TU WIEN Informatics

Towards Fair AI Systems

An Insurance Case Study to identify and mitigate Discrimination

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Data Science

eingereicht von

Annabel Resch
Matrikelnummer 11914287

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dr. Allan Hanbury

Wien, 1. Juni 2025

Annabel Resch	Allan Hanbury

TU WIEN Informatics

Towards Fair AI Systems

An Insurance Case Study to identify and mitigate Discrimination

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Data Science

by

Annabel Resch
Registration Number 11914287

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dr. Allan Hanbury

Vienna, June 1, 2025

Annabel Resch	Allan Hanbury

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Annabel Resch

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 1. Juni 2025

Annabel Resch

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Hanbury, for his valuable guidance throughout this thesis process. His willingness to take on a project outside his primary research area and his continuous support were of vital importance to the completion of this work. Additionally, I extend my gratitude to Dr. Moritz Zoppel from WU Vienna for his expert advice on legal regulations in Austria.

I am deeply grateful to my family for their unwavering belief and support in me. This degree is as much their achievement as it is mine. Finally, I thank my friends for their relentless emotional and intellectual support over the years. Their encouragement has been a constant source of strength throughout this journey.

Kurzfassung

Diese Arbeit untersucht geschlechts- und nationalitätsbasierte Diskriminierung in einem Machine-Learning-Modell einer österreichischen Versicherungsgesellschaft. Ziel des Modells ist die Identifikation von Schadensfällen mit signifikanten Kostensteigerungen von Entschädigungsansprüchen. Angesichts der Vorgaben des EU AI Act's und österreichischer Rechtsbestimmungen für diskriminierungsfreie algorithmische Systeme gewinnt die Sicherstellung fairer Modelle zunehmend an Bedeutung. Diese Arbeit prüft, ob ein von der Versicherung eingesetztes Light Gradient Boosting Machine (LightGBM)-Modell diskriminierende Eigenschaften aufweist und erforscht Ansätze zur Bias-Reduzierung.

Nach einer umfassenden Literaturanalyse des aktuellen Forschungsstands zu algorithmischer Fairness wurde ein Datensatz mit 450.000 Versicherungsschäden ausgewertet. Das methodische Vorgehen umfasste eine Fairness-Analyse des LightGBM-Modells, die qualitative Bewertung geeigneter Fairness-Metriken für den vorliegenden Anwendungsfall, sowie die Implementierung verschiedener Bias-Minderungsverfahren, darunter In-Processing-Techniken (FairGBM) und Post-Processing-Methoden (Reject Option Classification von AIF360, Threshold Optimizer von Fairlearn sowie Equalized-Odds Post-Processing von AIF360). Zuletzt wurde ein quantitativer Vergleich zwischen dem LightGBM-Modell und den Bias-Minderungsverfahren durchgeführt, basierend auf Fairness-Metriken und den Auswirkungen auf die Leistung des Vorhersagemodells.

Die Analyse des Baseline Modells deckte erhebliche Benachteiligungen weiblicher gegenüber männlichen Versicherungsnehmenden sowie nicht-österreichischer gegenüber österreichischen Versicherungsnehmenden auf. Darüber hinaus zeigte das Modell eine verminderte Leistung für Schadensfälle ohne eindeutige Geschlechts- oder Nationalitätszuordnung. Während die eingesetzten Bias-Minderungsverfahren die Fairness-Metriken erfolgreich verbesserten, gingen diese Verbesserungen erheblich zu Lasten der Vorhersageleistung des Machine-Learning-Modells.

Die Untersuchung verdeutlicht die Notwendigkeit systematischer Diskriminierungsprüfungen bei Machine-Learning-Modellen, insbesondere in kritischen Anwendungsbereichen wie dem Versicherungswesen. Obwohl Fairness-Optimierungstechniken diskriminierende Strukturen mathematisch addressieren können, erweisen sich die erheblichen Beeinträchtigungen bei der Vorhersageleistung als hinderlich für den produktiven Einsatz. Dies unterstreicht die fortbestehende Herausforderung, algorithmische Fairness mit betriebswirtschaftlichen Anforderungen in Einklang zu bringen.

Abstract

This thesis investigates potential gender and nationality-based discrimination in a real-world insurance machine learning model designed to identify claims likely to "explode" in compensation costs. With the EU AI Act and Austrian legal frameworks requiring non-discriminatory algorithmic systems, ensuring fairness in insurance claim prediction models has become critically important. The research examines whether a Light Gradient Boosting Machine (LightGBM) model used by an Austrian insurance company exhibits discriminatory behavior and explores methods to mitigate such bias.

Following a comprehensive literature review on algorithmic fairness state-of-the-art, this study analyzed a dataset of 450,000 insurance claims provided by an Austrian insurance company. Claims were classified as "explosive" if they required reserve increases exceeding €100,000 or if reserve amounts grew by a factor of ten or more within one month of initial reporting. The methodology included baseline fairness analysis of the current model, qualitative assessment of appropriate fairness metrics for this use case, implementation of various mitigation methods including in-processing (FairGBM) and post-processing techniques (reject option classification by AIF360, threshold optimizer by Fairlearn, and equalized odds post-processing by AIF360), and quantitative comparison of fairness improvements against predictive performance impacts.

The baseline analysis revealed significant discrimination against female claimants compared to male claimants and non-Austrian claimants compared to Austrian claimants. Additionally, claims with unknown gender or nationality group membership showed degraded prediction quality. While mitigation methods successfully improved fairness metrics, these improvements came at a severe cost to predictive performance.

This research demonstrates the critical importance of evaluating machine learning models for potential discrimination, particularly in high-stakes applications like insurance. Although fairness mitigation techniques can mathematically address discriminatory patterns, the substantial trade-off with predictive performance renders them impractical for real-world deployment in this case study, highlighting the ongoing challenge of balancing algorithmic fairness with business utility.

Contents

CHAPTER 1

Introduction

The pursuit of fairness and equality has been a central theme throughout human history, with profound roots in philosophy and psychology. For centuries, scholars and thinkers have grappled with questions of justice, discrimination, and bias in human decision-making processes [MMS+21]. In recent decades, these fundamental questions have found new relevance and urgency in the domain of machine learning and artificial intelligence [PS22]. As algorithms increasingly influence critical aspects of our lives, from loan approvals to hiring decisions, the pressing technological and ethical challenge arises to ensure these systems operate fairly.

The insurance industry represents a particularly critical domain for the application and evaluation of fair AI systems. On the one hand, insurance providers offer essential services to modern societies. Motor insurance facilitates mobility, health insurance provides access to medical care, and property insurance protects against catastrophic losses. These services fulfill essential social needs and contribute to broader social welfare.
On the other hand, most insurers are commercial entities with profit-oriented objectives. They must carefully calculate reserves, manage risk pools, and maintain financial viability to continue providing their services. This commercial imperative necessitates sophisticated risk assessment and pricing mechanisms, increasingly powered by AI and machine learning technologies. Fair AI systems in insurance must navigate this duality, ensuring non-discrimination while preserving the actuarial principles that enable insurers to operate sustainably.

The field of algorithmic fairness, which emerged approximately fifteen years ago, has rapidly evolved in response to the growing recognition of the potential harms that biased AI systems can inflict [LO24]. With the accelerating integration of algorithms into daily life, the adverse impacts of these technologies have moved from theoretical concerns to practical realities that demand immediate attention. A concrete example of algorithmic discrimination comes from the Austrian insurance sector, where AI systems used for risk assessment have been found to encode socioeconomic biases [FDCE19]. These

systems were found to discriminate against individuals from lower-income neighborhoods or with specific migration backgrounds when calculating premiums. Similarly, health insurance algorithms have been found to underestimate the healthcare needs of Black patients compared to White patients with similar health profiles [FDCE19]. Such examples highlight how algorithmic bias can perpetuate and even amplify existing social inequalities, even though they appear objective and data-driven.

The growing awareness of algorithmic bias has prompted significant regulatory responses, particularly in Europe. The recently introduced European Union's AI Act represents a landmark legislative framework specifically designed to address the risks associated with artificial intelligence systems, including those deployed in the insurance industry. Non-compliance with the European Union AI Act regulations carries substantial penalties, with up to €30 million or 6% of global annual turnover, whichever is higher [VB21]. These stringent penalties underscore the seriousness with which regulators view algorithmic fairness and indicate that insurance companies cannot afford to treat bias mitigation as optional. Ensuring fairness in AI systems is not only a technical necessity but also a regulatory imperative.

1.1 Problem Statement and Research Questions

Recent legal regulations, like the EU AI Act explicitly require companies to not only design new AI systems with fairness considerations but also to retrospectively evaluate existing systems for potential discrimination [aia]. In response to this challenge, an Austrian insurance company has provided a case study from 2019, including a baseline model and dataset, which is currently implemented in their business processes. This case study, called 'Explosive Claims', employs machine learning to identify insurance claims that ultimately result in significantly higher compensation costs than initially calculated. The detailed methodology and dataset characteristics of this case study are elaborated in Section 3.1.

Despite the clear regulatory mandate against discrimination in AI systems, as stipulated in frameworks such as the General Data Protection Regulation [GDP] and the European Union AI Act [aia], these regulations fall short of providing precise definitions of what constitutes discrimination in algorithmic contexts. This definitional ambiguity creates a substantial implementation gap: insurance companies are legally obligated to prevent discrimination but lack clear guidance on how to operationalize fairness in their AI systems.

The amount of published scientific literature on discrimination in AI systems has expanded considerably in recent years, yet significant challenges persist in developing solutions that are both robust and generalizable to real-world applications. A fundamental limitation in this field is the scarcity of comprehensive benchmark datasets that reflect the complexity of actual insurance operations [MMS+21]. The reliance on simplified datasets raises profound questions about the external validity of research findings [QRI+22]. The gap

between research environments and the multifaceted nature of real-world insurance data creates significant uncertainty about whether the developed fairness strategies will translate effectively to production environments.

The literature on algorithmic fairness presents numerous fairness metrics, each embodying different philosophical and mathematical conceptualizations of fairness [CCG+22]. The selection of inappropriate fairness metrics can lead to severe legal repercussions, as dramatically illustrated in the widely cited COMPAS[1] recidivism prediction case [KMR]. This case demonstrated how different fairness metrics led to contradictory conclusions about the same system's discriminatory impact. Even among computer scientists there remains a striking lack of consensus about which constraints are most appropriate [MMS+21]. This theoretical fragmentation creates significant barriers for insurance practitioners seeking to implement fairness-aware systems in compliance with regulations.

Another evaluation challenge is the methodological question about bias mitigation approaches. Insurance companies face the daunting task of selecting from numerous bias mitigation techniques without clear guidance on their comparative efficacy in insurance-specific contexts [PS22]. A critical shortcoming in current fairness literature is the frequent neglect of model performance considerations. In real-world AI systems, particularly in the insurance sector where predictive accuracy directly impacts business outcomes and customer experiences, maintaining high model performance is non-negotiable. The literature frequently presents fairness as a constraint to be satisfied, initially resulting in accuracy degradation [LV22]. Commonly used strategies typically minimize discrimination by adding penalty terms to loss functions, which restricts the model's ability to use all available information for prediction [CBJ+23, ZVGRG19, KAAS12]. Scientific work has advanced to better understanding these tradeoffs, with some research challenging the assumption that fairness necessarily reduces accuracy [DWY+20]. Despite this progress, it remains unclear what constitutes the optimal balance between fairness and accuracy objectives in real-world applications.

In conclusion, it is unclear whether the 'Explosive Claims' algorithm (unintentionally) currently discriminates, how discrimination can be measured, and which measures can be taken to mitigate discrimination.

From this problem statement, the overall goal and research questions of the master's thesis are derived:

Goal: Introducing fairness by mitigating discrimination in a real-world model from the insurance industry.

Overall research question: How can fairness in the case study model be measured and mitigated?

To successfully and efficiently address the above-mentioned problem, the following research questions are answered:

[1] https://www.kaggle.com/, visited on 05/11/2025

RQ1: To what extent does the current baseline model discriminate in the context of this case study?

1.1 Which fairness metrics are most suitable to capture discrimination in the case study?

1.2 What are the subgroups that are discriminated against in the case study?

RQ2: To what extent can fairness be improved through discrimination mitigation techniques in comparison to the baseline?

2.1 To what extent can fairness be improved by implementing in-processing mitigation techniques?

2.2 To what extent can fairness be improved by implementing post-processing mitigation techniques?

RQ3: To what extent does the predictive performance decrease by introducing the above-mentioned mitigation techniques, compared to baseline?

The solution will be regarded as successful if a model is found that effectively maximizes fairness without compromising the baseline model performance. Additionally, it is essential that all legal requirements are met throughout this process.

1.2 Structure of the Work

The diploma thesis structure addresses the specified research questions through the following organization: Chapter 2 presents a comprehensive literature research on algorithmic fairness, bias, and an in-depth examination of the term *discrimination*. Various definitions, types, and potential sources of discrimination are explored. The complex legal framework governing discrimination in AI systems under European and Austrian law is thoroughly analyzed, and the concept of 'protected attributes' is explained within its legal context. In addition, specific work on algorithmic fairness in AI systems within the insurance domain is analyzed in detail. The subsequent section discusses quantitative measures for discrimination in AI models. Definitions of fairness metrics are explained, and their limitations are reviewed. Following this, existing discrimination mitigation strategies proposed in literature are described for application when discrimination is detected in an AI model. These strategies are categorized into pre-processing, in-processing, and post-processing methods, each receiving a brief description and critical assessment.

Chapter 3 begins with a detailed description of the real-world case study. Data characteristics and protected attributes are examined, along with the subgroups present in the dataset. A brief exploratory data analysis demonstrates the label distribution and representation of subgroups. The baseline model is described in detail, including the underlying algorithm and the parameter settings. The subsequent section details how and which previously defined fairness metrics will be applied in the case study analysis.

Given the hundreds of mitigation techniques available in literature, the next section of the chapter explains in detail which mitigation methods will be implemented in this case study. This section provides thorough explanations of the selected strategies, their implementation approaches, and feature settings. The last section describes the metrics applied to capture the predictive performance of the baseline model compared to the mitigation methods.

Chapter 4 commences with a comprehensive fairness analysis of the baseline model to address research question 1. Next, the baseline model is compared against various mitigation strategies regarding both fairness (research question 2) and predictive model performance (research question 3).

Chapter 5 summarizes the findings and addresses the research questions. The research, legal framework, and case study analysis undergo critical assessment based on their limitations, and potential future work is identified. The final section discusses the crucial importance of this topic in the contemporary context, with references to current political and societal developments.

CHAPTER 2

Related Work

This chapter examines how discrimination manifests in artificial intelligence systems, particularly within the insurance sector, and analyzes the applicable legal frameworks. It then delves into various fairness metrics from the academic literature and breaks down the wide range of methods available for mitigating discrimination.

An artificial intelligence system is defined as "machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" [aia].

A machine learning model is defined as "an object (stored locally in a file) that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data" [Micnd].

2.1 Algorithmic Fairness and Bias

Computer scientists aim to mitigate discrimination by promoting algorithmic fairness and ensuring that AI applications operate impartially. To achieve this, fundamental questions must be addressed: What does fairness mean in the particular context? How can fair algorithms be developed? How can fairness be evaluated truthfully? These questions often lead to overlapping interpretations and misunderstandings, as the notion of fairness varies depending on the context [WXT+23].

There is no universal definition of fairness in artificial intelligence. Instead, there are multiple, and at times conflicting, definitions [BHN23]. In the context of decision-making, fairness is defined as the "absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" [SHD+19]. However,

different cultural norms and interpretations of fairness contribute to the development of diverse fairness frameworks [MMS+21]. Consequently, fairness must be evaluated on a case-by-case basis, as its meaning can differ depending on the specific circumstances. The existence of conflicting definitions underscores the complexity and challenges of addressing fairness in AI systems.

Two primary sources of unfairness are often identified: bias within the data itself and bias within the design of the algorithm [MMS+21]. Bias refers to the (un)intentional skewing of AI systems, which can arise at various stages, including biased training data, flawed algorithmic design, or the misinterpretation of model outcomes [OC20]. Bias is both a cause and a source of discrimination in AI systems [SMHP24].

Extensive research has been conducted on algorithmic fairness over the past decade [OC20, PS22]. However, the vast majority of studies propose generic, one-size-fits-all solutions without considering the specific regulatory frameworks, unique data characteristics, or protected attributes relevant to different domains [LO24]. These factors are essential in determining how fairness can be effectively implemented, as different sectors face distinct challenges, requiring tailored technical solutions. Figure 2.1 reveals the explosive growth in algorithmic fairness research during the past decade, highlighting a more than twofold increase in domain-specific studies within the last ten years.

The most studied domains are health, criminal justice, and employment. In contrast, the

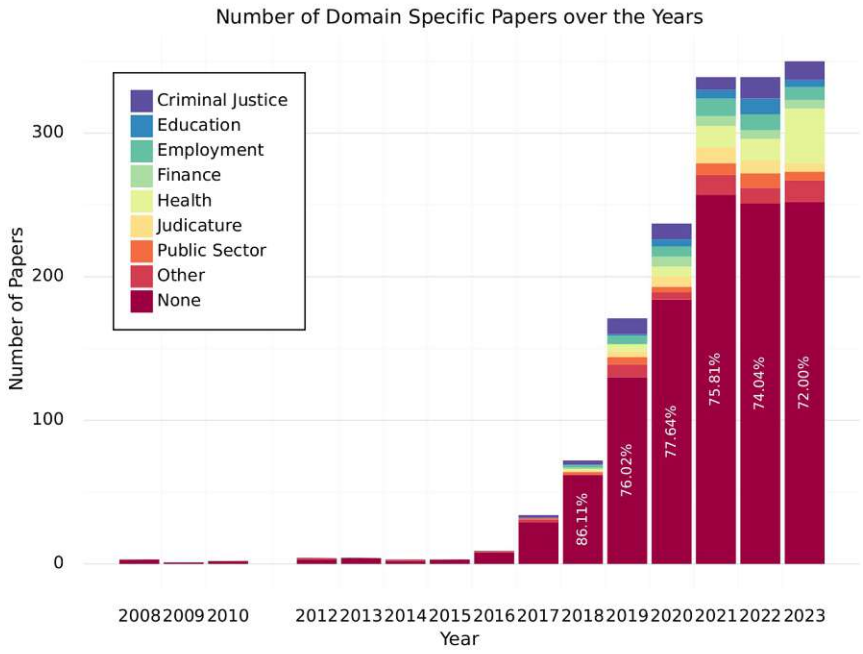


Figure 2.1: Number of domain-specific papers over the last 15 years, sourced from [LO24]

insurance sector falls under the *'Other'* category, with only three out of 1,570 papers addressing specific approaches for the insurance industry [LO24]. This highlights a significant research gap and underscores the need for domain-specific approaches that address both technological challenges and regulatory standards.

Most existing studies on algorithmic fairness in insurance focus on actuarial fairness, primarily investigating discrimination risks in underwriting [LRTW22][Xa24][CHR25]. Others explore how fairness principles from insurance pricing can be applied to machine learning, given the shared challenges of managing uncertainty, fairness, and accountability [FW24]. Additionally, moral trade-offs between non-discrimination and predictive accuracy in risk prediction have been examined, highlighting the complexity of balancing fairness with actuarial precision [LC21].
Beyond fairness in underwriting, further studies have analyzed broader challenges arising from AI-driven insurance models, including emerging regulatory concerns and technological shifts [OR22]. The dilemma of choosing a suitable fairness metric is demonstrated once in a fictional example about fraud detection in insurance claims [RD21].

In conclusion, although research on algorithmic fairness in data-driven insurance exists, most studies concentrate on risk classification rather than exploring the specific implications of other machine learning algorithms. Addressing this gap is essential for developing fair and context-aware AI-driven solutions within the insurance sector.

2.2 Establishing Discrimination

This section examines both the theoretical framework of discrimination as presented in scholarly literature and the formal legal definitions established in Austrian and EU legislation.

2.2.1 Types of Discrimination

The concept of discrimination is interpreted and perceived differently across disciplines, including philosophy, psychology, computer science, and ethics, as well as by the general public. Scientific literature [MMS+21] distinguishes between several kinds of discrimination:

- **Direct discrimination** occurs "where one person [or a group of people] is treated less favorably than another is, has been, or would be treated in a comparable situation on grounds of" a protected attribute [EUr00]. Protected attributes are defined by EU law and explained in more detail in Section 2.3.

  An extreme example of direct discrimination would be an insurer refusing motor insurance to Bulgarian citizens based solely on EU statistics showing Bulgaria has the highest road fatality rate[1].

[1] https://transport.ec.europa.eu/, visited on 05/05/25

- **Indirect discrimination** occurs when "an apparently neutral provision, criterion, or practice would put persons of a protected group at a particular disadvantage compared with other persons, unless that provision, criterion, or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary" [EUr00]. In other words, indirect discrimination occurs when decisions based on seemingly neutral attributes disproportionately disadvantage protected groups.

  In 2024, foreign-born residents constituted 45.4% of Vienna's total population[2]. Charging higher premium rates for residents with an 1150 zip code exemplifies indirect discrimination based on ethnicity, as Rudolfsheim-Fünfhaus (15th district) has the highest foreign-origin population at 56.1%.

- **Systemic/institutional discrimination** refers to "legal rules, policies, practices or predominant cultural attitudes in either the public or private sector which create relative disadvantages for some groups, and privileges for other groups" [UN 09].

  An example of systemic discrimination is the gender pay gap, a widely recognized and institutionalized form of inequality[3]. The gender pay gap may result in men having better insurance coverage than women due to their financial advantage.

- **Statistical discrimination** occurs when visible characteristics are used to assign individuals to demographic groups, with aggregate group traits replacing individual assessment [MMS+21]. This approach often perpetuates inequality between demographic groups [O'N16].

  Insurance companies typically charge young drivers higher motor insurance premiums based on this demographic's elevated accident rates. Consequently, even young drivers with clean records and safe driving habits face higher premiums solely due to their age group. This statistical discrimination penalizes individuals who defy group stereotypes. Such discrimination is particularly relevant to this thesis, as ML methods frequently leverage statistical correlations rather than causal relationships.

- **Explainable discrimination** occurs when differential treatment between groups can be justified and explained [MMS+21], rendering it legal.

  The previously mentioned example of higher motor insurance premiums for younger drivers exemplifies this concept. Insurers can lawfully incorporate age into risk assessment and pricing models when supported by actuarial or statistical data [CfJC+20].

- **Unexplainable discrimination** refers to discrimination against a group or an individual, which is not justifiable. It is often arbitrary, biased, or rooted in systemic inequities [MMS+21].

[2] https://www.wien.gv.at/, visited on 05/06/25
[3] https://ourworldindata.org/, visited on 05/07/25

In the Austrian insurance market, unexplainable discrimination could occur if an insurer charged higher premiums to residents of Vienna's 15th district (Rudolfsheim-Fünfhaus) compared to the 19th district (Döbling) for identical home insurance policies covering the same property values and risks, without any actuarial justification for the difference. Despite similar crime rates, fire risks, and flooding probabilities between these districts, the company might base this pricing disparity solely on the higher proportion of residents with migration backgrounds in Rudolfsheim-Fünfhaus. Since this pricing difference lacks statistical justification, it constitutes illegal, unexplainable discrimination based on ethnic origin.

The various types of discrimination are not mutually exclusive categories but often overlap and intersect. Consider this hypothetical scenario: An insurance company determines motor insurance premiums based on vehicle-specific factors, such as car brand, model, weight, manufacturing year, and engine capacity. In this hypothetical scenario, men typically pay higher premiums than women, despite gender not being an explicit factor in risk assessment, suggesting indirect discrimination. Further investigation demonstrates that men predominantly drive faster, more expensive, and newer vehicles compared to women. This indicates that seemingly neutral vehicle characteristics inadvertently serve as proxies for gender. Should the insurer eliminate all vehicle-related factors to mitigate legal exposure? Since these factors enable accurate risk assessment and support the company's profit-oriented business model, this decision-making process qualifies as explainable and justifiable discrimination. Conversely, applying blanket premium increases to all male policyholders solely based on gender would constitute direct, unjustifiable statistical discrimination.

This thesis focuses on discrimination in the legal sense, particularly within the framework of the European Union (EU) and Austrian insurance law.

2.2.2 Discrimination under EU and Austrian Law

Compared to other regions globally, the EU places significant emphasis on trustworthy AI, resulting in a comprehensive regulatory landscape [WMR21]. EU legislation explicitly prohibits two forms of discrimination: direct and indirect [EUr00]. As is typical in legal contexts, certain carefully defined exceptions exist, which we will analyze in the following section, with particular focus on the private insurance sector.

Protected Attributes
EU law establishes protected attributes (also known as sensitive attributes), which cannot legally be used as grounds for discrimination. Such discrimination remains illegal even when unintentional.
The European Union's non-discrimination directives collectively prohibit discrimination based on six protected attributes: **age, disability, gender, religion or belief, racial or ethnic origin, and sexual orientation** [EUr00, rig12, UNI04].

Insurance companies operate on the principle of risk classification to generate profit. Consequently, in the insurance sector, EU non-discrimination directives only strictly prohibit discrimination based on **gender and ethnicity**, without allowing any exceptions [gen12, EUr00].

The following regulations and laws are relevant in the EU and Austria:

General Data Protection Regulation[4](GDPR) [GDP]
According to Article 9 of the GDPR, processing special categories of personal data is generally prohibited. Those include attributes such as racial or ethnic origin, sex, religious or political beliefs, disability, age, biometric data, health data, or sexual orientation. However, an exception applies either if individuals provide explicit consent for the use of their sensitive data, if legal obligations mandate the data processing, or if the primary purpose of the processing is to monitor, detect, and mitigate discrimination. In addition, Article 22 of the GDPR prohibits completely automated decisions, which means decisions made by computers or algorithms without any human involvement. If protected attributes are included in data processing, justification must be clear, and the use of such data should be limited to what is strictly necessary to promote fairness.
In conclusion, fairness is not the main focus in the GDPR, but it calls for transparency and general exclusion of personal data in automated decision-making.

The Directives on Equal Treatment of the European Union
The EU Equal Treatment Law is a set of regulations, directives, and case laws for different areas of life, like employment or access to goods and services, with additional regulations for financial service providers, including insurance companies. Council Directive 2004/113/EC [UNI04] implements the **principle of equal treatment between men and women in the access to and supply of goods and services**. Following this, the **Guidelines on the application of Council Directive 2004/113/EC to insurance** [gen12] were issued, providing further detail on the principle's impact on the insurance premium calculation and establishing no exceptions for legal discrimination on the grounds of gender in the insurance business.
Furthermore, EU member states are subject to the Council Directive 2000/43/EC [EUr00], also called the 'Anti-Racism Directive', implementing the **principle of equal treatment between persons irrespective of racial or ethnic origin**. Again, there are no exceptions to discrimination on the grounds of race in the insurance business.
Last but not least, Article 21 of the **CHARTER OF FUNDAMENTAL RIGHTS OF THE EUROPEAN UNION 2012/C 326/02** [rig12] determines that non-discrimination is a fundamental right. Article 23 adds "equality between women and men" to the fundamental rights in the European Union.

[4]https://dsgvo-gesetz.de/, visited on 05/06/25

Gleichbehandlungsgesetz - GlBG [GlB]
While EU regulations apply directly within all member states, EU directives only represent minimum standards. Member states are required to implement those standards in their own legal systems with the freedom to decide how the regulatory aim is to be achieved [DECG23, WXT+23].
Austria embedded the equality rulings derived from the EU directives in the **Austrian Equal Treatment Act**, also specifying the difference between direct and indirect discrimination.

EU Artificial Intelligence Act (EU AI Act) [aia]
The EU AI Act aims to prevent AI systems from discriminating based on characteristics such as race, gender, age, disability, religion, or other protected categories. It focuses on ensuring that AI systems are fair, transparent, and do not cause harm. The AI Act introduces a risk-based framework to ensure that AI systems are used responsibly. Depending on the risk category, AI systems are subject to specific rules, especially in sensitive areas like employment, education, and insurance. Like the GDPR, the EU AI Act does not mainly focus on fairness but implies transparent regulations on AI systems. AI systems designed for evaluating risks and determining pricing strategies for individuals in life and health insurance contexts are classified as high-risk under Annex III of the regulatory framework. This designation subjects these systems to additional regulatory requirements and operational limitations.

Ethics Guidelines for Trustworthy AI [DG19]
These guidelines are not legally binding but emphasize the importance of diversity, non-discrimination, and fairness in AI systems. The guidelines advocate for transparency in AI operations, enabling stakeholders to understand and challenge decisions, thereby promoting fairness and preventing discrimination. By adhering to these principles, the EU aims to foster AI developments that uphold ethical standards and protect fundamental human rights.

In conclusion, EU legislation explicitly mandates that insurance AI systems must not discriminate based on protected attributes like gender and ethnicity while requiring transparency and continuous monitoring of these systems. However, the regulatory framework notably lacks quantifiable metrics or mathematical definitions of fairness [WMR21]. This absence creates a significant implementation gap when translating legal anti-discrimination principles into algorithmic contexts. Developers and insurers must therefore navigate between imprecise legal standards and the mathematical precision required for technical implementation in AI systems.

2.3 Algorithmic Fairness Metrics

To address the absence of clear instructions on quantifiable fairness, data scientists and researchers have developed various mathematical fairness metrics that attempt to formalize anti-discrimination requirements into algorithmic constraints. When properly

designed, algorithmic systems can enhance our ability to detect discriminatory patterns [KLMS19], as they enable the application of well-defined metrics that formalize and quantify fairness [MMS+21].

A thoughtful approach to fairness implementation demands critical consideration of which metrics to employ, rather than defaulting to whichever measurement proves most convenient or straightforward to compute. A systematic literature review identified 33 objective metrics for ethical AI [PCA24]. Thus, only a conscientious review can identify which metrics are most suitable. This process is comparable to selecting a suitable performance metric. Imagine a dataset with high class imbalance. If only 1% of observations fall into class A, an AI model might achieve almost perfect accuracy by predicting class B for all observations. In scenarios like fraud detection, it is however crucial that the model does not miss the few, yet important, observations from class A. Therefore, recall or precision needs to be evaluated as well.
Metrics that describe the fairness performance of an AI system are referred to as fairness definitions or non-discrimination criteria [BHN23]. We will therefore use the terms "fairness metric" and "fairness definition" synonymously.

Because of the vast amount of available fairness metrics in literature, they are commonly categorized into group (or statistical) fairness, individual (or similarity-based) fairness, and subgroup fairness [MMS+21, PCA24, CCG+22]. The following section describes the respective definitions in more detail.

2.3.1 Fundamentals

For the remaining analysis, we assume the following notation:

- Let $S \in \{a, b\}$ be a binary sensitive (or protected) attribute. $S = a$ denotes the group membership for the privileged group A (also called majority group) and $S = b$ the unprivileged group B (also called minority group).
- Let $Y \in \{0, 1\}$ be the binary target variable.
  $Y = 0$ represents a negative outcome, whereas $Y = 1$ represents a positive outcome.
- Let $\hat{Y} \in \{0, 1\}$ be the binary predicted outcome variable by a machine learning model.
- Let $R \in [0, 1]$ be the prediction score of a machine learning model.

The joint distribution of those random variables allows us to explicitly decide whether or not a fairness definition is satisfied [BHN23].

Table 2.1 describes possible prediction outcomes and respective notations, which are commonly used to evaluate the fairness of ML models [RD21].

The following statistical measures are derived from the confusion matrix:

True Class	Predicted Class	Notation
$Y = 1$	$\hat{Y} = 1$	True Positive (TP)
$Y = 1$	$\hat{Y} = 0$	False Negative (FN)
$Y = 0$	$\hat{Y} = 0$	True Negative (TN)
$Y = 0$	$\hat{Y} = 1$	False Positive (FP)

Table 2.1: Tabular representation of possible model prediction outcomes

Prevalence, also called positive base rate, represents the proportion of actual positives relative to the total dataset.

$$Prevalence = \frac{TP + FN}{TN + FP + TP + FN} \tag{2.1}$$

True Positive Rate, also called recall or sensitivity, describes the proportion of correctly predicted positives relative to all actual positives.

$$TPR = \frac{TP}{TP + FN} \tag{2.2}$$

False Negative Rate, also called type 2 error, describes the proportion of actual positives, which were misclassified as negative, relative to all actual positives. FNR and TPR are complementary, as they together account for all actual positives.

$$FNR = \frac{FN}{TP + FN} = 1 - TPR \tag{2.3}$$

True Negative Rate, also called specificity, describes the proportion of correctly predicted negatives relative to all actual negatives.

$$TNR = \frac{TN}{TN + FP} \tag{2.4}$$

False Positive Rate, also called type 1 error, describes the proportion of actual negatives, which were misclassified as positive, relative to all actual negatives. TNR and FPR are complementary, as they together account for all actual negatives.

$$FPR = \frac{FP}{TN + FP} = 1 - TNR \tag{2.5}$$

False Omission Rate describes the proportion of actual positives, that were incorrectly predicted as negatives, relative to all predicted negatives.

$$FOR = \frac{FN}{TN + FN} \tag{2.6}$$

Negative Predictive Value describes the proportion of correctly predicted negatives, relative to all predicted negatives. FOR and NPV are complementary, as they together account for all predicted negatives.

$$NPV = \frac{TN}{TN + FN} = 1 - FOR \tag{2.7}$$

False Discovery Rate describes the proportion of actual negatives, that were incorrectly predicted as positives, relative to all predicted positives.

$$FDR = \frac{FP}{TP + FP} \tag{2.8}$$

Positive Predicted Value, also called precision, describes the proportion of correctly predicted positives, relative to all predicted positives. FDR and PPV are complementary, as they together account for all predicted positives.

$$PPV = \frac{TP}{TP + FP} = 1 - FDR \tag{2.9}$$

2.3.2 Group Fairness

Group fairness aims to treat different groups, which are defined by one or several distinct values of a sensitive attribute, equally [SMHP24]. Therefore, the above-described classification performance metrics need to be equal across all groups [PCA24].

The comparison of metrics across all groups can be done using either (absolute) differences or ratios [KPB+24, PCA24]. Ratios and relative values allow to determine the direction of the discrimination and to account for extreme class imbalances. For any of the above-mentioned classification performance metrics, the respective ratio R is calculated as follows:

$$\text{Ratio R} = \frac{rate_{S=b}}{rate_{S=a}} \tag{2.10}$$

Throughout this analysis, ratios are consistently calculated with Group A (majority group) in the denominator and Group B (minority group) in the numerator.

The direction of discrimination revealed by these metrics depends critically on their contextual interpretation. When $TPRR < 1$, a clear disadvantage exists for the minority group, the model identifies their positive outcomes less accurately than those of the majority group. Conversely, the implications of $FPRR < 1$ require nuanced evaluation. Whether a lower score FPR for the minority group represents preferential treatment or systematic disadvantage hinges entirely on the specific domain context and the consequences of false classifications within that environment. $R = 1$ represents perfect parity, therefore perfect fairness [PS22, CDPF+17].

The large number of group fairness metrics found in literature is commonly divided into three non-discrimination criteria: independence, separation, and sufficiency [BHN23, RD21, Žl17]. Table 2.2 illustrates the respective conditional independence statements per non-discrimination criteria.

Criteria	**Independence**	**Separation**	**Sufficiency**
Conditional independence statement	$\hat{Y} \perp S$	$\hat{Y} \perp S \mid Y$	$Y \perp S \mid \hat{Y}$

Table 2.2: The three non-discrimination criteria from [BHN23]

The statements are to be understood as follows:

- Independence: Sensitive attribute S is unconditionally independent of the prediction $\hat{Y}$. This means that the probability of receiving any particular prediction should be the same across different demographic groups, regardless of the underlying distribution of true outcomes.
- Separation: Sensitive attribute S is conditionally independent of the prediction $\hat{Y}$, given the true output value Y. This means that among all individuals who share the same true label, the probability of receiving any particular prediction should be equal across demographic groups.
- Sufficiency: Sensitive attribute S is conditionally independent of the true output value Y, given the resulting prediction $\hat{Y}$. This means that among all individuals who receive the same prediction from the model, the probability of having any particular true label should be equal across demographic groups.

The following sections analyze the mathematical properties of fairness metrics for each criterion, followed by an examination of their inherent trade-offs, which mathematically prohibit the simultaneous satisfaction of all three non-discrimination criteria.

1. Independence

Fairness metrics that fall into the independence category only take into consideration the sensitive attribute S and the model prediction $\hat{Y}$. In other words, the true output value Y is irrelevant for determining whether or not algorithmic fairness is given.
The following metrics all examine the acceptance rate proportion $P(\hat{Y} = 1)$ across different sensitive groups. Each metric approaches this analysis from a unique perspective, illuminating various dimensions of potential disparities [BHN23]. The acceptance rate is the proportion of all predicted positives relative to the total dataset.

$$\text{Acceptance Rate} = \frac{TP + FP}{TP + FP + TN + FN} \tag{2.11}$$

The resulting conditional probability for independence is:

$$P(\hat{Y} = 1 | S = a) = P(\hat{Y} = 1 | S = b) \tag{2.12}$$

Several fairness metrics employ acceptance rate as their foundation, yet they diverge in their specific computational approaches and mathematical formulations.

Disparate Impact is a popular fairness definition, which descends from the corresponding law of the United States [U.S21]. It defines the "80% rule", which says that the acceptance rate of one (unprivileged) group must not be less than 80% of the other (privileged) group.
In other words, the likelihood of a positive outcome should be the same for group A as for group B [MMS+21].

$$\frac{P(\hat{Y} = 1|S = b)}{P(\hat{Y} = 1|S = a)} \geq 1 - \epsilon \tag{2.13}$$

Where $\epsilon = 0.2$. For perfect fairness, the disparate impact needs to equal 1 with $\epsilon = 0.0$ [PS22, SMHP24].

Demographic Parity, also known as **statistical parity**, seeks to establish equivalent acceptance rates across groups A and B [RD21]. While conceptually similar to disparate impact, this metric quantifies disparity through an arithmetic difference rather than a proportional ratio.

$$P(\hat{Y} = 1|S = a) = P(\hat{Y} = 1|S = b) \tag{2.14}$$

Conditional Statistical Parity is a relaxation of demographic parity. It allows a set of legitimate factors L to affect the prediction [RD21, CDPF+17]. The definition is satisfied if members in both subgroups have equal acceptance rates while controlling for a set of legitimate factors L [MMS+21]. The set of legitimate factors L consists of variables that are considered ethically acceptable and legally permissible to influence decisions.

$$P(\hat{Y} = 1|L = 1, S = a) = P(\hat{Y} = 1|L = 1, S = b) \tag{2.15}$$

Conditional Statistical Parity (CSP) proves particularly suitable in scenarios such as bank loan evaluations across racial demographics when adjusting for legitimate determinants like income and credit history. A model adhering to CSP ensures that any observed disparities in approval rates between racial groups stem exclusively from differences in these economically relevant factors rather than discriminatory practices. Omitting these legitimate variables from consideration could paradoxically introduce unfairness by ignoring genuinely predictive financial indicators that legitimately influence creditworthiness assessments.

Equal Selection Parity focuses on equal absolute numbers of favorable predictions [RD21, SMHP24]. In other words, the objective is to have the same absolute number of positive predictions for groups A and B, independent of their group sizes [JSeS+24].

Limitations of independence

Fairness metrics that fall into the independence category have the clear advantage of being easy to compute, making them a popular choice among researchers [BHN23]. Additionally, independent metrics are meaningful in real-world scenarios where true labels depend on the sensitive attributes, but the origin of this inequality is discrimination [SMHP24].

However, all the metrics discussed above completely ignore the error rates of a machine learning model and focus solely on the acceptance rate. As a consequence, they may encourage the generation of false positives, which can negatively impact model performance when optimizing for fairness.

Moreover, these metrics assume an equal claim to acceptance. Yet, the very need for fairness checks suggests that researchers acknowledge some level of heterogeneity in relevant factors. Consider a scenario where groups A and B represent male and female applicants for 20 open nursery teaching positions. To satisfy independence metrics, an equal number of male and female applicants would need to be accepted. If all 10 female applicants are highly qualified, whereas only 5 out of 10 male applicants meet the qualifications, this approach would lead to an increased number of unqualified hires among male applicants, ultimately harming their track record.

2. Separation

To overcome the above-mentioned weaknesses, the separation criterion extends independence to be conditional on the actual target value Y [PS22]. Separation requires that prediction $\hat{Y}$ be conditionally independent of sensitive attribute S given the actual target variable Y [BHN23]. In other words, the separation criterion is met when the proportion of correct predictions (TN and TP) is the same for all groups [RD21], which inherently demands error rate parity.

There are several fairness metrics found in literature that fall into the separation category.

Equalized Odds aims for equal true positive rates (TPR) and equal false positive rates (FPR) for both groups [HPS16]. This means that the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for individuals independent of their group membership.

$$P(\hat{Y} = 1|S = a, Y = y) = P(\hat{Y} = 1|S = b, Y = y), y \in 0, 1 \tag{2.16}$$

By definition, the equalized odds metric requires a binary outcome: a model either satisfies the metric or it does not. **Average Odds Difference** can be used to calculate how far a model deviates from achieving equalized odds.

$$\frac{1}{2}(P(\hat{Y}=1 \mid S=b, Y=1,) - P(\hat{Y}=1 \mid S=a, Y=1) + P(\hat{Y}=1 \mid S=b, Y=0) - P(\hat{Y}=1 \mid S=a, Y=0)) \quad (2.17)$$

Equal Opportunity is a relaxation of equalized odds and only requires the groups to have equal true positive rates ([VR18, MMS+21].

$$P(\hat{Y}=1|S=a, Y=1) = P(\hat{Y}=1|S=b, Y=1) \quad (2.18)$$

Mathematically, if a classifier has equal $TPRs$ for both groups, the $FNRs$ are also equal [PCA24].

Predictive Equality focuses on the other condition of equalized odds. Here, only the type 1 error (FPR) rate needs to be equal among both groups. [CDPF+17, RD21].

$$P(\hat{Y}=1|S=a, Y=0) = P(\hat{Y}=1|S=b, Y=0) \quad (2.19)$$

Mathematically, if a classifier has equal $FPRs$ for both groups, the $TNRs$ are also equal [PCA24].

Balance, also called mean difference, describes a metric that uses the predicted probability score instead of the class label and compares the average score for both groups per class. It aims to reveal steadily lower scores in one (unpriviliged) group, which might go unnoticed in the binary classification tasks [RD21].

$$E[\hat{Y}|S=a] = E[\hat{Y}|S=b] \quad (2.20)$$

Where E represents the mean of prediction values $\hat{Y}$.

Treatment Equality is achieved when the ratio of false negatives to false positives is equal for both groups [PS22].

$$\frac{FN_{S=a}}{FP_{S=a}} = \frac{FN_{S=b}}{FP_{S=b}} \quad (2.21)$$

Limitations of separation

A classifier with perfect accuracy will inevitably satisfy the first three separation metrics and, consequently, be considered entirely fair under the measure of equalized odds. This suggests that improving fairness can also lead to improved accuracy [PS22]. Nevertheless, separation metrics are often criticized for assuming that both groups have representative and bias-free base rates (prevalence). However, an optimal predictive model may not necessarily yield equal error rates across all groups, particularly when the prevalence

value differs between them. Naturally, the group with fewer true positives is more likely to have a higher number of false negatives, as the classifier may underpredict positives due to their scarcity. Enforcing equal error rates in such cases can lead to a model with worse predictive performance for one group than it could otherwise achieve.
Such disparities in base rates often stem from the historical marginalization and discrimination of certain groups, particularly minorities [BHN23]. By refining these separation metrics, researchers and practitioners may be incentivized to collect more representative, bias-free data.

3. Sufficiency

Sufficiency metrics guarantee that the predicted probability of an outcome remains independent of group membership. This implies that individuals sharing the same prediction score R should exhibit similar actual probability of the outcome across different groups [VR18]. Thus, for each predicted score, the outcome is independent of the group membership [PS22]. This is often formalized as calibration within groups [BHN23]. At the individual level, this ensures that individuals from different groups, but with identical predictions, have equal likelihoods of receiving the correct label [RD21]. However, sufficiency does not guarantee that individuals with the same prediction always have the same actual outcome, only that their probabilities align correctly across groups.

Conditional Use Accuracy Equality (CUAE) ensures that the accuracy of predictions within each predicted class (both positive and negative) is the same across groups. In other words, both positive predicted value (PPV) and negative predictive value (NPV) need to be equal for group A and B [RD21].

$$\begin{aligned} & P(Y = 1 \mid S = a, \hat{Y} = 1) = P(Y = 1 \mid S = b, \hat{Y} = 1) \\ \wedge \quad & P(Y = 0 \mid S = a, \hat{Y} = 0) = P(Y = 0 \mid S = b, \hat{Y} = 0) \end{aligned} \tag{2.22}$$

Predictive Parity is a relaxed version of CUAE, which only conditions the PPV to be equal for both groups [RD21].

$$P(Y = 1|S = a, \hat{Y} = 1) = P(Y = 1|S = b, \hat{Y} = 1) \tag{2.23}$$

Test Fairness, also called **equal calibration** or calibration by group [SMHP24, PS22], aims for similar $PPVs$ for both groups for any predicted probability value. Therefore, disparities in predictions based solely on the sensitive attribute are prevented for individuals with the same true outcome.

$$P(Y = 1|S = a, R = r) = P(Y = 1|S = b, R = r), \forall r \in \mathcal{R} \tag{2.24}$$

This formulation states that the probability of the positive outcome given a prediction score should be independent of the sensitive attribute. Equal calibration is similar to predictive parity, except that it considers the fraction of correct positive predictions for any value of the prediction score S. Test fairness focuses on fairness in predictions $\hat{Y}$ given the true outcome Y, whereas predictive parity focuses on fairness in true outcomes Y given the predictions $\hat{Y}$.

Limitations of sufficiency
Equal calibration and predictive parity can conflict when the base rates (prevalence) differ across groups. In such cases, sufficiency can result in a higher FPR for one group or an increased FNR for another.

Limitations of Group Fairness

In conclusion, the three categorically different criteria equalize either the acceptance rate, the error rates, or the outcome frequency of the prediction score [BHN23].
Data scientists without comprehensive knowledge of algorithmic fairness would intuitively seek a universal solution that satisfies all fairness definitions at once. However, it has been proven that it is impossible to satisfy more than one criterion simultaneously [RD21] without certain constraints in place [KMR, PS22].

Independence and sufficiency are generally mutually exclusive. In most real-world datasets, sensitive attribute S and target variable Y are not independent. Consequently, one group has a higher positive base rate than the other. Satisfying independence would result in a higher FPR for one group than the other. However, for the ML model to be fair (equally calibrated), FPR and FNR should be similar across the groups. As a result, sufficiency and independence cannot both hold [BHN23].

Independence and separation are also mutually exclusive when the target value Y is binary. In the real world, most groups have different base rates, thus different proportions of actual positives. In such cases, incompatibility can occur between satisfying equalized odds (measure for separation) and demographic parity (measure for independence) [PS22]. Satisfying demographic parity would result in a higher FPR for one group or a higher FNR for the other group. Equalized odds however, requires the same FPR and TPR for both groups. Consequently, independence and separation cannot both hold.

Separation and sufficiency cannot be satisfied simultaneously if all possible prediction outcomes (TP, TN, FP, FN) are non-zero for both groups. Ensuring that both groups have the same error rate (satisfying treatment equality) might reduce how well predictions match reality and therefore violate sufficiency measures, such as predictive parity or test fairness [PS22].

Enforcing any two of the criteria simultaneously leads to degenerate solutions under too many constraints. The incompatibility of fairness definitions leads to a moral dilemma about which fairness definition is the correct one to evaluate and optimize. Deciding which fairness metric is relevant will be further discussed in section 2.3.5.

Group fairness metrics also receive criticism for being too simplistic, as they all build on the confusion matrix values. Assuming S, Y, and $\hat{Y}$ are binary, all metrics can be calculated using only 8 values. However, those do not disclose the causes and mechanisms that created the discrimination [BHN23].

Group fairness metrics, like equalized odds or statistical parity, require complete parity per definition [HPS16]. However, achieving complete parity in reality is very unlikely. Therefore, some researchers recommend determining a value for ϵ, which must not be exceeded. For example, for equal opportunity, this would mean:

$$\frac{P(\hat{Y} = 1|S = b, Y = 1)}{P(\hat{Y} = 1|S = a, Y = 1)} = 1 - \epsilon \tag{2.25}$$

The chosen value for ϵ depends on the specific context, regulatory requirements, and ethical considerations applicable to the specific use case. Domain experts and stakeholders must collaborate to establish acceptable fairness boundaries that balance technical capabilities with real-world implications. The determination of appropriate thresholds should be documented and justified transparently to ensure accountability in fairness assessments.

2.3.3 Individual Fairness

As the name implies, individual fairness focuses on individuals rather than on groups of individuals. The basic idea is that individuals with similar characteristics should receive similar predictions, independent of their group membership [PCA24, DHP+12]. Metrics that fall into the individual fairness category are also called similarity-based measures [VR18].

The need for individual fairness metrics arises from the limitation of group fairness metrics. As discussed, independence can lead to a negative track record for one group if an equal acceptance rate is forced despite an imbalance in base rates. Individuals with a positive true outcome might receive a negative prediction to satisfy fairness rates, whereas individuals with a negative true outcome generously receive a positive prediction. Group fairness definitions do not provide any indication as to which individuals should be selected [DHP+12].

Fairness Through Awareness (FTA) describes the idea that an algorithm is fair if it gives similar predictions to similar individuals by considering meaningful differences between them [DHP+12]. In other words, any two individuals who are similar with respect to a well-defined similarity metric should receive a similar outcome [PS22]. FTA employs the *Lipschitz condition* to guarantee that the disparity in model predictions between two individuals remains proportional to their measured similarity. This mathematical constraint enforces a bounded, controlled change in decision-making relative to differences in input data [DHP+12].

FTA requires a carefully chosen similarity metric, which defines what makes individuals similar and determines fair comparisons in a specific AI application.

The model predictions should not change drastically for small differences in input features, preventing unfair sensitivity to minor variations. Individual fairness emerges by ensuring that the difference in treatment between two individuals does not exceed a multiple of their own differences. By explicitly defining similarity, this approach ensures fair, justifiable, and stable decision-making while respecting meaningful individual differences.

Limitations of FTA [Fle21]
Fairness through awareness faces fundamental challenges related to individual fairness. Defining a similarity metric depends on human judgment, requiring decisions on relevant features and acceptable similarity thresholds, which introduces subjectivity and pre-existing moral choices about fairness. The dependence on human judgment also increases the risk of encoding human biases and prejudice into the fairness metric.
Additionally, measuring similarity across large datasets is computationally expensive, limiting its practicality. Moreover, ensuring similar treatment does not guarantee fairness, as a system could treat individuals equally yet still unjust.
Due to these challenges, FTA should not be seen as a definitive fairness measure but rather as one tool among multiple approaches to address algorithmic bias comprehensively.

Causal Fairness ensures fairness by modeling cause-and-effect relationships in the data using causal graphs (directed acyclic graphs) to identify whether a sensitive attribute has a direct or indirect influence on the model's decision. It is typically implemented through causal analysis and counterfactual reasoning, enabling the differentiation between fair and unfair dependencies [BHN23].

To address unfair dependencies, causal methods attempt to identify and disregard certain causal pathways that link sensitive attributes to decision outcomes using, for example, a causal-based framework to detect and mitigate both direct and indirect discrimination [ZWW17].

Counterfactual fairness is a specific approach within causal fairness that assesses fairness by evaluating whether an individual's prediction remains the same in a hypothetical world where only their protected attribute is changed, while all other relevant factors remain constant. This definition is based on the intuition that a decision is fair if it does not change when an individual is placed in a different demographic group [KLRS17]. However, generating meaningful counterfactuals depends on a well-defined structural model, which is hard to construct in practice.

Limitations of causal fairness
A key challenge in applying causal fairness is that it requires knowledge of the underlying causal structure, which is often unknown or difficult to infer accurately. Different causal models can lead to vastly different conclusions, making it challenging to establish a definitive measure of fairness [CCG+22].
In addition, causal fairness methods are impractical for datasets with a large number

of features, as modeling all possible attribute-value relationships becomes computationally expensive. Completely removing all causal links between a protected attribute and the decision may lead to significant loss in model utility and predictive accuracy [SMHP24].

2.3.4 Subgroup Fairness

Instead of comparing only generic demographic groups, such as men and women, subgroup fairness requires more granular subgroups, such as young women or middle-aged men. Subgroup fairness seeks to bridge the gap between group and individual fairness by addressing their respective limitations. The idea is to use group fairness metrics for evaluation while incorporating elements of individual fairness [MMS+21]. Therefore, subgroup fairness ensures that the constraint holds over a diverse set of subgroups. This approach helps detect hidden biases that may not be visible in traditional group fairness metrics.

Limitations of subgroup fairness
It might be difficult to identify relevant subgroups, especially in high-dimensional data where numerous overlapping subgroups exist. Consequently, it is computationally demanding to ensure fairness across all of them. Additionally, optimizing for fairness across many subgroups increases the risk of overfitting, where the model becomes too tailored to specific fairness constraints, potentially reducing the overall predictive model performance. Another major issue is the presence of conflicting fairness constraints, as different subgroups may have competing needs. Adjusting for fairness in one subgroup might unintentionally introduce bias against another, making it difficult to balance trade-offs effectively. Individual and group fairness can sometimes be incompatible as well. There's a trade-off between demographic parity and individual fairness, and they cannot be satisfied simultaneously except in trivial degenerate solutions [DHP+12].

2.3.5 Deciding which fairness metric is relevant

Each fairness definition has its advantages and drawbacks, and real-world applications often aim for at least similar outcomes across groups. A key consideration is whether the consequence of an incorrect positive prediction is punitive or supportive. If punitive, fairness measures should account for false positives to avoid wrongly penalizing individuals. If supportive, they should prioritize false negatives to ensure those in need are not denied assistance.

Maximizing profit requires companies to balance a reward for true positives and a cost for false positives [JSeS+24], while also considering the consequences of incorrectly classifying true positives as negatives. The choice of fairness metric ultimately depends on the preferred error rate, which in turn is influenced by the consequences of incorrect predictions. Decision trees provide a structured approach to operationalizing this choice [JSeS+24, RD21].

Figure 2.2 provides a visual overview of fairness metrics.

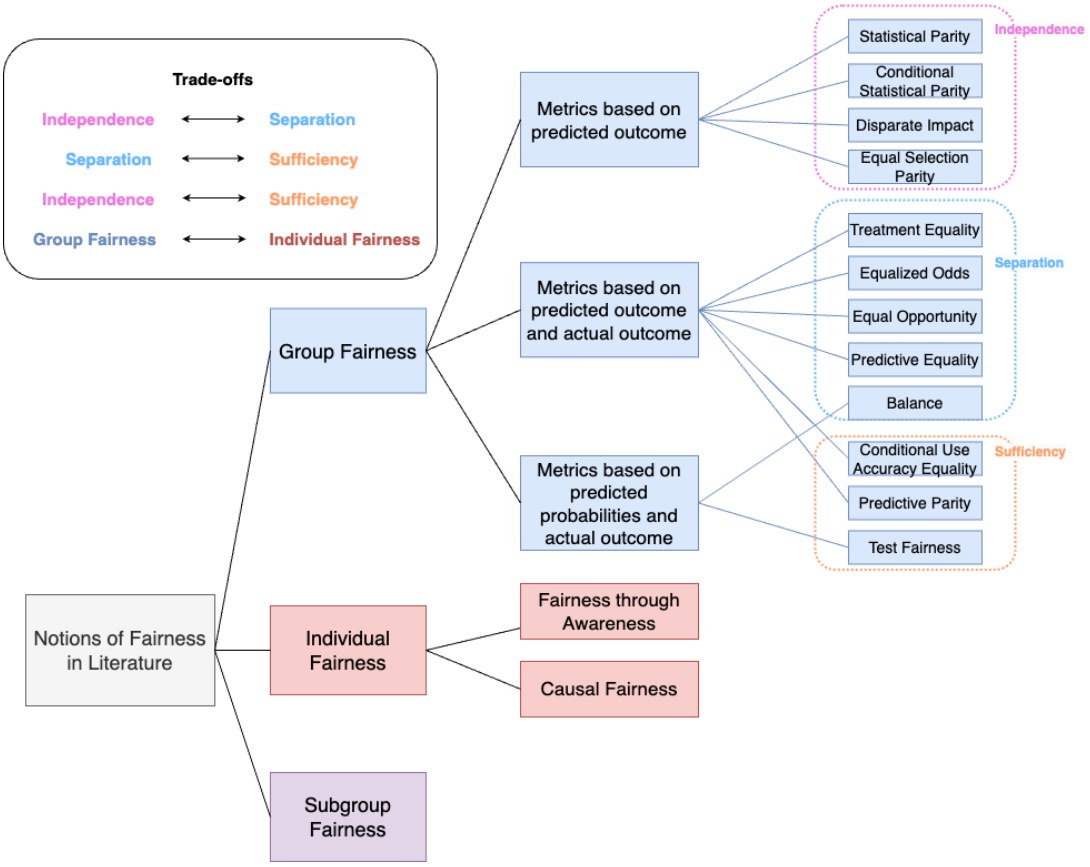


Figure 2.2: Group fairness metrics, adapted from [PS22]

Ultimately, fairness metrics must be evaluated within their legal, social, and ethical contexts [PS22].

2.4 Discrimination Mitigation Methods

This section analyzes how the fairness definitions can be optimized algorithmically. The general terms for improving fairness are discrimination mitigation or bias mitigation. Throughout this work, the terms method, technique, and strategy are used interchangeably to represent proposed solutions for fairness improvement. Research has identified over 340 publications until July 2022 addressing bias mitigation for ML classifiers [HCZ+24]. These methods are often domain-specific, with each technique targeting a different problem in different areas of machine learning [MMS+21].
Generally, algorithmic bias mitigation approaches can be categorized into three groups [BHN23]:

- **Pre-processing methods:** Transform the input data (feature space) to remove

underlying discrimination before model training.

- **In-processing methods:** Modify state-of-the-art ML algorithms to reduce discrimination during model training.
- **Post-processing methods:** Adjust the outputs of trained models to remove discriminatory patterns.

2.4.1 Pre-processing Methods

The main objective of pre-processing mitigation methods is to adjust the input data to be uncorrelated to the protected attribute(s). As a result, it ensures independence not only in the feature space but also for the training process [MMS+21].

Pre-processing techniques are further organized into the following categories [HCZ+24, CH24, DL19]:

- **Relabeling and Perturbation:** Changing the true target labels (relabeling) or the input features (perturbation).
- **(Re)sampling:** Changing the sample distribution by adding or removing samples or by adjusting their effect on training (e.g., reweighing).
- **Latent variables:** Augmenting the training data with additional features that are preferably unbiased.
- **Fair representations:** Learning a transformation of the training data so that bias is reduced while maintaining as much information as possible.

Limitations of pre-processing mitigation techniques
Empirical comparisons have demonstrated that pre-processing methods typically perform worse than in-processing and post-processing approaches [FSV+18]. Additionally, modifying training data requires access to raw data, which is not guaranteed and computationally expensive for complex ML models in real-world applications.

2.4.2 In-training Methods

In-processing methods aim to improve fairness directly during model training [WZLZ23]. By directly integrating fairness criteria into the optimization process and providing in-training feedback, they ensure the model inherently balances prediction accuracy with fairness considerations.

In-processing techniques are further organized into the following categories [HCZ+24, CH24, DL19]:

- **Regularization and Constraints:** Extending the learning algorithm's loss function to penalize unfair outcomes (regularization) or establish specific bias thresholds that cannot be exceeded during training (constraints).
- **Adversarial Learning:** Training a primary classifier that predicts ground truth values and a competing adversary designed to detect fairness issues at the same time. As a result, the classifier iteratively learns to make accurate predictions while preventing the adversary from identifying discriminatory patterns in its outputs.
- **Compositional:** Training separate classifiers for different demographic groups, maintaining high subgroup accuracy while achieving overall fairness. Predictions are generated either by using group-specific models or through ensemble techniques that combine outputs from multiple classifiers.
- **Adjusted Learning:** Modifying standard ML algorithms or developing entirely new algorithms that include fairness awareness considerations.

Limitations of in-processing mitigation techniques
In-processing techniques require direct access and permission to modify the ML model training process. In addition, certain approaches propose customized, model-specific implementations, limiting their compatibility. They also result in significant computational overhead, which increases training complexity and resource demands. Furthermore, some approaches, such as adversarial methods, can compromise model interpretability and reduce transparency in the decision-making process. Finally, in-processing methods can result in more significant accuracy-fairness trade-offs, especially when fairness constraints are strictly enforced throughout the training process [CBJ+23].

2.4.3 Post-processing Methods

Post-processing mitigation methods are applied after a classification model has been fully trained. The trained classifier or the classification output is adjusted so that detected discriminatory patterns in the outcome are reduced or eliminated. The main advantage is compatibility with any black-box classifier, eliminating the need to access the original data or the training pipeline. Methods in this category are often the only viable option when practitioners only have access to the trained model without control over the training or pre-processing [MMS+21].

Post-processing techniques are further organized into the following categories [HCZ+24, CH24, DL19]:

- **Input Correction:** Modifying input data to an already trained model through pre-processing techniques such as relabeling, perturbation, or representation learning. Input correction approaches apply these transformations as an additional layer before passing data through an already trained algorithm.

- **Classifier Correction:** Constructing another classifier related to the initial classifier, which yields fairer decisions. Because approaches in this category apply modifications directly to the model, they are also called intra-processing.
- **Output Correction:** Modifying the predicted outcome, either by changing the prediction threshold or predicted labels.

Limitations of Post-processing Methods
Post-processing techniques typically require access to protected attributes during inference, which becomes particularly problematic when working with discretized predictions and can paradoxically result in legal concerns with anti-discrimination regulations. In addition, these methods often lead to greater performance degradation compared to in-processing approaches, as they sacrifice model accuracy by altering outputs that were already optimized during training [CBJ+23]. Furthermore, post-processing techniques cannot address discrimination issues embedded in the model's learned representations or originating from biased training data. Finally, these approaches risk introducing new fairness problems when optimizing for specific metrics, and their uniform modifications often fail to account for individual nuances, potentially increasing unfairness at the individual level.

2.4.4 Choosing a suitable discrimination mitigation method

In conclusion, there is no one-size-fits-all solution for mitigating discrimination in AI systems. Figure 2.3 shows the number of publications per category as described above. Most research focuses on in-processing methods, while post-processing approaches have received the least attention.

Choosing a suitable mitigation method depends on several factors. These include access to the ML pipeline and protected attributes, the fairness definition in need of being optimized, the importance of explainability and interpretability, and what legal requirements must be met [SMHP24]. Finally, there is an inherent trade-off between fairness and accuracy that requires careful consideration of performance requirements when implementing fairness measures [KMR].

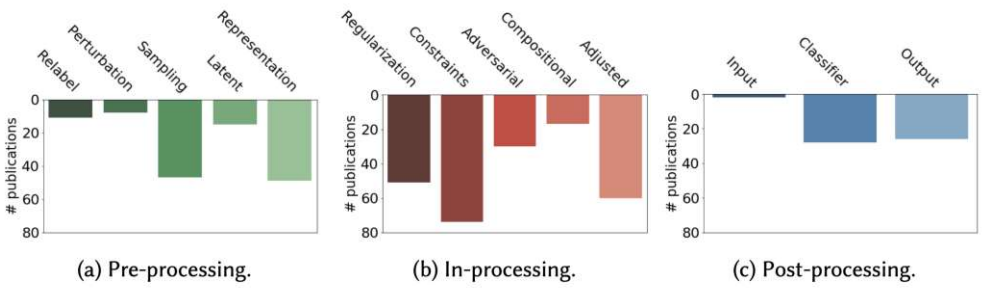


Figure 2.3: Number of publications per mitigation category, sourced from [HCZ+24]

CHAPTER 3

Methodology

This chapter describes the real-world case study from an Austrian insurance company in detail, including the dataset, baseline model, and model performance. The findings from the literature research in Section 2 are then analyzed for suitability in this case study. The implementation and interpretation of suitable fairness metrics and discrimination mitigation methods are discussed to provide a comprehensive understanding of how fairness is pursued in this thesis. Finally, the evaluation methods and metrics used to scientifically answer the research questions from Section 1.1 are explained.

3.1 Case Study Description: Explosive Claims

The case study is provided by an Austrian insurance company, where it was originally developed in 2022. However, judicial regulations (see Section 2.2.2) necessitate a revision of already implemented projects to ensure compliance with relevant anti-discrimination laws.

Private insurance companies need to continuously estimate and adjust their reserves to ensure sufficient payment liquidity to cover all claims compensations. In Motor Third Party Liability (MTPL) insurance, certain claims can lead to substantial payouts, especially when severe injuries result in long-term disability of insured people. The case study title 'Explosive Claims' refers to claims that meet either of two criteria: those requiring reserve increases exceeding €100,000 or those where the reserve amount grows by a factor of ten or more when compared to the cut-off date (which is defined as one month after the initial reporting date). When such a claim is identified, the reserves are adjusted, and the insured person is immediately provided with the best possible care to prevent long-term payouts.

This case study examines the development of an AI model to complement human decision-making for rapid and accurate identification of explosive claims. Early detection benefits

both parties: The insurance company can properly adjust reserves and maintain liquidity, while policyholders receive timely remedy. The AI model must therefore reliably identify explosive claims to enable prompt intervention.
At the same time, incorrectly identified explosive claims should be minimized to avoid wasting company resources and generating unnecessary expenses. In addition, uncalled-for supplementary checks for policyholders should be avoided, for example, when they are asked to answer further questions or provide additional documents.
Consequently, the optimal AI model prioritizes recall over precision, but only to an extent that maximizes operational efficiency while ensuring appropriate claims processing.

3.1.1 Data Characteristics and Protected Attributes

The Austrian insurance company provided the pre-processed dataset used to train the initial baseline model. The raw data included structured information about the insured person, as well as information about the accident in the form of unstructured claim reports (PDF documents). The raw data was subject to numerous pre-processing steps before reaching its final form for training. Since the raw data was not accessible for this thesis, the pre-processed dataset serves as the starting point. The pre-processed dataset contains encoded data and consists of 6,258 input features, also called feature space or input attributes. Thereof, 258 attributes contain structured information about the respective claim and policyholder, such as the date of the claim, the country where the claim occurred, and the gender and nationality of the policyholder. Besides that, the dataset includes 6,000 words or pairs of words (unigrams or bigrams) from the claim report, retrieved using a bag-of-words model. These uni- and bigram columns contain integer values representing the frequency of each word or word pair in the claim documentation. Previous analyses have identified the occurrence of certain uni- or bigrams, such as 'Krankenwagen' or 'schwer verletzt', as an indication of an explosive claim.

Section 2.2.2 describes the Austrian and European laws applicable to this use case and identifies relevant protected attributes. The analysis reveals that **gender and nationality** are both considered protected attributes and must not be grounds for discrimination. Consequently, these two attributes form the basis of the fairness analysis.

The dataset consists of claims reported between January 2010 and December 2021, comprising 406,981 rows in total, with each row representing one claim. The dataset was already split chronologically into training and test sets. Claims reported from 2010 through 2019 are used for training, while the remaining claims from 2020 and 2021 serve as test set. This division results in 348,904 (85.73%) claims in the training set and 58,077 (14.27%) claims in the test set.
A value of 1 of the binary target variable Y indicates an explosive claim, while 0 indicates a non-explosive claim.

Next, the data is explored with respect to the two protected attributes, label distribution, and representation of subgroups.

Protected Attribute: Gender

There are 3 distinct subgroups in the dataset: male, female, and unknown, encoded as 1.0, 2.0, and 3.0, respectively. The insurance company could not provide a clear explanation for why some claims are allocated to policy holders with 'unknown' gender designation. This may be attributed to database errors or legacy data entry processes with missing values. Additionally, Austrian legal changes have expanded gender categories: since January 2019, a 'diverse' gender entry has been permitted alongside 'female' and 'male'. Moreover, since September 2020, the options 'intersex', 'open', as well as deletion of the gender entry from civil status records have also become available[1].

Figure 3.1 demonstrates that both the training and test sets have similar gender distributions, with male policyholders accounting for 45% of all claims, while female policyholders and those with undisclosed gender each represent approximately 27% of claims.

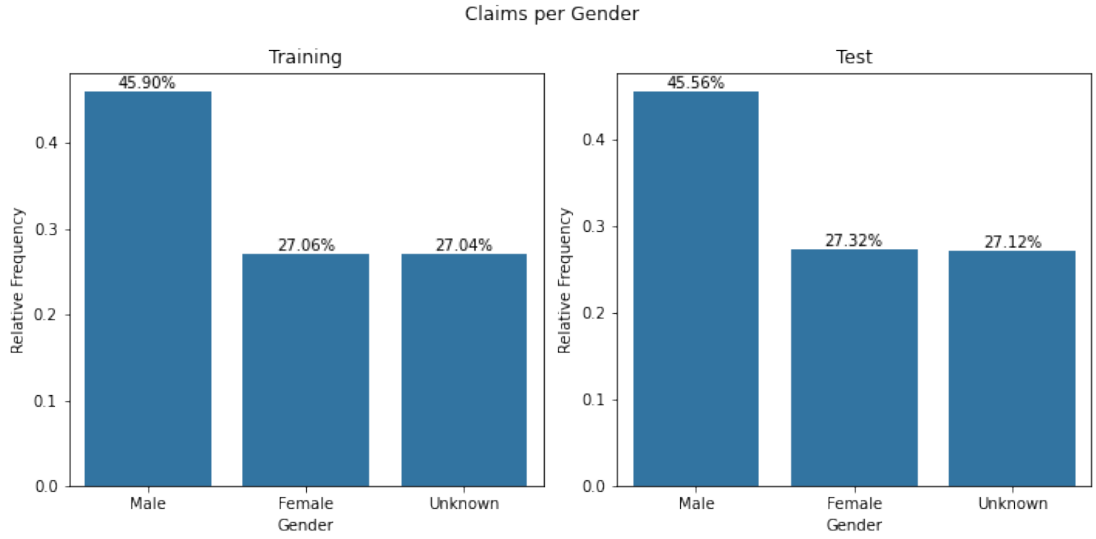


Figure 3.1: Gender distribution of insurance claims

Unfortunately, the overall distribution of policyholders or claims by gender was not disclosed. Therefore, it cannot be determined whether this distribution accurately represents the actual gender demographics within the insurance portfolio.

The dataset exhibits an extremely high class imbalance. The proportion of positive outcomes in the overall dataset is only 0.0009 (0.0909%). A positive outcome in this case study represents an explosive claim.
When comparing the label distributions by gender subgroup in Table 3.1, a similar class imbalance ratio for all groups is observed.

[1] https://www.wien.gv.at/menschen/queer/, visited on 05/28/25

Gender	Label: 0		Label: 1		Total
	Count	Percentage	Count	Percentage	
Male	186,422	99.90%	189	0.10%	186,611
Female	110,179	99.92%	90	0.08%	110,269
Unknown	110,010	99.92%	91	0.08%	110,101
Total	406,611	99.91%	370	0.09%	406,981

Table 3.1: Label distribution across gender subgroups

All subgroups reflect the overall high class imbalance ratio. Males have a marginally higher proportion of the positive label 1 (0.10%) compared to females and unknown (both 0.08%), but this difference is extremely small (0.02 percentage points).

Protected Attribute: Nationality

The dataset contains 102 distinct values for the nationality of the policyholder. However, as Table 3.2 shows, the large majority ($\sim$ 70%) of claims is reported by Austrian citizens in both training and test set. Only 1.65% of the claims in the training set and 2.41% in the test set are reported by non-Austrian policyholders. For 27% of the claims, the policyholder's nationality is unknown. The insurance company again provided no clear explanation for this issue, which may result from database errors or legacy data entry processes that failed to capture complete information.

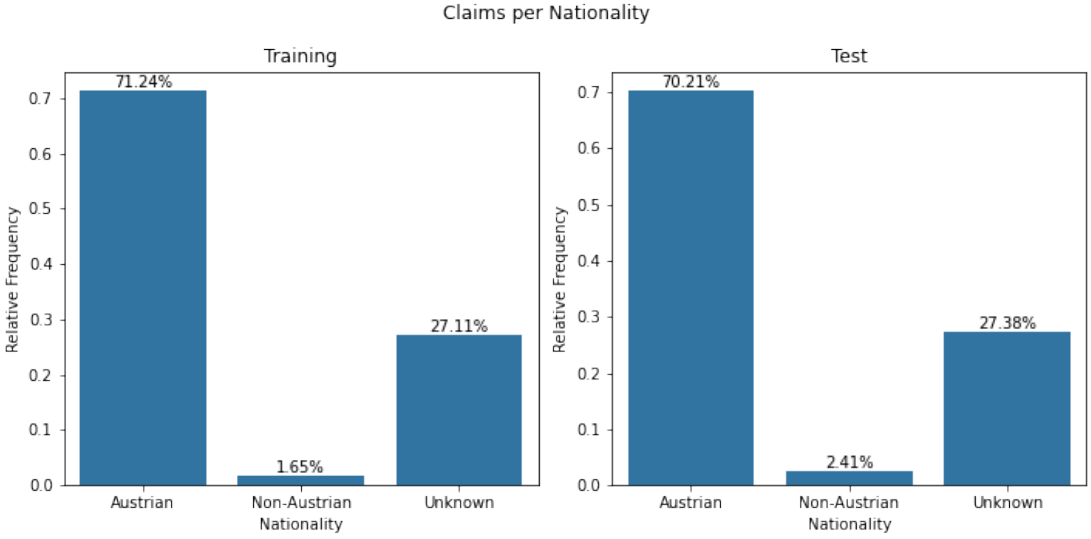


Figure 3.2: Nationality distribution of insurance claims

Again, it is unknown whether this reflects the true distribution of MTPL insurance policyholders by nationality.

Table 3.2 demonstrates that each nationality subgroup reflects the overall label distribution, with explosive claims consistently representing approximately 0.09% of all claims across categories.

Nationality	**Label: 0**		**Label: 1**		**Total**
	Count	Percentage	Count	Percentage	
Austrian	289,058	99.91%	273	0.09%	289,331
Non-Austrian	7,138	99.90%	7	0.10%	7,145
Unknown	110,415	99.92%	90	0.08%	110,505
Total	406,611	99.91%	370	0.09%	406,981

Table 3.2: Label distribution across nationality subgroups

Only 7,138 claims were reported by non-Austrian citizens, and thereof only 7 are explosive claims. Consequently, both model training and fairness evaluation might turn out to be challenging, depending on how many belong to the training and testing set, respectively.

In conclusion, the model will likely face challenges in accurately predicting explosive claims due to the significant class imbalance and the resulting bias toward predicting non-explosive claims. The scarcity of positive cases in the test set complicates thorough fairness evaluation. Importantly, the analysis reveals consistent label distribution patterns across gender and nationality subgroups, with no substantial differences observed between these demographic categories. The different group sizes (especially for nationality) might present an additional obstacle for fair model predictions.

3.1.2 Baseline model Description

The case study baseline model is a **Light Gradient-Boosting Machine**[2] (LGBM or LightGBM) [KMF+17] that predicts whether or not a claim will explode.
LGBM is an ensemble technique using multiple decision trees in a gradient boosting framework. That means that the algorithm iteratively creates models, where each new tree focuses on correcting errors made by previous trees by minimizing the negative gradient of the loss function. In LGBM, Gini impurity serves as a criterion for selecting split points of decision trees, measuring class distribution impurity to determine optimal splitting points that maximize information gain. The final result is a weighted average with shrinkage method lambda, where lambda acts as a regularization parameter that reduces each tree's contribution to prevent overfitting by scaling down predictions, thus improving generalization to unseen data.

LightGBM speeds up the training process of conventional Gradient Boosting Decision Trees (GBDT) by over 20 times while achieving almost the same accuracy and requiring lower memory usage [KMF+17].

These advantages stem from four main techniques:

[2] https://pypi.org/project/lightgbm/, visited on 05/11/25

1. **Faster training** through histogram-based splits (bin splitting). This approach sacrifices some accuracy for a significant gain in speed.

2. **Exclusive feature bundling** reduces feature count by combining mutually exclusive features. This speeds up tree building while slightly reducing accuracy for dense rows. Thus, LGBM is capable of handling large-scale data efficiently.

3. **Gradient-based one-side sampling** (GOSS) reduces loss functions. It sorts observations by gradient size and samples 20% from the largest gradients and 10% from the smallest. This splitting method results in new trees with only 30% of data, improving speed while maintaining accuracy.

4. **Leafwise tree growth** splits only one leaf at a time instead of depth-wise growth. It selects the leaf with the highest information gain. Without deliberate constraint, this creates unbalanced trees with varying branch depths until 100% purity is achieved.

Table 3.3 describes the parameter settings[3] used for the baseline model. The baseline model uses a decision threshold of 0.5 and achieves a recall score of 50%, and a precision score of 7.8%. The predictive performance is further discussed in Section 4.3.

3.2 Fairness Evaluation Metrics

This section presents the metrics applied during the fairness assessment and details the discrimination mitigation methods implemented to improve fair outcomes. The corresponding definitions were previously described in Sections 2.3 and 2.4.

As pointed out in Section 2.3, various definitions of fairness exist in scientific literature. There is no scientific consensus on which metric captures discrimination in AI systems best.
In order to scientifically answer RQ 1.1: "Which fairness metrics are most suitable to capture discrimination in the case study?", a broad variety of metrics is computed based on the baseline model classification output and evaluated for suitability. Critical factors include relevant legal implications, business use case context, and the particular objective of the AI model.

Section 2.3 not only described the metrics in detail but also pointed out their limitations. Based on these constraints, not all metrics will be incorporated into the fairness evaluation of the case study. Equal calibration metrics present significant challenges when applied to datasets with extreme class imbalance. When one class appears rarely, calibration becomes technically difficult due to insufficient samples for reliable probability estimation across groups. This can result in a misleading impression of fairness, as the metric may appear satisfied simply because the model rarely predicts the minority class for any

[3] https://lightgbm.readthedocs.io/, visited on 05/11/25

Parameter	Description	Value
boosting_type	Type of boosting algorithm	'gbdt'
colsample_bytree	Fraction of features used for each tree	0.84
importance_type	Method for feature importance	'split'
learning_rate	Rate at which each tree contributes to updates	0.151
max_depth	Maximum depth of trees to prevent overfitting	5
min_child_samples	Minimum samples required in a leaf node	4020
min_child_weight	Minimum sum of instance weights in a leaf	0.001
min_split_gain	Minimum gain required to split a node	0.0
n_estimators	Number of boosting iterations	100
n_jobs	Number of parallel threads used for training	-1
num_leaves	Maximum number of leaves in one tree	4
objective	Learning task and loss function	'binary'
reg_alpha	L1 regularization term to control overfitting	0.0
reg_lambda	L2 regularization term to control overfitting	0.0
subsample	Fraction of data used for each boosting iteration	1.0
subsample_for_bin	Number of samples used to construct bins	200000
subsample_freq	Frequency of subsampling during boosting	0
verbosity	Controls logging output, -1 for no output	-1
boost_from_average	Adjusts initialization using mean target value	False
feature_pre_filter	Indicates if features are filtered before training	False
lambda_l1	Strength of L1 regularization for feature selection	3.0
lambda_l2	Strength of L2 regularization to reduce overfitting	7.6
scale_pos_weight	Balances weight of positive and negative classes	131

Table 3.3: Baseline model parameters with descriptions from the official documentation.

group.
Furthermore, this thesis will not evaluate individual fairness metrics. Causal fairness approaches, including counterfactual fairness, face substantial implementation limitations when applied to large datasets without clear causal relationships. Without valid causal

assumptions, the effectiveness of these definitions becomes severely limited. Implementing 'fairness through awareness' in this case study would require extensive computational and human resources, which companies would rarely commit to in practical scenarios. Generally, the limitations of individual fairness outweigh any potential additional insights these metrics might provide.
Subgroup fairness will also remain outside the scope of this evaluation. The nationality subgroups already suffer from unequal representation in the dataset. Further subdividing the non-Austrian group would yield statistically insignificant results that could not support meaningful conclusions.

In conclusion, the following group fairness metrics will be computed and evaluated:

- Independence: Equal Selection Parity, Statistical Parity
- Separation: Treatment Equality, Balance, Equalized Odds, Equal Opportunity, Predictive Equality
- Sufficiency: Conditional Use Accuracy Equality, Predictive Parity

The previous section has revealed the extreme class imbalance in the dataset. Certain fairness metrics expose significant bias only after **normalization**, highlighting the impact of class imbalance [KPB+24]. For all basic statistical measures discussed in Section 2.3.1, their normalized variants are also computed. For any ratio R (see Equation 2.10), the normalization is computed as follows:

$$\text{Normalized Ratio} = \frac{R}{R+1} \tag{3.1}$$

The normalization of ratio values maps the original unbounded ratios to the interval $[0, 1]$, where 0.5 indicates equal rates between groups and values closer to the extremes indicate greater disparity in rates. It is important to note that the interpretation of normalized values depends on the specific metric being evaluated. The same normalized value (e.g., 0.8) can have opposite fairness implications depending on whether TPRR is examined (where higher values for the minority group may indicate better performance) versus FORR (where higher values for the minority group may indicate harmful bias). This context-dependent interpretation must be considered when analyzing results across different fairness metrics.

The proposed normalization is mathematically equivalent to a logistic transformation of the log-odds ratio. Specifically:

$$\frac{R}{R+1} = \frac{1}{1+\frac{1}{R}} = \frac{1}{1+e^{-\log(R)}} \tag{3.2}$$

Equation 3.2 holds for R $>$ 0, which is the case for all fairness ratios in this case study. This connection to the logistic function, commonly used in machine learning

for probability calibration[4], provides theoretical grounding for the transformation. The normalized ratio can be interpreted as a measure of the relative magnitude of disparity between groups on a standardized scale.

While this normalization does not eliminate the need for metric-specific interpretation of fairness implications, it provides a consistent framework for measuring and comparing the magnitude of rate disparities across different algorithmic fairness metrics in the context of extreme class imbalance.
In Section 4, both normalized and unnormalized fairness ratios will be systematically compared and evaluated for their analytical suitability and statistical significance. The bounded nature of normalized ratios enables more stable statistical comparisons across metrics with different baseline rates. In addition, comparing both normalized and unnormalized approaches allows for validation that observed fairness violations are robust to the choice of measurement scale, thereby strengthening the reliability of the conclusions about algorithmic bias in severely imbalanced datasets.

Some fairness metrics rely on absolute differences rather than ratios. Again, the normalized variants are calculated to account for the extreme class imbalance. For example, predictive equality requires the same FPR values across groups (see Equation 2.19). The normalized variant is calculated as follows [KPB+24]:

$$\text{Normalized Predictive Equality} = \frac{|\text{FPR}_{S=b} - \text{FPR}_{S=a}|}{\max\{\text{FPR}_{S=b}, \text{FPR}_{S=a}\}} \tag{3.3}$$

Another critical matter is the **fairness threshold** from which a metric is considered unfair. For this case study, the Austrian insurance company did not impose specific fairness thresholds. Therefore, the fairness metrics are evaluated and compared without tolerance threshold, as originally defined in literature [HPS16, BHN23].

3.3 Discrimination Mitigation Techniques

Discrimination mitigation methods in real-world business scenarios should be carefully chosen to maximize profit for the company [BHN23]. Unfortunately, the Austrian insurance company did not indicate quantitative amounts for the cost of false positives or the reward for true positives. Therefore, it is difficult to calculate the optimal trade-off threshold.
In order to decide which mitigation technique is most suitable, it is necessary to analyze the meaning of false positives and negatives in the case study once more. False positives, as well as false negatives, represent tangible costs to satisfy popular notions of algorithmic fairness, which need to be weighed against each other. In this case study, false positives mean that the model unnecessarily flags a low-risk claim as high-risk. This might lead to undesirable actions, such as avoidable costs for medical treatments and more workload

[4]https://scikit-learn.org/stable/modules/calibration.html, visited on 06/01/25

for insurance employees. The policyholder enjoys better caretaking on the other hand. False negatives mean that the model fails to identify a high-risk claim that is likely to explode. This results in high, unforeseen costs for the insurance company and worse long-term conditions for the insured person.
In conclusion, false negatives result in worse consequences for both insurance company and policyholder. Thus, the mitigation techniques should result in equal (low) false negative rates, which means similar (high) true positive rates.

Although pre-processing mitigation methods fall outside the scope of this thesis, the 'fairness through unawareness' (FTU) approach is incorporated into the analysis. FTA describes a simple pre-processing approach that involves removing protected attributes and their obvious proxies from the dataset before model training [MMS+21]. The approach is based on the assumption that if a classifier cannot access protected attribute values, it cannot directly discriminate based on them. While the baseline LGBM model utilizes both gender and nationality as input features, all models in the fairness analysis exclude these protected attributes from the training data. To distinguish between the effects attributable to FTU and those resulting from other mitigation techniques, it is essential to evaluate FTU independently. Solely using this method for discrimination mitigation is widely criticized. It fails to prevent indirect discrimination, as models can still infer the eliminated protected attributes from other seemingly neutral features [DHP+12, SMHP24].

3.3.1 In-processing Mitigation Methods

In-processing methods aim to improve fairness directly during the model training [WZLZ23]. There are more than 100 publications that discuss in-processing methods, each introducing different approaches [HCZ+24]. There is no one-size-fits-all solution, and each ML model requires individual analysis to identify suitable methods. Relevant criteria are the type of prediction task, the ML model, the fairness definition, computational constraints, the importance of model accuracy, and interpretability.

Based on those criteria, the following in-processing method was implemented and finally evaluated against the baseline model to answer RQ 2.1: "To what extent can fairness be improved by implementing in-processing mitigation techniques?"

Fair Gradient-Boosting Machine

Fair Gradient-Boosting Machine (FairGBM or FGBM) extends traditional gradient boosting by incorporating fairness constraints directly into the gradient boosting model training [CBJ+23].

FGBM solves a constrained optimization problem using the classic game-theoretic approach:

$$\min_{\theta \in \Theta} \mathcal{L}(\theta) \quad \text{subject to} \quad c_i(\theta) \leq 0, \quad i = 1, \ldots, m \tag{3.4}$$

Where $\mathcal{L}(\theta)$is the prediction loss and $c_i(\theta)$ represents fairness constraints.

This is reformulated as a minimax problem:

$$\theta^* = \arg\min_{\theta\in\Theta} \max_{\lambda\in\mathbb{R}^m_+} L(\theta, \lambda) \tag{3.5}$$

where:

- $L(\theta, \lambda) = \mathcal{L}(\theta) + \sum_{i=1}^{m} \lambda_i c_i(\theta)$, a Lagrangian function that combines the original objective function with fairness constraints.
- θ^* = Set of optimal parameters.
- $\arg\min_{\theta\in\Theta}$ = Finding the parameter values θ that minimize the following expression.
- $\max_{\lambda\in\mathbb{R}^m_+}$ = Maximum value with respect to λ.
- λ = Vector of non-negative Lagrange multipliers that control the trade-off between minimizing error and satisfying the fairness constraints.
- m = Number of constraints.

In other words, the constrained optimization aims to minimize the prediction error while satisfying fairness constraints across different demographic groups. The minimax formulation allows the algorithm to balance model performance with fairness requirements.

The FGBM algorithm uses an iterative approach with interleaved steps of gradient descent on parameters θ and gradient ascent on the Lagrange multipliers λ. This approach is known as dual ascent learning framework. By alternating between these steps, the algorithm converges to a solution that satisfies both the performance objective and the fairness constraints. This dual ascent approach allows FGBM to effectively navigate the trade-off between prediction accuracy and fairness requirements. Algorithm 3.1 describes the training process of the FairGBM model.

FGBM addresses the challenge of non-differentiable fairness metrics by using differentiable proxy functions. These proxies approximate the step-wise fairness constraints, making them compatible with gradient-based optimization. This 'proxy-Lagrangian' formulation allows the algorithm to simultaneously optimize for both model performance and fairness requirements through standard gradient methods. It effectively overcomes the primary obstacle in fair machine learning, where traditional fairness constraints lack the smoothness needed for gradient-based training. The following fairness constraints are available[5]:

- Equalize FNR, which is equivalent to equalizing TPR and therefore promotes equal opportunity.

[5]https://github.com/feedzai/fairgbm, visited on 05/13/25

Algorithm 3.1: FairGBM training pseudocode, adapted from [CBJ+23]

Input: $T \in \mathbb{N}$, number of boosting rounds
$\mathcal{L}, \tilde{\mathcal{L}} : \mathcal{F} \times \mathbb{R}_+^m \to \mathbb{R}$, Lagrangian and proxy-Lagrangian
$\eta_f, \eta_\lambda \in \mathbb{R}_+$, learning rates

Let $h_0 = \arg\min_{\gamma \in \mathbb{R}} \tilde{\mathcal{L}}(\gamma, 0)$; `// Initial constant "guess"`
Initialize $f \leftarrow h_0$;
Initialize $\lambda \leftarrow 0$;
for $t \in \{1, \ldots, T\}$ **do**
  Let $g_i = \frac{\partial \tilde{\mathcal{L}}(f,\lambda)}{\partial f(x_i)}$; `// Gradient of proxy-Lagrangian w.r.t. model`
  Let $\Delta = \frac{\partial \mathcal{L}(f,\lambda)}{\partial \lambda}$; `// Gradient of Lagrangian w.r.t. multipliers`
  Let $h_t = \arg\min_{h_t \in \mathcal{H}} \sum_{i=1}^{N} (-g_i - h_t(x_i))^2$; `// Fit base learner`
  Update $f \leftarrow f + \eta_f h_t$; `// Gradient descent`
  Update $\lambda \leftarrow (\lambda + \eta_\lambda \Delta)_+$; `// Projected gradient ascent`
end
return $h_0, \ldots, h_T$

- Equalize FPR, which is equivalent to equalizing TNR and therefore promotes predictive equality.
- Equalize FNR and FPR simultaneously, therefore promoting equalized odds.

A great advantage of FGBM is the possibility to implement global constraints alongside fairness constraints in the model. When dealing with class-imbalanced datasets where achieving high overall accuracy might be misleading, particular performance targets can be specified instead. For instance, in this case study the aim is to achieve equal opportunity (equal TPR rates) while keeping false positives under control. This allows building models that balance overall performance metrics with group-specific fairness considerations in a comprehensive approach.

Other advantages of FGBM are:

- Tree-based models ensure preservation of interpretability.
- Often achieves better fairness-performance tradeoffs than post-processing methods.
- No extra memory requirements compared to standard LGBM.
- Any core LGBMClassifier parameter[6] can be used with FairGBM as well[7].

[6] https://lightgbm.readthedocs.io/, visited on 05/14/25
[7] The only exception is the *objective* parameter, which has to be set to constrained cross entropy loss function. Any other standard non-constrained objective would result in using standard LGBM.

Though FGBM requires roughly double the training time of standard LGBM, it performs substantially faster than alternative fair ML approaches like Fairlearn[8] [CBJ+23].

FairGBM is particularly suitable for this case study because of its specialized design for GBDT models. This specialization makes it valuable for tabular data in high-stakes domains where both accuracy and fairness are critical, like in the insurance case study. The algorithm is supposed to achieve this dual objective remarkably well, maintaining strong predictive performance while simultaneously satisfying fairness constraints.

For the fairness assessment, FGBM was applied (1) using the same parameters as the baseline, only with additional fairness constraint parameters, and (2) using hyperparameter tuning (HPT) to find optimal parameters to maximize fairness.

FairGBM can only constrain one protected attribute. Therefore, the model was run separately for each protected attribute.

(1) FairGBM using same parameters as the baseline model
FairGBM praises itself for being especially suitable for imbalanced datasets [CBJ+23]. To test the statement, the baseline LGBM model was replaced with the FairGBM model. While all core LGBM parameters from Table 3.3 remain unchanged, FairGBM requires the definition of certain constraint parameters[9], as described in Table 3.4.

Parameter	Description	Value
constraint_type	Type of fairness constraint to use	FNR
global_target_fpr	Target rate for the global FPR constraint	0.01
global_constraint_type	Type of global equality constraint to use	FPR

Table 3.4: FairGBM Parameters with Descriptions

In other words, the fairness constraint is to optimize TPR ratio with at most a 1% FPR ratio. The constraint parameters were chosen in good faith to achieve the fairest model possible.

The method can handle any number of distinct values in the protected attribute groups. Therefore, both gender and nationality attributes are divided into the three subgroups that were previously described in Subsection 3.1, and the fairness constraint aims to ensure fairness between all three groups simultaneously.

(2) FairGBM using hyperparameter tuning
To investigate whether different parameter settings result in even better fairness and accuracy trade-off, comprehensive hyperparameter tuning (HPT) and selection were conducted.
The process was suggested by the authors of FairGBM[10] and adapted for this case study. The HPT ran for 4 hours, separately for gender and nationality. While the algorithm

[8] https://fairlearn.org/, visited on 05/14/25
[9] https://github.com/feedzai/fairgbm, visited on 05/14/25
[10] https://pypi.org/project/hyperparameter-tuning/, visited on 05/13/25

was tasked to maximize the overall recall score, it was simultaneously penalized if the TPRR (comparison of recall per subgroup) was below a generous threshold of 0.5.
The optuna hyperparameter optimization resulted in the optimal parameters listed in Table 3.5. Optuna[11] is an open-source hyperparameter optimization framework that uses efficient sampling algorithms to automatically search for optimal hyperparameter configurations [ASY+19]. Further descriptions of the optimized parameters can be found in Tables 3.3 and 3.4.

Parameter	Value for Gender	Value for Nationality
boosting_type	gbdt	gbdt
learning_rate	0.4764	0.1618
max_depth	47	58
min_child_samples	90	3513
n_estimators	2154	1374
num_leaves	18	17
reg_alpha	0.0348	0.0869
reg_lambda	0.0264	0.0219
verbosity	-1	-1
scale_pos_weight	425.35	271.21
enable_bundle	False	False
multiplier_learning_rate	0.3875	0.4790
constraint_type	FNR	FNR

Table 3.5: FairGBM HPT optimal Parameters

3.3.2 Post-processing Mitigation Methods

This section gives a comprehensive description of the post-processing mitigation methods applied in this thesis. Post-processing methods aim to adjust the outputs of trained models to remove discriminatory patterns (see Section 2.4).

The post-processing discrimination mitigation methods are implemented and compared to the baseline model to answer RQ 2.2: "To what extent can fairness be improved by implementing post-processing mitigation techniques?"

Threshold Optimizer

The threshold optimizer[12] is a component of the Fairlearn library[13] for mitigating fairness-related harms in ML systems. Threshold optimization is a post-processing technique for binary classification that addresses fairness concerns by applying different

[11] https://optuna.org/, visited on 05/30/25
[12] https://fairlearn.org/v0.12/user_guide/mitigation/postprocessing.html, visited on 05/30/25
[13] https://fairlearn.org/, visited on 05/14/25

decision thresholds to different demographic groups. The technique allows for balancing performance metrics and fairness constraints [WDE+23].

The threshold optimizer allows improving the equal opportunity metric [HPS16] after model training by post-processing the output of a scikit-learn[14] compatible estimator. In addition, it also allows optimizing various other fairness criteria, including demographic parity and equalized odds.

The modified classifier is obtained by applying group-specific thresholds to the provided estimator. For machine learning model M and protected attribute S, the optimizer identifies a set of thresholds $\{t_s\}$ where $s \in S$. These thresholds are selected to optimize a performance metric P subject to fairness constraints C. Performance objectives can be accuracy, balanced accuracy, recall, or true negative rate.

Algorithm 3.2 explains the individual steps of the classifier.

Algorithm 3.2: Threshold Optimizer

Data: Machine learning model M, protected attribute S, performance objective P, fairness constraints C
Result: Optimal thresholds T for each demographic group
Generate prediction scores $R \leftarrow M(\text{data})$;
Initialize threshold set $T \leftarrow \{\}$;
foreach *demographic group s defined by S* **do**
    Select subset of predictions R_s for group s;
    Find optimal threshold t_s that maximizes P subject to C;
    $T \leftarrow T \cup \{t_s\}$;
end
return T;

For the case study, the following implementation settings apply: Find optimal thresholds that maximize the recall (performance objective) of the LGBM baseline model while subject to equalized opportunities (fairness constraint).

Reject Option Classification

Reject option classification (ROC) is a post-processing technique for bias mitigation implemented in the AI Fairness 360 (AIF360) toolkit[15]. It promotes improving fairness while maintaining classification accuracy [KKZ12].

ROC operates on the premise that classifier predictions near the decision boundary are more likely to contribute to discrimination. By identifying instances in this critical region and applying different decision rules based on sensitive attributes, the algorithm reduces

[14] https://scikit-learn.org/stable/, visited on 05/14/25
[15] https://github.com/AIF360/algorithms/postprocessing/reject_option_classification.py, visited on 05/30/25

discrimination while preserving overall accuracy. In other words, it modifies predictions of a pre-trained classifier for instances near the decision boundary based on protected attributes.

Mathematically, the ROC works as follows: Given a classifier f with predictions $\hat{y} \in [0, 1]$, a protected attribute $s \in \{a, b\}$ (where b denotes the unprivileged group), and true labels $y \in \{0, 1\}$, ROC defines a critical region around the decision boundary:

$$\text{Critical region: } |\hat{y} - \theta| \leq \theta_c$$

where θ is the classification threshold, and θ_c is the critical region width parameter.

The modified prediction function $\hat{y}'$ is defined as:

$$\hat{y}' = \begin{cases} 1 & \text{if } \hat{y} > \theta + \theta_c \\ 1 & \text{if } |\hat{y} - \theta| \leq \theta_c \text{ and } s = b \\ 0 & \text{if } |\hat{y} - \theta| \leq \theta_c \text{ and } s = a \\ 0 & \text{if } \hat{y} < \theta - \theta_c \end{cases}$$

In other words, the formulation favors the unprivileged group in the critical region by assigning the favorable outcome. The ROC algorithm consists of the steps described in Algorithm 3.3.

Algorithm 3.3: Reject Option Classification

Data: Training data, test data with protected attributes s
Result: Modified predictions $\hat{y}'$

```
Train a classifier f on training data;
Compute prediction scores ŷ for each instance in the test set;
Determine optimal classification threshold θ and critical region width θ_c;
for each instance (x, s) in test data do
    Compute classifier score ŷ = f(x);
    if ŷ > θ + θ_c then
        Assign favorable outcome: ŷ' = 1;
    else if ŷ < θ − θ_c then
        Assign unfavorable outcome: ŷ' = 0;
    else
        if s = b then
            Assign favorable outcome: ŷ' = 1;
        else
            Assign unfavorable outcome: ŷ' = 0;
        end
    end
end
```

The main advantages are:

- It only modifies predictions for instances in the critical region and therefore minimizes accuracy loss compared to methods that modify all predictions.
- It explicitly favors the underprivileged group in this region.
- Allows control over the fairness-accuracy trade-off through the critical region width parameter.

The ROC approach has some limitations. First, it can only handle binary protected attributes. Second, its optimization process is computationally expensive, as it requires evaluating predictions for every instance in the dataset across multiple threshold configurations, leading to longer processing times, especially with large datasets. In addition, the AIF360 framework requires very specific dataset formatting.

Table 3.6 shows the parameter setting of the AIF360 ROC implementation in the case study.

Parameter	**Description**	**Value**
low_class_thresh	Smallest classification threshold to use	0.01
high_class_thresh	Highest classification threshold to use	0.99
num_class_thresh	Number of thresholds to consider during optimization	100
metric_name	Fairness metric to optimize	Equal opportunity difference
metric_ub	Upper bound for the fairness metric	0.05
metric_lb	Lower bound for the fairness metric	-0.05

Table 3.6: ROC Parameter setting

These settings configure a post-processing fairness algorithm that evaluates 100 different classification thresholds between 0.1 and 0.99 to optimize model predictions. The baseline default threshold is 0.5. The algorithm specifically targets equal opportunity as its fairness metric, ensuring that true positive rates across different protected groups remain within a ±5% difference. A score of 0 represents perfect fairness as the TPR difference is measured, not the TPR ratio.

Equalized odds post-processing

This bias mitigation technique solves a linear program to find a probabilistic transformation that maps the original classifier predictions to new predictions. The transformed predictions should satisfy the equalized odds constraints while maximizing accuracy

[HPS16, PRW+17]. Like ROC, it is also implemented in the AI Fairness 360 (AIF360) toolkit[16].

A classifier satisfies the equalized odds criterion if the predictions $\hat{y}$ are conditionally independent of the protected attribute S given the true label y. This requires that both true positive rates and false positive rates are equal across the protected groups (see Equation 2.16).

The equalized odds algorithm learns a probabilistic transformation represented by parameters $p_{s,y}$ and $q_{s,y}$ for each group s and true label y:

$$P(\tilde{Y} = 1|\hat{Y} = 1, A = a, Y = y) = p_{a,y} \tag{3.6}$$

$$P(\tilde{Y} = 1|\hat{Y} = 0, A = a, Y = y) = q_{a,y} \tag{3.7}$$

where $\tilde{Y}$ represents the transformed predictions.

The optimization problem is formulated as:

$$\min_{p_{s,y},q_{s,y}} \quad \sum_{s,y} P(S = s, Y = y) \cdot \mathbb{E}[L(\tilde{Y}, Y)|S = s, Y = y] \tag{3.8}$$

$$\text{s.t.} \quad P(\tilde{Y} = 1|S = s, Y = y) = P(\tilde{Y} = 1|S = s', Y = y) \quad \forall s, s', y \tag{3.9}$$

where L is a loss function measuring the prediction error.

The algorithm solves the above optimization problem by computing the group-specific TPR and FPR of the original classifier. A linear program aims find the transformation parameters that satisfy equalized odds. The learned transformation is then applied to create fair predictions. The resulting process is described in Algorithm 3.4.

The main advantage of the equalized odds post-processing method is that it works on any pre-trained classifier without requiring retraining. The method finds the optimal (most accurate) transformation that satisfies equalized odds. In addition, it can be adjusted to optimize for different cost functions.
In conclusion, unlike ROC or Threshold optimizer, equalized odds post-processing works by learning group-specific randomized transformations. It can simultaneously equalize both false positive and true positive rates and does not require access to the features or model internals, only the predictions. The method requires no specific parameter setting.

3.3.3 Performance Evaluation

Due to the highly imbalanced nature of the dataset (0.01% positive cases), standard accuracy can be misleading as it may appear artificially high by simply predicting the majority class for all instances. Instead, Table 3.7 presents comprehensive performance metrics agreed upon with the insurance company, including both standard accuracy for completeness and alternative metrics more appropriate for imbalanced dataset assessment.

[16] https://aif360.readthedocs.io/, visited on 05/14/25

Algorithm 3.4: Equalized odds post-processing

Data: Classifier f, training data with protected attributes (X, S, Y), cost function c
Result: Fair predictor $\tilde{f}$ satisfying equalized odds

Train classifier f on training data or use pre-trained model;
Compute prediction scores $\hat{Y} = f(X)$;
for *each group* $a \in A$ **do**
    $\text{TPR}_s \leftarrow P(\hat{Y} = 1 | S = s, Y = 1)$;
    $\text{FPR}_s \leftarrow P(\hat{Y} = 1 | S = s, Y = 0)$;
end
Define variables $p_{s,y}$ and $q_{s,y}$ for each group s and label y;
Construct linear program that minimizes expected cost while ensuring equal TPR and FPR across groups;
Solve linear program to obtain optimal transformation parameters $p^*_{s,y}$ and $q^*_{s,y}$;
foreach *instance* (x, s) **do**
    Compute original prediction $\hat{y} = f(x)$;
    if $\hat{y} = 1$ **then**
        $\tilde{y} \leftarrow \text{Bernoulli}(p^*_{s,y})$
    end
    else if $\hat{y} = 0$ **then**
        $\tilde{y} \leftarrow \text{Bernoulli}(q^*_{s,y})$
    end
end
return *fair predictor* $\tilde{f}$;

Balanced accuracy is the average of true negative rate (TNR) and true positive rate (TPR), which provides equal weight to performance on both classes regardless of their prevalence. Recall (TPR) is the proportion of actual positives correctly identified. Precision (PPV) is the proportion of positive predictions that are actually correct. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both considerations.

Metric	**Calculation**
Accuracy	$\frac{TN+TP}{TN+TP+FP+FN}$
Balanced Accuracy	$\frac{TNR+TPR}{2}$
Recall	TPR
Precision	PPV
F1 score	$\frac{2 \cdot PPV \cdot TPR}{PPV+TPR}$

Table 3.7: Baseline model performance evaluation metrics

CHAPTER 4

Results

This chapter describes the results of the fairness assessment on the explosive claims case study. First, the baseline model is analyzed, and relevant fairness metrics, as well as disadvantaged subgroups, are determined. Afterwards, the implementation of different discrimination mitigation methods is evaluated based on the fairness metrics. Finally, the impact of fairness strategies on the predictive performance is analyzed. The complementary Python scripts, including all relevant code, are available on GitHub[1] for reproducibility.

4.1 Baseline Model

The dataset contains two protected attributes: gender and nationality. Thus, the fairness assessment is conducted separately for each attribute and their respective subgroups. While the focus of the gender analysis is on the female vs. male comparison, the claims with 'unknown' gender designation will also be evaluated (see Section 3.1.1 for further description of the subgroups). To ensure no discrimination based on race or ethnicity, the nationality subgroups (Austrian, non-Austrian, and unknown) should experience the same level of fairness.

4.1.1 Gender

The confusion matrix values are analyzed separately for each subgroup in Table 4.1. Section 3.1.1 already describes the extreme class imbalance, which might be even more severe when the claims are split into training and test set.

The results show that all groups have at least some true explosive claims in the test set. But since the group sizes and base rates are imbalanced, fairness metrics are essential to evaluate parity.

[1] https://github.com/AnnabelRe/FairnessAnalysis, visited on 05/15/25

	Absolute frequencies				Relative frequencies (%)			
	TN	FP	FN	TP	TN	FP	FN	TP
Male	26,269	167	10	13	99.13	0.63	0.04	0.05
Female	15,766	90	6	4	99.39	0.57	0.04	0.03
Unknown	15,639	96	9	8	99.31	0.61	0.06	0.05

Table 4.1: Baseline model performance on gender test set

To decide which metrics are problematic in terms of discrimination, it is necessary to compare them group-wise. Figure 4.1 shows the statistical measure ratios for the two subgroups, male vs. female. The blue graph describes the original ratios, where values close to 1 indicate parity. The green values describe the normalized ratios, where values close to 0.5 indicate parity. The dotted graphs describe the respective parity.

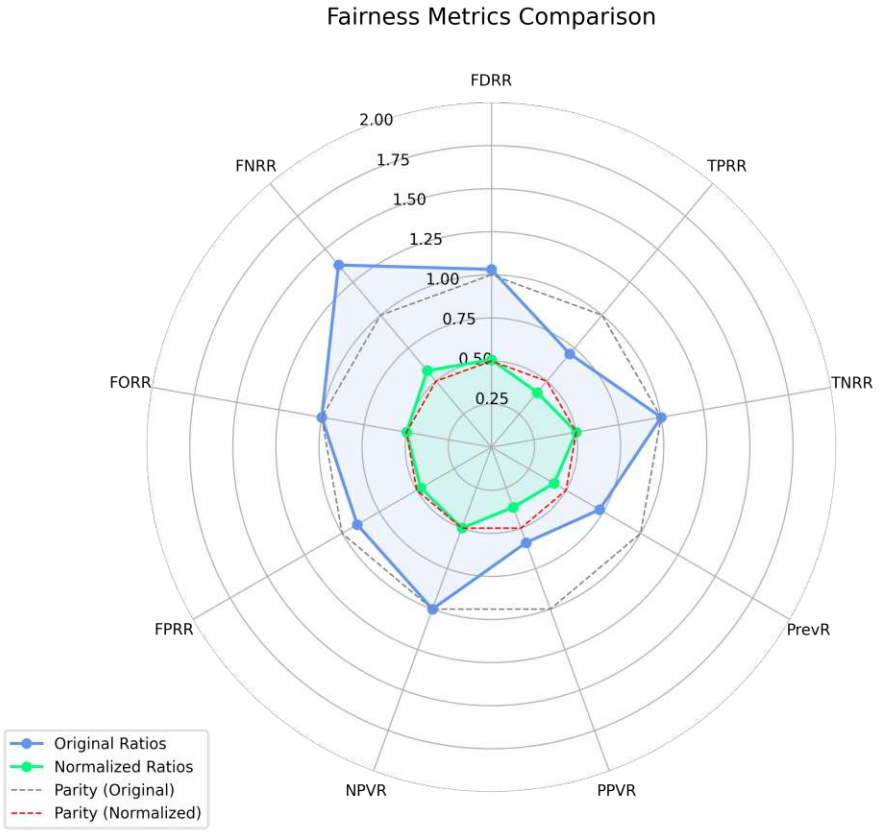


Figure 4.1: Fairness ratios male vs. female, baseline model

The key findings are:

- PPVR, which compares the precision per group, reveals potential discrimination against the female subgroup. That means that the proportion of positive predictions

that are actually correct is smaller for females.

- FNRR, which compares the proportion of false negatives per group, is much higher for the female subgroup, which means that actual positives are much more likely to be misclassified as negatives.
- TPRR, which describes the correctly classified positives, is lower for the female subgroup. As FNR and TPR are complementary, this means that the model fails to predict actual positives for the female subgroup as well as for the male subgroup.
- The male subgroup has a higher prevalence (positive base rate) than the female subgroup in the test set, which indicates an imbalance in the underlying distribution of positives.

Figure 4.2 analyzes the impact on the chosen fairness metrics.

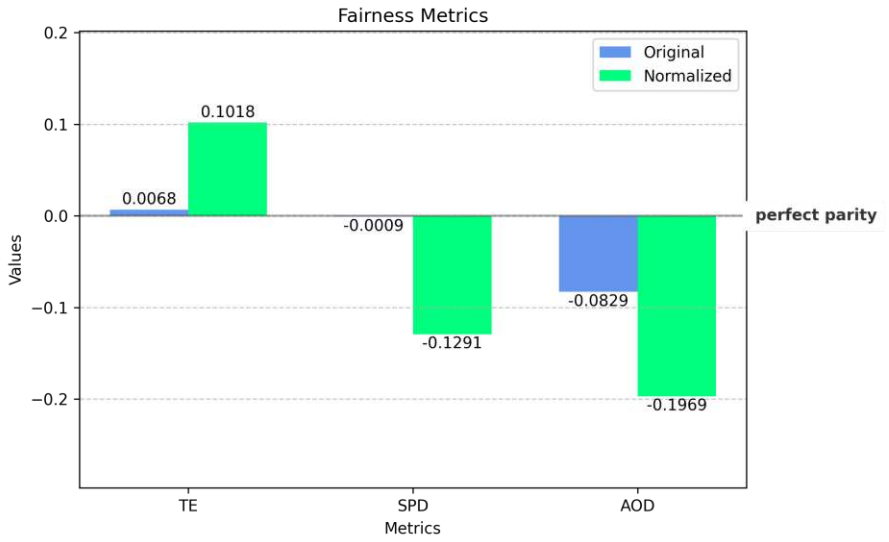


Figure 4.2: Comparison of treatment equality, statistical parity difference, and average odds difference for male vs. female, baseline model

The key findings are:

- The normalization of values significantly emphasizes the disparity for all three metrics.
- The positive treatment equality value shows that the female group receives more false negatives than the male group.
- The original statistical parity difference is insignificant. Only when normalizing the values does a moderate disparity show. It indicates that the female group has a

13% lower probability of receiving a favorable outcome compared to the male group. This suggests that overall decision outcomes are skewed in favor of the male group.

- The average odds difference of -0.0829 indicates that females have a lower chance of receiving a true or false positive prediction compared to males.

The equal selection parity is -86.0, which means that the male subgroup has a total of 86 more positively predicted claims. This conforms with the previous findings, like disparity in base rates.

The baseline model does not meet the strict fairness definitions that demand parity across all groups, as shown in Table 4.2.

Metric	**Original Ratio**	**Normalized Ratio**
Equal Opportunity	×	×
Predictive Equality	×	×
Equalized Odds	×	×
Conditional Use Accuracy Equality	×	×
Predictive Parity	×	×

Table 4.2: Evaluation of fairness metrics, male vs. female, baseline model

The comparison between the male and female groups remains the focus of this thesis. The unknown subgroup should not experience a disadvantage compared to the other groups. Figures 4.3 and 4.4 show both comparisons.

The key findings are:

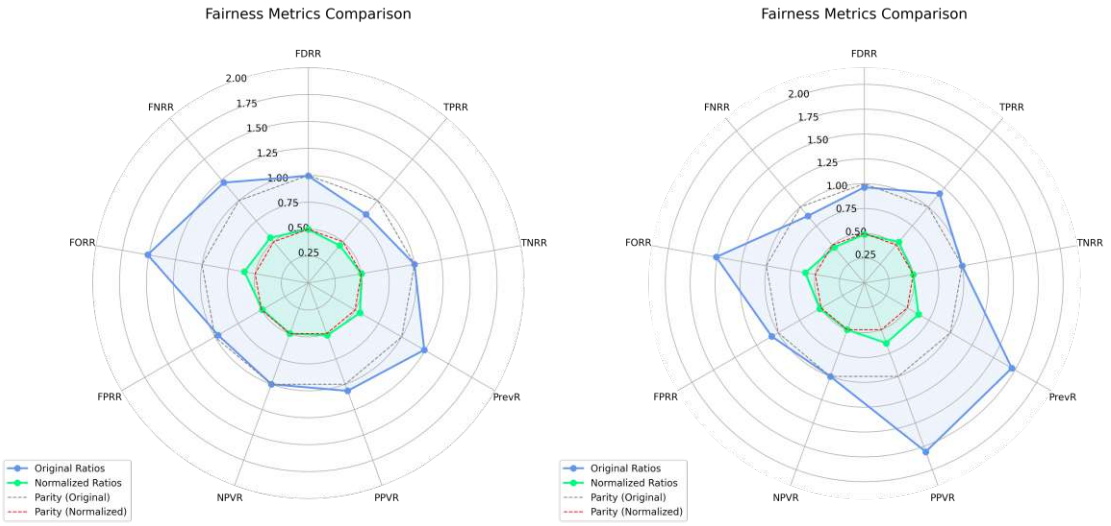
- Both male and female groups have a lower prevalence score than the unknown subgroup. This indicates a substantial imbalance in the underlying distribution between the groups in the dataset. Especially the female subgroup experiences a disadvantage, with the unknown subgroup having 71% higher representation in the positive class.
- The model correctly classifies male positives better than unknown positives, which however have an 18% higher true positive rate than the female group. The normalized value indicates this is actually a relatively minor deviation from parity, which might not warrant as much concern as the raw percentage suggests. The disparity is reinforced by the FNRR finding, as claims from the female group are more likely to be incorrectly classified as negative when they are actually positive.
- The FORR shows a strong, significant disparity among all groups. The unknown group has a 51% higher false omission rate than both female and male groups. This means that when predicted as negative, claims from the unknown group are much more likely to actually be positive cases compared to the other groups.

- The PPVR shows the most significant disparity among female vs. unknown metrics. The unknown group has an 81% higher PPV (precision) compared to the female group. That means that positive predictions of the unknown group are much more likely to actually be positive cases.

In conclusion, the baseline model appears to be under-identifying positive cases for the unknown group compared to the male group, despite the unknown group having a higher prevalence of positive cases in the dataset. The model seems to systematically miss positive cases for both the female and unknown groups, which could result in denied benefits for the groups. Again, none of the fairness definitions from Table 4.2 are satisfied.

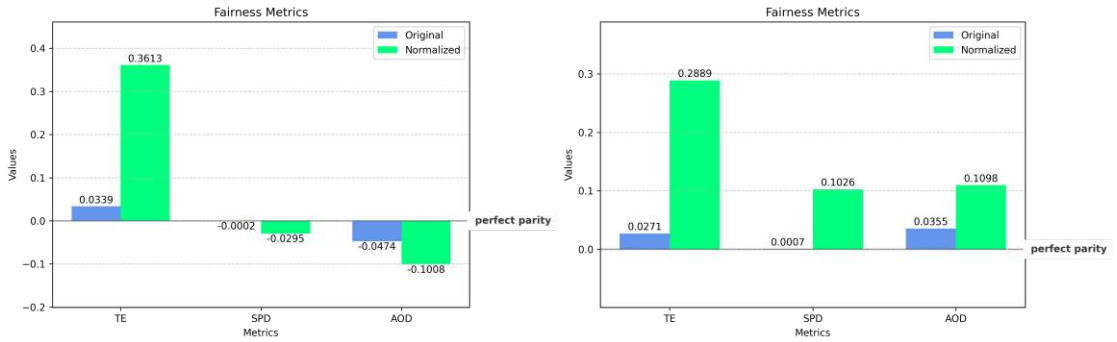
The equal selection parity (ESP) of male vs. unknown is -76.0, indicating that the male group receives significantly more positive predictions than the unknown group in absolute numbers. This probably reflects the underlying population size differences between the groups rather than necessarily indicating algorithmic bias. The unknown group receives 10 more absolute positive predictions than the female group, which is significant considering they have very similar population sizes and generally only very few positives.

Figure 4.4 shows minor differences for both comparisons of the original treatment equality (TE). However, the normalized TE shows a significant imbalance in the ratio of false negatives to false positives between the groups. The unknown group experiences a notably different error distribution pattern. The large normalized TE values suggest significantly different error patterns between groups, which could lead to disparate impacts. This aligns with the findings of the ratio comparisons in Figure 4.3. TE normalization reveals



(a) Male vs. unknown (b) Female vs. unknown

Figure 4.3: Fairness ratios unknown vs. male and female, baseline model



(a) Male vs. unknown (b) Female vs. unknown

Figure 4.4: Comparison of treatment equality, statistical parity difference, and average odds difference for unknown vs. male and female, baseline model

a disparity over 10 times larger than the original value indicates. This dramatic difference highlights how crucial normalization is for understanding the true magnitude of error distribution imbalances.

The SPD for male vs. unknown is insignificant. However, the normalized value for female vs. unknown reveals a substantial 10 percentage point difference in positive prediction rates favoring the unknown group. This is a critical insight that would be completely missed without normalization.

As Figure 4.3 hypothesizes, the male group has an advantage over the unknown group in terms of combined true positive and false positive rates, which is reflected in the AOD value. On the other hand, the chart reveals that the unknown group experiences better predictive performance than the female group.

4.1.2 Nationality

There are three subgroups for nationality: Austrian, non-Austrian, and unknown. Section 3.1 showed equally imbalanced label distribution for all subgroups. However, the fairness analysis only takes into consideration the test set.

	Absolute frequencies				**Relative frequencies (%)**			
	TN	FP	FN	TP	TN	FP	FN	TP
Austrian	40493	250	16	17	99.31	0.61	0.04	0.04
Non-Austrian	1392	7	0	0	99.50	0.50	0.00	0.00
Unknown	15789	96	9	8	99.31	0.60	0.06	0.05

Table 4.3: Baseline model performance on nationality test set

Table 4.3 reveals a significant challenge. The non-Austrian group has no actual positives in the test set and, consequently, also no false negatives. Since the training-test-split is done chronologically, there is no possibility to modify this. As a consequence, a

number of basic statistical measures and fairness definitions that rely on those values cannot be calculated and evaluated. A positive base rate of 0 alone indicates a strong disadvantage for the non-Austrian group and that it suffers from significant representation disadvantages compared to the other groups.

Since meaningful group fairness metrics like equalized odds cannot be calculated for the Austrian vs. non-Austrian comparison, the comparison between the Austrian and unknown groups remains the focus of the nationality fairness assessment. Figures 4.5 and 4.6 show the result of the fairness analysis.

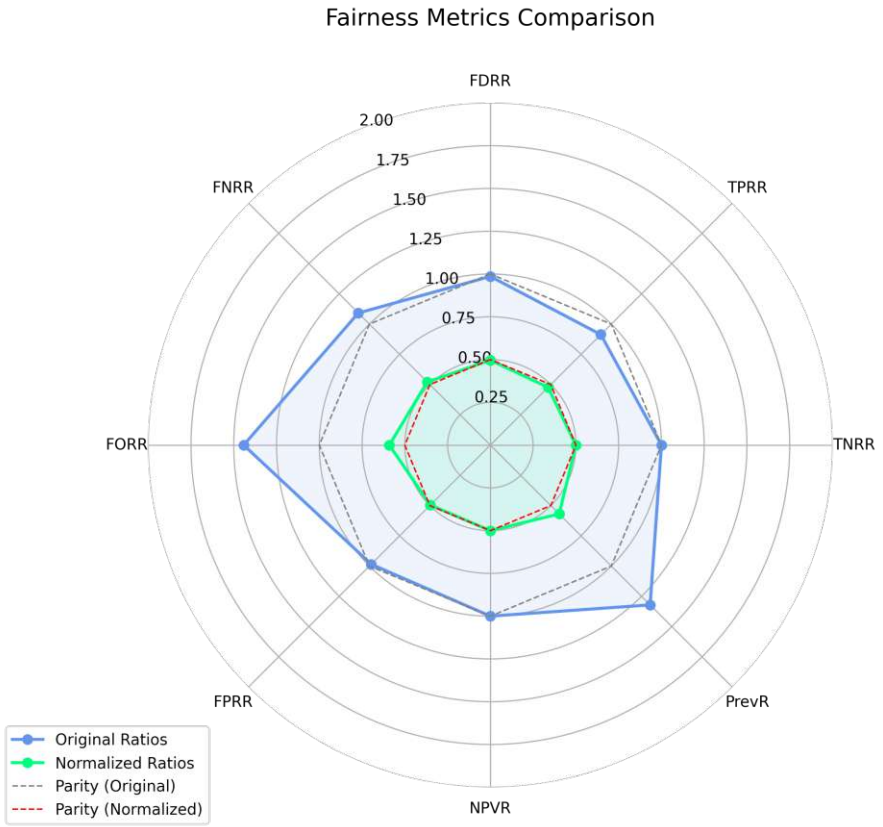


Figure 4.5: Fairness ratios Austrian vs. unknown, baseline model

The key findings are:

- The unknown group has a significant 32% higher representation in the positive class. This leads to a 21% higher positive predictive value.
- The unknown group has a 44% higher false omission rate than the Austrian group. This means when predicted as negative, claims from the unknown group are much more likely to actually be positives.

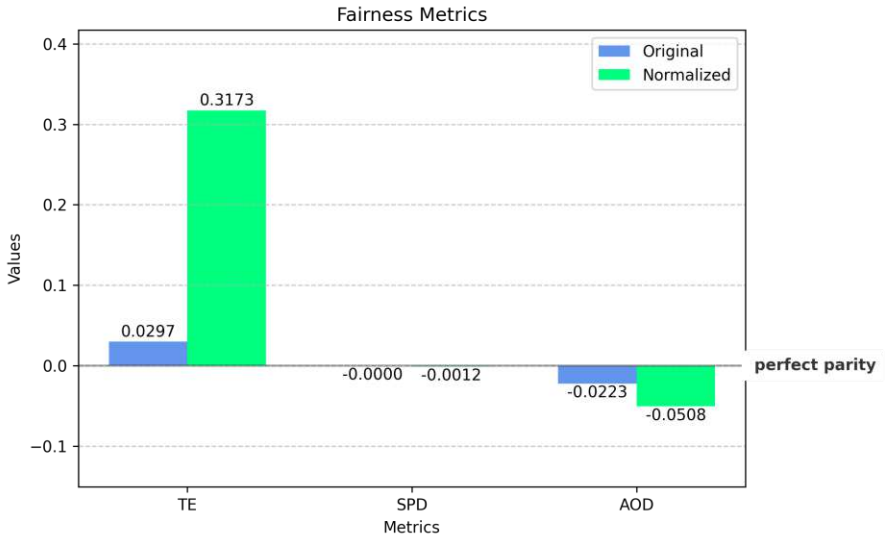


Figure 4.6: Comparison of treatment equality, statistical parity difference, and average odds difference for Austrian vs. unknown, baseline model

- Despite the unknown group having a higher prevalence in the positive class, the model is less likely to correctly identify positives of the unknown group, leading to a higher false negative rate.
- While there's a large absolute difference in positive predictions (ESP = -163), this doesn't translate to a proportional difference (SPD ≈ 0), suggesting the difference might be due to group size variations.
- The large normalized TE value (0.32) is especially concerning, indicating substantially different error patterns between groups that could lead to disparate impacts (e.g., disparate FORR).

Again, none of the strict fairness definitions in Table 4.4 are met.

Metric	Original Ratio	Normalized Ratio
Equal Opportunity	×	×
Predictive Equality	×	×
Equalized Odds	×	×
Conditional Use Accuracy Equality	×	×
Predictive Parity	×	×

Table 4.4: Fairness metrics Austrian vs. unknown, Baseline model

4.1.3 Summary

In conclusion, the fairness assessment of the baseline model has revealed substantial disparities among the subgroups of protected attributes, gender and nationality.

More precisely, the female subgroup suffers great disadvantages compared to the male and unknown groups. The model misclassifies actual positives as negatives for the female group at a larger rate than for the other groups, which results in a lower TPRR (and higher FNRR). The nationality subgroups also exhibit significant fairness challenges. The non-Austrian group experiences a disadvantage in representation and therefore cannot even be reliably compared to the other nationality subgroups. This situation highlights a fundamental limitation of group fairness metrics when dealing with extreme imbalances or zero base rates in one group.

Even though the unknown subgroup shows an advantage in terms of positive base rate and positive predicted value in both gender and nationality analysis, the model seems to make better predictions for the other groups. That results in a significant disparity in treatment equality (error rates) and especially the FORR.

The baseline model evaluation demonstrates that fairness metrics based on absolute differences, such as equal selection parity and balance, prove inadequate for this particular case study. Due to the substantial disparity in group sizes across the dataset, these metrics produce potentially misleading conclusions about algorithmic fairness.

In conclusion, none of the strict fairness definitions are satisfied for any group comparison. The following chapter analyzes the impact of various mitigation methods on the fairness scores.

4.2 Discrimination Mitigation Methods

After evaluating the baseline model, establishing appropriate fairness metrics, and identifying disadvantaged subgroups, various discrimination mitigation methods are implemented and evaluated. This section aims to find significant results to answer research question 2: "To what extent can fairness be improved through discrimination mitigation techniques in comparison to baseline?"
First, in-processing methods are implemented by replacing the baseline LGBM model with appropriate fair models. Afterwards, post-processing methods are evaluated.

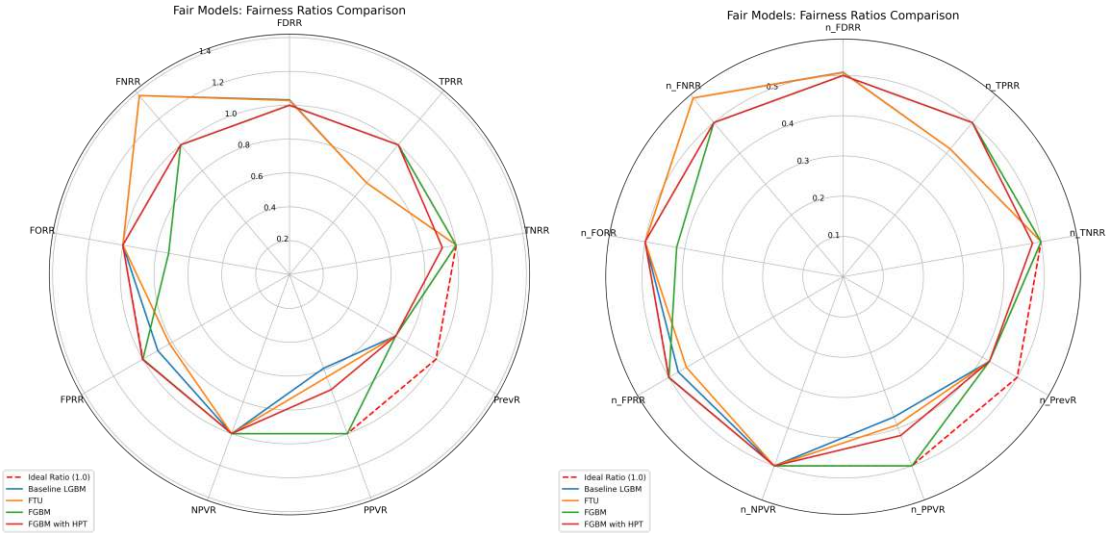
4.2.1 In-Processing Mitigation Methods

The broad concept of in-processing mitigation methods are explained in Section 2.4. The detailed technical implications of the applied methods are described in Section 3.3. The following section describes the key findings from the comparison of the baseline model and the applied in-processing mitigation methods. The aim is to improve fairness considerations by directly integrating fairness criteria into the optimization process and providing in-training feedback.

Besides the selected in-processing method, the comparison also includes one pre-processing mitigation method: Fairness through unawareness (FTU), see Section 3.3. The comparison is essential for a comprehensive understanding of the different strategies, since all of the mitigation methods are based on the FTU outcome.

Gender

This analysis aims to explore potential interventions that could reduce or eliminate the documented algorithmic bias against the female demographic group.
Figure 4.7 provides an overview of the adjusted fairness ratios per mitigation method. It shows that especially the FairGBM model with hyperparameter tuning achieves great fairness scores close to the perfect parity value of 1.0 for unnormalized values, and 0.5 for normalized values. Generally, the normalized values highlight the improved performance of all fair models compared to the baseline model. The FTU method shows only very little variance from the baseline model, thus their graphs are overlapping.



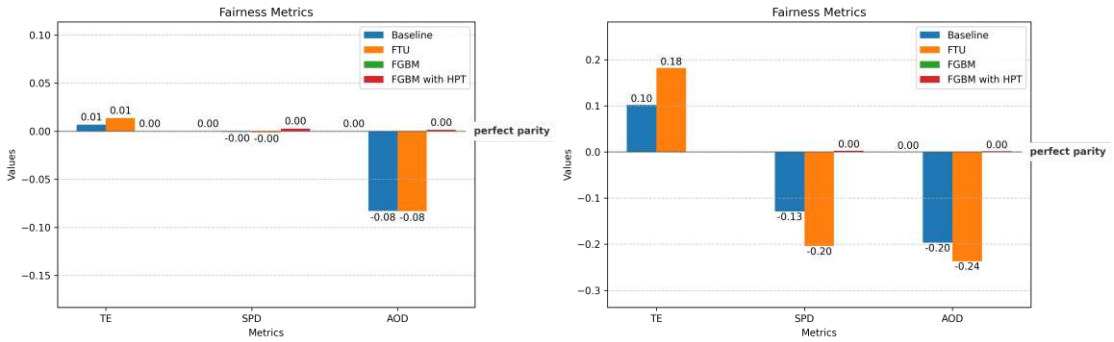
(a) Original fairness ratios (b) Normalized fairness ratios

Figure 4.7: Fairness ratios male vs. female, in-processing mitigation methods

Figure 4.8 confirms that both FGBM versions perform significantly better than the FTU mitigation, which barely shows any improvement from the baseline model.

Table 4.5 gives the impression that the FairGBM model with baseline parameters is the best model in terms of fairness.

The result is the same for the normalized fairness values. The only exception is the FGBM model without HPT, for which normalized metrics cannot be calculated due to divisions by zero. This result highlights the importance of a thorough analysis, including the evaluation of normalized values. What seems to be a successful mitigation comes with hidden side effects, which will be further analyzed in Sections 4.3 and 5.1. It



(a) Original fairness differences

(b) Normalized fairness differences

Figure 4.8: Comparison of treatment equality, statistical parity difference, and average odds difference for male vs. female, in-processing mitigation methods

Metric	FTU	FairGBM	FairGBM with HPT
Equal Opportunity	×	✓	✓
Predictive Equality	×	✓	×
Equalized Odds	×	✓	×
Conditional Use Accuracy Equality	×	✓	×
Predictive Parity	×	✓	×

Table 4.5: Fairness metrics male vs. female, in-processing mitigation methods

is crucial to acknowledge that fairness metrics, particularly ratio-based measures, can achieve seemingly perfect scores while concealing fundamental performance deficiencies.

In conclusion, when considering fairness metrics alone (such as ratios and differences), both versions of FGBM demonstrate significantly superior performance compared to the baseline model. Thus, the in-processing mitigation strategies yield great success.

While the main focus remains on comparing the male subgroup to the female subgroup, both groups are individually compared to the unknown subgroup as well. In summary, the main findings are:

- Both group comparisons (male vs. unknown and female vs. unknown) result in the same conclusion: The FGBM model without HPT yields the best fairness results, whereas the FTU model shows no significant difference from the baseline model.
- The FGBM model with HPT shows by far the worst fairness results. The model predicts disproportionately more positives for the male and female groups. Consequently, the unknown group has a much higher TNR and a lower FPR and FDR. Both original and normalized values reveal that the model fails to predict any true positives for the unknown group. See Figures 1 and 4 in Annex A for more details.

In conclusion, it is important to recognize that a single mitigation strategy may not treat all demographic groups equitably, often producing varying results across different group comparisons. This heterogeneity highlights the necessity of a comprehensive evaluation including a variety of fairness metrics and comparison approaches.

Nationality

As established in the baseline analysis, the non-Austrian subgroup has no true positives in the test set. Consequently, only a fraction of fairness metrics can be calculated. Nevertheless, the FGBM model without HPT achieves successful results again, as displayed in Table 4.6. More fairness metrics and comparison ratios are detailed in Annex A, Figure 6.

Metric	FTU	FairGBM	FairGBM with HPT
Equal Opportunity	×	✓	×
Predictive Equality	×	✓	×
Equalized Odds	×	✓	×
Conditional Use Accuracy Equality	×	✓	×
Predictive Parity	×	✓	×

Table 4.6: Fairness metrics Austrian vs. non-Austrian, in-processing mitigation methods

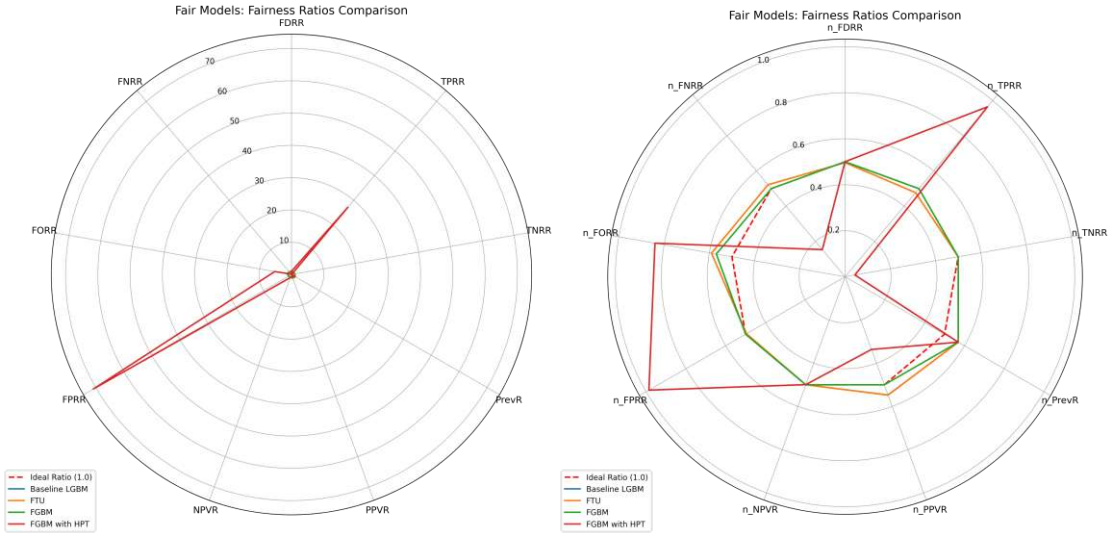
The comparison of the Austrian vs. unknown group described in Figures 4.9 and 4.10 results in the following key findings:

- The FTU model shows no significant difference from the baseline model.
- The FGBM model with HPT predicts disproportionately more positives for the unknown group than for the Austrian group. Consequently, the unknown group has a much higher TPR but also a higher FPR.
- The FGBM model without HPT shows slightly better fairness results than FTU and the baseline model, except for the declined false omission rate ratio.

These values result in the same satisfaction of the fairness definitions as shown in Table 4.6. Again, the normalized metrics for FGBM without HPT cannot be calculated due to divisions by zero.

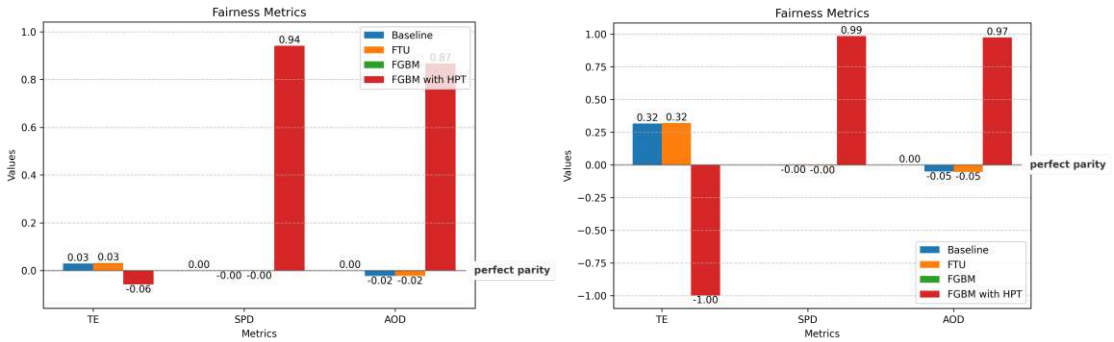
The results for the non-Austrian vs. unknown group comparison are inconclusive due to the unfavorable positive base rate.

In conclusion, the fairness analysis of the nationality subgroups proves to be very challenging, but the FGBM model without HPT successfully delivers fair results (except for FORR).



(a) Original fairness ratios (b) Normalized fairness ratios

Figure 4.9: Fairness ratios Austrian vs. unknown, in-processing mitigation methods



(a) Original fairness differences (b) Normalized fairness differences

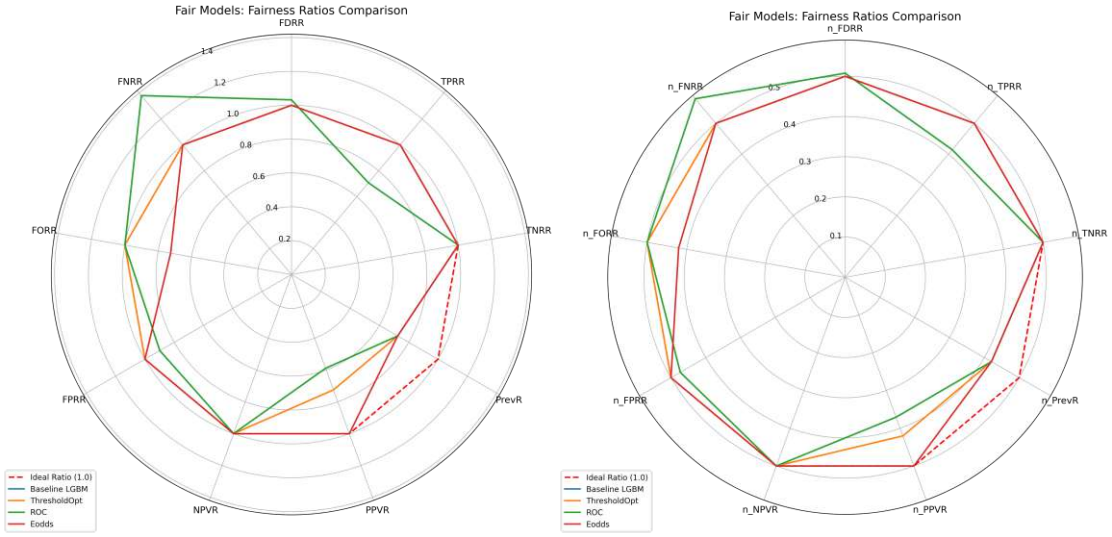
Figure 4.10: Comparison of treatment equality, statistical parity difference, and average odds difference for Austrian vs. unknown, in-processing mitigation methods

4.2.2 Post-Processing Mitigation Methods

This section summarizes the findings of the implemented post-processing mitigation methods. Unlike the strategies evaluated in the section before, post-processing methods address fairness concerns only after the model training. The outputs of trained models are modified so that fairness is improved. The following strategies are compared: threshold optimizer, reject option classifier (ROC), and equalized odds post-processing (EOdds).

Gender

Both Figures 4.11 and 4.12 reveal that the ROC did not change the baseline model prediction at all. The only metric that differs when comparing the two outcomes is balance, simply because the baseline LGBM model outputs prediction scores contrary to the binary labels of the ROC output.
Furthermore, both EOdds and the threshold optimizer show fairer results compared to baseline. Especially the statistical parity difference of 0 shows no favoring of one group regarding predicting the positive label. The average odds difference is also 0, which means that both groups receive the same ratio of TPR and FPR. However, EOdds results in a declined FORR compared to the baseline.



(a) Original fairness ratios (b) Normalized fairness ratios

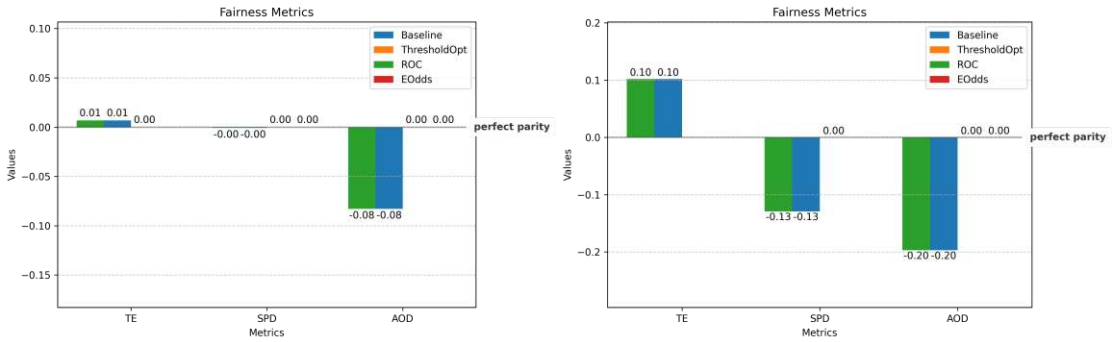
Figure 4.11: Fairness ratios male vs. female, post-processing mitigation methods

These findings result in the fairness metrics depicted in Table 4.7.

Metric	Threshold Optimizer	Reject Option Classification	Equalized Odds
Equal Opportunity	✓	×	✓
Predictive Equality	✓	×	✓
Equalized Odds	✓	×	✓
Conditional Use Accuracy Equality	×	×	✓
Predictive Parity	×	×	✓

Table 4.7: Fairness metrics, male vs. female, post-processing mitigation methods

The calculation of fairness metrics using normalized values presents a technical challenge for EOdds post-processing metrics, as they cannot be computed due to zero denominators



(a) Original fairness differences (b) Normalized fairness differences

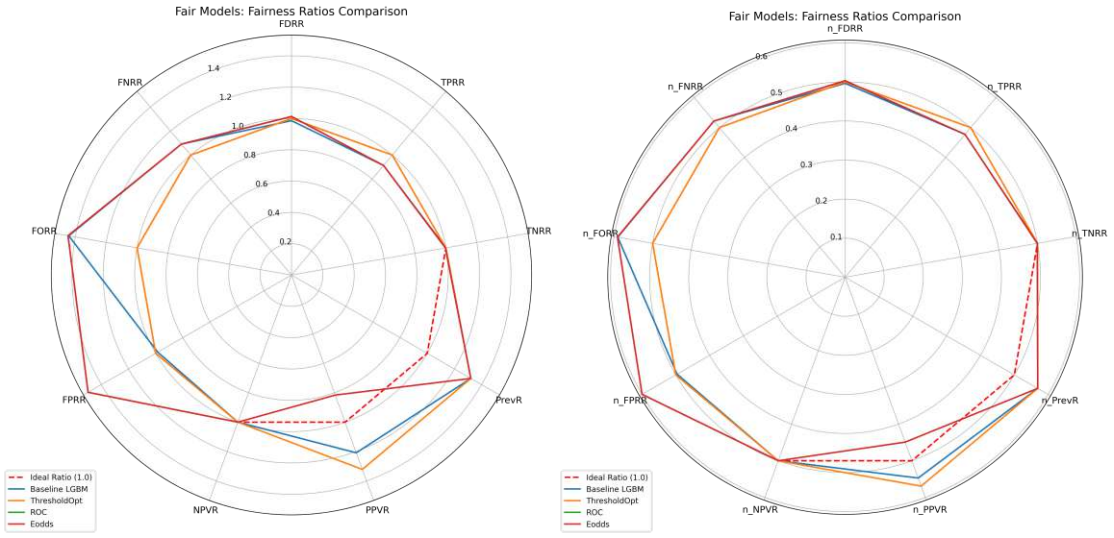
Figure 4.12: Comparison of treatment equality, statistical parity difference, and average odds difference for male vs. female, post-processing mitigation methods

in their formulas. This computational limitation is significant and suggests that despite seemingly positive outcomes, there are underlying complexities worth investigating during the model prediction evaluation phase in Section 4.3.

The evaluation of the unknown group is similar. The ROC reflects the same results as the baseline model, whereas the threshold optimizer and EOdds show an improvement in fairness.

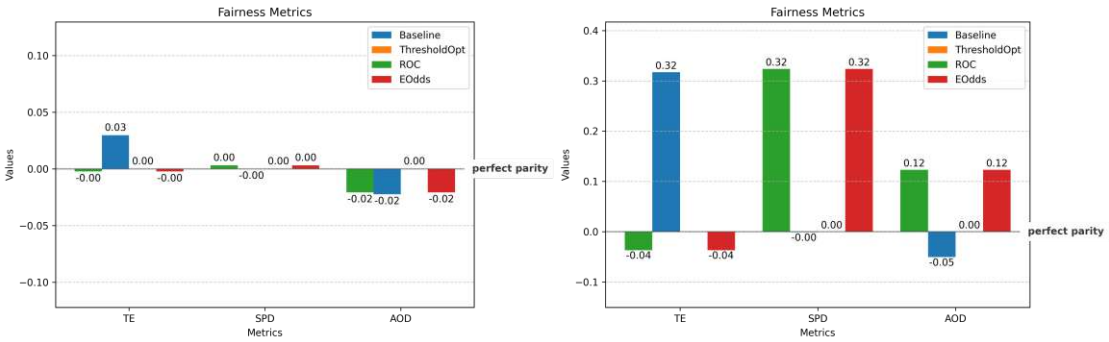
Nationality

Figures 4.13 and 4.14 summarize the findings of the post-processing mitigation strategies.



(a) Original fairness ratios (b) Normalized fairness ratios

Figure 4.13: Fairness ratios Austrian vs. unknown, post-processing mitigation methods



(a) Original fairness differences (b) Normalized fairness differences

Figure 4.14: Comparison of treatment equality, statistical parity difference, and average odds difference for Austrian vs. unknown, post-processing mitigation methods

The key findings are:

- ROC produces similar results to the baseline model, except for the FPRR, which indicates a much higher amount of false positives for the Austrian group.
- Threshold optimizer achieves TPRR parity and perfect fairness difference metrics.
- All mitigation methods improve the treatment equality compared to the baseline model, which means that the error rate is more equal.
- ROC and EOdds show a declined value for SPD and AOD, which is likely due to a proportionally large amount of incorrectly predicted positives (FP) for the Austrian subgroup.

These findings result in the fairness metrics depicted in Table 4.8.

Metric	Threshold Optimizer	Reject Option Classification	Equalized Odds
Equal Opportunity	✓	×	×
Predictive Equality	✓	×	×
Equalized Odds	✓	×	×
Conditional Use Accuracy Equality	×	×	×
Predictive Parity	×	×	×

Table 4.8: Fairness metrics Austrian vs. unknown, post-processing mitigation methods

For the Austrian vs. non-Austrian and unknown vs. non-Austrian comparison, both ROC and EOdds failed to complete the task. This is due to the missing true positives, which are a strict requirement for the algorithms. The threshold optimizer at least improves

the baseline model in terms of predictive equality, which means that both groups have the same FPR (equal type 1 error rate).

4.3 Performance Evaluation

This section compares the predictive output of the baseline LGBM model to the modified outputs of the mitigation strategies. The aim is to identify the effect of the improvement in fairness on the predictive performance. All performance metrics described in Section 3.3.3 range from 0 to 1, with 1 describing the best possible performance.

The baseline model employs a classification threshold of 0.5, where prediction scores above this value yield a positive label (1), while scores at or below this threshold result in an negative label (0). The baseline model achieves an accuracy score of 99.35%, which is not necessarily informative in datasets with high class imbalance, though. Thus, the other performance metrics, like balanced accuracy, recall, precision, and F1 score, provide better insights. Especially the recall score should not drop lower than the baseline model result of 50%, since already half of the very few positives are misclassified as negative. Only 6.61% of predicted positives of the baseline model are actually positive, indicating many false positives and a very low precision score. Consequently, the F1 score also shows poor overall performance (11.68%). However, in this case study, false negatives impose higher costs on the insurance company than false positives. Section 3.1 provides further details on the consequences of FNs and FPs.
In conclusion, despite the high overall accuracy, the model's performance on the positive class is poor, as evidenced by the modest balanced accuracy, low precision, and particularly weak F1 score. This serves as an indication that the model struggles with the extreme class imbalance.

Figure 4.15 shows the predictive performance of the mitigation strategies applied to optimize fairness between the male and female gender groups. Both FTU and ROC perform identically or at least very similarly to the baseline model. The FGBM and EOdds models both show a recall of 0, which indicates no correctly predicted positives. In fact, both models predict the negative class for all claims, which results in almost perfect accuracy but very poor recall, precision, and F1 score.
The threshold optimizer applies a different solution to the fairness optimization by predicting all claims as explosive. Thus, it is the only strategy that achieves perfect recall for correctly predicting all true positives. The trade-off is a low (balanced) accuracy, precision, and F1 score.
The FGBM with HPT also achieves a desirable higher recall score than the baseline model, but with the same negative trade-offs as the threshold optimizer.

ROC and EOdds show different results for the comparison of the unknown group because those strategies require the sensitive attribute to have binary values. Thus, the post-processing methods were optimized individually per pair of subgroups. ROC shows the same results as the baseline model again, whereas EOdds performs better by at least predicting some positives, but still even worse than the baseline.

Figure 4.15: Performance metrics by model type for gender

Figure 4.16 shows the predictive performance of the mitigation strategies applied to optimize fairness within the nationality groups. Since ROC and EOdds fail to modify the algorithm for the non-Austrian group, Figure 4.16 shows the results of the comparison of the Austrian vs. unknown group for those two mitigation strategies.

The findings are similar to the gender evaluation. FTU, ROC, and EOdds result in the same performance scores as the baseline model described above. FGBM without HPT fails to predict any positives and therefore has a recall score of 0. The threshold optimizer predicts the positive class for all claims, thus correctly predicting 100% of positives. Due to the high class imbalance, this still results in an accuracy score of 0 and also very poor scores for the other performance metrics. FairGBM with HPT fails to exceed the baseline recall score and generally performs worse than the baseline model.

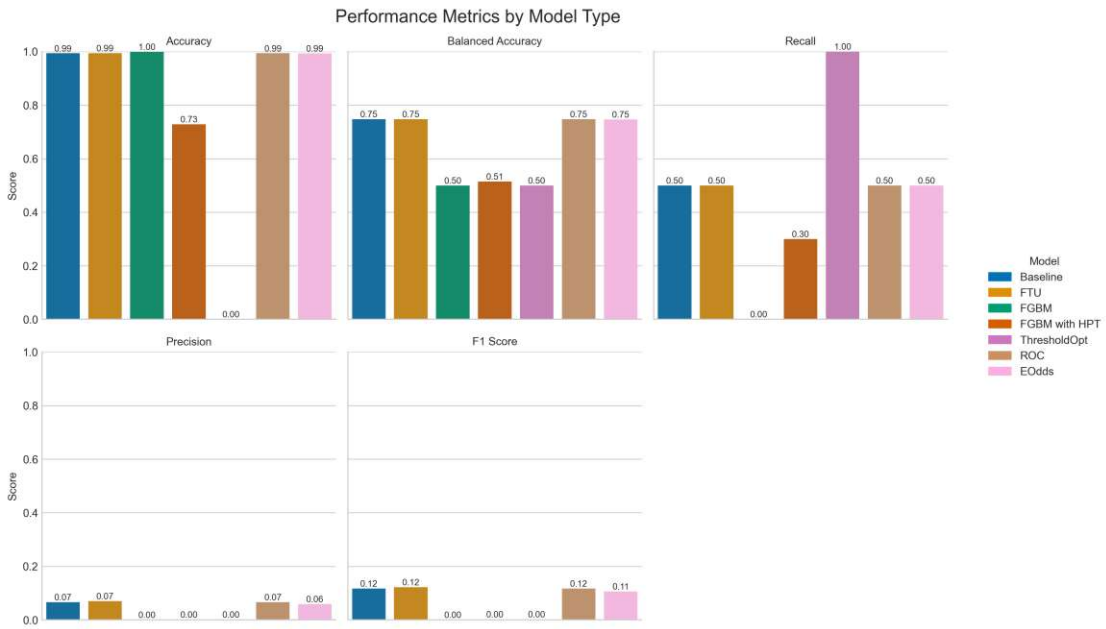


Figure 4.16: Performance metrics by model type for nationality

CHAPTER 5

Summary

The last chapter summarizes and critically evaluates the findings from Chapter 4 to address the research questions established in Chapter 1. In addition, the constraints inherent in the case study methodology, research design, and implemented mitigation strategies are examined. The chapter concludes by underscoring the fundamental significance of algorithmic fairness within AI systems in contemporary society.

5.1 Conclusion

In this thesis, the subject of algorithmic fairness and discrimination was discussed, and a real-world machine learning model from the insurance industry was tested for discrimination. To do so, suitable fairness metrics had to be identified, and various discrimination mitigation strategies were evaluated.
The research questions from Section 1.1 were answered in the course of this fairness analysis.

RQ1: To what extent does the current baseline model discriminate in the context of this case study?

A thorough literature research was conducted to establish what discrimination means in a legal and technical context. In addition, existing insurance domain-specific research was analyzed for relevance and related work. It was found that the case study is complex, both in terms of dataset and legal framework. The legal analysis revealed that both gender and nationality are protected attributes that must not be grounds for discrimination. In addition, the EU AI Act requires transparent documentation of the design and implementation of the AI model.

RQ 1.1: Which fairness metrics are most suitable to capture discrimination in the case study?

This research question requires a qualitative evaluation of the respective fairness metrics and their suitability for the use case. The comprehensive literature research resulted in dozens of potential fairness metrics, each with a different focus. On the one hand, there are group fairness metrics, which are further divided into separation, independence, and sufficiency. On the other hand, individual fairness metrics try to put the focus on the individual fairness rather than the group fairness. Subgroup fairness metrics aim to combine both approaches to overcome the respective limitations.
For this case study, measuring individual fairness and subgroup fairness was found to be too computationally expensive and complex. The dataset includes over 400,000 insurance claims, each having over 6,000 columns.
Out of the 11 group metrics introduced in Section 2.3, the following nine were used for the baseline model fairness evaluation:

- Independence: Equal Selection Parity, Statistical Parity.
- Separation: Treatment Equality, Balance, Equalized Odds, Equal Opportunity, Predictive Equality.
- Sufficiency: Conditional Use Accuracy Equality, Predictive Parity.

Due to the high class imbalance, the normalized value was calculated for all relative metrics in addition and proved very valuable. The main goal of the thesis was to successfully predict the few positives equally for all subgroups, thus satisfying equalized opportunity while maintaining a relatively low FN and FP rate.

The baseline evaluation demonstrated that metrics based on absolute differences, such as balance and equal selection parity, prove inadequate when analyzing datasets characterized by significant class imbalance. Furthermore, binary fairness criteria that demand exact parity (like equalized odds or conditional use accuracy equality) yield only binary satisfaction states, limiting their interpretability and actionability. As a result, this research found that relative measures expressed as ratios or proportional differences provide the most interpretable and practically applicable framework for assessing algorithmic fairness in this context, as they enable direct comparison of disparities across different demographic groups regardless of baseline rate variations.

RQ 1.2: What are the subgroups that are discriminated against in the case study?

The fairness analysis of the baseline model reveals that the female group has a lower probability of receiving a favorable outcome and a higher probability of incorrectly classifying positives as negatives compared to the male and unknown groups. In addition, the female subgroup shows the lowest prevalence score and SPD, which means that it has the least amount of actual and predicted positives.

The distribution of claims in the nationality subgroup is highly imbalanced. The non-Austrian subgroup is barely represented and has no actual explosive claims in the test

set. Thus, it is impossible for the non-Austrian subgroup to achieve fairness levels equal to the other groups.
Even though the unknown subgroup has a higher prevalence score and PPV than the Austrian group, the model fails to correctly predict positives as well as for the Austrian group.

RQ2: To what extent can fairness be improved through discrimination mitigation techniques in comparison to baseline?

This research question aims to identify to what extent the implementation of in-processing and post-processing mitigation methods affects the fairness metrics. The chosen mitigation methods should eliminate the disadvantages that were identified in the context of the previous RQ.

RQ 2.1: To what extent can fairness be improved by implementing in-processing mitigation techniques?

The FGBM model with baseline parameters successfully modifies the predictions during model training so that the most valuable fairness metrics, like equalized opportunity, are satisfied for all 6 group pairings of this fairness analysis. When hyperparameter tuning is applied to the FGBM, equalized opportunity is only satisfied for the male vs. female comparison. In addition, fairness metrics that measure proportional differences generally showed a much closer value to 0 for both FGBM versions when optimized for the gender attribute.

RQ 2.2: To what extent can fairness be improved by implementing post-processing mitigation techniques?

Three different post-processing mitigation methods were implemented to modify the baseline model output to improve the group fairness metrics. The reject option classification generally produces no significantly different predictions from the baseline model. EOdds outperforms the baseline model for all gender comparisons, except for a drop in the FORR performance. The threshold optimizer managed to satisfy some metrics like predictive equality for all group comparisons. ROC and EOdds failed to optimize the equal opportunity metric for non-Austrian comparisons due to the missing actual explosive claims in the non-Austrian test subset.

Tables 5.1 and 5.2 present the outcomes of various bias mitigation strategies, evaluated using binary fairness criteria that yield only two possible states: satisfied or violated. The original LGBM baseline model exhibited fairness violations across all binary metrics assessed.

RQ3: To what extent does the predictive performance decrease by introducing the above-mentioned mitigation techniques, compared to baseline?

The baseline model has a recall score of 50%, which should not decrease any further by implementing fairness measures, according to the insurance company. Section 4.3 describes the predictive performance after implementing the respective fairness mitigation

Metric	FTU	FGBM	FGBM with HPT	Threshold Optimizer	ROC	EOdds
Equal Opportunity	×	✓	✓	✓	×	✓
Predictive Equality	×	✓	×	✓	×	✓
Equalized Odds	×	✓	×	✓	×	✓
CUAE	×	✓	×	×	×	✓
Predictive Parity	×	✓	×	×	×	✓

Table 5.1: Fairness metrics male vs. female for all mitigation methods

Metric	FTU	FGBM	FGBM with HPT	Threshold Optimizer	ROC	EOdds
Equal Opportunity	×	✓	×	✓	×	×
Predictive Equality	×	✓	×	✓	×	×
Equalized Odds	×	✓	×	✓	×	×
CUAE	×	✓	×	×	×	×
Predictive Parity	×	✓	×	×	×	×

Table 5.2: Fairness metrics Austrian vs. unknown for all mitigation methods

methods. Both gender and nationality comparisons yield the same conclusion: None of the mitigation methods lead to an improvement in the predictive performance. At most, they do not worsen. FTU and ROC show no difference in performance from the baseline model, as well as EOdds for nationality. The threshold optimizer and FGBM with HPT exceed the baseline model in terms of recall score, but even a higher recall score does not outweigh the significantly worse accuracy score. FGBM without HPT outperforms the baseline model in terms of accuracy, but the recall score of 0% eliminates the method from further pursuit.

Overall goal: Introducing fairness to a real-world model from the insurance industry

In conclusion, the overall goal was successfully accomplished. Mitigation methods were found that significantly improve the fairness and mitigate discrimination. However, the negative impact on the predictive performance was too serious to select a strategy for the future. Nevertheless, the FTU mitigation method should definitely be implemented, even if there was no change in fairness or performance metrics. Using protected attributes during model training should be avoided.

To create a comprehensive summary, the advantages and disadvantages of each mitigation method are briefly described below.

FTU

The model did not show significant changes in both fairness metrics and performance metrics. I recommend applying FTU in the future to avoid using protected attributes for

model training.

FGBM without HPT
The FGBM model with baseline parameters predicts the negative class for all claims. Although it therefore satisfies some fairness metrics, the negative impact on the predictive performance metrics is too significant to further pursue this strategy.

FGBM with HPT
The FGBM model with HPT succeeds in improving some fairness metrics compared to the baseline model. However, the model incorrectly predicts too many positives and therefore yields a significantly worse accuracy score than the baseline model. This model demonstrates the greatest versatility among all mitigation strategies, as it enables the optimization of various fairness metrics alongside global constraints while maintaining full optimization of all LGBM parameters. Therefore, further hyperparameter tuning should be conducted until a satisfying trade-off between fairness and recall is found.

Threshold Optimizer
The Fairlearn threshold optimizer perfectly complies with its initial requirement: Find optimal thresholds that maximize recall (performance objective), while subject to equalized opportunities (fairness constraint). Unfortunately, it does not offer the option to set an additional constraint on a maximal false positive rate and consequently simply predicts the positive class for all claims. This mitigation method has the great advantage that it can be applied to any model outcome. Therefore, I recommend extending the available code from the Fairlearn library with additional fairness constraints to ensure generalization for any ML model.

Reject Option Classifier
The ROC mitigation method does not significantly change the fairness or performance metrics compared to the baseline model. The algorithm has the great disadvantage that it has to iterate through every single prediction and decide whether or not the prediction label should be flipped. That process is time-consuming compared to the other mitigation strategies, which often need less than a second to produce output. I therefore do not recommend using this strategy for any future fairness analysis.

Equalized Odds post-processor
This mitigation method does not offer tailored parameter settings. Therefore, trying to optimize equalized odds in this case study is a self-fulfilling prophecy, and the algorithm simply predicts the negative class for all claims, which results in equal true and false positive rates of 0. I do not recommend this mitigation method for the future, due to its simplified approach and the limited choice of fairness metric to be optimized.

5.2 Limitations and Future Work

Implementing fairness in AI systems presents significant real-world challenges, particularly in balancing stakeholder objectives, profitability, ethical considerations, and legal requirements. Maintaining human oversight remains crucial, as fairness determinations

are inherently subjective and metric-based assessments cannot fully capture contextual fairness.

The dataset characteristics represent significant challenges throughout this thesis. The subdivision of demographic groups requires careful methodological consideration. The nationality-based analysis encountered limitations due to the small sample size of non-Austrian claimants and absence of explosive claims in the test set. Continuous monitoring and analysis of new claims data would provide more robust insights over time. In addition, the high class imbalance makes accurate interpretation of fairness and performance metrics difficult.

This thesis initiates fairness assessment post-preprocessing, as access to raw claim reports was unavailable. Future work could extend this analysis to earlier stages, employing fair NLP word embedding methods [MMS+21] to eliminate potential discrimination patterns within the original documentation. Selecting appropriate mitigation strategies proved challenging, as no universal solution exists. The process involves considerable trial and error, compounded by the inherent trade-offs between different fairness metrics—optimizing for one metric often compromises another. Future research could focus more intensively on comparative evaluation of mitigation techniques and methods for balancing these trade-offs. A promising direction for future research could involve the combination of in-processing and post-processing mitigation strategies. While this thesis focused on comparing each approach independently against the baseline model, combining a fairness-aware training methodology with sophisticated post-processing techniques could potentially enhance fairness outcomes while maintaining an acceptable trade-off with predictive performance. Nevertheless, it is important to acknowledge that such a combined approach would necessitate substantial human effort in implementation and fine-tuning. Given the additional computational and operational overhead, the practical adoption of such hybrid approaches in real-world scenarios might be limited by efficiency constraints and resource considerations.

This research did not define acceptable tolerance ranges for fairness metrics like equalized odds. As a result, many of the metrics demanded perfect parity, which is rarely achievable in real-world scenarios. For future work, I recommend establishing reasonable thresholds around the ideal parity score of 1, allowing fairness criteria to be considered satisfied even when results show minor deviations from perfect equality. While resource-intensive approaches like individual fairness assessment may be impractical in business contexts, group fairness metrics should be evaluated with appropriate caution. The insurance industry's potential to impact individuals' financial and health outcomes makes non-discrimination particularly critical, not only ethically but also financially, given potential regulatory penalties.

Generally, the scope of this thesis was rather extensive. Future studies could benefit from a more focused approach, perhaps examining a single protected attribute or concentrating only on baseline model fairness evaluation, which would permit a more thorough analysis. Similarly, a dedicated investigation into discrimination mitigation techniques alone would enable more precise parameter optimization and potentially yield superior results.

A key critique emerges from the current legislative situation in Austria. Laws prohibiting discrimination require greater specificity in the AI context. Current regulations such as GDPR and the AI Act remain vague regarding concrete fairness requirements. Future regulations should provide clear thresholds, metrics, or guidelines for measuring and ensuring fairness in algorithmic systems. The development of fair AI systems requires bridging the conceptual and linguistic divide between legal expertise and computer science. While lawyers provide the ethical framework necessary for AI development, technical understanding remains crucial for crafting effective regulations.

Only through interdisciplinary collaboration can we bridge the significant divide between legal, technical, and organizational approaches to algorithmic fairness. At minimum, companies deploying AI systems should be able to clearly identify which fairness definitions and mitigation strategies will ensure compliance with legal regulations. The EU's existing requirements remain excessively context-dependent, subjective, and susceptible to varying judicial interpretations, which makes them resistant to automated verification approaches [WMR21].

5.3 The Critical Importance of Fairness in AI Systems

We currently face a critical juncture: will AI be leveraged to reinforce existing inequalities, or will it serve as a tool for creating a more equitable society? This choice carries profound implications. The path forward requires a proactive rather than reactive approach to algorithmic fairness, embedding equity considerations from the earliest stages of development rather than attempting retrofits. Understanding the historical context of discrimination and its social manifestations is essential, as we cannot address future challenges without comprehending our past.

Europe currently has the unique opportunity to lead in ethical AI development, especially as the United States faces political resistance to diversity and inclusion initiatives. Recent developments have shown concerning trends, such as the elimination of terms related to equity and justice from federal communications and educational materials[1]. These include words such as "equal opportunity", "gender", "race", "discrimination", "bias", "underrepresented", "systemic", "social justice", "inequality", "minorities", and "disparity".
The urgency of addressing these challenges cannot be overstated. The democratization of AI, which means ensuring its benefits reach all of society, is not merely a technical priority but a moral imperative that will shape our collective future.

Data scientists must extend their focus beyond eliminating statistical discrimination to acknowledge and address how machine learning systems can perpetuate and amplify structural discrimination. While statistical bias can be measured and mitigated through technical means (as shown in this thesis), structural discrimination operates through deeper societal patterns that may be inadvertently encoded into training data, model

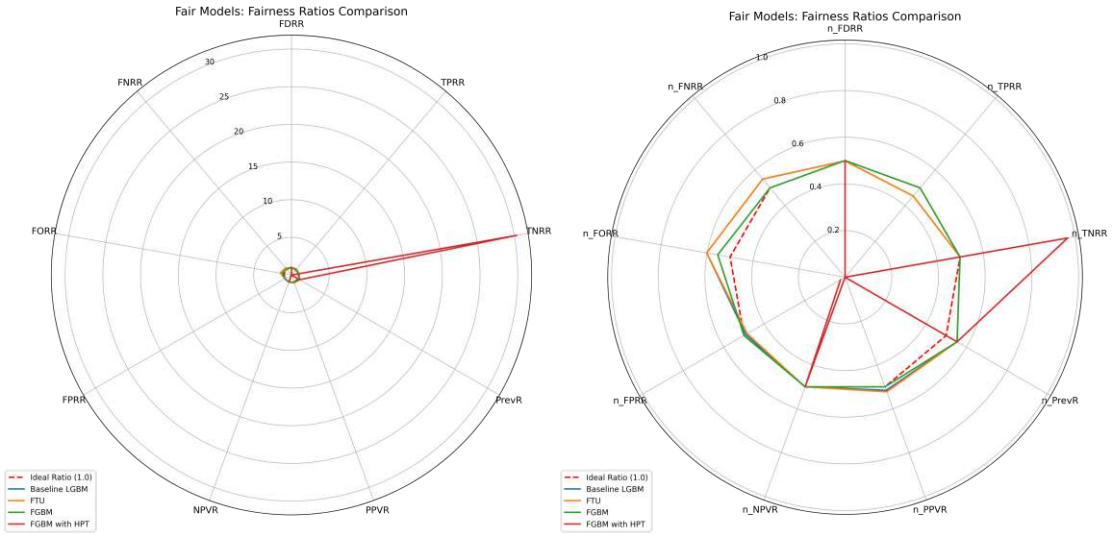
[1] https://www.nytimes.com/, visited on 05/16/25

architectures, and deployment practices. The responsibility extends beyond traditional fairness metrics to encompass a broader understanding of how ML systems interact with and potentially reshape social structures. This shift represents not just a technical evolution but a fundamental reimagining of the data scientist's ethical responsibilities in society.

This call for a more holistic approach to AI ethics is not isolated but part of a growing movement toward responsible technology development. The Digital Humanism Initiative[2] at TU Wien exemplifies this shift, advocating for technology that serves human values and dignity. The establishment of dedicated academic venues further demonstrates this momentum: the ACM Conference on Fairness, Accountability, and Transparency (FAccT)[3], founded only five years ago, has rapidly become a premier forum for examining fairness, accountability, and transparency in socio-technical systems. Its European counterpart, the European Workshop on Algorithmic Fairness (EWAF)[4], reflects the global nature of these concerns.
These efforts collectively signal a paradigm shift toward viewing AI development as inherently intertwined with questions of justice, equity, and human flourishing.

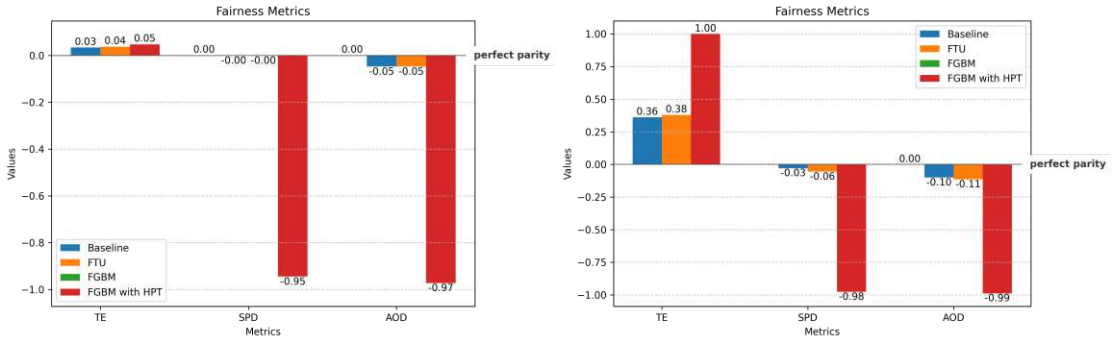
[2]https://caiml.org/dighum/, visited on 05/30/25
[3]https://facctconference.org/, visited on 05/30/25
[4]https://2025.ewaf.org/, visited on 05/30/25

Annex A



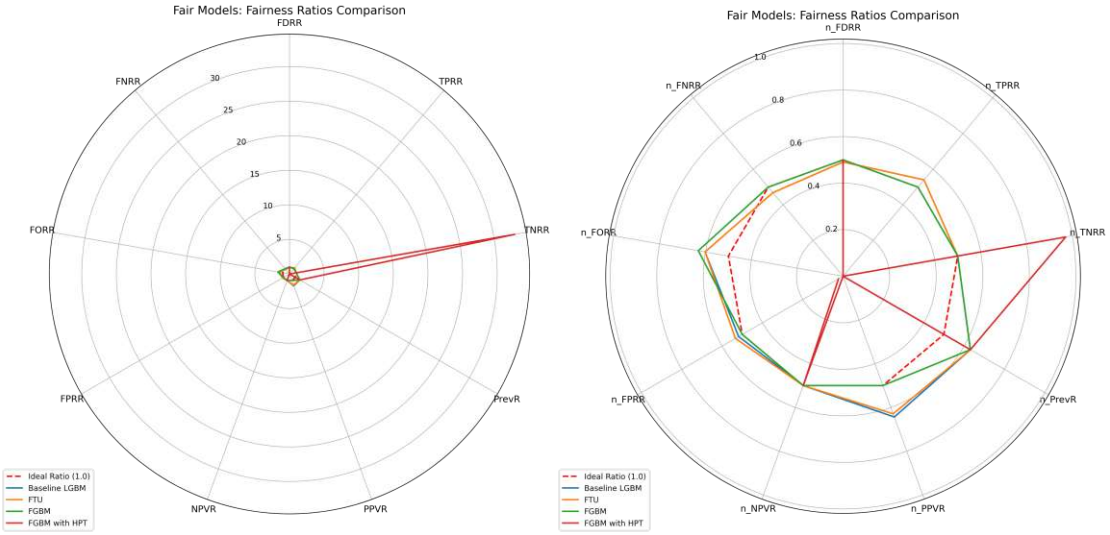
(a) Original fairness ratios (b) Normalized fairness ratios

Figure 1: Fairness ratios male vs. unknown, in-processing mitigation methods



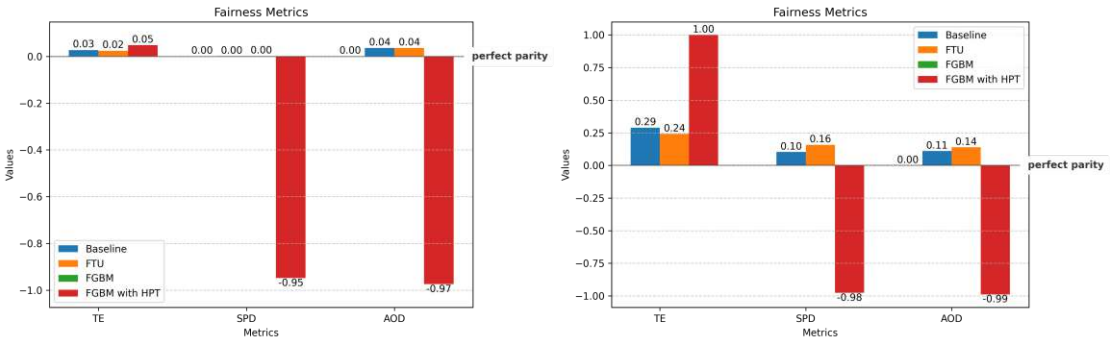
(a) Original fairness differences (b) Normalized fairness differences

Figure 2: Comparison of treatment equality, statistical parity difference, and average odds difference for male vs. unknown, in-processing mitigation methods



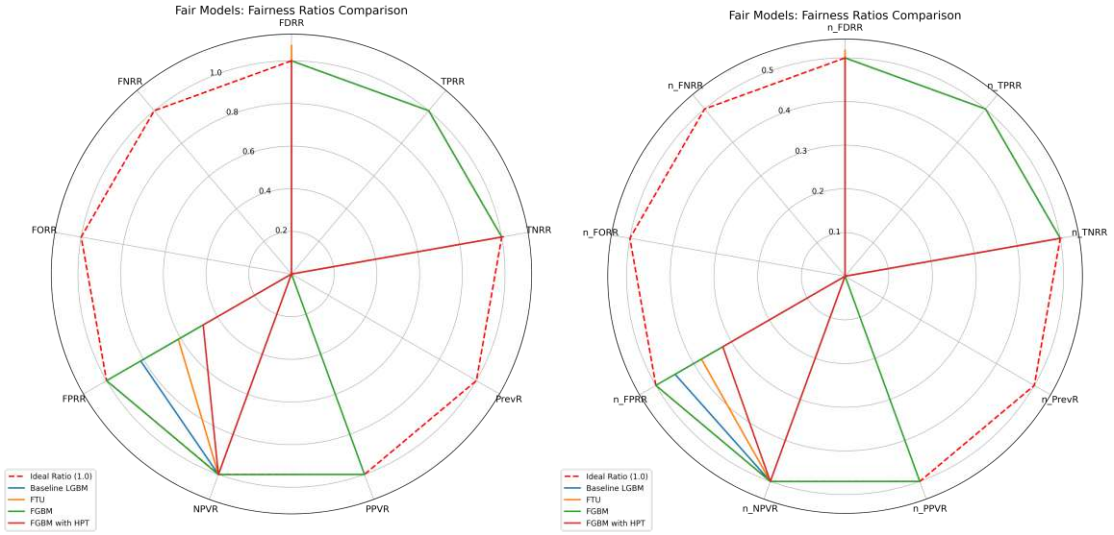
(a) Original fairness ratios (b) Normalized fairness ratios

Figure 3: Fairness ratios female vs. unknown, in-processing mitigation methods



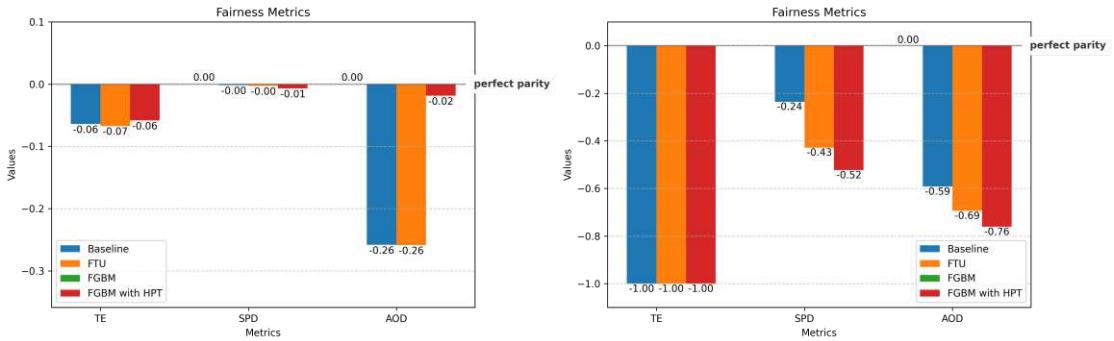
(a) Original fairness differences (b) Normalized fairness differences

Figure 4: Comparison of treatment equality, statistical parity difference, and average odds difference for female vs. unknown, in-processing mitigation methods



(a) Original fairness ratios (b) Normalized fairness ratios

Figure 5: Fairness ratios Austrian vs. non-Austrian, in-processing mitigation methods



(a) Original fairness differences (b) Normalized fairness differences

Figure 6: Comparison of treatment equality, statistical parity difference, and average odds difference for Austrian vs. non-Austrian, in-processing mitigation methods

Overview of Generative AI Tools Used

In the context of research, writing, and analysis of this thesis, the following generative AI tools were employed to enhance productivity and ensure quality:

- **Claude Sonnet 4**[5] by Anthropic: Used for the creation of LaTeX code, rephrasing statements to ensure grammatical correctness and eliminate spelling mistakes, as well as generating Python code for data visualizations and analysis.
- **ChatGPT GPT-4.1 mini**[6] by OpenAI: Utilized for rephrasing statements to ensure grammatical correctness and eliminate spelling mistakes.
- **NotebookLM**[7] by Google: Used in its browser version to generate podcast-style audio summaries of research papers for enhanced comprehension and review.

Despite the use of these generative AI tools, they were employed solely in a supportive capacity for inspiration and assistance, with all output thoroughly reviewed and adjusted. All ideas presented in this thesis are either original contributions from me or properly attributed to their respective sources through accurate citations.

[5] https://claude.ai, visited on 05/30/25
[6] https://chat.openai.com, visited on 05/30/25
[7] https://notebooklm.google.com, visited on 05/30/25

List of Figures

List of Tables

List of Algorithms

Bibliography

[aia] Regulation (eu) 2024/1689 of the European Parliament and of the Council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Published in the Official Journal on 12 July 2024.

[ASY+19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery.

[BHN23] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[CBJ+23] André F. Cruz, Catarina Belém, Sérgio Jesus, João Bravo, Pedro Saleiro, and Pedro Bizarro. FairGBM: Gradient boosting with fairness constraints. In *The Eleventh International Conference on Learning Representations*, 2023.

[CCG+22] 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. *Scientific Reports*, 12(1):4209, March 2022.

[CDPF+17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, August 2017.

[CfJC+20] European Commission, Directorate-General for Justice, Consumers, European network of legal experts in gender equality, non discrimination, and E. Dewhurst. *Age discrimination law outside the employment field – 2020*. Publications Office, 2020.

[CH24] Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7):1–38, July 2024.

[CHR25] Arthur Charpentier, François Hu, and Philipp Ratz. Mitigating discrimination in insurance with wasserstein barycenters. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 161–177. Springer Nature Switzerland, 2025.

[DECG23] Directorate-General for Justice and Consumers (European Commission), European network of legal experts in gender equality and non-discrimination, Isabelle Chopin, and Catharina Germaine. *A comparative analysis of non-discrimination law in Europe 2022: the 27 EU Member States, Albania, Iceland, Liechtenstein, Montenegro, North Macedonia, Norway, Serbia, Turkey and the United Kingdom compared*. Publications Office of the European Union, 2023.

[DG19] Directorate-General for Communications Networks, Content and Technology (European Commission) and Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji. *Ethics guidelines for trustworthy AI*. Publications Office of the European Union, 2019.

[DHP+12] 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*, pages 214–226. Association for Computing Machinery, January 2012.

[DL19] Jannik Dunkelau and Michael Leuschel. Fairness-aware machine learning: An extensive overview. Technical report, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, 2019. Version 01, October 17, 2019.

[DWY+20] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[EUr00] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, June 2000.

[FDCE19] Maddalena Favaretto, Eva De Clercq, and Bernice Simone Elger. Big Data and discrimination: perils, promises and solutions. A systematic review. *Journal of Big Data*, 6(1):12, February 2019.

[Fle21] Will Fleisher. What's Fair about Individual Fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21,

pages 480–490, New York, NY, USA, July 2021. Association for Computing Machinery.

[FSV+18] Sorelle A. Friedler, Carlos Eduardo Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018.

[FW24] Christian Fröhlich and Robert C. Williamson. Insights From Insurance for Fair Machine Learning. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 407–421, New York, NY, USA, June 2024. Association for Computing Machinery.

[GDP] Regulation (eu) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[gen12] Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats). *Official Journal of the European Union*, C 11:1–11, January 2012. (2012/C 11/01).

[GlB] RIS - Gleichbehandlungsgesetz - Bundesrecht konsolidiert, Fassung vom 14.10.2024.

[HCZ+24] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM J. Responsib. Comput.*, 1(2), June 2024.

[HPS16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[JSeS+24] Sérgio Jesus, Pedro Saleiro, Inês Oliveira e Silva, Beatriz M. Jorge, Rita P. Ribeiro, João Gama, Pedro Bizarro, and Rayid Ghani. Aequitas flow: Streamlining fair ml experimentation. *Journal of Machine Learning Research*, 25(354):1–7, 2024.

[KAAS12] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 35–50. Springer, 2012.

[KKZ12] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929, 2012.

[KLMS19] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, 10:113–174, April 2019.

[KLRS17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[KMF+17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[KMR] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[KPB+24] Parameswaran Kamalaruban, Yulu Pi, Stuart Burrell, Eleanor Drage, Piotr Skalski, Jason Wong, and David Sutton. Evaluating fairness in transaction fraud models: Fairness metrics, bias audits, and challenges. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 555–563, New York, NY, USA, 2024. Association for Computing Machinery.

[LC21] Michele Loi and Markus Christen. Choosing how to discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philosophy & Technology*, 34(4):967–992, December 2021.

[LO24] Daphne Lenders and Anne Oloo. 15 years of algorithmic fairness - scoping review of interdisciplinary developments in the field. *CoRR*, abs/2408.01448, 2024.

[LRTW22] M. Lindholm, R. Richman, A. Tsanakas, and M. V. Wüthrich. DISCRIMINATION-FREE INSURANCE PRICING. *ASTIN Bulletin: The Journal of the IAA*, 52(1):55–89, January 2022.

[LV22] Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. *Computational Management Science*, 19:513–537, 2022.

[Micnd] Microsoft Learn. What is a machine learning model? `https://learn.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model`, n.d. Accessed: May 12, 2025.

[MMS+21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[OC20] Luca Oneto and Silvia Chiappa. *Fairness in Machine Learning*, pages 155–196. Springer International Publishing, Cham, 2020.

[O'N16] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA, 2016.

[OR22] Torsten Oletzky and Armin Reinhardt. Herausforderungen der Regulierung von und der Aufsicht über den Einsatz Künstlicher Intelligenz in der Versicherungswirtschaft. *Zeitschrift für die gesamte Versicherungswissenschaft*, 111(4):495–513, December 2022.

[PCA24] Guilherme Palumbo, Davide Carneiro, and Victor Alves. Objective metrics for ethical AI: a systematic literature review. *International Journal of Data Science and Analytics*, April 2024.

[PRW+17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[PS22] Dana Pessach and Erez Shmueli. A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3):51:1–51:44, February 2022.

[QRI+22] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, 12(3), May 2022.

[RD21] Boris Ruf and Marcin Detyniecki. Towards the right kind of fairness in AI. *CoRR*, abs/2102.08453, 2021.

[rig12] Charter of fundamental rights of the European Union. *Official Journal of the European Union*, C 326:391–407, October 2012. (2012/C 326/02).

[SHD+19] 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 *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 99–106, New York, NY, USA, 2019. Association for Computing Machinery.

[SMHP24] Janine Strotherm, Alissa Müller, Barbara Hammer, and Benjamin Paaßen. *Fairness in KI-Systemen*, page 163–183. Springer Fachmedien Wiesbaden, 2024.

[UN 09] UN Committee on Economic, Social and Cultural Rights. General comment no. 20: Non-discrimination in economic, social and cultural rights (art. 2, para. 2, of the international covenant on economic, social and cultural rights), July 2009.

[UNI04] THE COUNCIL OF THE EUROPEAN UNION. Council directive 2004/113/ec of 13 december 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, 2004.

[U.S21] U.S. Department of Justice, Civil Rights Division. *Title VI Legal Manual*, 2021.

[VB21] Michael Veale and Frederik J. Zuiderveen Borgesius. Demystifying the draft eu artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22:97 – 112, 2021.

[VR18] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7. Association for Computing Machinery, May 2018.

[WDE+23] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.

[WMR21] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review*, 123(3):735, 2021.

[WXT+23] Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 805–816, June 2023.

[WZLZ23] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Trans. Knowl. Discov. Data*, 17(3):35:1–35:27, March 2023.

[Xa24] Xi Xin and Fei Huang and. Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, 28(2):285–319, 2024.

[ZVGRG19] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: a flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[ZWW17] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3929–3935, 2017.

[Žl17] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, July 2017.