΔCFlips ↓			Privileged			Unprivileged			ΔnDCCF ↓		
Dataset	S_i	model	@10	@50	@100	@10	@50	@100	@10	@50	@100	@10	@50	@100	@10	@50	@100	@10	@50	@100
Adult	*gender*	**LR**	12.332	10.886	10.212	66.353	72.932	77.165	54.021	62.046	66.953	0.8678	0.8849	0.886	0.3522	0.2913	0.2497	0.5156	0.5936	0.6363
		DT	8.721	9.442	9.563	67.553	73.179	74.152	58.832	63.737	64.589	0.911	0.9067	0.8988	0.3371	0.284	0.2685	0.5739	0.6227	0.6303
		SVM	6.752	7.533	7.742	77.095	80.973	81.372	70.343	73.44	73.63	0.9306	0.9258	0.9171	0.2474	0.2042	0.1948	0.6832	0.7216	0.7223
		LGBM	9.195	8.541	8.781	65.918	76.605	79.697	56.723	68.064	70.916	0.9049	0.9124	0.9049	0.3611	0.2633	0.2272	0.5438	0.6491	0.6777
		XGB	10.011	8.788	9.07	64.796	76.243	79.512	54.785	67.455	70.442	0.8968	0.9088	0.9014	0.3708	0.2677	0.2298	0.526	0.6411	0.6716
		RF	7.18	7.226	7.577	68.926	77.217	80.578	61.746	69.991	73.001	0.9246	0.9269	0.9181	0.3296	0.2527	0.2164	0.595	0.6742	0.7017
		MLP	8.787	8.53	9.135	68.991	78.262	80.487	60.204	69.732	71.352	0.9071	0.9129	0.9025	0.3355	0.2453	0.2163	0.5716	0.6676	0.6862
		ADV	30.046	34.488	34.968	36.11	38.694	43.041	6.064	**4.206**	**8.073**	0.7016	0.6668	0.6537	0.6427	0.6199	0.5812	0.0589	**0.0469**	**0.0725**
		LFERM	31.459	28.632	24.965	31.764	47.464	57.47	**0.305**	18.832	32.505	0.6857	0.7062	0.7314	0.6864	0.5632	0.4701	**0.0007**	0.143	0.2613
		FairC	58.841	60.464	56.68	17.891	22.141	27.338	40.95	38.323	29.342	0.4135	0.3981	0.4219	0.8238	0.789	0.7415	0.4103	0.3909	0.3196
AdultDeb	*gender*	**LR**	8.438	10.838	13.192	54.816	57.521	57.047	46.378	46.683	43.855	0.9239	0.9012	0.8736	0.464	0.4332	0.4303	0.4599	0.468	0.4433
		DT	6.334	14.796	17.298	31.092	49.802	54.639	24.758	35.006	37.341	0.9451	0.8723	0.8398	0.7224	0.5539	0.4936	0.2227	0.3184	0.3462
		SVM	11.937	16.377	17.379	31.305	33.869	35.385	**19.368**	**17.492**	**18.006**	0.8871	0.8468	0.8295	0.6661	0.6616	0.6449	**0.221**	**0.1852**	**0.1846**
		LGBM	4.596	9.384	12.817	66.779	74.088	73.366	62.183	64.704	60.549	0.958	0.9185	0.8818	0.3744	0.2879	0.2804	0.5836	0.6306	0.6014
		XGB	1.803	3.152	6.523	81.289	88.9	84.48	79.486	85.748	77.957	0.9804	0.9711	0.9386	0.2183	0.1378	0.1599	0.7621	0.8333	0.7787
		RF	3.616	6.067	8.473	71.126	81.269	81.498	67.51	75.202	73.025	0.9652	0.9452	0.9186	0.31	0.2151	0.201	0.6552	0.7301	0.7176
		MLP	0.402	0.935	1.934	92.918	96.854	96.337	92.516	95.919	94.403	0.9951	0.9917	0.9766	0.0889	0.0452	0.0435	0.9062	0.9465	0.9331
		ADV	16.369	20.067	23.123	44.722	52.645	57.043	28.353	32.578	33.92	0.8493	0.8119	0.7787	0.5803	0.4998	0.4536	0.269	0.3121	0.3251
		LFERM	8.943	13.316	16.561	47.036	54.87	55.83	38.093	41.554	39.269	0.9248	0.8809	0.8452	0.5618	0.4791	0.4584	0.363	0.4018	0.3868
		FairC	1.326	2.723	4.359	80.127	85.728	88.23	78.801	83.005	83.871	0.9864	0.976	0.9556	0.1921	0.1533	0.1293	0.7943	0.8227	0.8263
Crime	*race*	**LR**	2.857	3.429	3.667	75.286	81.943	85.143	72.429	78.514	81.476	0.9688	0.9656	0.9568	0.2659	0.2011	0.1678	0.7029	0.7645	0.789
		DT	7.2	6	6.32	65.211	75.239	79.254	58.011	69.239	72.934	0.9258	0.9376	0.9289	0.3648	0.2738	0.2321	0.561	0.6638	0.6968
		SVM	6.25	5.917	5.63	73.239	80.789	84.493	66.989	74.872	78.863	0.938	0.94	0.9359	0.2868	0.2149	0.1776	0.6512	0.7251	0.7583
		LGBM	5.652	5.913	5.696	74.571	80.143	83.55	68.919	74.23	77.854	0.9424	0.9407	0.9357	0.2875	0.2215	0.1854	0.6549	0.7192	0.7503
		XGB	5	5.455	5.045	73.38	80.141	83.613	68.38	74.686	78.568	0.9492	0.9467	0.9427	0.2938	0.2214	0.1851	0.6554	0.7253	0.7576
		RF	8.261	7.478	7.652	65.417	73.972	77.918	57.156	66.494	70.266	0.9181	0.9245	0.9171	0.3813	0.2907	0.2477	0.5368	0.6338	0.6694
		MLP	5.263	6.211	6.053	73.562	79.014	82.452	68.299	72.803	76.399	0.9509	0.9401	0.9337	0.2899	0.231	0.1953	0.661	0.7091	0.7384
		ADV	7.576	7.03	7.273	69	77.571	80.643	61.424	70.541	73.37	0.9295	0.9295	0.9209	0.3396	0.2519	0.2164	0.5899	0.6776	0.7045
		LFERM	3.913	6.174	6.696	64.412	71.588	75.103	60.499	65.414	68.407	0.9592	0.9408	0.929	0.3695	0.305	0.2686	0.5897	0.6358	0.6604
		FairC	23.256	22.093	22.289	30.769	41.731	49.218	**7.513**	**19.638**	**26.929**	0.7654	0.7742	0.7689	0.7248	0.6189	0.5455	**0.0406**	**0.1553**	**0.2234**
German	*gender*	**LR**	31.364	35.455	39.081	27.500	27.500	27.250	3.864	7.955	11.831	0.6977	0.6582	0.6228	0.7302	0.7306	0.7267	0.0325	**0.0724**	0.1039
		DT	27.273	26.727	29.455	33.333	48.667	51.333	6.06	21.94	21.878	0.7486	0.7391	0.71	0.7246	0.5633	0.5217	0.024	0.1758	0.1883
		SVM	22.5	31	32	30.000	70.000	70.000	7.5	39	38	0.8009	0.7185	0.6944	0.7722	0.4044	0.3569	0.0287	0.3141	0.3375
		LGBM	3.333	4.333	4.25	17.500	16.500	17.000	14.167	12.167	12.75	0.9706	0.9598	0.9528	0.8383	0.8379	0.8273	0.1323	0.1219	0.1255
		XGB	30	35.091	36.727	26.667	25.000	28.667	**3.333**	10.091	**8.06**	0.6891	0.6563	0.6364	0.7945	0.7686	0.7271	0.1054	0.1123	**0.0907**
		RF	32	33.2	35.046	40.000	46.000	47.303	8	12.8	12.257	0.6957	0.6751	0.6537	0.6779	0.5816	0.5530	**0.0178**	0.0935	0.1007
		MLP	23	29.4	30.097	0^*	0^*	0^*	23	29.4	30.097	0.7726	0.7184	0.7039	0^*	0^*	0^*	0.7726	0.7184	0.7039
		ADV	5	7.667	6.583	15.714	14.857	17.286	10.714	**7.19**	10.703	0.9573	0.9315	0.9295	0.8299	0.8458	0.8243	0.1274	0.0857	0.1052
		LFERM	15	19	24	0^*	0^*	0^*	15	19	24	0.8408	0.8156	0.7676	0^*	0^*	0^*	0.8408	0.8156	0.7676
		FairC	24	27	29.019	0.000	4.000	5.000	24	23	24.019	0.7679	0.734	0.7124	1.0000	0.9705	0.9505	0.2321	0.2365	0.2381

- Considering the CFlips metric we can notice that for the *unprivileged* group we generally have higher values. This indicates that samples belonging to the unprivileged group need to take the characteristics of the privileged group to reach a positive outcome (i.e., $f(\mathbf{c_x}) = 1$). This is confirmed by the nDCCF metric where we can see how the (most similar) counterfactuals of the unprivileged group should take on the characteristics of the privileged group to pass to a favourable prediction.

- From a group-based point of view, we are interested in evaluating the difference between the two proposed metrics between the privileged and unprivileged group (i.e., ΔCFlips and ΔnDCCF) and see for which model the value of the Δ is near to zero. We point out that for models with highly accurate sensitive feature classifiers (i.e., Adult and Crime), debiasing models seem to perform best in Fairness and thus succeed in fulfilling their debiasing task. This is confirmed by both standard

Table 5.7 (KDtree) CFlip and nDCCF results at different $|k|$ number of Counterfactuals for each negatively predicted Test set sample, (0* there are no negative predicted *unprivileged* samples which result in no CF samples for the unprivileged group). We mark the best-performing model for each fairness metric in bold font.

			CFlips@$\|k\|$ (%)									nDCCF@$\|k\|$								
			Privileged			Unprivileged			ΔCFlips ↓			Privileged			Unprivileged			ΔnDCCF ↓		
Dataset	S_i	model	@10	@50	@100	@10	@50	@100	@10	@50	@100	@10	@50	@100	@10	@50	@100	@10	@50	@100
Adult	*gender*	**LR**	7.754	8.584	9.11	81.273	84.21	85.137	73.519	75.626	76.027	0.9229	0.9162	0.9049	0.2053	0.1701	0.1573	0.7176	0.7461	0.7476
		DT	6.971	6.927	7.371	76.419	82.602	84.523	69.448	75.675	77.152	0.9296	0.9305	0.9207	0.2567	0.1941	0.1707	0.6729	0.7364	0.75
		SVM	6.306	6.427	7.158	80.971	85.209	86.517	74.665	78.782	79.359	0.9343	0.9351	0.9232	0.2088	0.1632	0.1465	0.7255	0.7719	0.7767
		LGBM	6.55	7.077	7.663	82.009	84.902	86.064	75.459	77.825	78.401	0.9352	0.9307	0.9192	0.1889	0.1602	0.1469	0.7463	0.7705	0.7723
		XGB	6.761	7.14	7.666	81.677	84.796	86.076	74.916	77.656	78.41	0.9334	0.9299	0.9189	0.192	0.1616	0.1472	0.7414	0.7683	0.7717
		RF	6.44	7.578	8.076	80.218	83.325	84.767	73.778	75.747	76.691	0.9362	0.9268	0.9155	0.2161	0.1795	0.1627	0.7201	0.7473	0.7528
		MLP	7.153	7.464	7.94	81.455	84.484	85.888	74.302	77.02	77.948	0.93	0.9266	0.9161	0.1965	0.1654	0.1498	0.7335	0.7612	0.7663
		ADV	25.411	18.854	17.847	52.22	66.941	71.783	26.809	48.087	53.936	0.7368	0.7946	0.8029	0.5015	0.37	0.3159	0.2353	0.4246	0.487
		LFERM	15.141	11.82	11.882	66.288	77.278	80.16	51.147	65.458	68.278	0.8385	0.8715	0.8689	0.3719	0.262	0.2255	0.4666	0.6095	0.6434
		FairC	37.337	27.402	25.323	37.832	55.442	61.712	**0.495**	**28.04**	**36.389**	0.6137	0.7005	0.7208	0.6442	0.4916	0.4238	**0.0305**	**0.2089**	**0.297**
AdultDeb	*gender*	**LR**	17.578	19.789	20.026	42.592	44.051	45.165	25.014	24.262	25.139	0.8422	0.8066	0.8018	0.5742	0.5612	0.5473	0.268	0.2454	0.2545
		DT	25.93	42.627	41.962	31.724	44.805	55.245	**5.794**	**2.178**	**13.283**	0.779	0.6248	0.6037	0.6637	0.5775	0.4877	**0.1153**	**0.0473**	**0.116**
		SVM	15.283	16.551	16.574	42.549	43.343	44.088	27.266	26.792	27.514	0.8427	0.8211	0.8184	0.5705	0.5621	0.5524	0.2722	0.259	0.266
		LGBM	9.811	14.834	15.608	45.253	69.258	75.566	35.442	54.424	59.958	0.907	0.8626	0.8457	0.568	0.3686	0.2978	0.339	0.494	0.5479
		XGB	8.955	12.959	14.183	49.229	70.872	76.837	40.274	57.913	62.654	0.9149	0.88	0.8601	0.5303	0.3479	0.2805	0.3846	0.5321	0.5796
		RF	10.354	14.39	15.675	45.517	65.841	73.207	35.163	51.451	57.532	0.9024	0.8645	0.8469	0.5564	0.3953	0.3239	0.346	0.4692	0.523
		MLP	5.241	8.797	7.754	71.167	85.204	89.1	65.926	76.407	81.346	0.9414	0.917	0.9141	0.3445	0.194	0.1534	0.5969	0.723	0.7607
		ADV	8.608	14.501	13.277	51.118	72.431	78.922	42.51	57.93	65.645	0.917	0.8695	0.8663	0.4993	0.3248	0.2613	0.4177	0.5447	0.605
		LFERM	7.028	10.736	13.216	65.974	73.594	73.763	58.946	62.858	60.547	0.933	0.9022	0.8737	0.3571	0.2864	0.2755	0.5759	0.6158	0.5982
		FairC	7.191	12.942	11.435	64.482	78.892	85.475	57.291	65.95	74.04	0.9266	0.8833	0.8827	0.4124	0.2565	0.1892	0.5142	0.6268	0.6935
Crime	*race*	**LR**	5.714	8.762	9.143	85.139	87.722	88	79.425	78.96	78.857	0.9489	0.9207	0.9077	0.1603	0.1319	0.1257	0.7886	0.7888	0.782
		DT	10.4	12.96	14.6	75.07	80.113	80.789	64.67	67.153	66.189	0.8765	0.8695	0.8512	0.258	0.2126	0.2007	0.6185	0.6569	0.6505
		SVM	9.167	10.667	11.75	82.877	86.466	86.959	73.71	75.799	75.209	0.9011	0.8937	0.8788	0.1716	0.1424	0.1351	0.7295	0.7513	0.7437
		LGBM	4.348	8.522	9.783	81.806	85.444	86.236	77.458	76.922	76.453	0.9528	0.9221	0.9033	0.1892	0.1551	0.1443	0.7636	0.767	0.759
		XGB	5.455	8.727	9.727	83.288	86.329	86.63	77.833	77.602	76.903	0.9308	0.9149	0.8992	0.1767	0.1463	0.1396	0.7541	0.7686	0.7596
		RF	10.87	13.043	14.261	78.75	81.694	82.486	67.88	68.651	68.225	0.8861	0.8736	0.857	0.2214	0.1926	0.1822	0.6647	0.681	0.6748
		MLP	4.5	8.4	10	80.972	84.389	85.347	76.472	75.989	75.347	0.9611	0.926	0.9028	0.194	0.164	0.1526	0.7671	0.762	0.7502
		ADV	8.485	9.818	10.667	81.389	83.889	84.861	72.904	74.071	74.194	0.9134	0.9034	0.8895	0.1912	0.1685	0.1577	0.7222	0.7349	0.7318
		LFERM	9.13	14	15.174	71.857	77.971	78.886	62.727	63.971	63.712	0.9119	0.8709	0.8516	0.291	0.2371	0.2224	0.6209	0.6338	0.6292
		FairC	33.023	32.512	34.186	39.808	48.885	52.577	**6.785**	**16.373**	**18.391**	0.6808	0.6787	0.6607	0.6207	0.5395	0.4981	**0.0601**	**0.1392**	**0.1626**
German	*gender*	**LR**	23.182	26.545	26.818	45.000	56.000	62.000	21.818	29.455	35.182	0.761	0.7407	0.732	0.5683	0.4618	0.4006	0.1927	0.2789	0.3314
		DT	28.182	27.818	28.273	60.000	57.333	61.333	31.818	29.515	33.06	0.7362	0.7272	0.7174	0.4351	0.4332	0.3958	0.3011	0.294	0.3216
		SVM	23.333	26.333	27.417	40.000	56.000	65.000	16.667	29.667	37.583	0.7731	0.7459	0.7295	0.6805	0.5024	0.4122	0.0926	0.2435	0.3173
		LGBM	26.667	29.167	29.417	37.500	50.000	58.250	10.833	20.833	28.833	0.7616	0.7249	0.7124	0.6265	0.5235	0.4519	0.1351	0.2014	0.2605
		XGB	22.727	27.818	27.909	46.667	53.333	60.167	23.94	25.515	32.258	0.7976	0.7424	0.7301	0.5489	0.4807	0.4211	0.2487	0.2617	0.309
		RF	23	27.6	26.6	55.000	58.000	64.000	32	30.4	37.4	0.7951	0.7414	0.7379	0.5127	0.4424	0.3874	0.2824	0.299	0.3505
		MLP	30	29.6	30	0*	0*	0*	30	29.6	30	0.726	0.7145	0.7024	0*	0*	0*	0.726	0.7145	0.7024
		ADV	28.333	30	29.833	40.000	56.286	61.429	11.667	26.286	31.596	0.74	0.7128	0.7027	0.6109	0.4703	0.4155	0.1291	0.2425	0.2872
		LFERM	20	23	23.5	0*	0*	0*	20	23	23.5	0.7702	0.7609	0.7567	0*	0*	0*	0.7702	0.7609	0.7567
		FairC	27	29.4	28.7	20.000	50.000	60.000	**7**	**20.6**	**31.3**	0.745	0.7174	0.7127	0.7760	0.5517	0.4571	**0.031**	**0.1657**	**0.2556**

fairness metrics and our new proposed metrics (ΔCFlips and ΔnDCCF). In contrast, debiasing models rarely give the best results when considering datasets with low correlation with sensitive information (i.e., Adult-debiased and German - with worst sensitive classifiers performance). As a matter of fact, for Adult-debiased, SVM performs better than other debiasing models. Indeed, ΔCFlips and ΔnDCCF, differently from standard fairness metrics, reward SVM as the best-performing model. A similar trend can be seen in German. However, since there are no correctly predicted female samples for MLP and LFERM models, we have 0 CFlips for the *unprivileged* group. Thus, the small size of the dataset made the evaluation of the metrics impracticable.

- Table 5.7 reports the results of our metrics with counterfactuals generated with the KDtree strategy. KDtree generates counterfactuals by choosing from samples that already belong to the dataset. Therefore, unlike Genetic, which generates new samples, KDtree does not explore an unknown space. This means that each counterfactual is chosen from the known data space (i.e., $\mathbf{c}_\mathbf{x} \in \mathcal{D}$), so KDtree measures how similar samples behave. We can see that the trend is similar to the Genetic strategy but with higher CFlips and nDCCF. In this case, enlarging the number of counterfactuals for each sample worsens the metric values since more distant counterfactual samples are chosen.

<u>Final comments.</u> *In the various plots emerges that the unprivileged samples, to achieve favorable decisions, must take on the characteristics of privileged samples. The results demonstrate that counterfactual reasoning effectively discovers decision biases and complements SOTA fairness metrics.*

5.4.1 Ablation

In Section 5.4, we have evaluated the performance of the models based on our proposed metrics. However, these metrics have been evaluated only with k equal to 10, 50, and 100 and only with XGB as the sensitive feature classifier. Now, we study the effect of the number of generated counterfactuals on different sensitive feature classifiers (i.e., RF and MLP). Furthermore, nDCCF can be affected by the ranking criterion of the counterfactual generator. This point deserves a broader discussion as follows.

As mentioned in Section 4.2, a counterfactual sample $\mathbf{c}_\mathbf{x}$ is a deviation from a starting vector $\mathbf{x}$ of a quantity ϵ that is computed k times (the number of counterfactuals). For each sample, using a function $g(\cdot)$, we generate a set of counterfactuals $\mathcal{C}_\mathbf{x}$ such that $g(\mathbf{x}) = \mathcal{C}_\mathbf{x}$. This set is ranked according to a model-specific utility function $u(\cdot)$ (e.g., euclidean distance or absolute distance of the counterfactual sample from the original one[8]). Indeed, $g(\mathbf{x})$ returns a set of counterfactuals such that $\mathcal{C}_\mathbf{x} = \langle \mathbf{c}_\mathbf{x}^1, \mathbf{c}_\mathbf{x}^2, \ldots, \mathbf{c}_\mathbf{x}^k \rangle$ with $u(\mathbf{c}_\mathbf{x}^i) > u(\mathbf{c}_\mathbf{x}^j)$ and $i < j$. However, the ranking of this set depends on the strategy used by the *counterfactual generator*. In this regard and to be totally agnostic from that strategy, we reranked the list of counterfactual samples based on the absolute difference between the expected model prediction of a counterfactual sample and that of the original sample (i.e., $u(\mathbf{c}_\mathbf{x}) = -|\mathbb{E}[f(\mathbf{x})] - \mathbb{E}[f(\mathbf{c}_\mathbf{x})]|$, s.t. $\forall \mathbf{c}_\mathbf{x}^i, \mathbf{c}_\mathbf{x}^j \in \mathcal{C}_\mathbf{x}$, with $i < j$, $u(\mathbf{c}_\mathbf{x}^i) > u(\mathbf{c}_\mathbf{x}^j)$). We narrow our analysis to the Adult dataset (see Figure 5.2).

- Comparing the different sensitive feature classifiers, it is evident that the metrics can be considered stable also due to the high performance of each classifier. Furthermore, considering the ΔCFlips and ΔnDCCF, we can notice that for each Decision Maker,

[8] A counterfactual that is closer to the original sample has greater utility than one further away.

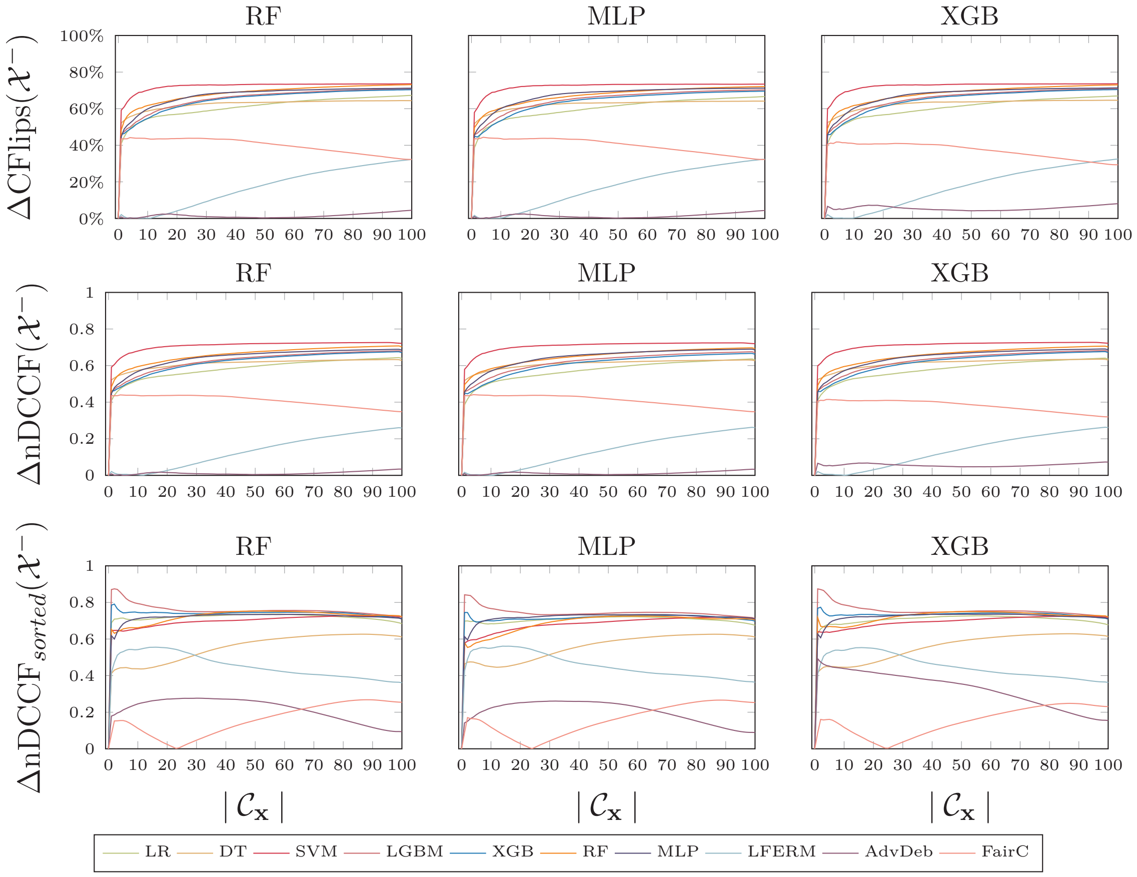


Fig. 5.2 Ablation study at a different number of generated CF (i.e. $|\mathcal{C}_\mathbf{x}|$) for each sample and with three different sensitive feature classifiers $f_{S_i}(\cdot)$ (i.e., RF, MLP, and XGB). The result refers to Adult dataset, with gender as sensitive information and Genetic as counterfactual generation strategy.

we reach a stable result after 20 generated counterfactuals except for LFERM. LFERM seems to increase the value of each metric by enlarging the number of counterfactual samples. Investigating motivations from a distance perspective may be a viable option, but it is a current challenge and limitation of fairness research [77]. Another interesting point is the similarity between the two metrics. It seems that the trend of the ΔCFlips with the increasing of $|\mathcal{C}_\mathbf{x}|$ is consistent with the one of ΔnDCCF.

- Considering the sorted version of ΔnDCCF, we can observe three different trends: (i) the Decision Maker has a behaviour similar to the sorted version meaning that the discrimination is also in the proximity of the positive decision boundary side, (ii) models like LFERM and FairC, instead, have an opposite behaviour in the proximity of the positive decision boundary, and (iii) AdvDeb starts with greater discrimination for counterfactual samples closer to the decision boundary and then becomes fairer with the distance increasing.

Final comments. *The findings show that counterfactual reasoning is appropriate for identifying decision-making biases and enhancing SOTA fairness indicators. Studying classifiers' decision boundaries can also reveal further information about the discrimination behaviour of the model.*

Chapter 6

An Explainable and Responsible Pipeline grounded in Counterfactual Reasoning

This section underscores the critical role of Explainable AI (XAI) and Counterfactual Reasoning in the context of loan decision-making, focusing on the detection of proxy features and the provision of transparent explanations. As AI systems increasingly play a pivotal role in financial services, the need for transparency and fairness in decision-making becomes paramount. The detection of proxy features is essential in identifying and mitigating biased decision outcomes, ensuring equitable access to credit and adhering to anti-discrimination regulations. Explainable AI techniques contribute to making these proxy feature detection mechanisms accessible to both stakeholders and end-users.

In this context, Counterfactual Reasoning emerges as a powerful tool for explaining loan decisions. By generating counterfactuals, one can highlight the specific factors that led to an adverse or favourable loan outcome, offering a clear and interpretable narrative for stakeholders and applicants. These counterfactual explanations not only aid in building trust but also provide actionable insights to improve financial literacy and empower individuals to make more informed financial decisions.

The importance of this research lies in its potential to enhance fairness, accountability, and transparency in loan decision-making processes. By combining Explainable AI and Counterfactual Reasoning, this study contributes to the ongoing efforts to address the challenges of algorithmic bias, ensuring that loan decisions are based on justifiable and non-discriminatory criteria. Furthermore, it advances our understanding of the role of XAI and counterfactual explanations in promoting responsible AI adoption in the financial sector, ultimately fostering trust and equity among all stakeholders.

6.1 Counterfactual Explanation in Machine Learning

In logic and philosophy, a counterfactual is an event that did not actually occur and is, therefore, only assumed or imagined. The Counterfactual explanation is another type of explanation technique that, since it does not require opening the black box, can be considered as a *model-agnostic explanator* based on example. A counterfactual explanation describes a causal situation in the form: "*If X had not occurred, Y would not have occurred*". For example, in a loan application case, following the example of Watcher et al. [165] if the user's loan application has been rejected for inadequate income, a counterfactual explanation can be as *"You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan"*. Multiple counterfactuals are possible since multiple desirable outcomes can be achievable in different ways. However, the counterfactual is considered as *"the smallest change to the feature values that changes the prediction to a predefined output"*.

According to McGrath et al. [121], a counterfactual explanation can be described as a generic causal situation:

> *"Score y was returned because variables x had values* $(x_1, x_2, ...)$ *associated with them. If x instead had values* $(x'_1, x'_2, ...)$*, and all other variables had remained constant, score* y' *would have been returned."*

The counterfactual explanation does not depend on the type of the model, so every model f can be treated as a black box. The Counterfactual explanation is generated by calculating the smallest possible change (Δx) that can be made to the input x to flip the outcome from y to y'.

As proposed by Wachter et al. [165], the optimization of the loss function $\mathcal{L}$ is formulated as:

$$\mathcal{L}(x, x', y', \lambda) = \lambda(\hat{f}(x') - y')^2 + d(x, x')$$

$$\arg\min_{x'} \max_{\lambda} \mathcal{L}(x, x', y', \lambda)$$

where x is the actual input feature vector, x' is the changed and counterfactual vector, y' is the desired output, $\hat{f}(...)$ is the trained model, λ is a weight balancing the counterfactual between racing the exact desired output and making the smallest possible change to the input vector x, and $d(x, x')$ is the distance metric that measures Δx. The distance metric $d(x, x')$ can be written as:

$$d(x,x') = \sum_{j=1}^{p} \frac{\mid x_j - x'_j \mid}{MAD_j}$$

with MAD

$$MAD_j = median_{i \in \{1,\ldots,n\}} \left(\left| x_{i,j} - median_{l \in \{1,\ldots,2\}}(x_{l,j}) \right| \right),$$

The work of McGrath et al. [121] suggests two different uses of counterfactual explanations in the Loan application scenario. The first one is the *Positive Counterfactuals*, which is interpreted as a *safety margin* from the opposite decision boundary to consider in case of acceptance. It can also respond to the question *"How much was I accepted by?* and is considered as tolerance or better knowledge for future loan applications. The second one is the *Weighted Counterfactuals* that consider each feature differently; for example, there is a case in which some feature must be fixed or immutable. It can be divided into two different strategies: Global feature importance and the Nearest Neighbors approach. With Global feature importance, they use the analysis of variance (ANOVA F-values) between each feature and target to obtain a smaller set of features for the explanation by creating a weight vector that promotes highly discriminative features. The K-Nearest Neighbors find cases that are near the sample considered, but that achieve the desired results.

Hashemi et al. [98], in their work, propose a model criticism and explanation framework based on adversarial generated counterfactual examples. PermuteAtack, the proposed algorithm, uses a gradient-free optimisation based on genetic algorithms. The algorithm generates sensible and realistic counterfactual examples using permutation as adversarial perturbations. The perturbations keep the range, and the distribution of each feature the same as the training set data. Counterfactual samples using Adversarial perturbations seem to be useful to increase the model behaviour understanding and offer helpful feedback to the customers to improve the credit scores.

Table 6.1 Demonstrative example of ρ computation based on $\mathcal{E}$ and Δ for a numerical, ordinal, and categorical feature of the Adult-debiased dataset.

	numeric	ordinal	Category (*workclass*)				*gender*
	capital gain	*education-num*	*Private*	*Public*	*Unemployed*		$\mathbb{E}[f_s(\mathbf{x})\|\mathbf{x}]$
$\mathbf{c}_{\mathbf{x}_1}$	5000	6	1	0	0	$f_s(\mathbf{c}_{\mathbf{x}_1})$	0.7
$\mathbf{x}_1$	2000	2	0	0	1	$f_s(\mathbf{x}_1)$	0.1
$\epsilon_{\mathbf{x_1}}$	3000	4	1	0	-1	$\delta_{\mathbf{x}_1}$	0.6
$\mathbf{c}_{\mathbf{x}_2}$	2800	5	0	1	0	$f_s(\mathbf{c}_{\mathbf{x}_2})$	0.3
$\mathbf{x}_2$	600	5	1	0	0	$f_s(\mathbf{x}_2)$	0.7
$\epsilon_{\mathbf{x_2}}$	2200	0	-1	1	0	$\delta_{\mathbf{x}_2}$	-0.4
		...					
$\epsilon_{\mathbf{x_3}}$	1200	-1	0	1	-1	$\delta_{\mathbf{x}_3}$	-0.6
$\rho(\mathcal{E},\Delta)$	0.91	0.93	0.78	-0.99	-0.36		

6.2 Detecting Proxy Features Through Counterfactuals Reasoning (RQ4)

In RQ1, we highlighted how it is possible to determine if a dataset contains proxy features. Here, we define a strategy to identify them in the dataset. Specifically, we want to identify the features that drive the Decision Maker to a positive outcome and that, at the same time, lead to a Flip in the sensitive information. Thus, the following section aims to answer the following research question:

RQ4

Is it possible to define a strategy for identifying the proxy features?

To answer the mentioned question, we propose a methodology grounded in counterfactual reasoning. Thus, we propose to study the relationship that can occur between feature change and sensitive feature classifier flip exploited in Chapter 5. Thus, we investigate the Pearson correlation between the perturbation ϵ (i.e., $\epsilon = \mathbf{c}_\mathbf{x} - \mathbf{x}$) and the distance δ as the difference between the posterior conditional probability of predicting a counterfactual sample and the original sample as belonging to the privileged group (i.e., $\delta = \mathbb{P}[f_s(\mathbf{c}_\mathbf{x}) = 1|\mathbf{c}_\mathbf{x}] - \mathbb{P}[f_s(\mathbf{x}) = 1|\mathbf{x}]$). For a numerical or ordinal feature i, ϵ_i can be expressed as the difference between the counterfactual and the feature of the sample $c_{x_i} - x_i$. For a categorical feature j, ϵ_j can be expressed in a *one-hot encoding* form as -1 to the category that is removed and 1 to the category that is engaged. We identify with $\mathcal{E} = \{\epsilon_i|\forall \mathbf{x}_i \in \mathcal{X}^-\}$ and $\Delta = \{\delta_i|\forall \mathbf{x}_i \in \mathcal{X}^-\}$, respectively, the set of all perturbations of $\mathcal{C}_\mathbf{x}$ and the difference between all conditional probability that we can

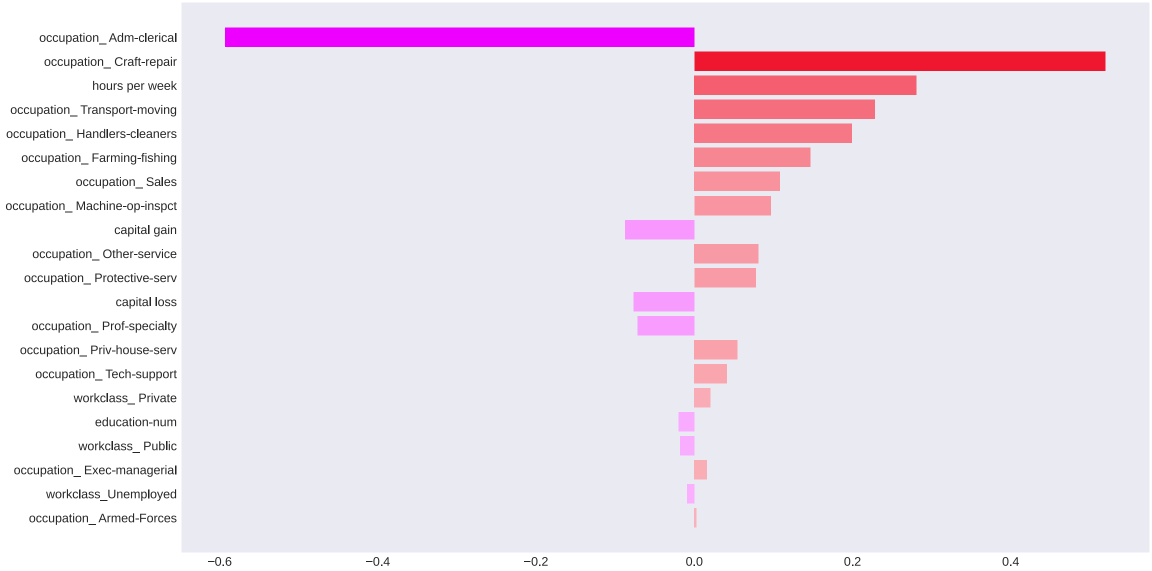


Fig. 6.1 Features correlation rank with a *gender* Flip (i.e., $\rho(\mathcal{E},\Delta)$ on the x-axis) on the Adult-debiased dataset with Genetic strategy as $g(\cdot)$, MLP as $f(\cdot)$ and XGB $f_s(\cdot)$ for only sample, and, thus, counterfactuals, belonging to $\mathcal{X}^-$.

express with the expected values of $\mathbb{E}[f_s(\mathbf{c_x})]$ and $\mathbb{E}[f_s(\mathbf{x})]$, in both cases $\forall \mathbf{x} \in \mathcal{X}^-$. Thus, it is possible to identify the most influential features for $f_s(\cdot)$ evaluating the Pearson correlation between $\mathcal{E}$ and Δ: $\rho(\mathcal{E},\Delta)$ (a demonstrative example in Table 6.1).

- In Figure 6.1 we can find the rank of features correlation with a Flip in $f_s(\cdot)$ with MLP as $f(\cdot)$ decision boundary for the generation of $\mathbf{c_x}$ and XGB as $f_s(\cdot)$ for the Adult-debiased dataset. The analysis is restricted to only samples negatively predicted in order to specifically quantify the *proxy-features* that lead to a positive prediction with also a change in the sensitive information. In detail, a negatively correlated feature (e.g., *Adm-Clerical*) is a feature that has an opposite direction with respect to $\mathbb{E}[f_s(\mathbf{x}) \mid f(\mathbf{x}=0)]$ while a positively correlated one (e.g., *Craft-repair*) has the same direction.

- For a binary classification task a negative correlation between $\mathcal{E}$ and Δ can be considered a positive correlation w.r.t. $-\Delta$. Considering a single δ, each conditional probability of the counterfactual sample, as well as of the dataset sample, can be rewritten as $\mathbb{P}[f_s(\mathbf{c_x}) = 1|\mathbf{c_x}] = 1 - \mathbb{P}[f_s(\mathbf{c_x}) = 0]$. This means that $\delta = (1 - \mathbb{P}[f_s(\mathbf{c_x}) = 0]) - (1 - \mathbb{P}[f_s(\mathbf{x}) = 0])$. It is easy to demonstrate that $-\delta = \mathbb{P}[f_s(\mathbf{c_x}) = 0] - \mathbb{P}[f_s(\mathbf{x}) = 0]$ where the class 0 in the second term correspond to the opposite class. For instance, in Figure 6.1 the ones negatively correlated with the privileged group are positively correlated with the unprivileged group. This shows how our approach is straightforward, effective, and flexible.

Final comments. *Counterfactual reasoning not only can accurately detect and quantify biases in the decision process but also can quantify the contribution of each feature with a positive, or negative, outcome.*

6.3 Explaining a Loan Decision

Several definitions are provided in the literature on what *explainable* means when we talk about a Machine Learning algorithm. The most relevant one for our purpose is provided by Bracke et al. [21] *"explanations can answer different kinds of questions about a model's operation depending on the stakeholder they are addressed to"*. This definition introduces an interesting characteristic of the explanation that has to consider the point of view of a specific stakeholder. Thus, the following section aims to answer the following research questions:

RQ5

Can counterfactual reasoning be useful to explain a loan decision?

RQ6

What is the most suitable explanation strategy depending on each stakeholder in the loan domain?

Accordingly, in a credit score scenario, for example, the explanation for a given decision might be different if addressed to customers rather than to the risk management functions. From the customer's point of view, which is the most interesting in our analysis, the explanation should describe the motivations behind a decision in a way that is easy to understand. Naturally, as mentioned above, the decisions are made by algorithms. Thus, it is crucial to know how these algorithms work. The ML algorithms belong to two main classes: interpretable and uninterpretable. More specifically, the former implements a *white-box* model, the latter a *black-box* one. On this perspective, Sharma et al. [155] distinguish *model-agnostic* and *model-specific* explanations. Model-agnostic methods provide an explanation that is not dependent on the ML model adopted and are geneally used for *black-box* models. A *surrogate* model is thus implemented with the aim of *simulating* the behaviour of the original algorithm.

Today, explaining how a black box model works is still a challenging task. However, several methods have been proposed to explain black-box models. Two of the most

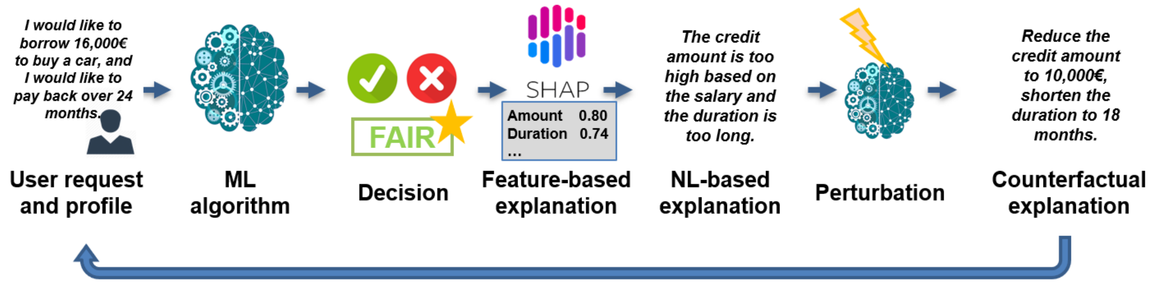

I would like to borrow 16,000€ to buy a car, and I would like to pay back over 24 months.
User request and profile
ML algorithm
FAIR
Decision
SHAP
Amount 0.80
Duration 0.74
...
Feature-based explanation
The credit amount is too high based on the salary and the duration is too long.
NL-based explanation
Perturbation
Reduce the credit amount to 10,000€, shorten the duration to 18 months.
Counterfactual explanation

Fig. 6.2 Workflow for generating explanation and counterfactual explanation for loan application

important are LIME and SHAP. LIME trains local surrogate models explaining single data[146]. It generates a perturbation of initial data, creating a new dataset and observing how the prediction changes through training an interpretable model. The analysis of the outcome of the perturbated data allows us to interpret the original model. SHAP [115] is inspired by the cooperative game theory based on the Shapley Values. Each feature is considered a player that contributes differently to the outcome (i.e., the algorithm decision).

SHAP does not compute all the possible combinations between all the features but performs only a random set of combinations for efficiency constraints. SHAP provides a ranked list of the features that contributed to the outcome ordered from the most to the least important. However, this explanation probably is not so clear for a customer who does not have experience with how an algorithm works. For this reason, if we want to improve the user's trust and general user experience with the system, we need to make the explanation more understandable.

In that direction, we guess that an effective solution could be to transform the output produced by software like LIME or SHAP into a natural language sentence. We propose the pipeline described in Figure 6.2. Customer characteristics are the input, and then the algorithm makes a decision, e.g. the computation of the CS, and shows, using SHAP, the features that contributed the most to the decision. At this point, another module takes as input the decision and the SHAP output and generates a natural-language explanation: e.g. *Dear Giulio, your loan application has been rejected since you don't have an account with us, the credit amount you asked for is too high compared to your income, and the duration is too long.*

An interesting opportunity in this context could be provided by a counterfactual explanation that explains how the output of the algorithm could be changed [157]. For example, the system can add: *In the case you decide to open an account with us, to reduce the credit amount to 10,000$, and to reduce the duration to 12 months,* *the application will probably be accepted.* Conversely, model-specific explanations are based on the analysis of the structural information and the internal components of the algorithm that should be interpretable natively. From a technical perspective, these algorithms are easier to explain, but in this case, as well, most users will not be able to understand them. Therefore, the scenario is quite similar to the previous one, and here, the exploitation of natural language can improve comprehensibility.

6.3.1 A Recommendation Lending case (RQ5)

A further interesting contribution in this direction is provided by a counterfactual analysis obtained by a feature perturbation step (see Section 6.3). This explanation shows how to modify the loan request to get the loan accepted [157]. For example, the system can add: *Reduce the credit amount to 10,000€, shorten the duration to 18 months, . . . , and the loan request will probably be accepted.*

However, how can we generate this kind of natural language explanation? In the next section, we propose a template-based formal model able to transform the SHAP values into a natural language sentence. The model we designed for generating Natural Language explanations is inspired by Musto et al. [130].

The principal insight is that our natural language explanation can be generated by exploiting a template composed of some slots that can be filled with features, adverbs, and adjectives according to the output produced by SHAP. We remember that the SHAP output consists of a set of couples *<feature, score>* (e.g., <income, 0.8>).

Let us consider the example in Figure 6.2: *The credit amount is too high based on the salary and the duration is too long.* In that case, the template for the explanation is *<feature> <verb> <adverb> <adjective> <motivation>* followed by a new set of *<feature> <verb> <adverb> adjective>* without motivation. The problem is to properly fill each slot and compose the whole explanation.

In the above-mentioned example, the number of features taken into account for generating the explanation are three: *the credit amount*, *the salary*, and *the duration* each of which is associated with adverbs and/or adjectives (e.g., too high, too long, etc.). The number of features used for generating the explanation can be set as desired. However, since the explanation has to be as useful as possible, too many features can, in some cases, cause a loss of effectiveness and efficiency.

In our model, the generation of the natural language explanation exploits a set of rewriting rules using the Back-Naur Form (BNF) as described in the following. Even though these templates and rules can be exploited also in other domains, the terminal symbols (e.g., the credit amount, the duration, long, short, etc.) are specific to a loan application.

```
<explanation> ::= <sentence> | <explanation> <conjunction> <sentence>
<statement> ::= <feature> <verb> <adverb> <adjective>
<sentence> ::= <statement> <motivation>
<motivation> ::= <motivation> <conjunction> <motivation>
<motivation> ::= <adverbial phrase> <feature>
<adverbial phrase> ::= ‘based on’ | (etc.)
<adverb> ::= ‘too’ | ’so’ | ’few’ | ’almost’ | ’enough’ (etc.)
<adjective> ::= ‘high’ | ’long’ | ’short’ | ’little’ | (etc.)
<conjunction> ::= ‘and’ | ’but’ | , |(etc.)
<feature> ::= ‘the credit amount’ | ’the duration’ | ’the salary’ | (etc.)
<verb> ::= ‘is’ | ’are’ | ’has’ | ’have’ | ’is not’| (etc.)
```

These rewriting rules can be applied for generating, for example, the explanation *The credit amount is too high based on the salary and the duration is too long.*

A further problem is the choice of adverbs and adjectives. For the adverbs, we defined a matching between value intervals and the *intensity* of the adverb. As an example, if the SHAP value of a feature is 0.8 (the highest interval), the corresponding *<adverb>* will be ’too’ emphasizing how this feature has a strong impact on the loan application decision. Obviously, the association between the *<feature>* and the type of *<adjective>* is not arbitrary, but it depends on the type of *<feature>* is considered. Therefore, for each feature, we defined a vocabulary of compatible adjectives.

6.3.2 Counterfactual explanation

In the previous subsection, we have described how a loan recommendation platform can generate an explanation for each decision given by a provider.

To make our explanation more effective, we propose to the user some indications useful for revising her request and getting the loan application accepted. This is obtained through a *counterfactual explanation.*

The counterfactual explanation consists of a set of corrective actions to the characteristics of the requested loan, based on the results of a counterfactual analysis. Providing a counterfactual explanation is an opportunity for the loan provider that provide an additional service to enhance customer satisfaction and make the customer aware of his or her chances of getting a loan. This service will result in a Responsible and Trustworthy use of AI systems towards customers.

The counterfactual analysis performs a *perturbation* on the feature space of the customer's loan application. The perturbation will generate a new sample that will be considered as a new application. Subsequently, the counterfactual analysis will detect the new nearest sample to the original one that the ML algorithm will accept. The result of this analysis will consist of detecting the change in the loan's customer characteristics and recommending corrective actions.

The approach we adopted for generating the counterfactual explanation is the same one described in the previous section, namely a set of BNF rewriting rules.

Following the previous example, a counterfactual explanation can be: *"Reduce the credit amount to 10,000€, shorten the duration to 18 months"*.
The BNF template is:

<counterfactualexplanation>::= <sentence>|<counterfactualexplanation>
<conjunction> <counterfactualexplanation>
<sentence>::= <action><feature><value>
<action> ::= 'reduce'|'expand'|'shorten'|etc.
<feature> ::= 'the credit amount'|'the duration'|etc.
<value> ::= '10,000€'|'18 months'|
<conjunction> ::= 'and' | 'but' | , |(etc.)

The counterfactual explanation has a small set of rules, in fact, it includes a feature, the corrective actions, and optionally the desirable new feature value. Since the counterfactual analysis works by perturbing all the features of a determined instance, the recommended actions should impact the minimum set of features that allow a change of the algorithm decision.

The action is chosen according to the relation between the old and the new feature value. For example, if the old value for the feature *duration* was 24 and the new value after the perturbation is 18, the verb (action) chosen will be *reduce*. Regarding the values, if the new value is equal to the original one, the respective feature will not be included in the explanation since there is no corrective action to be done, otherwise, the new perturbed value will be shown in the explanation.

6.3.3 A General Responsible Pipeline for the Credit Score and Credit Behavior Domain (RQ6)

In the last few years, the research has been focused more on applying ML models to predict customer creditworthiness rather than predicting whether the customer will effectively repay the loan. In this section, we propose a general Creditworthiness-Assessment Platform pipeline that deals with both tasks. Furthermore, this pipeline is characterised by an explainer module that can handle stakeholder-specific explanations (RQ6).

In Figure 6.3, the Creditworthiness Assessment Platform prediction depicts the general architecture of the platform. First, the system can exploit two kinds of user data: static and dynamic. The former does not change over time (or change slowly). Data such as demographics, income, and gender belong to this first group. It is worth noticing that some of those features are considered *sensitive* by the legislation. The second group belongs to features that change very frequently over time. In this scenario, these features are user transactions. Static and dynamic features are the input to two distinct modules: the Credit Scoring Model (CSM) and the Early Warning Detector (EWD).

The CSM is a binary classifier. Given a set of static user characteristics, it can decide whether that user will be able to repay its debt. The decision of the bank to grant or not the loan to a specific customer depends on this module. The task addressed by this module is particularly crucial because, as stated earlier, some user features are considered sensitive. Indeed, the last EU Commission regulation (April 21, 2021) considered the financial domain one of the most regulated [46]. The law proposal presents a pyramidal division of risk-based application of AI systems from minimal risk to unacceptable risk. Financial applications of AI systems (e.g., credit scoring) are considered *high-risk applications.* AI systems should comply with key ethical and trustworthy requirements since they need to pass different assessment steps. Therefore, it is really important that the algorithm implemented by the CSM does not put in place any kind of discrimination.

The second key component is the EWD. The Early Warning Monitoring System will rely only on accepted credit requests. Customer card transactions periodically feed the system (e.g., daily, weekly, or at predefined intervals) that models customer trends in terms of expenses and the available balance ratio for each transaction. Given the transaction, the model will predict a potential future bankruptcy. Once the EWS has triggered the Business Intelligence team; they will analyse the model's output and decide if it corresponds to a False Positive or a True Positive situation. More in detail,

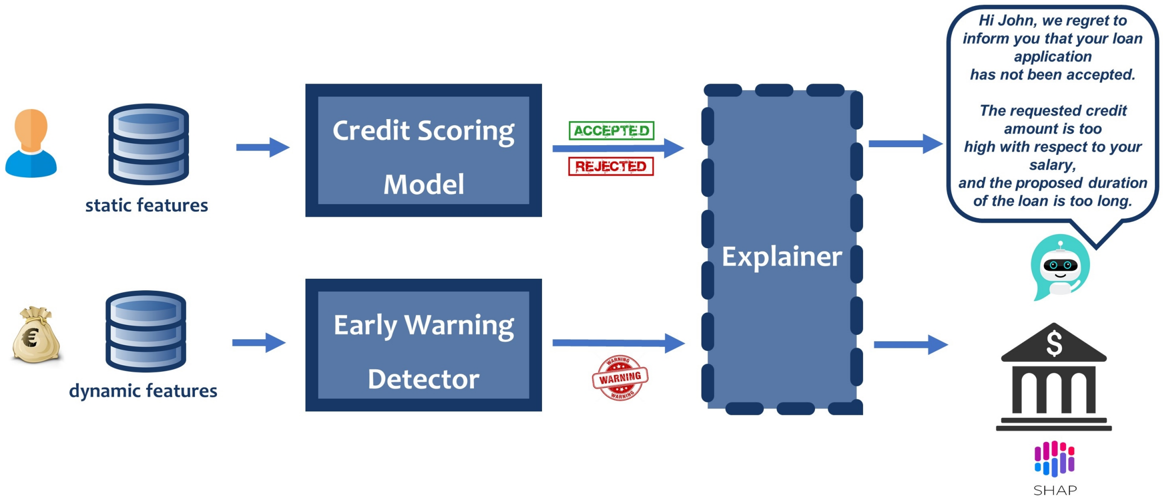

static features
dynamic features
Credit Scoring Model
Early Warning Detector
ACCEPTED
REJECTED
WARNING
Explainer
Hi John, we regret to inform you that your loan application has not been accepted.
The requested credit amount is too high with respect to your salary, and the proposed duration of the loan is too long.
SHAP

Fig. 6.3 Creditworthiness Assessment Platform prediction pipeline

the input of this module consists of the user transactions. When a customer makes a payment, purchase, or whatever financial transaction, this module checks whether that action could in some way jeopardise her ability to repay the debt. In contrast to the output of the Credit Scoring Model, which generally does not change over time, the prediction of the EWD is extremely fickle. In this case, EWD considers all the customer's history. Thus, the decision to trigger or not the warning depends on all the actions the user has done so far. The more the warning is true and early, the more the component is effective.

To comply with current regulations, the Business Intelligence team can use explanatory tools (e.g., Shapley values) to understand better which transactions have been responsible for this warning. Indeed, regulations require the bank to be aware of potential poor decisions and perform human-controlled actions in critical and life-changing situations.

The last component is the Explainer. Once the model performs the prediction task, the customer will be provided with an explanation, especially in case of rejection. In previous work [52, 60, 62], we provide different pipelines for generating natural language-based explanations, using both Shapley values and Counterfactual reasoning. As a game theory approach, the Shapley values give ranked feature importance of the most discriminating features for the decision task. It corresponds to the first stage of a user-friendly explanation. Shapley values theory is recognized as an effective tool for unveiling complex model decisions and a useful business intelligence analysis tool.

In contrast, counterfactual reasoning is used to discover a polarity between attribute and feature values to generate a natural-language-based explanation. Furthermore, the counterfactual exploration could provide plausible actions to receive the required credit. Again, in this case, the legislation plays a crucial role. More specifically, in the EU, the GDPR sets off the *right to explanation*: users have the right to ask for an explanation about an algorithmic decision made about them. In the UK, the Financial Conduct Authority (FCA) requires firms to explain why a more expensive mortgage has been chosen if a cheaper option is available. The G20 has adopted the OECD AI Principles for a trustworthy AI, which underlined that users should understand AI outcomes and be able to challenge them.

These are the motivations behind putting this component in the architecture. For this component, the shape is dashed since we propose only a possible implementation depicted in the previous section (i.e., Section 6.3.2). The module will be able to provide two kinds of explanation based on the type of stakeholder faced: a technical explanation for bank professionals, and a user-friendly explanation for the customer. The former is based on SHAP which is inspired by the cooperative game theory based on the Shapley Values [116]. Each feature is considered a player that contributes differently to the outcome (i.e., the algorithm decision). SHAP provides a ranked list of the features that contributed the most to the least to the outcome. In this case, the bank analyst can understand what are the features that impacted the most algorithm decisions.

The second form of explanation is in natural language. An effective solution could be to transform the output produced by SHAP into a natural language sentence. The natural language generation might be based on a set of rules that transform the Shapley values into natural language sentences.

In conclusion, the platform provides different steps that cope with Fairness and Explainability requirements. Considering the Creditworthiness Assessment step, the model should provide evidence of fair decisions based on a specific metric of fairness before being placed on the market. Several metrics can be used to evaluate the algorithm's fairness [66]. However, choosing which one to optimize is a complex task since each metric can belong to different statistical criteria (i.e., Independence, Separation, and Sufficiency) and to different fairness concepts (e.g., group fairness, individual fairness, sub-group fairness). Choosing the right fairness metrics remains a challenging task [27].

6.3.4 Limitations and Future Works

While the counterfactual reasoning and fairness auditing framework presented in this thesis provides a robust mechanism for bias detection and model explanation, it has several limitations that need to be addressed.

One key limitation of the current methodology lies in the interpretability of counterfactual explanations. Although counterfactuals offer actionable insights by presenting alternative scenarios, they are often limited by the need for domain knowledge or expertise to fully understand their implications. The complexity of the generated explanations may be challenging for non-technical users, particularly when dealing with high-dimensional or non-linear models [165]. Additionally, the current model focuses primarily on structured data, leaving out unstructured text or multimedia inputs that are increasingly prevalent in machine learning systems. The reliance on manually predefined sensitive attributes and the challenge of defining proxy features also limits the generalization of the approach to real-world applications, where biases are often hidden in complex relationships within the data [8]. Finally, the reliance on static counterfactual generation methods limits the dynamic nature of the explanation. Real-world scenarios often require explanations to adapt to evolving data streams, user interactions, and context shifts, which are not fully addressed in the current pipeline [73].

Recent advances in Natural Language Processing (NLP), particularly with the development of Large Language Models (LLMs) such as GPT-4 and BERT, offer promising avenues for improving the accessibility and usability of explanations [23]. LLMs have demonstrated exceptional capability in generating natural language explanations that are coherent, contextually relevant, and easy to understand by non-experts [11].

In the context of fairness and bias auditing, LLMs can be leveraged to generate natural language counterfactuals, where instead of altering structured features like income or age, the system could generate narrative alternatives describing different hypothetical scenarios [88]. These LLM-driven explanations could help bridge the gap between technical and non-technical stakeholders, offering more user-friendly explanations of model behaviour. Moreover, LLM-based summarization techniques can be applied to reduce the verbosity and complexity of explanations by providing concise, high-level insights into why certain decisions were made by the model [147].

There are several promising directions for future research to extend the current framework:

i. **Natural Language Generation (NLG) for Counterfactual Explanations:** NLG techniques can be explored to automatically generate explanations in plain language that adapt to the needs of different user groups, including non-experts. Future work could investigate the integration of NLG techniques to enhance user comprehension of complex decisions in domains such as finance or healthcare [161].

ii. **User-Centric Evaluation:** A user study would be invaluable to evaluate the effectiveness of counterfactual explanations generated by the model. Future research could include designing experiments to assess how well users understand these explanations, their perceived fairness, and whether the explanations foster trust in AI systems [79].

iii. **Interactive Explanations:** Another area for future research is the development of interactive explanation systems. These systems would allow users to query the model or adjust the generated counterfactuals to explore alternative explanations interactively. Such systems could leverage the capabilities of LLMs to provide real-time dialogue-based interactions with users, fostering greater transparency and engagement [100].

iv. **Fairness in Unstructured Data:** Extending the current fairness auditing framework to handle unstructured data such as text, images, or audio would significantly increase its applicability. Combining structured data counterfactual reasoning with multimodal LLMs could open up avenues for analyzing bias in more complex datasets, such as those used in social media platforms or recommendation systems [163].

By addressing these limitations and incorporating the advances in NLP and interactive user feedback, future work can help develop more transparent, fair, and accountable AI systems.

Chapter 7

Closing Remarks

This dissertation started with an analysis of the four most mentioned topics of our work: regulation, fairness under unawareness, counterfactual reasoning, and explainability (i.e., Chapter 1, Chapter 2, and Chapter 3). We ended up synthesizing all of them, proposing approaches that leverage and combine all of them. The bunch of problems that motivated this three-year work has given us a chance not only to study and propose some potential solutions but also to unveil new opportunities.

In Chapter 4, we present a novel methodology to detect bias in *Decision-Making* models, when regulation requires an unawareness setting, by analyzing the sensitive behaviour of (counterfactual) samples belonging to the positive side of the decision boundary. In detail, we exploit two different counterfactual generation strategies to do so. Specifically, the newly generated counterfactuals could belong to another sensitive group, thus suggesting potential discrimination in the decision process. To comprehensively assess the proposed methodology, in Chapter 5 we conducted experiments with ten decision makers (including state-of-the-art debiasing models) and three sensitive-feature classifiers. To measure the extent of the discriminatory behaviour, we introduce a new metric to find how many counterfactuals belong to another sensitive group. The contribution of this dissertation section is manifold: (i) we demonstrate that *fairness under unawareness* assumption is not sufficient to mitigate bias, (ii) we propose a methodology for the bias auditing task, and (iii) we show that counterfactual reasoning is an effective methodology to unveil the bias,

In the future, we plan to define a strategy to generate fair and actionable counterfactual samples to develop a model that could be effectively fair in the context of *fairness under unawareness.*

Chapter 6, as the last part of this dissertation, proposes the use of counterfactual reasoning as the main methodology to quantify proxy features and explain ML models.

Specifically, we define a procedure to identify proxy features leveraging counterfactual reasoning and we define a model to generate a natural language explanation for ML decisions in the context of loan recommendation platforms through a counterfactual analysis. This results in a set of corrective actions to be performed by the user.

The defined model finds a straightforward application in a scenario of a conversational recommender system. The user expresses her request in natural language, and the platform compares the different offers and provides an explanation for each of them. The user can thus ask for help on how to modify her request for getting the loan. Eventually, the platform, thanks to the counterfactual analysis and explanation, can provide a set of actions for getting the application accepted. However, the conversational system should preserve from discovering the complete set of decision criteria and avoid adverse action from unfair users. Furthermore, the type of explanation can depend on both business purposes and the stakeholders targeted. Shapley's values, even if very informative on the features that most influence the decision of ML models could not help customers, especially in the loan domain, to have a quantitative idea about why a certain decision occurred based on his characteristics. Another problem is the construction of the natural language-based explanation which polarity could not be easily extrapolated. All these issues can be easily overcome through the use of counterfactual reasoning.

In future works, we plan to implement the whole pipeline and conversational environment. Then, extensive experimental evaluations and user studies have to be carried out to assess the effectiveness of the model both in terms of the capability of generating NL explanations and in terms of improved user/stakeholder experience.

Bibliography

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. “Controlling popularity bias in learning-to-rank recommendation.” In: *Proceedings of the eleventh ACM conference on recommender systems*. 2017, pp. 42–46.

[2] Sray Agarwal and Shashin Mishra. *Responsible AI*. Springer, 2021.

[3] Kenneth J Arrow, Orley Ashenfelter, and Albert Rees. *The theory of discrimination*. Princeton University Press, 2015.

[4] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy Artificial Intelligence*. Tech. rep. European Commission, 2019.

[5] S. Baden and European Commission. Directorate-General for Development. *Gender Issues in Financial Liberalisation and Financial Sector Reform*. BRIDGE Report. BRIDGE, Institute of Development Studies at the University of Sussex, 1996.

[6] Ricardo Baeza-Yates. “Bias on the web.” In: *Commun. ACM* 61.6 (2018), pp. 54–61.

[7] Mislav Balunovic, Anian Ruoss, and Martin T. Vechev. “Fair Normalizing Flows.” In: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org, 2019.

[9] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. “Consumer-lending discrimination in the FinTech era.” In: *Journal of Financial Economics* (2021).

[10] Rachel K. E. Bellamy et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Oct. 2018.

[11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *FAccT*. ACM, 2021, pp. 610–623.

[12] George J Benston. “Discrimination in Financial Services: What Do We Not Know?” In: *Discrimination in Financial Services*. Springer, 1997, pp. 209–213.

[13] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. “Fairness in criminal justice risk assessments: The state of the art.” In: *Sociological Methods & Research* 50.1 (2021), pp. 3–44.

[14] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. "Sex Bias in Graduate Admissions: Data from Berkeley." In: *Science* 187.4175 (1975), pp. 398–404. DOI: 10.1126/science.187.4175.398. eprint: https://www.science.org/doi/pdf/10.1126/science.187.4175.398.

[15] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Tech. rep. MSR-TR-2020-32. Microsoft, May 2020.

[16] Arpita Biswas and Suvam Mukherjee. "Ensuring Fairness under Prior Probability Shifts." In: *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021.* Ed. by Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan. ACM, 2021, pp. 414–424. DOI: 10.1145/3461702.3462596.

[17] Kenneth A. Bollen. *Structural Equations with Latent Variables.* New York, NY: Wiley, 1989. ISBN: 978-0471011712.

[18] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. "A Training Algorithm for Optimal Margin Classifiers." In: *COLT.* ACM, 1992, pp. 144–152.

[19] Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, and Ed Snelson. "Counterfactual reasoning and learning systems: the example of computational advertising." In: *J. Mach. Learn. Res.* 14.1 (2013), pp. 3207–3260.

[20] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising." In: *Journal of Machine Learning Research* 14.11 (2013).

[21] Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. "Machine learning explainability in finance: an application to default risk analysis." In: 816 (Aug. 2019).

[22] Sandra Braunstein and Carolyn Welch. "Financial literacy: An overview of practice, research, and policy." In: *Fed. Res. Bull.* 88 (2002), p. 445.

[23] Tom B. Brown et al. "Language Models are Few-Shot Learners." In: *NeurIPS.* 2020.

[24] Consumer Financial Protection Bureau. "Using publicly available information to proxy for unidentified race and ethnicity: a methodology and assessment." In: (2014).

[25] Toon Calders and Szymon Jaroszewicz. "Efficient AUC Optimization for Classification." In: *Knowledge Discovery in Databases: PKDD 2007.* Ed. by Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 42–53. ISBN: 978-3-540-74976-9.

[26] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. "Building Classifiers with Independency Constraints." In: *ICDM Workshops.* IEEE Computer Society, 2009, pp. 13–18.

[27] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. “A clarification of the nuances in the fairness metrics landscape.” In: *Scientific Reports* 12.1 (Mar. 2022), p. 4209. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07939-1.

[28] Simon Caton and Christian Haas. “Fairness in machine learning: A survey.” In: *arXiv preprint arXiv:2010.04053* (2020).

[29] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. “Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees.” In: *FAT*. ACM, 2019, pp. 319–328.

[30] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. “How algorithmic confounding in recommendation systems increases homogeneity and decreases utility.” In: *RecSys*. ACM, 2018, pp. 224–232.

[31] Jiahao Chen. “Fair lending needs explainable models for responsible recommendation.” In: *FATREC’18 Proceedings of the Second Workshop on Responsible Recommendation*. Vancouver, British Columbia, Canada, Oct. 2018. arXiv: 1809.04684 [cs.LG].

[32] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. “Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved.” In: *FAT*. ACM, 2019, pp. 339–348.

[33] Robert S. Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. “Robust Optimization for Non-Convex Objectives.” In: *NIPS*. 2017, pp. 4705–4714.

[34] Alexandra Chouldechova. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” In: *Big Data* 5.2 (2017), pp. 153–163.

[35] Alexandra Chouldechova and Aaron Roth. “A snapshot of the frontiers of fairness in machine learning.” In: *Commun. ACM* 63.5 (2020), pp. 82–89.

[36] Evgenii Chzhen and Nicolas Schreuder. “A minimax framework for quantifying risk-fairness trade-off in regression.” In: *arXiv preprint arXiv:2007.14265* (2020).

[37] Giovanni Luca Ciampaglia, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini. “How algorithmic popularity bias hinders or promotes quality.” In: *Scientific reports* 8.1 (2018), pp. 1–7.

[38] Ethan Cohen-Cole. “CREDIT CARD REDLINING.” In: *The Review of Economics and Statistics* 93.2 (2011), pp. 700–713. ISSN: 00346535, 15309142.

[39] European Commission. *Protection of Personal Data*. Directive 95/46/EC. 1995.

[40] European Commission. *EU Charter of Fundamental Rights*. Official Journal of the European Communities, C 364/1, 18 December 2000. 2000.

[41] European Commission. *Racial Equality Directive*. Directive 2000/43/EC. 2000.

[42] European Commission. *Gender Equality Directive*. Directive 2006/54/EC. 2006.

[43] European Commission. *Consumer Credit Directive*. Directive 2008/48/EC. 2008.

[44] European Commission. *General Data Protection Regulation*. Regulation (EU) 2016/679. 2016.

[45] European Commission. *White Paper on Artificial Intelligence - A European approach to excellence and trust.* Tech. rep. European Commission, 2020.

[46] European Commission. *Proposal for a Regulation laying down harmonized rules on artificial intelligence.* COM(2021) 206 final. 2021.

[47] United States Congress. *Fair Housing Act.* 42 U.S.C. §§ 3601-3631. 1968.

[48] United States Congress. *Truth in Lending Act.* 15 U.S.C. §§ 1601-1667f. 1968.

[49] United States Congress. *Equal Credit Opportunity Act.* 15 U.S.C. §§ 1691-1691f. 1974.

[50] Sam Corbett-Davies and Sharad Goel. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." In: *CoRR* abs/1808.00023 (2018).

[51] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic Decision Making and the Cost of Fairness." In: *KDD.* ACM, 2017, pp. 797–806.

[52] *Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio. "Auditing fairness under unawareness through counterfactual reasoning." In: *Information Processing & Management* 60.2 (2023), p. 103224. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2022.103224.

[53] Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio. "Auditing fairness under unawareness through counterfactual reasoning." In: *Information Processing & Management* 60.2 (2023), p. 103224. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2022.103224.

[54] *Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "A General Architecture for a Trustworthy Creditworthiness-Assessment Platform in the Financial Domain." In: *Annals of Emerging Technologies in Computing (AETiC)* 7.2 (2023).

[55] *Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Reasoning for Decision Model Fairness Assessment." In: *Companion Proceedings of the ACM Web Conference 2023.* WWW '23 Companion. Austin, TX, USA: Association for Computing Machinery, 2023, pp. 229–233. ISBN: 9781450394192. DOI: 10.1145/3543873.3587354.

[56] Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Fair Opportunity: Measuring Decision Model Fairness with Counterfactual Reasoning." In: *CoRR* abs/2302.08158 (2023).

[57] *Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Reasoning for Bias Evaluation and Detection in a Fairness under Unawareness setting." In: *ECAI.* Vol. 372. Frontiers in Artificial Intelligence and Applications. IOS Press, 2023, pp. 477–484. DOI: 10.3233/FAIA230306.

[58] *Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, and Eugenio Di Sciascio. "Counterfactual Reasoning for Responsible AI Assessment." In: *Ital-IA*. Vol. 3486. CEUR Workshop Proceedings. CEUR-WS.org, 2023, pp. 347–352.

[59] *Giandomenico Cornacchia, Francesco M. Donini, Fedelucio Narducci, Claudio Pomo, and Azzurra Ragone. "Explanation in Multi-Stakeholder Recommendation for Enterprise Decision Support Systems." In: *Advanced Information Systems Engineering Workshops - CAiSE 2021 International Workshops, Melbourne, VIC, Australia, June 28 - July 2, 2021, Proceedings*. Ed. by Artem Polyvyanyy and Stefanie Rinderle-Ma. Vol. 423. Lecture Notes in Business Information Processing. Springer, 2021, pp. 39–47. DOI: 10.1007/978-3-030-79022-6_4.

[60] *Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. "A General Model for Fair and Explainable Recommendation in the Loan Domain (Short paper)." In: *KaRS/ComplexRec@RecSys*. Vol. 2960. CEUR Workshop Proceedings. CEUR-WS.org, 2021.

[61] Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. "A General Model for Fair and Explainable Recommendation in the Loan Domain (Short paper)." In: *KaRS/ComplexRec@RecSys*. Vol. 2960. CEUR Workshop Proceedings. CEUR-WS.org, 2021.

[62] *Giandomenico Cornacchia, Fedelucio Narducci, and Azzurra Ragone. "Improving the User Experience and the Trustworthiness of Financial Services." In: *Human-Computer Interaction – INTERACT 2021*. Ed. by Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen. Cham: Springer International Publishing, 2021, pp. 264–269. ISBN: 978-3-030-85607-6.

[63] *Giandomenico Cornacchia, Giulio Zizzo, Kieran Fraser, Muhammad Zaid Hamed, Ambrish Rawat, and Mark Purcell. "MoJE: Mixture of Jailbreak Experts, Naive Tabular Classifiers as Guard for Prompt Attacks." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, 2024. arXiv: 2409.17699.

[64] Federal Deposit Insurance Corporation. *2015: FDIC National Survey of Unbanked and Underbanked Households*. Census, Federal Deposit Insurance Corporation, 2014.

[65] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. "Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification." In: *Big Data* 5.2 (2017), pp. 120–134.

[66] Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, Jared Katzman, Krishnaram Kenthapadi, Pedro Larroy, Pinar Yilmaz, and Muhammad Bilal Zafar. "Fairness Measures for Machine Learning in Finance." In: *The Journal of Financial Data Science* 3.4 (2021), pp. 33–64. ISSN: 2640-3943. DOI: 10.3905/jfds.2021.1.075. eprint: https://jfds.pm-research.com/content/3/4/33.full.pdf.

[67] Sanjiv Ranjan Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, Jared Katzman, Krishnaram Kenthapadi, Pedro Larroy, Pinar Yilmaz, and Bilal Zafar. "Fairness Measures for Machine Learning in Finance." In: *The Journal of Financial Data Science*. 2021.

[68] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. "A Survey of Research on Fair Recommender Systems." In: *arXiv preprint arXiv:2205.11127* (2022).

[69] George F DeMartino. "The confounding problem of the counterfactual in economic explanation." In: *Review of Social Economy* (2020), pp. 1–11.

[70] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. "Image counterfactual sensitivity analysis for detecting unintended bias." In: *arXiv preprint arXiv:1906.06439* (2019).

[71] L. Dlabay, J.L. Burrow, and B. Kleindl. *Intro to Business*. Cengage Learning, 2008. ISBN: 9780538445610.

[72] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. "Empirical Risk Minimization Under Fairness Constraints." In: *NeurIPS*. 2018, pp. 2796–2806.

[73] Finale Doshi-Velez and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning." In: *arXiv: Machine Learning* (2017).

[74] Julia Dressel and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism." In: *Science Advances* 4.1 (2018), eaao5580. DOI: 10.1126/sciadv.aao5580. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aao5580.

[75] Miroslav Dudík, John Langford, and Lihong Li. "Doubly Robust Policy Evaluation and Learning." In: *ICML*. Omnipress, 2011, pp. 1097–1104.

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

[77] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. "Fairness through awareness." In: *ITCS*. ACM, 2012, pp. 214–226.

[78] Organisation for Economic Co-operation and Development. *OECD AI Principles for a Trustworthy AI*. Online. 2019.

[79] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. "Automated rationale generation: a technique for explainable AI and its effects on human perceptions." In: *IUI*. ACM, 2019, pp. 263–274.

[80] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. "Fairness in Information Access Systems." In: *Foundations and Trends® in Information Retrieval* 16.1-2 (2022), pp. 1–177.

[81] Marc Elliott, Allen Fremont, Peter Morrison, Philip Pantoja, and Nicole Lurie. "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity." In: *Health services research* 43 (May 2008), pp. 1722–36. DOI: 10.1111/j.1475-6773.2008.00854.x.

[82] Marc N. Elliott, Peter A. Morrison, Allen M. Fremont, Daniel F. McCaffrey, Philip M Pantoja, and Nicole Lurie. “Using the Census Bureau’s surname list to improve estimates of race/ethnicity and associated disparities.” In: *Health Services and Outcomes Research Methodology* 9 (2009), pp. 69–83.

[83] Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. “Measuring Fairness Under Unawareness of Sensitive Attributes: A Quantification-Based Approach.” In: *J. Artif. Intell. Res.* 76 (2023), pp. 1117–1180. DOI: 10.1613/jair.1.14033.

[84] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. “Algorithmic fairness datasets: the story so far.” In: *Data Min. Knowl. Discov.* 36.6 (2022), pp. 2074–2152. DOI: 10.1007/s10618-022-00854-z.

[85] Michael Fay and Lesley Williams. “Gender bias and the availability of business loans.” In: *Journal of Business Venturing* 8.4 (1993), pp. 363–376.

[86] Federal Reserve Board. *The Truth in Lending Act.* Law. 1968.

[87] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. “Certifying and Removing Disparate Impact.” In: *KDD.* ACM, 2015, pp. 259–268.

[88] Norman E. Fenton, Martin Neil, and Anthony C. Constantinou. “The Book of Why: The New Science of Cause and Effect, Judea Pearl, Dana Mackenzie. Basic Books (2018).” In: *Artif. Intell.* 284 (2020), p. 103286.

[89] Roberta Ferrario. “Counterfactual Reasoning.” In: *CONTEXT.* Vol. 2116. Lecture Notes in Computer Science. Springer, 2001, pp. 170–183.

[90] Ronald A Fisher. “On the mathematical foundations of theoretical statistics.” In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222.594-604 (1922), pp. 309–368.

[91] Batya Friedman and Helen Nissenbaum. “Bias in Computer Systems.” In: *ACM Trans. Inf. Syst.* 14.3 (1996), pp. 330–347.

[92] Matthew L. Ginsberg. “Counterfactuals.” In: *Artif. Intell.* 30.1 (1986), pp. 35–79.

[93] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. “Provider fairness across continents in collaborative recommender systems.” In: *Inf. Process. Manag.* 59.1 (2022), p. 102719.

[94] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. “The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making.” In: *Proceedings of the NIPS Symposium on Machine Learning and the Law.* Vol. 1. 2016, p. 2.

[95] Maurício Holler Guntzel. “Fairness in machine learning: an empirical experiment about protected features and their implications.” In: (2022).

[96] Michaela Hardt et al. “Amazon SageMaker Clarify: Machine learning bias detection and explainability in the cloud.” In: *KDD 2021.* 2021.

[97] Moritz Hardt, Eric Price, and Nati Srebro. “Equality of Opportunity in Supervised Learning.” In: *NIPS.* 2016, pp. 3315–3323.

[98] Masoud Hashemi and Ali Fathi. “PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards.” In: *CoRR* abs/2008.10138 (2020).

[99] Hans Hofmann. *UCI Statlog (German Credit Data) Data Set.* UCI Machine Learning Repository, 2000.

[100] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. “Causability and explainability of artificial intelligence in medicine.” In: *WIREs Data Mining Knowl. Discov.* 9.4 (2019).

[101] The White House. *National Strategic Research and Development Plan for Artificial Intelligence.* Tech. rep. Executive Office of the President, 2019.

[102] Guy B. Johnson. “The Economics of Discrimination. By Gary S. Becker. Chicago: University of Chicago Press, 1957.137 pp. $3.50.” In: *Social Forces* 37.2 (Dec. 1958), pp. 180–181. ISSN: 0037-7732. DOI: 10.2307/2572813. eprint: https://academic.oup.com/sf/article-pdf/37/2/180/6504252/37-2-180a.pdf.

[103] Jungseock Joo and Kimmo Kärkkäinen. “Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation.” In: *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia.* 2020, pp. 1–5.

[104] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. “Decision Theory for Discrimination-Aware Classification.” In: *ICDM.* IEEE Computer Society, 2012, pp. 924–929.

[105] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. “Exploiting reject option in classification for social discrimination control.” In: *Inf. Sci.* 425 (2018), pp. 18–33.

[106] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. “Fairness-Aware Classifier with Prejudice Remover Regularizer.” In: *ECML/PKDD (2).* Vol. 7524. Lecture Notes in Computer Science. Springer, 2012, pp. 35–50.

[107] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” In: *ITCS.* Vol. 67. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017, 43:1–43:23.

[108] Ronny Kohavi and Barry Becker. *UCI Adult Data Set.* UCI Machine Learning Repository, May 1996.

[109] Anton Korikov, Alexander Shleyfman, and J. Christopher Beck. “Counterfactual Explanations for Optimization-Based Decisions Fin the Context of the GDPR.” In: *IJCAI.* ijcai.org, 2021, pp. 4097–4103.

[110] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. “Fairness in Credit Scoring: Assessment, Implementation and Profit Implications.” In: *European Journal of Operational Research* (2021).

[111] Alex Kulesza and Ben Taskar. “Determinantal Point Processes for Machine Learning.” In: *Found. Trends Mach. Learn.* 5.2-3 (2012), pp. 123–286.

[112] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. “Counterfactual Fairness.” In: *NIPS.* 2017, pp. 4066–4076.

[113] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. "Delayed Impact of Fair Machine Learning." In: *IJCAI*. ijcai.org, 2019, pp. 6196–6200.

[114] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. "The Variational Fair Autoencoder." In: *ICLR*. 2016.

[115] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In: *NIPS*. 2017, pp. 4765–4774.

[116] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In: *NIPS*. 2017, pp. 4765–4774.

[117] *Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, and Tommaso Di Noia. "On Popularity Bias of Multimodal-Aware Recommender Systems: A Modalities-Driven Analysis." In: MMIR '23 (2023), pp. 59–68. DOI: 10.1145/3606040.3617441.

[118] Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, Felice Antonio Merra, Tommaso Di Noia, and Eugenio Di Sciascio. "Formalizing Multimedia Recommendation through Multimodal Deep Learning." In: *ACM Transaction on Recommender System* (Apr. 2024). DOI: 10.1145/3662738.

[119] *Daniele Malitesta, Giandomenico Cornacchia, Claudio Pomo, and Tommaso Di Noia. "Disentangling the Performance Puzzle of Multimodal-aware Recommender Systems." In: *EvalRS@KDD*. Vol. 3450. CEUR Workshop Proceedings. CEUR-WS.org, 2023.

[120] Natalia Martinez and Martin Bertran. "Blind Pareto Fairness and Subgroup Robustness." In: *International Conference Machine Learning*. 2021.

[121] Rory McGrath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lécué. "Interpretable Credit Application Predictions With Counterfactual Explanations." In: *CoRR* abs/1811.05245 (2018).

[122] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A Survey on Bias and Fairness in Machine Learning." In: *CoRR* abs/1908.09635 (2019).

[123] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning." In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–35.

[124] Agnieszka Mikolajczyk, Michal Grochowski, and Arkadiusz Kwasigroch. "Towards Explainable Classifiers Using the Counterfactual Approach - Global Explanations for Discovering Bias in Data." In: *J. Artif. Intell. Soft Comput. Res.* 11.1 (2021), pp. 51–67.

[125] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences." In: *Artif. Intell.* 267 (2019), pp. 1–38.

[126] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. "Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds." In: *FAccT*. ACM, 2021, pp. 386–400.

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

[128] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. “Explaining machine learning classifiers through diverse counterfactual explanations.” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 607–617.

[129] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations.” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 607–617. ISBN: 9781450369367. DOI: 10.1145/3351095.3372850.

[130] Cataldo Musto, Fedelucio Narducci, Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. “ExpLOD: A Framework for Explaining Recommendations Based on the Linked Open Data Cloud.” In: *Proceedings of the 10th ACM Conference on Recommender Systems*. RecSys ’16. Boston, Massachusetts, USA: Association for Computing Machinery, 2016, pp. 151–154. ISBN: 9781450340359. DOI: 10.1145/2959100.2959173.

[131] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. “Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries.” In: *Frontiers Big Data* 2 (2019), p. 13.

[132] Luca Oneto and Silvia Chiappa. “Fairness in Machine Learning.” In: *CoRR* abs/2012.15816 (2020).

[133] Steven Ongena, Alexander Popov, et al. “Take care of home and family, honey, and let me take care of the money. Gender bias and credit market barriers for female entrepreneurs.” In: *Social Science Research Network. Accessed September* 20 (2013), p. 2014.

[134] Osonde Osoba and W. Welser. “An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence.” In: 2017.

[135] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. “FairLens: Auditing black-box clinical decision support systems.” In: *Inf. Process. Manag.* 58.5 (2021), p. 102657.

[136] Francesco Paolo Schena, Vto Walter Anelli, Giandomenico Cornacchia, Tommaso DI Noia, Maria Stangou, Aikaterina Papagianni, and Rosanna Coppo. “FC048: New Tool to Predict the Clinical Course and Renal Failure in Patients with Immunoglobulin a Nephropathy.” In: *Nephrology Dialysis Transplantation* 37.Supplement_3 (May 2022), gfac105.004. ISSN: 0931-0509. DOI: 10.1093/ndt/gfac105.004. eprint: https://academic.oup.com/ndt/article-pdf/37/Supplement_3/gfac105.004/43535905/gfac105_004.pdf.

[137] Judea Pearl. “Causation, Action and Counterfactuals.” In: *ECAI*. John Wiley and Sons, Chichester, 1994, pp. 826–828.

[138] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. “Discrimination-aware data mining.” In: *KDD*. ACM, 2008, pp. 560–568.

[139] Stephen R. Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H. Shah. “Counterfactual Reasoning for Fair Clinical Risk Prediction.” In: *MLHC*. Vol. 106. Proceedings of Machine Learning Research. PMLR, 2019, pp. 325–358.

[140] Edmund S Phelps. “The statistical theory of racism and sexism.” In: *The american economic review* 62.4 (1972), pp. 659–661.

[141] Flávio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. “Optimized Pre-Processing for Discrimination Prevention.” In: *NIPS*. 2017, pp. 3992–4001.

[142] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. “Fairness in rankings and recommendations: an overview.” In: *VLDB J.* 31.3 (2022), pp. 431–458.

[143] John Platt. “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.” In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74.

[144] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. “On Fairness and Calibration.” In: *NIPS*. 2017, pp. 5680–5689.

[145] Michael Redmond. *UCI Communities and Crime Data Set*. UCI Machine Learning Repository, 1995.

[146] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier.” In: *KDD*. ACM, 2016, pp. 1135–1144.

[147] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier.” In: *KDD*. ACM, 2016, pp. 1135–1144.

[148] R Tyrrell Rockafellar, Stanislav Uryasev, et al. “Optimization of conditional value-at-risk.” In: *Journal of risk* 2 (2000), pp. 21–42.

[149] Daniele Rossini, Danilo Croce, Sara Mancini, Massimo Pellegrino, and Roberto Basili. “Actionable Ethics through Neural Learning.” In: *AAAI*. AAAI Press, 2020, pp. 5537–5544.

[150] Boris Ruf and Marcin Detyniecki. “Active Fairness Instead of Unawareness.” In: *CoRR* abs/2009.06251 (2020).

[151] Ms Ratna Sahay and Mr Martin Cihak. *Women in Finance: A Case for Closing Gaps*. International Monetary Fund, 2018.

[152] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. “Aequitas: A bias and fairness audit toolkit.” In: *arXiv preprint arXiv:1811.05577* (2018).

[153] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. “How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness.” In: *AIES*. ACM, 2019, pp. 99–106.

[154] Ram Shanmugam. “Causality: Models, Reasoning, and Inference : Judea Pearl; Cambridge University Press, Cambridge, UK, 2000, pp 384, ISBN 0-521-77362-8.” In: *Neurocomputing* 41.1-4 (2001), pp. 189–190.

[155] Rudrani Sharma, Christoph Schommer, and Nicolas Vivarelli. “Building up Explainability in Multi-layer Perceptrons for Credit Risk Modeling.” In: *DSAA*. IEEE, 2020, pp. 761–762.

[156] Kacper Sokol and Peter A. Flach. “Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety.” In: *SafeAI@AAAI*. Vol. 2301. CEUR Workshop Proceedings. CEUR-WS.org, 2019.

[157] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. “A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence.” In: *IEEE Access* 9 (2021), pp. 11974–12001.

[158] Harini Suresh and John V Guttag. “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle.” In: *arXiv preprint arXiv:1901.10002* (2019).

[159] Harini Suresh and John V. Guttag. “A Framework for Understanding Unintended Consequences of Machine Learning.” In: *CoRR* abs/1901.10002 (2019).

[160] Adith Swaminathan and Thorsten Joachims. “Batch learning from logged bandit feedback through counterfactual risk minimization.” In: *J. Mach. Learn. Res.* 16 (2015), pp. 1731–1755.

[161] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. “Counterfactual Explainable Recommendation.” In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM ’21. Virtual Event, Queensland, Australia: Association for Computing Machinery, 2021, pp. 1784–1793. ISBN: 9781450384469. DOI: 10.1145/3459637.3482420.

[162] Maryam Tavakol. “Fair Classification with Counterfactual Learning.” In: *SIGIR*. ACM, 2020, pp. 2073–2076.

[163] Hugo Touvron et al. “LLaMA: Open and Efficient Foundation Language Models.” In: *CoRR* abs/2302.13971 (2023).

[164] Sahil Verma and Julia Rubin. “Fairness definitions explained.” In: *FairWare@ICSE*. ACM, 2018, pp. 1–7.

[165] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR.” In: *CoRR* abs/1711.00399 (2017).

[166] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. “Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences.” In: *CHI*. ACM, 2020, pp. 1–14.

[167] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. “The What-If Tool: Interactive Probing of Machine Learning Models.” In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 56–65. DOI: 10.1109/TVCG.2019.2934619.

[168] Robert C. Williamson and Aditya Krishna Menon. “Fairness risk measures.” In: *ICML*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6786–6797.

[169] Yongkai Wu, Lu Zhang, and Xintao Wu. “Counterfactual Fairness: Unidentification, Bound and Algorithm.” In: *IJCAI*. ijcai.org, 2019, pp. 1438–1444.

[170] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. “FairGAN: Fairness-aware Generative Adversarial Networks.” In: *IEEE BigData*. IEEE, 2018, pp. 570–575.

[171] Samuel Yeom, Anupam Datta, and Matt Fredrikson. “Hunting for discriminatory proxies in linear regression models.” In: *Advances in Neural Information Processing Systems* 31 (2018).

[172] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. “Fairness Constraints: Mechanisms for Fair Classification.” In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, 20–22 Apr 2017, pp. 962–970.

[173] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. “Learning Fair Representations.” In: *ICML (3)*. Vol. 28. JMLR Workshop and Conference Proceedings. JMLR.org, 2013, pp. 325–333.

[174] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. “Mitigating Unwanted Biases with Adversarial Learning.” In: *AIES*. ACM, 2018, pp. 335–340.

[175] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. “Mitigating Unwanted Biases with Adversarial Learning.” In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’18. New Orleans, LA, USA: Association for Computing Machinery, 2018, pp. 335–340. ISBN: 9781450360128. DOI: 10.1145/3278721.3278779.

[176] Lu Zhang, Yongkai Wu, and Xintao Wu. “A Causal Framework for Discovering and Removing Direct and Indirect Discrimination.” In: *IJCAI*. ijcai.org, 2017, pp. 3929–3935.

[177] Yukun Zhang and Longsheng Zhou. “Fairness assessment for artificial intelligence in financial industry.” In: *arXiv preprint arXiv:1912.07211* (2019).

[178] Dávid Zibriczky. “Recommender Systems meet Finance: a Literature Review.” In: *FINREC*. Vol. 1606. CEUR Workshop Proceedings. CEUR-WS.org, 2016, pp. 3–10.