Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Master in Data Science

Master Thesis

Trustworthy Machine Learning: Mitigating Bias and Promoting Fairness in Automated Decision Systems

Author: Stephan Wolters

Madrid. July, 2023

Master Thesis
Máster en ***Ciencia de Datos***

Title: **Trustworthy Machine Learning: Mitigating Bias and Promoting Fairness in Automated Decision Systems**

July, 2023

Author: **Stephan Wolters**

Supervisor:
Juan Antonio Fernández del Pozo de Salamanca
ETSI Informáticos
de Inteligencia Artificial
UPM

Abstract

The widespread adoption of Artificial Intelligence (AI) and Machine Learning (ML) algorithms in recent years across many domains has led to an increased reliance on automated decision making. While these algorithms have shown tremendous promise in improving decision-making efficiency and accuracy, they are not immune to errors and biases. Consequently, there is a growing concern about the trustworthiness of automated decision-making systems (AD-MS).

Trustworthiness refers to the degree to which an AD-MS can be relied upon to produce accurate, fair, robust, transparent, inclusive, and empowering results. Ensuring the trustworthiness of AD-MS is crucial in several domains, among many others, healthcare, finance, criminal justice, and human resources. For instance, biased or inaccurate automated credit scoring systems can result in unfair denial of loans to certain individuals, while biased recruitment systems can perpetuate discrimination in the workplace. Consequently, there is a need for tools and approaches to improve the trustworthiness of AD-MS.

This project aims to explore the challenges and opportunities in achieving trustworthy automated decision-making. Specifically, it seeks to investigate the limitations of machine learning algorithms, the importance of trustworthiness, and the tools and approaches for improving trustworthiness. The project also aims to examine the ethical, legal, and social implications of AD-MS and present case studies that illustrate the practical application of the proposed tools and approaches. The project concludes by discussing future directions for research in this field.

Table of Contents

List of Acronyms

AAA	Algorithmic Accountability Act
ADA	Americans with Disabilities Act
AD-MS	Automated Decision-Making Systems
AIA	Artificial Intelligence Act
AI	Artificial Intelligence
AOD	Average Odds Difference
BBM	Black-box Models
CNN	Convolutional Neural Networks
CDD	Conditional Demographic Disparity
CF	Counterfactual Fairness
CNN	Convolutional Neural Networks
COMPAS	Correctional Offender Mgt. Profiling for Alternative Sanctions
CPD	Chicago Police Department
DNN	Deep Neural Networks
DPR	Demographic Parity Ratio
DSA	Digital Services Act
EAA	Education Amendments Act
ECOA	Equal Credit Opportunity Act
EED	Employment Equality Directive
EOD	Equalized Odds Difference
EOR	Equalized Odds Ratio
FHA	Fair Housing Act
FNRD	False Negative Rate Difference
FPRD	False Positive Rate Difference
FTC	Federal Trade Commission
FTU	Fairness Through Unawareness
GAM	Generalized Additive Models
GDPR	General Data Protection Regulation
GEI	Generalized Entropy Index
KLD	Kullback-Leibler Divergence
LGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LAPD	Los Angeles Police Department
ML	Machine Learning
NMI	Normalized Mutual Information
OAP	Overall Accuracy Parity
PCA	Principal Component Analysis
PDP	Partial Dependence Plot
PED	Predictive Equality Difference
PER	Predictive Equality Ratio
PIPL	Personal Information Protection Law

PIT	Probability Integral Transform
PPP	Positive Predictive Parity
PredPol	Predictive Policing
RED	Race Equality Directive
RNN	Recurrent Neural Networks
SCM	Structural Causal Models
SPD	Statistical/Demographic Parity Difference
SHAP	SHapley Additive exPlanations
SME	Small and Medium Enterprises
SMOTE	Synthetic Minority Oversampling Technique
TED	Treatment Equality Difference
TER	Treatment Equality Ratio
t-SNE	t-Distributed Stochastic Neighbor Embedding
WBM	White-box Models
XAI	Explainable Artificial Intelligence

List of Figures

List of Tables

List of Equations

List of Algorithms and Techniques

AdaBoost
Additive Counterfactually Fair Estimator
Adversarial Debiasing
CatBoost
Cox Proportional Hazards
Counterfactual Data Augmentation
Deep Neural Network
Disparate Impact Remover
Equalized Odds Postprocessing
Exponentiated Gradient Reduction
Fair k-Means Clustering
GerryFair Classifier
Grid Search Reduction
HistGradientBoost
Learning Fair Representations
Light Gradient Boosting Machine (LGBM)
Logistic Regression
Meta-Fair Classifier
Optimized Pre-processing
Platt Scaling
Prejudice Remover
Reject Object Based Classification
ResNet-50
Reweighing
Structural Causal Model
SensitiveNet
Synthetic Minority Oversampling Technique (SMOTE)
XGBoost

1 Introduction

1.1 Background and Context

Artificial intelligence (AI) has come a long way since its inception in the 1950s. and can be traced back to the Dartmouth Conference of 1956, where the term "Artificial Intelligence" was coined (McCarthy, J., 1955). In the following decades, researchers made significant progress in the development of AI technologies, with notable breakthroughs such as the introduction of expert systems in the 1970s and the development of neural networks in the 1980s. However, progress was slow, and it was not until the early 2000s that AI algorithms began to gain widespread adoption.

Today, the field of AI is rapidly evolving, with advances in deep learning, natural language processing, and computer vision, and AI and machine learning (ML) algorithms are used in decision-making processes across many domains, including finance, healthcare, criminal justice, and human resources. These automated decision-making systems (AD-MS)[1] have the potential to enhance efficiency, accuracy, and speed, leading to better outcomes. However, they are not safe from biases and discrimination, leading to a growing concern about the trustworthiness of these systems. There are limitations to these systems, such as the inability to explain their decision-making processes fully. This lack of transparency can lead to distrust and limit the adoption of these systems.
It is essential to ensure that AD-MS are fair, transparent, robust, and inclusive, especially when used to make decisions about individuals that may impact their lives.

Bias and unfairness in AD-MS can result from the use of protected attributes such as gender, race, ethnicity, age, or disability in decision-making processes. A series of domains with a myriad of known cases of bias and discrimination are briefly mentioned without being an exhaustive overview:

Recruiting[2]
AD-MS have been used in the recruitment process to screen resumes and applications, reducing the workload of recruiters. However, these systems can perpetuate gender or racial biases if they use historical data that reflects those

[1] Trustworthy AD-MS, AI systems and ML systems are sometimes used interchangeably in this project. However, the primary differentiating factor among trustworthy AD-MS, AI systems, and ML systems lies in their focus and underlying principles. Trustworthy AD-MS prioritize transparency, fairness, and accountability in decision-making processes, specifically targeting the need for explainability and interpretability. On the other hand, AI systems encompass a broader set of technologies that emulate human intelligence, encompassing perception, reasoning, and decision-making capabilities. ML systems, as a subset of AI, primarily leverage statistical models and learning algorithms to extract patterns from data and make predictions or decisions (e.g. Varshney, K., 2022).

[2] In 2016, Amazon created an experimental AI-driven hiring tool designed to identify top talent from a pool of job applicants. However, it was discovered that the algorithm was biased against female candidates, as the system had been trained on resumes submitted to Amazon over a 10-year period, which were predominantly from men (Dustin, J., 2018).

biases. For example, if historical data shows that men are more likely to be hired for particular jobs, the automated system may replicate that bias. This can lead to discrimination against women or minorities, impacting their career opportunities (E. Kazim, A., 2021).

Credit Scoring
Automated credit scoring systems use ML algorithms to evaluate the creditworthiness of individuals. These systems consider various factors such as credit history, employment, and income. However, the use of protected attributes such as race, ethnicity, or gender in credit scoring can result in discrimination against certain groups. For example, studies have shown that automated credit scoring systems are more likely to deny loans to individuals from minority communities (Kozodoi, N., et al., 2022 / Pagano, T., et al., 2023).

Pre-trial Recidivism Screening[3]
AD-MS have been used in pre-trial recidivism screening to predict the likelihood of an individual committing a crime again in the future. However, these systems can perpetuate racial biases if they use historical data that reflects those biases. For example, if historical data shows that individuals from certain racial or ethnic groups are more likely to reoffend, the automated system may replicate that bias. This can lead to discrimination against individuals from those communities (Mitchell, S., et al., 2021).

Salary and Promotion[4]
AD-MS have been used in determining salaries and promotions in organizations. However, these systems can perpetuate gender or racial biases if they use historical data that reflects those biases. For example, if historical data shows that men are more likely to receive higher salaries and promotions, the automated system may replicate that bias. This can lead to discrimination against women or minorities, impacting their career opportunities.

Healthcare Health Insurance [5]
AD-MS have been used in healthcare to suggest treatments for patients. However, these systems can perpetuate biases if they use historical data that reflects those biases. For example, if historical data shows that certain

[3] The COMPAS case involves an algorithm used to predict the likelihood of a defendant committing future crimes, which was used by judges in several U.S. states to inform decisions about pretrial detention, sentencing, and parole. In 2016, investigative news organization ProPublica published a study that found that the algorithm, called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), was biased against African American defendants, as it was more likely to incorrectly predict that they would commit future crimes (Angwin, J., et al., (ProPublica), 2016).

[4] Most articles refer to the US Census “Adult” dataset from 1994 to investigate algorithmic biases (e.g., Li, S., et al., 2022) and unfortunately, the datasets of known cases are not publicly available. However, one of the many cases is the Goldman Sachs Case: In 2021, a group of Goldman Sachs employees filed a lawsuit alleging that the company's automated system for setting salaries and bonuses was biased against women. The lawsuit claims that the system systematically undervalued women's contributions to the company and paid them less than their male colleagues. The trial commences on June 5, 2023 (Chen-Oster v. Goldman Sachs class action, 2018).

[5] Most authors use the Medical Expenditure Panel Survey (MEPS) dataset to investigate algorithmic bias in healthcare and health insurance, probably also due to the fact that most datasets are not publicly available.

treatments are less effective for individuals from certain racial or ethnic groups, the automated system may replicate that bias. This can lead to unequal health outcomes for individuals.
AD-MS have also been used in health insurance to evaluate risk and set premiums. However, the use of protected attributes such as gender, age, or health status in determining premiums can result in discrimination against certain groups. For example, if an automated system considers age as a factor, older individuals may have to pay higher premiums, leading to discrimination (e.g., Creedon, T., et al., 2022).

Computer Vision and Object Recognition in Images[6]
AD-MS have been used in computer vision and object recognition in images. However, these systems can perpetuate biases if they use historical data that reflects those biases. For example, if historical data shows that certain objects are more likely to appear in images from certain regions or demographics, the automated system may replicate that bias. This can lead to inaccurate object recognition and biased decisions based on image analysis.

Therefore, it becomes evident that these developments cause a major impact on how our societies have been constructed, and it seems doubtful that these high-impact problems will be solved in a self-regulatory manner by the agents who adopted automated decision-making processes based on AI, i.e., there is a need for legal frameworks that govern the use of AD-MS. To name just a few of them, in the European Union, the General Data Protection Regulation (GDPR) sets rules for data protection, including automated decision-making. The GDPR requires that individuals have the right to contest automated decisions and to receive information about how those decisions are made. In the United States, the Equal Credit Opportunity Act (ECOA) prohibits credit discrimination based on protected attributes. The ECOA requires that creditors provide applicants with the reasons for credit denials.

In a technical context, how fairness and bias mitigation are handled relies heavily on a precise definition of fairness metrics, which in turn are essential in evaluating the performance of AD-MS. Demographic parity, equalized odds, and opportunity equality are some of the commonly used fairness metrics. **Demographic parity** refers to the proportion of individuals receiving a positive outcome, irrespective of their protected attribute status. **Equalized odds** refer to the equality of true positive rates and false positive rates across protected attributes. **Opportunity equality** ensures that individuals have an equal chance of receiving a positive outcome, regardless of their protected attribute status.

These metrics, among many others, are used in a series of publicly available tools to analyze and mitigate bias and fairness in ML. There are several tools and frameworks that can analyze and mitigate bias, e.g., AIF360 is an open-source toolkit that provides algorithms for fairness metrics computation, bias

[6] In 2015, a Google Photos user reported that the app had automatically tagged a photo of him and a friend as "gorillas." This incident highlighted the potential for facial recognition algorithms to perpetuate racial biases and reinforce harmful stereotypes.
Google issued an apology and took immediate action to address the issue, including removing the "gorilla" tag from the app's lexicon and improving its image recognition algorithms to better recognize people of all races (Dougherty, C., NYT, 2015).

mitigation, and fairness visualization. FairLearn is another open-source toolkit that provides a variety of ML algorithms for fair classification. TensorFlow Responsible AI is a framework that provides tools for building and deploying responsible AI models. Aequitas is an open-source bias audit toolkit that evaluates ML models for disparate impact, all of which are analyzed in detail in section 3.6 Tool-based Bias Mitigation.

Piecing together the previous building blocks, this project aims to investigate the challenges and opportunities in achieving trustworthy automated decision-making in ML. By exploring the legal frameworks, fairness metrics, and bias mitigation tools available for ML algorithms, this project seeks to provide practical solutions for mitigating bias and ensuring fairness in AD-MS. By addressing bias and ensuring fairness in AD-MS, this project aims to contribute to the development of trustworthy systems that can serve the needs of all individuals.

1.2 Rationale and Significance

AD-MS are a subset of AI technologies that leverage ML models to make predictions or decisions based on input data. These systems can be applied across various domains, often automating complex and time-consuming tasks that would otherwise require human intervention. The primary advantage of AD-MS is their ability to process large amounts of data quickly and efficiently, leading to more informed and data-driven decisions.

AI technologies have evolved significantly over the past few decades, with rapid advancements in ML techniques, computational power, and data availability. These developments have enabled AI systems to tackle more complex tasks and achieve unprecedented levels of performance, resulting in widespread adoption across various domains. From natural language processing to computer vision and robotics, AI technologies have permeated every aspect of modern society, transforming the way we live and work.

The emergence of AI technologies in high-stakes decision-making systems has brought both opportunities and challenges. For example, AI-based credit scoring models can process vast amounts of data to assess an individual's creditworthiness more accurately and efficiently than traditional methods. Similarly, AI-driven healthcare applications can help clinicians diagnose and treat patients more effectively, while AI-powered criminal justice tools can assist in predicting recidivism and allocating resources more efficiently. However, these applications also underscore the need for fairness, transparency, and accountability in AI systems, as the consequences of biased or untrustworthy decision-making can be dire.

Trustworthiness in AI systems is of paramount importance, as it directly impacts the public's perception, acceptance, and adoption of these technologies. Trustworthy AI systems must be designed with a focus on fairness, transparency, and accountability, ensuring that they do not perpetuate biases or exacerbate existing societal inequalities. A lack of trust in AI systems can hinder their adoption and limit the potential benefits that these technologies can bring to various domains.

In recent years, the rapid development and widespread application of ML and AI in general have significantly impacted various aspects of modern society.

From healthcare to finance and criminal justice, data-driven decision-making has become an integral part of our lives. This reliance on ML and AI systems, however, raises concerns about their fairness, transparency, and accountability, which are the focus of this project.

The consequences of using untrustworthy AI systems can be far-reaching and severe, particularly when these systems are deployed in high-stakes decision-making processes. Unfair and biased AI systems can lead to discriminatory outcomes, causing harm to individuals and marginalized communities. Such consequences can result in legal liabilities, reputational damage to organizations, and a loss of public trust in AI technologies.

One of the most significant assumptions in data-driven decision-making is the supposed objectivity of these decisions. It is often believed that machines, being free from human prejudices and biases, can make fair and unbiased decisions. However, this is far from the truth, as ML algorithms are trained on historical data, which may contain inherent biases and prejudices. These biases can be inadvertently learned and perpetuated by the algorithms, resulting in unfair and discriminatory outcomes.

Companies that deploy ML models often prioritize factors such as ease of implementation, runtime scalability, and low risk of failure over fairness and ethical considerations. This is primarily due to the competitive nature of the market and the need for businesses to maintain efficiency and profitability. However, such an approach can lead to the implementation of models that inadvertently perpetuate biases and discrimination, causing significant harm to marginalized communities.

Discrimination, in the context of ML and AI, refers to the unequal treatment of individuals or groups based on certain protected attributes, such as race, gender, and age. There are two primary forms of discrimination: disparate treatment and disparate impact. Disparate treatment occurs when a decision-making system explicitly treats individuals differently based on their protected attributes. This form of discrimination is illegal in many legislations, e.g., under the US Civil Rights Act or the EU Equality Treatment Framework Directive (2000/78/EC), among many others. Disparate impact, on the other hand, occurs when a decision-making system, even if unintentionally, has a disproportionate adverse effect on a protected group. Legal remedies for disparate impact are not as clear-cut, and it is a matter of ongoing debate.

The use of uninterpretable black-box models in ML further exacerbates the issues of fairness and accountability. These models, which include complex neural networks and ensemble methods, often produce highly accurate predictions but are difficult to interpret or explain. This lack of interpretability makes it challenging to identify and rectify potential biases, which can lead to unintended discriminatory consequences. White-box models, on the other hand, are inherently transparent and can be easily interpreted such as linear or logistic regression or simple decision tree models.

A common misconception is that the removal of protected attributes from the data used to train ML models will ensure fairness and prevent discrimination. However, this approach is often insufficient, as biases can still be introduced through correlated proxies. For example, if an ML model is trained on a dataset that excludes gender but includes occupation, the model may still learn gender biases if certain occupations are predominantly associated with one gender. This highlights the need for more robust bias mitigation measures to ensure fairness in ML models.

Considering the potential risks associated with untrustworthy AI systems, there is a pressing need for research in this area. Developing and implementing effective fairness metrics and bias mitigation measures can help ensure that AI systems are designed and deployed responsibly, minimizing the potential for harm, and maximizing the benefits of these technologies. Furthermore, research in this area can contribute to the establishment of guidelines and best practices for the development, deployment, and monitoring of AI systems, promoting transparency and accountability.

In conclusion, this project aims to address the crucial issue of trustworthy automated decision-making by exploring fairness metrics and bias mitigation measures in machine learning. By doing so, it seeks to provide a foundation for more ethical, transparent, and accountable AI systems, which will ultimately lead to better decision-making processes and greater trust in these technologies. The significance of this research is underscored by the increasing reliance on AI and ML in various sectors of society, as well as the potential consequences of perpetuating biases and discrimination through automated decision-making.

1.3 Research Questions and Objectives

In this master project the following main research questions and objectives are addressed:

1) How can methods and techniques be applied to develop AD-MS that **enhance transparency and interpretability**, thereby improving their explainability?
2) In what ways do **data-driven approaches unintentionally encode human biases** and introduce new ones, and what are the implications of these biases for fairness in AD-MS?
3) How can **fairness in AD-MS be effectively measured**, particularly considering the complex relationships between input features, protected attributes, and target variables?
4) What are the **key trade-offs between performance and fairness** in machine learning models, and how can these trade-offs be navigated in practice to balance optimal outcomes with fairness considerations?
5) What methods can be investigated and applied to **minimize the potential for AI systems to introduce and perpetuate discriminatory practices**, reproduce, reinforce, and exacerbate existing biases, and create feedback loops from deployed systems?
6) How can effective methods for **incorporating causality into fairness-aware AD-MS** be applied to mitigate bias and discrimination in decision-making processes?
7) How can **multimodal input features be handled in fairness-aware models**, and what strategies can be employed to mitigate non-apparent bias?

The research questions and objectives contribute to the overall field of AI and ML by addressing critical ethical and practical challenges that arise in the development and deployment of AD-MS. In particular, the focus on fairness metrics, tradeoffs, and potential biases in data-driven approaches aligns with the growing recognition of the importance of trustworthiness, transparency, and accountability in AI.

A key challenge in ensuring fairness in AD-MS is the tradeoff between performance and fairness. As models strive to minimize average error, they may inadvertently fit the majority population while neglecting the needs of minority groups. Furthermore, data-driven approaches can unintentionally encode human biases and introduce new ones, leading to the potential for AI systems to perpetuate discriminatory practices.

Overall, the research objectives will help advance the field of AI and machine learning by providing new insights into the measurement and mitigation of biases in AD-MS, as well as by offering practical guidance on how to balance fairness and performance in real-world applications. As AI technologies continue to evolve and play an increasingly prominent role in various domains, ensuring fairness and trustworthiness will be essential for realizing their full potential and addressing the ethical challenges they present.

1.4 Methodology

The methodology of this project is designed to provide a comprehensive understanding of the issues and challenges related to fairness and bias in algorithmic decision-making processes. It encompasses a combination of approaches, including theoretical frameworks, mitigation techniques, legal and ethical implications, and practical applications in the form of case studies. The methodology consists of the following steps:

1) Foundation of theoretical framework with state-of-the-art research on algorithmic fairness and bias:
   The study begins with a thorough review of current literature and research on algorithmic fairness and bias. This provides a solid theoretical foundation that is necessary to understand the context, challenges, and limitations of existing ML algorithms. The theoretical framework also includes a discussion on the importance of trustworthiness in automated decision-making and the concepts of trustworthy AI.

2) Mitigation of algorithmic bias, research, and tools:
   The second step involves identifying and evaluating various techniques and tools available to mitigate algorithmic bias. These approaches include:
   a) Automated data cleaning and correction
   b) Algorithm-agnostic methods
   c) Model interpretability and explainability
   d) Causality-based fairness methods
   e) Tool-based bias mitigation solutions such as AI-Fairness 360, What-if Tool, Fairness Flow, Aequitas, and Themis AI.

3) Legal and ethical implications:
   The study also examines the legal and ethical implications of implementing fairness and bias mitigation measures in machine learning. This includes exploring transparency, responsibility, preventing unintended consequences and bias, and balancing accuracy, fairness, and privacy concerns.

4) Assembling the building blocks in three case studies based on the CRISP-DM methodology:
   The project adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to structure the case studies. The CRISP-DM framework consists of six stages:
   a) Business understanding

b) Data understanding
c) Data preparation
d) Modeling
e) Evaluation
f) Deployment.

This allows for a systematic investigation of fairness and bias mitigation measures in three different real-world applications:

a) HR recruitment process,
b) Automated credit scoring,
c) Predictive policing and recidivism profiling (COMPAS).

The case studies take into account:

1) **Pre-processing** (e.g., relabel, reweight, resample): The data preprocessing stage involves techniques such as relabeling, reweighting, and resampling to reduce biases present in the data before feeding it to the machine learning algorithms.

2) **In-processing** (e.g., augmented cost function, fairness regularizer): In-processing methods are applied during the model training phase. They include techniques like augmented cost functions and fairness regularizers that modify the learning process to ensure fairness while optimizing for accuracy.

3) **Post-processing** (based on holdout set): Post-processing techniques are applied after the model has been trained, using a holdout set. These methods aim to adjust the algorithm's predictions or decisions to reduce unfair outcomes.

By employing this comprehensive methodology, the project aims to address the challenges of fairness and bias in algorithmic decision-making, provide practical solutions, and contribute to the development of more trustworthy AI systems.

1.5 Outline of the Project

In this project, the topic of trustworthy automated decision-making is investigated, with a particular focus on fairness and bias mitigation measures in ML. The following sections provide an overview of the different chapters and their contents:

Chapter 1 provides an introduction to the topic, discussing the background and context of the research, the rationale and significance of the study, and the research questions and objectives. The chapter also outlines the methodology used in the study, which includes a review of relevant literature and case studies.

In Chapter 2, the theoretical framework that underpins the study is presented. The chapter discusses the limitations of ML algorithms and the importance of trustworthy AI. Furthermore, the concepts of trustworthy AI and what it means for an AI system to be considered trustworthy are explored.

Chapter 3 examines the tools and approaches that can be used to improve the trustworthiness of AD-MS. The chapter covers automated data cleaning and correction, algorithm-agnostic approaches, model interpretability and explainability, causality-based fairness methods, and tool-based bias mitigation. The most widely used tool-based bias mitigation frameworks, such as IBM: AI-

Fairness 360, Google: What-if Tool, Facebook: Fairness Flow, Carnegie Mellon: Aequitas, and Themis AI are also presented in detail.

In Chapter 4, the ethical, legal, and social implications of trustworthy AD-MS are explored. The chapter discusses the importance of ensuring transparency and responsibility, preventing unintended consequences and bias, and balancing accuracy, fairness, and privacy.

Chapter 5 presents case studies that demonstrate the application of the tools and approaches discussed in previous chapters. A methodological approach is provided, which includes data collection and preprocessing, model selection and training, evaluation and validation metrics, and quantifying bias and fairness. The case studies focus on HR recruitment processes, automated credit scoring, and predictive policing and recidivism profiling, with a specific focus on the use of the COMPAS system.

In Chapter 6, the results of the case studies are discussed, and conclusions are drawn based on the findings. Recommendations for future research directions that can contribute to improving the trustworthiness of AD-MS are also provided.

Finally, Chapter 7 provides an overview of the main contributions of the study and the limitations of the research. Recommendations for practitioners and policymakers who are involved in the development and deployment of trustworthy AI systems are also presented.

The study aims to provide a comprehensive understanding of trustworthy AD-MS, with a particular emphasis on the role of fairness and bias mitigation measures in ML. The theoretical and practical contributions of the study can help inform the development and deployment of trustworthy AI systems in various domains.

2 Theoretical Framework

This chapter delineates the conceptual basis upon which this research is constructed. The chapter bifurcates into two primary domains.

1) "Understanding Shortcomings of AI Systems," scrutinizes the inherent limitations of Artificial Intelligence systems, including common sources and types of bias. This examination offers a comprehensive understanding of the pitfalls and challenges that encumber the pursuit of optimized AI systems, thereby offering a vantage point from which to approach their potential amelioration.
2) "Building Trustworthy AI: Principles and Approaches," delves into the constitution of trustworthy AI systems. Informed by the understanding of AI limitations and biases, this domain encompasses critical dimensions of trustworthiness, including transparency, fairness, and robustness, and expounds on the integral role trustworthiness plays in the effective deployment of AI systems. This chapter also elucidates essential tools for fostering trust in AI, namely fairness metrics and explainable AI, thus providing a practical dimension to the theoretical discourse.

The chosen theoretical framework represents a fusion of critical analysis of AI challenges with a forward-looking perspective on trustworthiness enhancement. This balance equips readers with a nuanced understanding of both the challenges inherent in AI and the potential solutions offered by transparent, fair, and robust AI systems.

2.1 Understanding Shortcomings of AI Systems

The journey to harness the potential of AI systems comes with several complexities and challenges. While these systems are powerful tools for processing large amounts of data and making predictions, they also have inherent limitations. The understanding of these limitations forms the foundation for the exploration of AI's trustworthy attributes.

2.1.1 Artificial Intelligence Systems and Their Limitations

AI systems, ranging from simple rule-based algorithms to complex deep learning models, can process vast volumes of data, uncover patterns, make predictions, and even simulate human-like decision-making processes.

However, along with their numerous advantages, AI systems inherently carry important limitations. Their functionality and efficiency are significantly reliant on the quality and the nature of the input data they are trained and tested on. Moreover, their ability to make predictions or decisions is confined to their training experiences and the patterns they have learnt, which may not always fully encapsulate the complexity of real-world situations. Consequently, any bias present in the training data, whether unintentional or deliberate, may result in biased predictions or decisions.

Moreover, the 'black box' nature of many complex AI models, particularly deep learning systems, can make their decision-making processes obscure and difficult to interpret. This lack of transparency undermines user trust and poses challenges to their wider acceptance and application. Another key limitation is that AI systems lack the ability to comprehend context or understand meaning in the same way humans do. They perform tasks based on their training, without an inherent understanding of the task's significance.

Without claiming to be exhaustive, table 2-1 is an attempt to categorize the most common limitations of AI systems which manifest in many different ways. It is beyond the scope of this project to delve into all these issues in detail, however, a short explanation is provided. The compilation and categorization are based on a myriad of sources, among others Suresh, H. et al. (2021), Thampi, A. (2022), and Varshney, K. (2022).

Subsection 2.1.2 Common Sources and Types of Bias and Their Limitations, on the other hand, summarizes the most commonly known biases as their understanding is a prerequisite for the subsequent analyses in the remaining sections.

Category	Specific Issues
Data-Related Issues	Data Bias, Sampling Bias, Measurement Bias, Availability Bias, Temporal Bias, Unrepresentative Features, Data Drift
Modeling Issues	Labeling Bias, Aggregation Bias, Misclassification, Overfitting, Underfitting, Overemphasis on Certain Features, Neglect of Correlated Variables, Model Drift, Lack of Robustness
User-Related Issues	Confirmation Bias, Automation Bias, Confirmation Bias in Interpretation, Lack of Trust, Cognitive Overload
Societal Impact	Prejudice Bias, Exclusion, Unfair Punishment, Economic Inequality, Digital Divide, Environmental Impact, Job Displacement, Information Manipulation
Design and Usability Issues	Lack of Transparency, Accessibility and Usability, Misinterpretation of AI outputs, Bias in Design, Complexity of AI systems
Privacy Concerns	Privacy Violations, Data Security, Inference Attacks
Ethics and Legal Issues	Ethical Considerations, Accountability and Liability, Regulation and Oversight

Table 2-1: AI Systems' Problems and Risks from Different Viewpoints

Data-Related Issues

Data Bias: This occurs when the dataset used to train an AI model is not representative of the real-world environment it is expected to operate in, skewing the model's performance and predictions. It can arise due to biases in data collection, labelling, and curation.

Sampling Bias: It arises when the data used for training the AI model is not randomly selected and does not adequately represent the broader population or phenomenon, which can lead to systematic error and skewed results.

Measurement Bias: This occurs when there are consistent errors in the way data is collected or measured, affecting the model's learning and prediction capabilities.

Availability Bias: It is the bias that results from over-relying on readily available or easily accessible data for model training, leading to an unrepresentative sample and skewed results.

Temporal Bias: This occurs when the data used to train an AI model does not adequately account for changes over time. It can cause a model to learn outdated patterns or miss emerging trends.

Unrepresentative Features: This occurs when the features used in the AI model do not adequately represent the problem space, leading to inaccuracies or biases in the model's outputs.

Data Drift: It refers to changes in input data distribution over time, which can result in a decrease in model performance as the data it was trained on no longer reflects the current environment.

Modeling Issues

Labeling Bias: This occurs when the labels used in supervised learning models are inaccurately assigned, often due to human bias, leading to errors in the model's learning and prediction capabilities.

Aggregation Bias: It arises when data from different groups are improperly aggregated, potentially hiding significant group-specific trends or characteristics.

Misclassification: This refers to errors in predictive models where instances are incorrectly assigned to classes. It can result from data or modeling issues and may disproportionately affect certain groups, leading to fairness concerns.

Overfitting: It refers to a modeling error that occurs when an AI model is excessively complex and captures noise or random fluctuations in the training data, leading to poor generalization performance on unseen data.

Underfitting: Describes a model that is too simple to capture the underlying structure of the data. An underfitted model has poor performance not only on the test data but also on the training data, as it fails to capture the complexity of the data distribution and relationships between variables.

Overemphasis on Certain Features: This occurs when an AI model assigns excessive importance to certain features, potentially leading to skewed predictions and overlooking important relationships with other variables.

Neglect of Correlated Variables: It occurs when an AI model fails to account for variables that are correlated, leading to an oversimplified model and potentially skewed results.

Model Drift: It is a phenomenon where the model's performance degrades over time because the statistical properties of the target variable, which the model is trying to predict, change.

Lack of Robustness: It refers to the vulnerability of AI systems to minor alterations in the input data, noise, outliers, or adversarial attacks, which can drastically impact their performance.

User-Related Issues

Confirmation Bias: This is a type of cognitive bias where users prefer information that confirms their existing beliefs, potentially leading to incorrect or biased decision-making when interpreting AI outputs.

Automation Bias: This occurs when users overly rely on AD-MS and ignore other sources of information, including their own judgment, leading to potential errors and over-trust in the AI system.

Lack of Trust: It arises when users are reluctant to rely on AI systems due to a variety of factors such as lack of transparency, perceived inaccuracy, fear of job loss, or privacy concerns.

Cognitive Overload: This occurs when an AI system provides too much information or too complex information, overwhelming users and potentially leading to decision paralysis or misuse of the system.

Societal Impact

Prejudice Bias: It refers to biases in AI outputs that can lead to unfair disadvantages or harm to certain groups based on characteristics such as race, gender, or age, reflecting and amplifying existing societal prejudices.

Exclusion: This refers to the unintentional marginalization of certain groups due to biases in AI system design, implementation, or outcomes, limiting their access to benefits or opportunities.

Unfair Punishment: This occurs when an AI system's biased or erroneous decisions result in unjust negative consequences for individuals or groups.

Economic Inequality: This can arise when the benefits and opportunities created by AI technologies are unevenly distributed, exacerbating existing economic disparities.

Digital Divide: This refers to the gap between those who have access to computers, the internet, and AI technologies and those who do not, leading to inequality in opportunities and benefits.

Environmental Impact: The energy consumption of large-scale AI systems contributes to environmental harm, while the use of AI in environmental monitoring and prediction can also have significant implications for sustainability.

Job Displacement: It relates to the potential loss of jobs due to automation and AI technologies, which can lead to societal and economic disruption.

Information Manipulation: This applies to the use of AI technologies in disseminating false or misleading information, affecting public opinion and trust.

Design and Usability Issues

Lack of Transparency: It pertains to the 'black box' nature of many AI systems, where the logic behind their decisions is not clear to users, undermining trust and accountability.

Accessibility and Usability: This refers to the ease with which diverse users can access and effectively use AI technologies, which can be impacted by factors such as design complexity, user literacy, and language capabilities.

Misinterpretation of AI outputs: This occurs when users, especially non-experts, misinterpret the results or recommendations of an AI system, potentially leading to incorrect decisions or actions.

Bias in Design: This concerns biases that can be introduced during the design of an AI system, which can affect the way it interacts with users and its overall performance and outcomes.

Complexity of AI systems: The complexity of AI models, particularly deep learning systems, can make them difficult for users to understand and use effectively, potentially limiting their adoption and impact.

Privacy Concerns

Privacy Violations: This relates to potential breaches of personal privacy due to data collection, storage, and processing practices in AI systems, which can expose sensitive information and lead to harm.

Data Security: This implies the measures in place to protect data used in AI systems from unauthorized access, alteration, or damage, which is critical for maintaining privacy and trust.

Inference Attacks: These are attacks where an adversary uses AI system outputs to infer sensitive information about the training data, representing a significant privacy risk.

Ethics and Legal Issues[7]

Ethical Considerations: These encompass a broad range of concerns, including respect for autonomy, fairness, accountability, and transparency, that are essential for the responsible development and use of AI technologies.

Accountability and Liability: These refer to the mechanisms in place to assign responsibility for the outcomes of AI systems, which can be complex due to the automated nature of these systems and the potential for unintended consequences.

Regulation and Oversight: This refers to the legal and regulatory frameworks governing the use of AI, including standards for privacy, fairness, and transparency, which can shape the development and deployment of AI technologies and their societal impact.

[7] cf. Section 4. Ethical, Legal, and Social Implications

After this short introduction to the great variety of limitations in AI systems, those which are most relevant to this project are described in detail in the following subsection.

2.1.2 Common Sources and Types of Bias in AI Systems

This section delves into the common sources and types of bias inherent in AI systems. Building upon the foundational understanding of AI system limitations established in the preceding section, this segment aims to explicate the various origins and categories of bias that can infiltrate and influence these systems. Seven critical types of bias, namely historical, representation, measurement, aggregation, learning, evaluation, and deployment biases, are discussed in detail. Each type of bias is examined from its genesis to its potential impact on AI systems. The objective is to provide an exhaustive understanding of how bias can manifest within AI systems, which forms the foundation for subsequent discussions on trustworthiness, fairness, and the mitigation of these biases in the development and application of AI. Figure 2-1 (Harini, S. et al., 2021) shows the first three biases all of which are explained below.

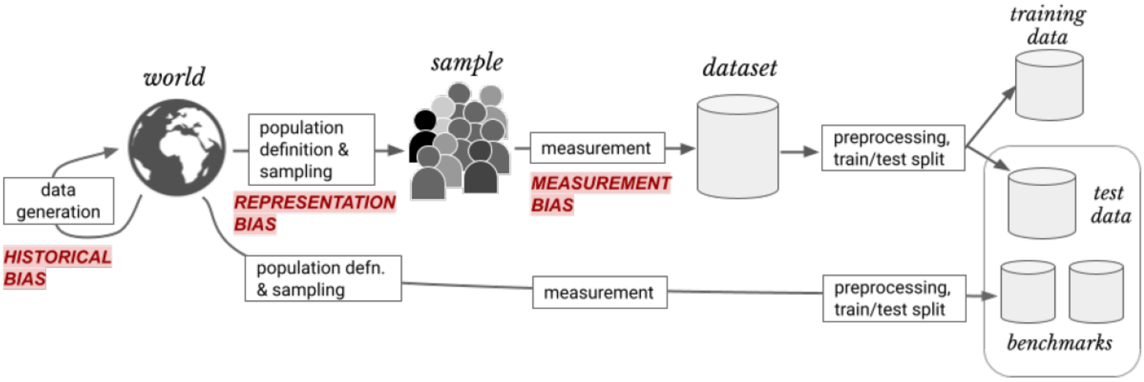


Figure 2-1: Bias Sources in the Data Generation Pipeline

Historical Bias: Historical bias manifests when the training data incorporates biases that have existed over time in the society or system it represents. The learned model, even when accurately reflecting the training data, may therefore perpetuate these biases and cause harmful outcomes.

An illuminating example of historical bias lies within predictive policing models. These systems are trained on past crime data, aiming to predict future crime hotspots and allocate resources accordingly. However, the training data, composed of historical police records, often reflects not just the actual crime rates but also the biases and discriminatory practices of the past law enforcement system (Lum, K. et al., 2016). For example, if a particular neighborhood was historically over-policed due to racial or socio-economic biases, the records will show an inflated crime rate in that area, and the predictive model would continue to target the same region disproportionately. This could cause a feedback loop, perpetuating a cycle of over-policing and exacerbating social inequities.

Researchers like Barocas et el. (2016) have highlighted this aspect of data mining in their work, arguing for a more careful approach to model construction, and the necessity of considering the societal context from which data originates.

Representation Bias: Representation bias is a type of bias that arises when the sample used to train an ML model does not adequately reflect the target population, resulting in a model that generalizes poorly to certain subgroups within the population. This bias can manifest in a variety of ways:

Inaccurate Target Population Definition: If the defined target population does not correspond to the actual user population, representation bias can arise. For example, a model trained using data representative of the population of Madrid may not perform well when applied to the population of Barcelona due to regional differences. Moreover, the temporal aspect is crucial as well, since data representing Madrids's population 30 years ago may not accurately reflect the current population, given demographic shifts, cultural changes, and other factors.

Underrepresented Groups within Target Population: The presence of minority groups within the target population that are underrepresented in the training data can lead to representation bias. For instance, if a medical dataset defines its target population as adults aged 18-40, and only 5% of this population consists of pregnant individuals, the model may perform poorly for this subgroup due to the lack of sufficient training data. Despite perfect sampling, the model's robustness could be compromised for these underrepresented individuals.

Flawed Sampling Methods: Even when the target population is accurately defined, biased sampling methods can lead to representation bias. For example, in modeling an infectious disease, the target population might be all adults, but the available medical data may only include individuals who were deemed severe enough for further screening. Consequently, the model's training data represents a skewed subset of the target population, a situation often referred to as sampling bias in statistical literature.

These instances demonstrate how representation bias can result in ML models that fail to deliver reliable or fair predictions for all sections of the population. Researchers have stressed the importance of mitigating representation bias to ensure more robust and equitable models. Zhang, J., et al., (2018), for instance, discuss methods to correct for sampling bias in causal inference, a concept that can be extended to a broader ML context.

An apt example of representation bias is presented in the realm of facial recognition technologies. Commercial gender classification systems have been found to perform poorly on darker-skinned and female faces, as a result of underrepresentation of these groups in the training data.

This issue is meticulously studied by Joy Buolamwini et al. (2018) in their paper, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In their research, they compared commercial gender classification systems from IBM, Microsoft, and Face++. They found that all three systems displayed lower accuracy rates for female faces compared to male faces, and lower accuracy rates for darker-skinned individuals compared to lighter-skinned individuals.

Moreover, the highest disparities were observed for darker-skinned females, the group that is least represented in many facial analysis benchmarks. For instance, the error rates for classifying gender of lighter-skinned males were less than 1% for all three systems, whereas for darker-skinned females, the error rates were as high as 34.7% in one system.

This example underscores how representation bias can lead to substantial disparities in model performance across different subgroups. It also highlights the necessity of ensuring diverse and representative training data in developing machine learning models to avoid unfair outcomes and discriminatory practices.

Measurement Bias: This form of bias arises when the selection, collection, or computation of features and labels in a ML model introduces errors or inconsistencies. This typically occurs when a proxy measurement is used to represent an abstract construct, leading to inaccuracies when the proxy poorly reflects the target construct or is generated differently across groups. This can occur under the following scenarios:

Oversimplified Proxies: In some instances, the proxy chosen oversimplifies a complex construct, leading to a lack of nuances that could impact the model's performance. For example, in predicting student success for college admissions, the algorithm designers might resort to using GPA as a proxy. However, the construct of a "successful student" encapsulates much more than academic performance alone, such as emotional intelligence, perseverance, and creativity. The use of a simplified proxy such as GPA ignores these varied indicators of success and can lead to a model that misrepresents the full complexity of student success (Duckworth, A. et al., 2015).

Varying Measurement Methods: When the method of measurement varies across groups, it can introduce bias. Consider a scenario where factory workers at different locations are monitored for the number of errors they commit, with the observed errors being a proxy for work quality. If one location is scrutinized more stringently or frequently, it may appear that workers at that location commit more errors. This could trigger a feedback loop, where this group is subject to increased monitoring due to the apparent higher error rate (Frey, C. B., et al., 2017).

Inaccurate Measurement Across Groups: There can be instances where the accuracy of measurement varies across groups, contributing to measurement bias. In medical contexts, "diagnosed with condition X" is often used as a proxy for "has condition X." However, structural discrimination can lead to systematically higher rates of misdiagnosis or underdiagnosis in certain groups (Smedley, B. D., et al., 2004). For instance, gender and racial disparities have been observed in the diagnosis of conditions involving pain assessment (Green, C. R. et al., 2003).

By recognizing and addressing measurement bias, we can work towards creating more equitable ML systems that better represent the constructs they aim to predict.

A representative example of measurement bias is apparent in healthcare systems, specifically in the context of predictive models for patient readmission. These models often use proxies for health status, such as previous healthcare utilization, to predict future healthcare needs.

One such study, conducted by Obermeyer et al. (2019), titled "Dissecting racial bias in an algorithm used to manage the health of populations", examined a commercial algorithm used in healthcare that predicts which patients will benefit from extra care management resources. The algorithm predicts future healthcare utilization using past healthcare costs as a proxy for health needs.

The study found that this algorithm displayed a significant bias against Black patients. Despite equal levels of risk as determined by the algorithm, Black patients had more chronic illnesses and worse health outcomes than White patients. The key issue was that the algorithm used past healthcare costs as a proxy for health needs, overlooking the fact that Black patients generally have lower healthcare costs due to systemic biases, including reduced access to care. As a result, the algorithm under-predicted the health needs of Black patients, contributing to disparities in healthcare provision.

This example demonstrates how measurement bias can result in unfair outcomes, even in widely used, seemingly objective algorithms. It underscores the importance of carefully selecting and validating proxies to ensure they accurately represent the intended construct across all groups.

2-2 ((Harini, S. et al., 2021) shows the following biases which are also described in detail below.

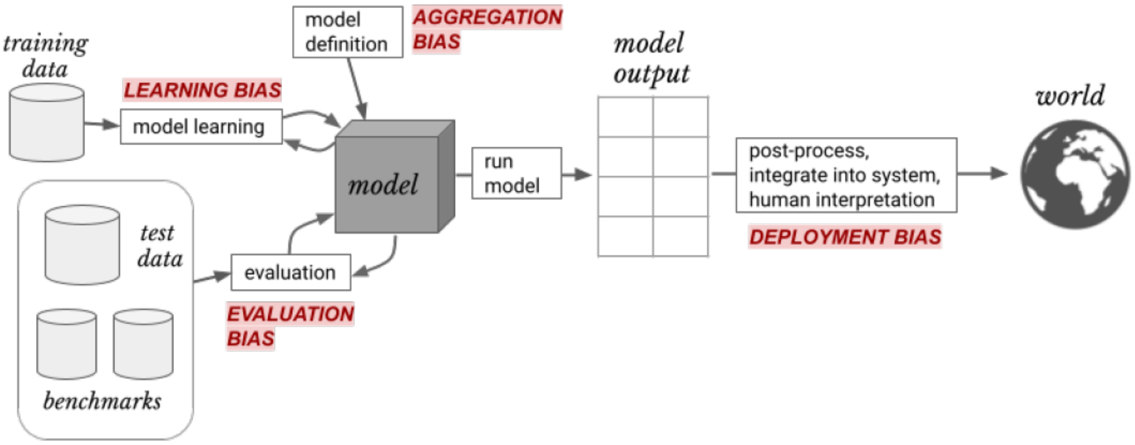


Figure 2-2: Bias Sources in the Model Building and Implementation Phases

Aggregation bias manifests when a uniform model is applied to a dataset where there exist distinct groups or types of examples that necessitate differential treatment. It arises from the erroneous assumption that the relationship between inputs and outputs is consistent across all subsets of data. In reality, this is often not the case, as individual datasets might represent diverse groups with differing backgrounds, cultures, or norms, thereby meaning that a given variable can have different implications across these groups. Aggregation bias can result in a model that is suboptimal for all groups, or a model that caters only to the dominant population, especially when representation bias is also present.

A classic example of aggregation bias is found in the field of social media analysis. Patton et al. (2017) conducted an analysis of Twitter posts by gang-involved youth in Chicago. Recognizing that a standard language model might not accurately interpret the nuances and subcultures in these tweets, they employed domain experts from the local community to interpret and annotate the tweets.

This strategy enabled them to identify several limitations of non-context-specific Natural Language Processing (NLP) tools. For instance, specific emojis or hashtags held particular meanings that a generic model trained on a larger Twitter corpus would likely miss. Similarly, words or phrases that may denote aggression in a different context were, in some cases, lyrics from a local rapper.

Neglecting this group-specific context in favor of a more general model designed for all social media data would likely result in harmful misclassifications of tweets from this population, potentially reinforcing stereotypes or misunderstanding critical communications within the community.

This example underscores the importance of recognizing and addressing aggregation bias in machine learning models to ensure accurate and fair outcomes across diverse groups.

Learning or algorithmic bias occurs when certain choices made during the modeling process exacerbate disparities in performance across various subsets of the data. A key aspect of this bias involves the selection of the objective function, which an ML algorithm seeks to optimize during training. However, certain problems may arise if the prioritization of one objective (such as overall accuracy) detrimentally impacts another (such as disparate impact).

A practical example of learning bias can be seen in the context of optimizing machine learning models for privacy or compactness. An instance of this is presented in a study by Bagdasaryan, E. et al. (2019), which investigated the implications of training models that preserve differential privacy. Differential privacy aims to prevent models from inadvertently revealing excessive identifying information about training examples during their use.

However, the researchers found that while differentially private training did enhance privacy, it also diminished the influence of underrepresented data on the model. This in turn led to decreased performance on that data compared to a model trained without differential privacy. In other words, an algorithmic choice intended to protect privacy inadvertently introduced a bias against underrepresented data, leading to worse outcomes for those groups.

Another study by Hooker, S., et al. (2020) showed a similar issue with models optimized for compactness, for instance, using techniques like pruning. The study demonstrated that the prioritization of compact models can amplify performance disparities on data with underrepresented attributes. This occurs because, given limited capacity, the model learns to retain information about the most prevalent features, often neglecting the less frequent, yet potentially important features.

These examples illustrate how learning or algorithmic bias can inadvertently produce models that perform poorly for certain groups or examples within the data, highlighting the importance of considering fairness and representation when making algorithmic decisions.

Evaluation bias arises when the benchmark data utilized to assess the performance of a model does not adequately represent the population in which the model is expected to be used. This form of bias operates on a larger scale than other bias types, as a misrepresentative benchmark can inadvertently encourage the development and deployment of models that perform well only on the data subset represented by the benchmark data.

The GLUE benchmark (General Language Understanding Evaluation) has been widely used in the evaluation of various NLP models, such as BERT, GPT-3, and RoBERTa, among others (Wang, A., et al., 2018). It includes a diverse set of resources that measure a model's ability to understand various aspects of language, such as sentiment analysis, question answering, and textual entailment.

However, the GLUE benchmark has been critiqued for not fully representing the diversity and complexity of natural language use. In a paper by Jia R. et al. (2017), they demonstrated how models that achieved high scores on standard evaluation sets, such as those included in the GLUE benchmark, were still prone to making errors when confronted with adversarial examples—modified inputs designed to induce model errors.

In this case, the evaluation bias arises because the GLUE benchmark, while diverse, might not fully capture the variety of language usage in real-world settings, and especially in adversarial situations. This leads to an overestimation of a model's capabilities when it comes to handling language understanding tasks, as the model might perform well on the benchmark but fail in more complex or unanticipated scenarios.

This situation suggests that it is important for evaluation datasets and metrics to be diverse and comprehensive enough to account for the complexity and variety of real-world applications. Careful consideration needs to be given to the ways in which the performance of ML models is evaluated to ensure that they can handle the breadth of scenarios they will encounter in practice.

Deployment bias refers to the mismatch between the problem the model is designed to solve and the manner in which it is actually used in the real world. Deployment bias often surfaces when a model, built and validated under certain assumptions, operates within a complex sociotechnical system influenced by institutional structures and human decision-makers. This bias can lead to unintended harmful consequences due to phenomena such as automation bias (over-reliance on automated decision-making) and confirmation bias (interpreting information in a way that confirms one's preexisting beliefs).

A pertinent example of deployment bias can be seen in the application of ML models in predictive policing. Ensign et al. (2018) discuss how deployment bias can manifest in this context. Predictive policing algorithms are designed to predict crime rates and help law enforcement allocate resources more effectively. However, these models can unintentionally perpetuate historical bias if the data they are trained on reflect past policing biases.

Furthermore, the use of these models can lead to a feedback loop: areas predicted to have high crime rates receive more police attention, which leads to more recorded crimes, which in turn strengthens the model's prediction of high crime in those areas. Despite the model's accuracy in replicating historical data, its deployment can reinforce existing biases and inequalities in policing practices. This highlights the risk and complexity of deploying machine learning systems in real-world settings, and the need for careful consideration of broader social and institutional contexts.

2.2 Building Trustworthy AI: Principles and Approaches

This section presents a comprehensive exploration of the core concepts, standards, and techniques that underpin the development and evaluation of trustworthy AI systems. The focal point of this section is the multidimensional construct of trustworthy AI, which is characterized by attributes of transparency, fairness, and robustness. These dimensions are further elaborated in Subsection 2.2.1, underscoring their critical role in fostering trust and reliability in AI systems.

Subsequent subsections delve into specific strategies and tools employed in the pursuit of trustworthiness. Subsection 2.2.3 introduces fairness metrics, sometimes also referred to as statistical fairness, which are quantitative measures that provide a means to evaluate and monitor the equity of AI systems. This framework serves as an essential reference point in identifying and mitigating biases in AI outputs.

Subsection 2.2.4 turns attention towards Explainable AI (XAI), a burgeoning field within AI research that addresses the transparency aspect of trustworthy AI. By enabling human users to understand, interpret, and consequently trust AI systems' decisions, XAI contributes significantly to the trustworthiness of these systems.

This section underscores that building trustworthy AI is not a singular, linear process but rather a multifaceted endeavor requiring attention to various principles and the application of specific approaches. It demonstrates that trustworthiness is achieved through a combination of transparency, fairness, robustness, quantitative fairness evaluation, and explainability of the AI systems.

2.2.1 Defining Trustworthy AI: Transparency, Fairness & Robustness

Trust, an intangible yet critical component of human relationships, finds its roots in a multitude of disciplines including psychology, sociology, economics, organizational management, and philosophy. Mayer, R. et al. (1995) succinctly define trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party". This definition elegantly captures the essence of trust in the context of human-machine interaction, providing a foundation for operationalizing trust in ML.

However, it is crucial to differentiate between 'trusted' and 'trustworthy' systems. Trustworthiness refers to the inherent properties of the system that make it deserving of trust. Still, a system being 'trustworthy' does not imply it is 'trusted'. The act of trusting, subject to cognitive biases and other factors, is a decision made by the trustor, which may or may not align with the system's trustworthiness (Hardin, R., 2001).

Trust and trustworthiness encompass a wide array of attributes in both human and technological contexts. For individuals, these traits include availability, competence, consistency, discreetness, fairness, integrity, loyalty, openness, promise fulfilment, and receptivity, among others (Dietz, G, et al., 2006). Information systems, on the other hand, may be considered trustworthy based on characteristics like correctness, privacy, reliability, safety, security, and survivability (Schneider, F., 1998).

In the realm of ML, trustworthiness may encompass diverse topics such as interpretability, adversarial examples, causality, fairness, privacy-preserving statistics, and robust statistics (Conference on Machine Learning, 2019). The European Commission's High-Level Expert Group on Artificial Intelligence (2019) has also identified lawful, ethical, and robust (both technically and socially) characteristics as being fundamental to trustworthy AI systems.

However, given the broad and disparate nature of these attributes, they are best viewed as rough guidelines. By distilling these characteristics, a set of distinct sub-domains can be identified that form a framework for trustworthiness, each of which can be examined in isolation.

- **Basic Performance:** For ML, competence can be equated to basic performance, such as the accuracy of a model. Effective performance, quantified based on the specifics of the problem and application, is a necessity for any real-world task.
- **Reliability:** This includes the safety, security, and fairness of ML models and systems. ML systems need to maintain good and correct performance across varying operating conditions. Different conditions could come from natural changes in the world or from human-induced changes.
- **Human Interaction:** This encompasses aspects of openness and human interaction with the ML system. It includes communication from the machine to the human through comprehensibility of models by people as well as transparency into overall ML system pipelines and lifecycles. It also includes communication from the human to the machine to supply personal and societal desires and values.
- **Aligned Purpose:** The alignment of the ML system's purpose with societal wants. The creation and development of ML systems is not independent of its creators. It is crucial for ML development to be intertwined with matters of societal concern and applications for social good, especially if the most vulnerable members of society are empowered to use ML to meet their own goals.

Given the complexity of these attributes, they may have entangled interrelationships, with some being trade-offs while others are not. Policymakers must reason about these relationships to decide a system's intended operations.

In sum, a trustworthy ML system can be defined as one that demonstrates sufficient basic performance, reliability, meaningful human interaction, and alignment with societal purposes. The focus should be on making ML systems worthy of trust rather than pursuing other means of making them trusted (Varshney, K. 2022).

The concept of trustworthiness as defined above presents a springboard towards understanding three integral attributes of trustworthy AI systems: robustness, fairness, and transparency. All three concepts are repeatedly referred to in this research, and therefore, the following explanations are meant as a first conceptual overview.

Robustness

Robustness embodies the reliability aspect of trustworthiness. A robust AI system consistently maintains its high-quality performance across various operating conditions, showing resilience against alterations in the environment or manipulations aimed at misleading the system (Tsipras, D. et al., 2018).

> **Distributional robustness** pertains to an AI system's resilience when the input data deviates from the training distribution. This type of robustness is critical as real-world data often vary due to changing environmental conditions or population dynamics. For instance, a facial recognition system trained on a particular demographic should be robust enough to accurately identify faces from a diverse range of demographics.
>
> **Adversarial robustness** involves an AI system's capability to withstand adversarial attacks that aim to manipulate its outputs by introducing carefully crafted perturbations to the input data. A striking example of the importance of adversarial robustness is found in autonomous driving, where a minor alteration to a stop sign (such as applying a sticker) can lead an AI model to misclassify it, leading to potentially severe consequences.
>
> **Model robustness** is related to the AI model's ability to provide reliable performance across different model configurations or settings. For instance, if a small change in a model's parameters leads to significantly different results, the model may lack robustness.
>
> Lastly, there's **robustness to concept drift**, which is the capability of an AI system to adapt to changes in the underlying concept or relationship between the input features and the target variable over time. This form of robustness is essential in dynamic environments where data distributions change over time, such as in predicting customer behaviour or market trends.

Fairness

Fairness, deeply connected with societal alignment, implies that AI systems should operate in a way that respects the norms and values of the society they serve. Fair AI ensures equitable treatment for all users and avoids discriminatory biases, thereby promoting social justice and inclusivity.

Algorithmic fairness, while embedded in the technical realm, extends into the broader sphere of social justice, making it a complex and often contentious aspect of AI systems. This topic is intertwined with concepts of justice, encompassing:

- **distributive justice** (equality in outcomes),
- **procedural justice** (sameness in decision-making processes),
- **restorative justice** (reparation of harm), and
- **retributive justice** (punishment of wrongdoers), (Rawls, J., 1971).

Fairness is inherently political, often reflecting power imbalances within society. As AI systems are employed in tasks that involve allocation of resources or decision-making, issues of fairness arise. It is generally accepted that AI systems can be discriminatory in their allocation, for example, in the distribution of health care resources to the more chronically ill patients. However, it becomes problematic when this allocation systematically privileges certain groups and disadvantages others. Such privilege is defined by groups that have historically been more likely to receive favorable outcomes in ML tasks, such as employment opportunities, loan approvals, and health care services (Dwork, C. et al, 2012).

Protected attributes such as ethnicity, gender, religion, and age often delineate privileged and unprivileged groups. These attributes are not universally fixed but are context-dependent and influenced by laws, regulations, and policies within specific jurisdictions and domains (Zliobaite, I., 2015).

Fairness in AI systems is typically categorized into two primary types: **group fairness** and **individual fairness**.

- **Group fairness** refers to the requirement for average classifier behavior to be the same across groups defined by protected attributes.
- **Individual fairness** stipulates that individuals with similar features should receive similar model predictions, with a special case being counterfactual fairness, where individuals differing only in one protected attribute should be treated the same (Kusner, M., et al., 2017).

Exploring the principles and approaches to building trustworthy AI systems, it is crucial to take into account both forms of fairness, despite the emphasis often placed on group fairness due to regulatory mandates. Balancing these two forms of fairness can guide the creation of more robust, equitable, and trustworthy AI systems.

Transparency

Transparency is closely related to the attribute of human interaction in trustworthy AI. Transparent AI systems provide clarity about their decision-making processes, helping humans understand and interpret their actions. This transparency fosters open communication between human operators and AI systems, leading to improved trustworthiness.

The principle of transparency in AI systems is a pivotal aspect of trustworthiness, providing a mechanism through which the decisions made by these systems can be understood, examined, and audited. The inclusion of transparency measures in AI systems can help mitigate potential harms and biases, while ensuring that these systems are accountable and reliable.

- **Factsheets**, inspired by the idea of nutrition labels on food products, offer one approach to improving transparency. They provide a standardized summary of a model's key characteristics, performance metrics, training data, and potential biases, among other relevant information (Hind, M. et al, 2018).
- **Quantitative testing** is a vital part of transparency, offering concrete and measurable insights into the system's performance across a variety of dimensions. This may involve testing for specific dimensions of trustworthiness such as fairness, reliability, and robustness, and evaluating performance on these dimensions under different conditions and scenarios (Doshi-Velez, F. et al., 2017).
- **Generating and testing edge cases**, the situations that test the limits of the system's capabilities, are also an integral part of transparency. They provide insights into how the system performs under extreme or uncommon circumstances and help identify potential weaknesses or points of failure.
- **Uncertainty quantification**, the process of estimating the likely errors or variances in AI predictions, plays a crucial role in transparency. It allows users to understand the confidence level of the system's outputs and adjust their trust in the system accordingly (Guo, C. et al., 2017).

- **Effective communication** of test results and uncertainty is paramount to transparency. This involves providing clear, understandable explanations of the system's performance and potential errors to the end-users, thus enabling them to make informed decisions based on the system's outputs.
- Lastly, **maintaining provenance**, the record of the system's development process including data sources, model specifications, and performance evaluations, is another important aspect of transparency. This documentation provides a comprehensive overview of the system's creation and validation, offering a transparent record of its life cycle and quality control measures (Varshney, K., 2022).

In the following sections, we delve into a more detailed exploration of these key attributes and their critical roles in building trustworthy AI systems.

2.2.2 The Importance of Trustworthiness in AI Systems

The trustworthiness of AI systems, particularly in terms of fairness, robustness, and transparency, holds significant bearing on broader societal issues. When these attributes are neglected, AI can inadvertently amplify existing human biases, exclude marginalized groups, unfairly punish individuals, and distort economic systems, among other impacts.

Three case studies are presented in section 5 in detail, however, the following three examples already illustrate the deep implications of neglecting trustworthy AI systems.

1) **Bias in Word Associations:** A study by Princeton University researchers demonstrated how AI systems could inadvertently propagate and amplify existing societal biases (Caliskan, A. et al., 2016). The team used a ML model to analyze and link 2.2 million words, revealing that the system associated European names with more positive attributes compared to African-American names. It also associated women and girls more with arts, while men and boys were more linked to science and math. Such biases, if used in applications like search engine ranking algorithms or auto-complete tools, could perpetuate and reinforce racial and gender biases in digital systems.
2) **Bias in Online Ads:** Harvard researcher Latanya Sweeney found that online search queries for African-American names were more likely to return ads from services offering arrest records compared to searches for white names (Sweeney, L., 2013). This bias extended to the targeting of financial products, with African-Americans offered higher-interest credit cards, despite comparable financial backgrounds to whites. Such disparities illustrate the potential unfairness that can arise in AI systems when biases in data or algorithms are not adequately addressed.
3) **Bias in Facial Recognition Technology:** The robustness of AI systems is crucial for their effectiveness, as illustrated by the research of MIT researcher Joy Buolamwini (2018). She found that commercially available facial recognition systems failed to accurately recognize darker-skinned complexions. The error rates for darker-skinned women reached up to 34 percent, compared to less than one percent overall. Such inaccuracies reflect the need for robustness in AI, ensuring that systems perform effectively across varied conditions and inputs, and do not disproportionately disadvantage certain groups.

The exacerbation of historic human bias, as witnessed in the AI word-association and online ad examples, can compound societal inequalities. Biased AI can propagate stereotypes, limit opportunities, and exacerbate exclusion of marginalized groups. Furthermore, unfair punishment may occur when biased AI systems are used in sensitive areas like law enforcement or judicial systems, potentially affecting individuals' livelihoods and freedoms based on flawed or biased data.

In terms of economic inequality, AI systems that lack robustness could lead to labor market distortions. For example, an AI system used in hiring processes that is not robust to variations in data could consistently favor certain demographics over others, contributing to unemployment or underemployment among certain populations.

Regarding privacy concerns, a lack of transparency in AI systems can lead to violations of privacy and unchecked data collection. Without transparency, users may be unaware of what data is collected, how it is used, and what measures are taken to protect it. This can lead to misuse of personal data, erosion of privacy, and potential exploitation.

Lack of transparency and fairness in AI design can also result in accessibility and usability issues. If an AI system is not transparent, users may not understand how it works, making it difficult for them to use it effectively or trust its outputs. Similarly, if an AI system is not designed with fairness in mind, it might not be usable by all individuals equally, creating accessibility barriers.

In summary, lack of trustworthiness in AI, characterized by deficits in fairness, robustness, and transparency, can lead to broad and significant societal repercussions. Achieving trustworthiness in AI is therefore not just about optimizing AI systems; it's about ensuring these systems contribute positively to society rather than leading to dystopian outcomes.

2.2.3 Statistical Fairness Metrics

Statistical fairness metrics serve as initial quantifiable approaches to assessing and mitigating biases within AI systems. These metrics, grounded in statistical theory, can provide useful insights into potential disparities across different demographics or groups, offering a method to discern the initial degree of fairness of an algorithm. However, it is crucial to recognize the inherent limitations of such metrics. Statistical fairness metrics, while valuable, offer only a simplified, reductionist view of the complex, multidimensional nature of fairness. Fairness is an intricate socio-technical construct, and its realization cannot be solely determined through mathematical formulations. Consequently, while statistical fairness metrics can contribute to the broader discourse on algorithmic fairness, they are not panaceas and should be contextualized within a more comprehensive approach that includes ethical, legal, and social considerations.

There are numerous fairness metrics used in ML and AI. Here, an overview of some of the most common metrics is provided, along with their explanations. This list is not exhaustive, as new fairness metrics continue to be developed and explored, but it gives a comprehensive introduction to the most common metrics which can be found in current academic literature.

The metrics are described and categorized into broader groups based on their focus, and a selection of the most used ones is presented, including examples and use cases within their respective contexts. This categorization is rather based on technical characteristics and serves as an overview, however, section *3.3 Post-Processing: Evaluation Metrics & Outcome Manipulation* introduces a combination of different criteria that are common in the academic AI fairness literature.

The following variables are used throughout the formulas below:

$\boldsymbol{TP}$: True positives - the number of instances correctly classified as positive.
$\boldsymbol{TN}$: True negatives - the number of instances correctly classified as negative.
$\boldsymbol{FP}$: False positives - the number of instances incorrectly classified as positive.
$\boldsymbol{FN}$: False negatives - the number of instances incorrectly classified as negative.

As expressed in the well-known confusion matrix:

		Predicted Condition	
	Total Population Positive + Negative	Positive (PP)	Negative (PN)
Actual Cond.	Positive (P)	$\boldsymbol{TP}$ True positives	$\boldsymbol{FN}$ False negatives
	Negative (N)	$\boldsymbol{FP}$ False positives	$\boldsymbol{TN}$ True negatives

$\boldsymbol{A}$: Protected attribute - a binary variable indicating membership in a protected demographic group (e.g., gender, race, or age group), with A = p (privileged group) or A = u (unprivileged group)
$\boldsymbol{Y}$: Target variable - a binary variable representing the true class label (1 = positive or 0 = negative) for an instance.
$\boldsymbol{\hat{Y}}$: Prediction variable - a binary variable representing the predicted class label (positive or negative) for an instance.

A. Metrics based on differences and ratios:

These metrics typically compare the difference or ratio of a specific outcome or error rate between different demographic groups. The difference usually ranges from -1 to 1 (inequality in both extremes), zero representing equality, and the ratios from 1 (equality) to infinity (the greater the ratio the greater the inequality), sometimes zero by swapping numerator and denominator, depending on the author's definition of the ratio. The most common metrics include:

A.1 Statistical/Demographic Parity Difference (SPD)

Difference in the probability of a positive outcome between two groups concerning a protected attribute (e.g., race).

$$\mathrm{SPD} = \mathrm{P}(\hat{\mathrm{Y}} = 1 | \mathrm{A} = \mathrm{p}) - \mathrm{P}(\hat{\mathrm{Y}} = 1 | \mathrm{A} = \mathrm{u})$$

Equation 2-1: Statistical/Demographic Parity Difference (SPD)

Where:
$P(\hat{Y} = 1 | A = p)$ is the probability of a positive outcome given that the individual belongs to the 'privileged' group,

$P(\hat{Y}=1|A=u)$ is the probability of a positive outcome given that the individual belongs to the 'unprivileged' group,
$\hat{Y}=1$ represents a favorable outcome,
$A=p$ indicates that the individual is from the privileged group,
$A=u$ indicates that the individual is from the unprivileged group.

The SPD is a **deterministic metric** that quantifies the difference in the probabilities of a positive outcome for the privileged and unprivileged groups. The **range** of the SPD is from -1 to 1, where 0 is the best value indicating perfect fairness (equal probabilities for both groups), -1 indicates maximum unfairness in favor of the unprivileged group, and 1 indicates maximum unfairness in favor of the privileged group. **Threshold** values can vary depending on the acceptable level of fairness for a given context. In many cases, absolute values close to 0 would be considered good as it indicates fairness.

It can also be expressed as (in-)equality between the two positive outcomes and is then referred to as statistical parity or demographic parity. Demographic parity is achieved when the probability of receiving a positive outcome is equal for all groups, irrespective of their protected attributes.

$$\mathrm{P}(\hat{Y}=1|A=p)=\mathrm{P}(\hat{Y}=1|A=u)$$

Example: In a hiring process, a company uses an AI system to screen job applicants. Demographic parity is achieved if the proportion of applicants from different gender groups who are selected for interviews is the same.

Context use case: It is used where the goal is to ensure equal representation of different groups in a positive outcome, such as hiring, college admissions, or loan approvals. However, it may not account for differences in qualifications or risk profiles. It measures the difference in selection rates across groups. This metric is appropriate when the goal is to ensure that the percentage of individuals selected for a particular outcome is roughly the same across different demographic groups (e.g., race, gender, age).

A.2 Demographic Parity Ratio (DPR): Ratio of the probability of a positive outcome between two groups which measures the ratio of selection rates across groups. This metric is similar to A.1 SPD, but accounts for differences in the base rate of the outcome across groups.

$$\mathrm{DPR}=\frac{\mathrm{p}(\hat{\mathrm{Y}}=1|\mathrm{A}=\mathrm{u})}{\mathrm{p}(\hat{\mathrm{Y}}=1|\mathrm{A}=\mathrm{p})}$$

Equation 2-2: Demographic Parity Ratio (DPR):

Where:

$P(\hat{Y}=1|A=u)$ is the probability of a positive outcome given that the individual belongs to the 'unprivileged' group,

$P(\hat{Y}=1|A=p)$ is the probability of a positive outcome given that the individual belongs to the 'privileged' group,

$\hat{Y}=1$ represents a favorable outcome,

$A = p$ indicates that the individual is from the privileged group,

$A = u$ indicates that the individual is from the unprivileged group.

The range of the DPR is from 0 to infinity, where 1 is the best value indicating perfect fairness (equal probabilities for both groups), values below 1 indicate a bias in favor of the privileged group, and values above 1 indicate a bias in favor of the unprivileged group. However, what is considered an acceptable deviation from 1 can depend on the specific context and the degree of fairness required.

Example: A healthcare provider uses a ML model to predict whether a patient should be recommended for a particular treatment ($\hat{Y}$=1) or not ($\hat{Y}$=0). The provider wants to ensure that the model is not biased against any demographic group (e.g., race, gender, age). Suppose the privileged group is patients with private insurance, and the unprivileged group is patients with public insurance. The healthcare provider can calculate DPR to evaluate whether both groups have a similar likelihood of being recommended for the treatment. A DPR close to 1 would suggest that the model treats both groups fairly.

A.3 Equalized Odds Difference (EOD): Difference in true positive rates and false positive rates between two groups. Equalized odds require that the true positive rate (TPR) and the false positive rate (FPR) are equal for all groups. In other words, the probability of a positive outcome given the true label and the protected attribute should be equal for all groups.

$$\mathrm{P}\big(\hat{Y} = 1\big|Y = 1, A = p\big) = \mathrm{P}\big(\hat{Y} = 1\big|Y = 1, A = u\big)$$

$$\mathrm{P}\big(\hat{Y} = 1\big|Y = 0, A = p\big) = \mathrm{P}\big(\hat{Y} = 1\big|Y = 0, A = u\big)$$

Or alternatively, as difference:

$$\begin{aligned}&\mathrm{EOD}\\ &= \big|P\big(\hat{Y} = 1\big|Y = 1, A = p\big) - P\big(\hat{Y} = 1\big|Y = 1, A = u\big)\big|\\ &- \big|P\big(\hat{Y} = 1\big|Y = 0, A = p\big) - P\big(\hat{Y} = 1\big|Y = 0, A = u\big)\big|\end{aligned}$$

Equation 2-3: Equalized Odds Difference (EOD)

Where:

$P\big(\hat{Y} = 1\big|Y = 1, A = p\big)$ is the probability of a TP for the privileged group.

$P\big(\hat{Y} = 1\big|Y = 1, A = u\big)$ is the probability of a TP for the unprivileged group.

$P\big(\hat{Y} = 1\big|Y = 0, A = p\big)$ is the probability of a FP for the privileged group.

$P\big(\hat{Y} = 1\big|Y = 0, A = u\big)$ is the probability of a FP for the unprivileged group.

The range of the EOD is from -2 to 2, where 0 is the best value indicating perfect fairness (equal odds for both groups). Sometimes only the TPR or the FPR are compared in which case the range is from -1 to 1. Positive values indicate bias in favor of the privileged group, and negative values indicate bias in favor of the unprivileged group. What is considered an acceptable deviation from 0 can depend on the specific context and the degree of fairness required.

Example: A healthcare provider uses an AI system to predict the likelihood of patients developing a certain medical condition. Equalized odds are achieved if the true positive rate and false positive rate of the model's predictions are equal for different racial groups.

Context use case: It is used where it is essential to minimize both false positives and false negatives for all groups, such as medical diagnoses, fraud detection, or criminal risk assessments.

A.4 Equalized Odds Ratio (EOR): A fairness metric that compares the ratio of true positive rates (TPR) to false positive rates (FPR) between privileged and unprivileged groups. A value of EOR close to 1 indicates that the classifier's performance is similar across both groups, which implies fairness.

$$EOR = \frac{p(\hat{Y} = 1 | Y = 1, A = u)/p(\hat{Y} = 1 | Y = 0, A = u)}{p(\hat{Y} = 1 | Y = 1, A = p)/p(\hat{Y} = 1 | Y = 0, A = p)}$$

Equation 2-4: Equalized Odds Ratio (EOR)

Where the different probabilities are the same as in the previous equation for EOD. The range of EOR is from 0 to ∞. A value of 1 is the best, indicating equal odds for both groups (perfect fairness). A value less than 1 indicates bias towards the unprivileged group, while a value greater than 1 indicates bias towards the privileged group. The degree of fairness required depends once again on the specific context.

Example: In criminal recidivism prediction, the criminal justice system can assess whether the model's performance, in terms of both true positive rates (correctly identifying individuals who will reoffend) and false positive rates (incorrectly identifying individuals as reoffenders when they will not reoffend), is similar across both racial groups. An EOR close to 1 would indicate that the model treats both white (privileged) and minority individuals fairly, without bias. A biased model could disproportionately affect minority individuals by over-predicting their likelihood of reoffending, leading to unfair treatment in parole decisions, sentencing, or rehabilitation programs.

A.5 Treatment Equality Difference (TED): Measures the difference in positive predictive value across groups.

$$TED = p(Y = 1 | \hat{Y} = 1, A = p) - p(Y = 1 | \hat{Y} = 1, A = u)$$

Equation 2-5: Treatment Equality Difference (TED)

Where:

$p(Y = 1 | \hat{Y} = 1, A = u)$ is the Positive Predictive Value (PPV), or precision, for the unprivileged group.

$p(Y = 1 | \hat{Y} = 1, A = p)$ is the PPV, or precision, for the privileged group.

The range of TED is from -1 to 1. A value of 0 is the best, indicating equal PPVs for both groups (perfect fairness). A value less than 0 indicates higher PPV for the unprivileged group, while a value greater than 0 indicates higher PPV for the privileged group. Again, the desired degree of fairness depends on the context.

Example: A financial institution uses a ML model to predict whether a transaction is fraudulent (Y=1) or legitimate (Y=0). The institution wants to ensure that the model is not biased against customers from specific demographic groups (e.g., race, age, gender). Using TED, the institution can evaluate whether the model is equally accurate in detecting fraudulent transactions for both privileged and unprivileged groups. A low TED value would indicate that the model is treating both groups fairly and accurately identifying fraudulent transactions.

Context use case: This metric is appropriate when the goal is to ensure that the probability of a positive prediction being correct is roughly the same across different demographic groups.

A.6 Treatment Equality Ratio (TER): Measures the ratio of positive predictive value across groups. This metric is similar to A.5 TED, but accounts for differences in the base rate of the outcome across groups.

$$TER = \frac{p\left(Y = 1 \middle| \hat{Y} = 1, A = u\right)}{p\left(Y = 1 \middle| \hat{Y} = 1, A = p\right)}$$

Equation 2-6: Treatment Equality Ratio (TER)

Where the probabilities of the PPV are the same as in the previous example for TED. TER measures the ratio of positive true outcomes (Y=1) between unprivileged and privileged groups, given that the classifier predicted a positive outcome ($\hat{Y}$=1). The ideal value of TER is 1, which indicates that both privileged and unprivileged groups have the same PPV. A value less than 1 indicates a higher PPV for the privileged group, and a value greater than 1 indicates a higher PPV for the unprivileged group.

A.7 Predictive Equality Difference (PED): Measures the difference in positive predictive value and false discovery rate across groups.

$$\begin{aligned} &PED \\ &= \left|p\left(Y = 1 \middle| \hat{Y} = 1, A = p\right) - p\left(Y = 1 \middle| \hat{Y} = 1, A = u\right)\right| \\ &- \left|p\left(Y = 0 \middle| \hat{Y} = 1, A = p\right) - p\left(Y = 0 \middle| \hat{Y} = 1, A = u\right)\right| \end{aligned}$$

Equation 2-7: Predictive Equality Difference (PED)

Where the probabilities of the PPV are the same as before and:

$p\left(Y = 0 \middle| \hat{Y} = 1, A = u\right)$ is the False Discovery Rate (FDR) for the unprivileged group.

$p\left(Y = 0 \middle| \hat{Y} = 1, A = p\right)$ is the FDR for the privileged group.

The ideal value of PED is 0, which indicates equal PPV and equal FDR between the privileged and unprivileged groups. Positive values indicate bias favoring the privileged group, while negative values indicate bias favoring the unprivileged group. However, mostly both PPV and FDR are compared separately across groups.

Context use case: This metric is appropriate when the goal is to ensure that the probability of a positive prediction being correct and the probability of a negative prediction being incorrect are roughly the same across different demographic groups.

Example: An insurance company uses a ML model to predict whether a customer is likely to file a claim (Y=1) or not (Y=0). The model takes into account various factors, such as the customer's age, driving history, location, and other relevant factors. The insurance company wants to ensure that the model is not biased against any demographic group (e.g., race, gender, age) based on historical data. Using PED, the insurance company can assess whether the model is treating customers from different demographic groups fairly when it comes to predicting the likelihood of filing a claim. A low PED value would suggest that the model is treating customers from different demographic groups fairly in terms of both PPV (correctly predicting claims) and FDR (correctly predicting no claims).

A.8 Predictive Equality Ratio (PER): Measures the ratio of positive predictive value and false discovery rate across groups. This metric is similar to A.7 PED, but accounts for differences in the base rate of the outcome across groups.

$$\text{FDR Ratio} = \frac{p\left(Y=0\middle|\hat{Y}=1, A=u\right)}{p\left(Y=0\middle|\hat{Y}=1, A=p\right)} \quad \text{PPV Ratio} = \frac{p\left(Y=1\middle|\hat{Y}=1, A=u\right)}{p\left(Y=1\middle|\hat{Y}=1, A=p\right)}$$

Equation 2-8: Predictive Equality Ratio (PER) via FDR and PPV Ratios

Where the PPV and FDR are the same as in the previous equation for PED. The ideal value of PER is 1, indicating equal PPVs and FDRs between the privileged and unprivileged groups. Values above 1 indicate bias favoring the privileged group, and values below 1 indicate bias favoring the unprivileged group.

Example: An online advertising platform uses a ML model to predict which users are likely to click on a specific ad (y=1) or not (y=0). The model considers various factors, such as user's browsing history, interests, demographics, and other relevant factors. The advertising platform wants to ensure that the model is not biased against any demographic group (e.g., race, gender, age) based on historical data. Using PER, the advertising platform can assess whether the model is treating users from different demographic groups fairly when it comes to predicting the likelihood of clicking on an ad. A PER value close to 1 would suggest that the model is treating users from different demographic groups fairly in terms of their PPV, meaning that the proportion of true positive predictions (clicks) among the predicted positive instances (ad impressions) is similar for both groups.

B. Aggregate metrics:

These metrics provide an overall assessment of fairness by considering multiple aspects of the model's performance:

B.1 Average Odds Difference (AOD): The Average Odds Difference measures fairness with respect to the separation criterion. It is the average of the differences in true positive rates (TPR) and false positive rates (FPR) between two groups (p - privileged and u - unprivileged) concerning a protected attribute (e.g., race).

$$AOD = \frac{1}{2}\left[\left(\frac{TP_p}{TP_p + FN_p} - \frac{TP_u}{TP_u + FN_u}\right) + \left(\frac{FP_p}{FP_p + TN_p} - \frac{FP_u}{FP_u + TN_u}\right)\right]$$

Equation 2-9: Average Odds Difference (AOD)

Where all variables are already explained in the introductory part of this section. The range of the AOD can be from -1 to 1. An AOD of 0 indicates perfect fairness. An AOD greater than 0 indicates a bias in favor of the privileged group, while an AOD less than 0 indicates a bias in favor of the unprivileged group.

Example: In a credit approval process, a financial institution uses a ML model to predict the likelihood of a loan applicant defaulting. The AOD metric is used to check whether the model is fair across different racial groups. A lower AOD value indicates that the model has a more similar performance across the groups in terms of both true positive rates and false positive rates.

Context use case: The AOD metric is useful in situations where the goal is to ensure that the model performs equally well for different groups with respect to both positive outcomes (correctly approving creditworthy individuals) and negative outcomes (correctly identifying individuals who are likely to default). It is related to the separation area of fairness, which focuses on whether the model's predictions are equally accurate across different groups. The AOD metric takes into account both true positive rates and false positive rates and provides a single value to assess fairness. This metric is particularly relevant in high-stakes decision-making contexts such as credit approval, medical diagnosis, and fraud detection, where both types of errors are important to consider.

B.2 Conditional Demographic Disparity (CDD): The Conditional Demographic Disparity measures fairness with respect to the separation criterion. It is the difference in demographic disparity between two groups (privileged and unprivileged) concerning a protected attribute (e.g., race) conditional on a specific outcome or error rate. In other words, CDD measures how the disparity between the groups changes when considering different outcomes or error rates.

$$CDD(Y = y, A = a) = P(\hat{Y} = y | A = p, Y = a) - P(\hat{Y} = y | A = u, Y = a)$$

Equation 2-10: Conditional Demographic Disparity (CDD)

Where:

$P(\hat{Y} = y | A = p, Y = a)$ is the probability of a certain predicted outcome $\hat{Y} = y$ given the actual outcome is $Y = a$ and the attribute is privileged (A=p).

$P(\hat{Y} = y | A = u, Y = a)$ is the probability of a certain predicted outcome $\hat{Y} = y$ given the actual outcome is $Y = a$ and the attribute is unprivileged (A=u).

$Y = y$ represents a specific outcome or error rate.

The range of the CDD can be from -1 to 1. A CDD of 0 indicates perfect fairness. A CDD greater than 0 indicates a bias in favor of the privileged group, while a CDD less than 0 indicates a bias in favor of the unprivileged group.

Example: In a hiring process, a company uses a ML model to screen job applicants. CDD can be used to evaluate the fairness of the model with respect to different demographic groups, considering both positive outcomes (invitations for interviews) and negative outcomes (rejections) for each group. By calculating CDD for different outcomes, the company can better understand how the model's performance varies across groups and make adjustments if needed.

Context use case: The CDD metric is useful when the goal is to understand how demographic disparities change depending on different outcomes or error rates. It is related to the separation area of fairness, which focuses on whether the model's predictions are equally accurate across different groups. The CDD metric helps in identifying potential sources of bias in the model's predictions and can be used to guide adjustments to the model to improve fairness. This metric is particularly relevant in contexts where the impact of different outcomes or error rates on different groups needs to be assessed, such as hiring, college admissions, or loan approvals.

C. Metrics based on probabilistic predictions:

These metrics assess the reliability and consistency of the model's predicted probabilities across different groups:

C.1 Calibration: Calibration can be used as a fairness metric by assessing it separately within each demographic group defined by the protected attributes. This is sometimes referred to as "group calibration" (Barocas, S. et al., 2019). When considering fairness, the principle is that the model's prediction probabilities should be well-calibrated not just over the entire population, but also when considering each demographic group individually. In other words, for any particular group defined by the protected attributes, when the model predicts an event with a probability of p, that event should happen about p percent of the time within that group.

$$E\left[\hat{Y}\middle|Y = y, A = a\right] = y$$

Equation 2-11: Calibration

Where:

$E[\quad]$ denotes the expectation,

$\hat{Y}$ is the predicted label,

Y is the true label,

$Y = y$ is a specific value of the label, $A = a$ a specific value of the attribute.

This formula expresses that, for each subgroup defined by the protected attribute and for each level of the predicted probability, the average true outcome should be equal to the predicted probability.

Violations of group calibration can indicate that the model is systematically over- or under-predicting certain outcomes for certain groups, which could be considered unfair.

It is worth noting that achieving perfect group calibration can be challenging or even impossible in practice, especially if the base rates of the outcome variable differ across groups or if the model's predictive performance varies across groups.

Example: A weather forecasting system uses an AI model to predict the probability of rain for different regions. Calibration is achieved if, for each region and predicted probability, the proportion of days with actual rain matches the predicted probability.

Context use case: Used in contexts where ensuring equal confidence in the model's predictions across different groups or regions is important, such as weather forecasting, financial risk assessments, or demand forecasting.

C.2 Probability Integral Transform (PIT): The Probability Integral Transform is a measure of calibration, which evaluates how well the predicted probabilities of a model are aligned with the actual observed probabilities. PIT compares the distribution of predicted probabilities to a uniform distribution. A well-calibrated model should have a PIT distribution that closely follows a uniform distribution.

PIT is defined for a model with predicted probabilities p_i for each instance i, and corresponding binary outcomes y_i:

$$PIT_i = \frac{1}{N}\sum_{j=1}^{N} I\left(p_j \leq p_i\right)$$

Equation 2-12: Probability Integral Transform (PIT)

Where:

N is the total number of instances,

p_i is the predicted probability of the positive class for instance i,

I is the indicator function, equal to 1 if $p_j \leq p_i$ and 0 otherwise.

PIT is a probabilistic metric where PIT values should follow a uniform distribution in $[0,1]$ for a well-calibrated model. The average value should be around 0.5 for a well-calibrated model. Deviation from 0.5 indicates a miscalibrated model. There are no specific thresholds, but the closer the PIT distribution is to a uniform distribution, the better. To interpret this metric, one needs to understand the idea of calibration in probability predictions and how it impacts model fairness. A plot of the PIT values (a PIT histogram) can be helpful in visually assessing model calibration. This metric is not specific to any protected attribute.

Context use case: The PIT metric is useful in situations where the goal is to assess the reliability of a model's predicted probabilities, ensuring that they are well-aligned with the actual observed probabilities. It is related to the calibration area of fairness, which focuses on the consistency of the model's predictions across different demographic groups. The PIT metric can be applied in various contexts, such as weather forecasting, disease prediction, customer churn prediction, or any other application where probabilistic predictions are important. In these cases, having well-calibrated probabilities is crucial for making informed decisions and managing risks.

C.3 Normalized Mutual Information (NMI): The Normalized Mutual Information is a measure of the mutual dependence between the predicted probabilities and the true outcomes, normalized by the entropy of the true outcomes. NMI evaluates the degree of shared information between the predicted probabilities and the actual outcomes, with higher NMI values indicating a stronger relationship between the predictions and the true outcomes.

NMI is defined as:

$$NMI = \frac{I(Y;\hat{Y})}{\sqrt{H(Y)H(\hat{Y})}}$$

Equation 2-13: Normalized Mutual Information (NMI)

Where $I(Y;\hat{Y})$ is the mutual information between the true outcomes Y and the predicted probabilities $\hat{Y}$ and $H(Y)$ and $H(\hat{Y})$ are the entropies of the true outcomes and predicted probabilities, respectively.

The NMI is also a probabilistic metric which **ranges** from 0 (no mutual information, worst) to 1 (complete mutual information, best). The average value indicates the degree of shared information between the predicted probabilities and the actual outcomes. A higher average value indicates a stronger relationship. There is no universal **threshold**, but closer to 1 is generally better. This metric is not specific to any protected attribute but evaluates the overall quality of the model's probabilistic predictions and is therefore often additionally used in a fairness context.

Example: The AI model is trained to predict whether a customer would default or not on a loan, and based on that, the bank decides to grant or deny a loan. The true outcomes Y are whether the customers actually defaulted or not and the predicted outcomes $\hat{Y}$ are the predictions made by the model. If the model is fair, it should make the same predictions for individuals who are alike except for the protected attribute (e.g., race, gender). In terms of mutual information, this means that the mutual information between Y and $\hat{Y}$ should not be affected by the protected attribute. Therefore, for a fair model, the NMI should be similar across all groups defined by the protected attribute.

Context use case: The NMI metric is useful in situations where the goal is to assess the strength of the relationship between a model's predicted probabilities and the true outcomes, and to determine how much information is shared between them. This metric can be applied in various contexts, such as medical diagnosis, customer churn prediction, or any other application where the quality of probabilistic predictions is important. In these cases, having a high NMI value can indicate that the model's predictions are more informative and better aligned with the true outcomes, which can lead to better decision-making and improved performance.

D. Metrics based on information theory and distance measures:

These metrics quantify the divergence or distance between the distributions of predicted outcomes or features for different demographic groups:

D.1 Generalized Entropy Index (GEI): The Generalized Entropy Index is a measure of inequality that quantifies the dispersion of predicted outcomes or features within and between different groups. GEI is particularly useful for evaluating fairness in the sufficiency area, as it helps assess the informativeness of the model's predictions for each group defined by the protected attribute.

The Generalized Entropy Index is defined as:

$$GEI(\alpha) = \frac{1}{n(\alpha-1)(1-\alpha)} \sum\nolimits_{i=1}^{n} \left[\left(\frac{p_i}{\mu} \right)^{\alpha-1} - 1 \right]$$

Equation 2-14: Generalized Entropy Index (GEI)

where n is the number of instances, p_i is the predicted outcome or feature value for instance i, μ is the mean of the predicted outcomes or features, and α is a parameter that controls the sensitivity of the index to differences in predicted outcomes or features. When $\alpha = 2$, the GEI becomes the widely-used **Theil index.**

This metric is not probabilistic, the range is $[0, +\infty)$ with the best value being 0 (indicating perfect equality) and the worst value being +∞ (indicating maximum inequality). The average value indicates the level of inequality in the predicted probabilities across instances. Threshold values are domain-specific and would depend on what level of inequality is considered acceptable. This metric is usually interpreted in the context of the specific application, the protected attribute(s), and the domain-specific knowledge of what constitutes acceptable levels of inequality.

Context use case: The GEI metric is useful when the goal is to assess the dispersion of predicted outcomes or features within and between different groups, particularly in the context of fairness and sufficiency. It is relevant in various applications, such as education, health care, or any other situation where understanding inequality and potential biases in predicted outcomes is crucial. By identifying disparities in the model's predictions, the GEI metric can help guide adjustments to the model to improve fairness and informativeness for different demographic groups.

D.2 Kullback-Leibler Divergence (KLD): The KLD is a measure of the divergence between the distributions of predicted outcomes or features for different demographic groups. KLD can be useful for evaluating fairness in the independence area, as it helps assess the discrepancy between the model's predictions for each group defined by the protected attribute.

KLD is defined as:

$$KLD(P \parallel Q) = \sum\nolimits_{i} P(i) \log \frac{P(i)}{Q(i)}$$

Equation 2-15: Kullback-Leibler Divergence (KLD)

where $P(i)$ is the probability distribution of the predicted outcomes or features for the privileged group, and $Q(i)$ is the probability distribution of the predicted outcomes or features for the unprivileged group.

This metric is not probabilistic. The range is $[0, +\infty)$, where 0 is the best value (indicating that P and Q are the same distribution), and $+\infty$ is the worst value (indicating that P and Q are completely different). The average value can be understood as the average difference between the two distributions in terms of information content. There are no fixed threshold values, as they would be domain-specific and depend on the acceptable level of divergence between the distributions. To interpret this metric, it's necessary to have context on the specific application, the protected attribute(s), and the understanding of what level of divergence between distributions for different groups is considered acceptable.

Example: In a recommendation system for online platforms, a model predicts the probability of users engaging with certain content. The KLD can be used to measure the divergence between the predicted engagement probabilities for different demographic groups (e.g., gender, race, age). By calculating KLD, platform developers can identify potential disparities in the predicted outcomes and address any biases in their recommendation algorithms.

Context use cases: The KLD metric is useful when the goal is to assess the discrepancy between the distributions of predicted outcomes or features for different demographic groups, particularly in the context of fairness and independence. It is relevant in various applications, such as recommendation systems, marketing, or any other situation where understanding the divergence in predicted outcomes for different demographic groups is crucial. By identifying disparities in the model's predictions, the KLD metric can help guide adjustments to the model to improve fairness and reduce biases for different demographic groups.

E. Metrics based on algorithmic fairness:

These metrics focus on the fairness properties of the algorithm itself, often by incorporating fairness constraints or assumptions into the model:

E.1 Counterfactual Fairness (CF): Counterfactual Fairness is a fairness metric that requires the model's prediction to remain the same if the protected attribute were counterfactually changed (not be confounded with attribute flipping). CF is particularly useful for evaluating fairness in the independence area, as it aims to ensure that the model's predictions are not influenced by the protected attribute.

While Counterfactual Fairness does not have a single formula, it is based on the idea of counterfactuals. A model is counterfactually fair if for any individual i and any values of the protected attribute $A = p$ and $A = u$:

$$P\left(\widehat{Y_i^p} = y \mid X_i, A_i = p\right) = P\left(\widehat{Y_i^u} = y \mid X_i, A_i = u\right)$$

Equation 2-16: Counterfactual Fairness (CF)

where $\widehat{Y_i^p}$ and $\widehat{Y_i^u}$ represent the counterfactual predictions for individual i if their protected attribute were p (privileged group) and u (unprivileged group), respectively, and X_i denotes the observed features for individual i. This metric is not probabilistic. The range is not applicable in the traditional sense because this metric does not produce a numerical value, but rather a boolean evaluation of whether the model is counterfactually fair or not. The best and worst values would be that the model is or is not counterfactually fair, respectively. There is no specific threshold for this metric, as it is a binary (fair/unfair) assessment. The context needed to interpret this metric includes understanding the nature of the protected attribute, the possible values it can take, and the counterfactual scenarios considered.

To implement CF in practice, one needs to have a causal graph that captures the relationships between variables, including the protected attribute, proxy variables, and the target variable. This causal graph can be used to compute counterfactual predictions under different values of the protected attribute while accounting for potential correlations with proxies. By doing so, CF helps to ensure that the model's predictions are not influenced by the protected attribute, either directly or indirectly through proxies.

Example: If a recidivism model (cf. case study 5.4) predicts that a black individual with a specific set of features (age, number of past offenses) is likely to reoffend, to test counterfactual fairness, a counterfactual world would be considered where this individual is white. In this counterfactual world, due to different systemic interactions, the number of past offenses might be lower for this individual. If this feature is adjusted accordingly in the counterfactual scenario, the model's prediction should ideally remain the same to satisfy counterfactual fairness.

Context use cases: The CF metric is useful when the goal is to ensure that a model's predictions are not influenced by the protected attributes, particularly in the context of fairness and independence. It is relevant in various applications, such as credit scoring, hiring, medical diagnosis, or any other situation where understanding and mitigating the impact of protected attributes on predicted outcomes is crucial. By adhering to the principle of Counterfactual Fairness, decision-makers can help reduce biases in their models and promote fair treatment of individuals across different demographic groups.

E.2. Fairness Through Unawareness (FTU): Fairness Through Unawareness is a naive approach to achieving fairness by removing the protected attributes from the model's input features. FTU is based on the idea that if a model is not provided with the protected attribute, it cannot discriminate based on that attribute. However, this approach often falls short in addressing fairness, as it fails to account for the potential correlations between the protected attribute and other input features, which can still lead to biased predictions.

FTU does not have a specific formula. Instead, it involves training a model on a modified dataset where the protected attribute, A, is removed from the input features:

$$\text{Model: } f(X_{-A}) \rightarrow \hat{Y}$$

Equation 2-17: Fairness Through Unawareness (FTU)

where X_{-A} represents the set of input features without the protected attribute, and $\hat{Y}$ is the predicted outcome.

Context use cases: While FTU is a simple and intuitive approach to fairness, it is often insufficient for effectively addressing biases in a model, particularly in cases where there are correlations between the protected attribute and other input features. It can be considered as a first step in fairness-aware modeling but should generally be supplemented with more advanced fairness techniques that account for correlations with proxies and indirect discrimination. In various applications such as hiring, credit scoring, or medical diagnosis, it is crucial to use more comprehensive fairness techniques to ensure fair treatment of individuals across different demographic groups.

E.3 Indirect Bias (Dwork et al, 2012): is based on the notion that a model is fair if its predictions are not influenced by the protected attribute, either directly or indirectly through other correlated features. This concept, also known as "fairness under composition," aims to ensure that biases do not accumulate or propagate through a sequence of decision-making processes that involve multiple models or stages.

Dwork's Indirect Bias does not have a specific formula but is rather a conceptual approach to fairness. It involves evaluating the influence of the protected attribute on the model's predictions, taking into account the correlations with other features, and examining the potential for indirect discrimination.

Context use cases: Dwork's Indirect Bias is relevant when considering fairness in complex, multi-stage decision-making processes, or when there are strong correlations between the protected attribute and other input features. This approach helps to ensure that biases do not accumulate or propagate through the decision-making process, leading to fair outcomes for individuals across different demographic groups. Applications may include college admissions, hiring pipelines, or any other situation involving multiple models or stages where fairness is a concern. To effectively address Dwork's Indirect Bias, it is essential to combine this approach with other fairness techniques that account for correlations with proxies and indirect discrimination.

E.4 Fairness Constraints (Hardt et al., 2016): A family of fairness constraints that can be incorporated into the learning process to ensure that the model's predictions satisfy specific fairness criteria, such as equalized odds or demographic parity. These constraints can be used to balance fairness and accuracy objectives during model training. Hardt's Fairness Constraints are a part of the independence area, as they aim to reduce the relationship between the protected attribute and the model's predictions.

The idea behind the Fairness Constraints is to explicitly incorporate fairness criteria into the optimization problem during the training process. The specific constraints depend on the fairness metric being considered. For example, for Demographic Parity (DP), the constraint would be:

$$\left|P\left(\hat{Y}=1\middle|A=p\right)-P\left(\hat{Y}=1\middle|A=u\right)\right|\leq\epsilon_1$$

Equation 2-18: Fairness Constraints

where $A = p$ and $A = u$ represent the privileged and unprivileged groups, respectively, and ϵ_1 is a predefined tolerance level for fairness violations.

For Equalized Odds (EO), the constraints would be:

$$\left|P(\hat{Y}=1|Y=1,A=p)-P(\hat{Y}=1|Y=1,A=u)\right|\leq\epsilon_1$$

And:

$$\left|P(\hat{Y}=1|Y=0,A=p)-P(\hat{Y}=1|Y=0,A=u)\right|\leq\epsilon_2$$

where ϵ_1 and ϵ_2 are predefined tolerance levels for fairness violations.

Context use cases: Fairness Constraints are useful when the goal is to balance fairness and accuracy during model training, particularly in cases where the protected attribute may be related to the model's predictions. By incorporating these constraints into the learning process, decision-makers can promote fair treatment of individuals across different demographic groups while maintaining good predictive performance. Applications may include hiring, credit scoring, medical diagnosis, or any other situation where fairness is a concern and the model's predictions should satisfy specific fairness criteria.

F. Metrics focused on specific error rates:

These metrics focus on the differences or parity in specific error rates between demographic groups:

F.1 False Negative Rate Difference (FNRD): The False Negative Rate Difference is a fairness metric that measures the difference in false negative rates between two demographic groups. It falls into the separation category, as it compares error rates across different groups without considering the relationship between the protected attribute and other input features.

$$\text{FNRD}=\left|\frac{FN_p}{FN_p+TP_p}-\frac{FN_u}{FN_u+TP_u}\right|$$

Equation 2-19: False Negative Rate Difference (FNRD)

where FN_p and FN_u represent the number of false negatives for the privileged and unprivileged groups, respectively, and TP_p and TP_u represent the number of true positives for the p - privileged and u-unprivileged groups, respectively.

Context use cases: The False Negative Rate Difference is relevant when it is crucial to monitor and minimize the differences in false negative rates between different demographic groups. It is particularly useful in situations where the consequences of false negatives are significant, such as medical diagnosis, fraud detection, or public safety. This metric helps ensure that the model's predictions do not disproportionately impact certain demographic groups, promoting fairness and reducing the potential for discrimination.

F.2 False Positive Rate Difference (FPRD): The False Positive Rate Difference is a fairness metric that measures the difference in false positive rates between two demographic groups. It falls into the separation category, as it compares error rates across different groups without considering the relationship between the protected attribute and other input features.

$$\text{FPRD} = \left|\frac{FP_p}{FP_p + TN_p} - \frac{FP_u}{FP_u + TN_u}\right|$$

Equation 2-20: False Positive Rate Difference (FPRD)

where FP_p and FP_u represent the number of false positives for the privileged and unprivileged groups, respectively, and TN_p and TN_u represent the number of true negatives for the privileged and unprivileged groups, respectively.

Context use cases: The False Positive Rate Difference is relevant when it is crucial to monitor and minimize the differences in false positive rates between different demographic groups. It is particularly useful in situations where the consequences of false positives are significant, such as credit scoring, hiring, or college admissions. This metric helps ensure that the model's predictions do not disproportionately impact certain demographic groups, promoting fairness and reducing the potential for discrimination.

F.3 Positive Predictive Parity (PPP) or Predictive Parity: The Positive Predictive Parity is a fairness metric that measures the difference in positive predictive values (PPV) between two demographic groups. It falls into the separation category, as it compares the predictive performance across different groups without considering the relationship between the protected attribute and other input features.

$$\text{PPP} = \left|\frac{TP_p}{TP_p + FP_p} - \frac{TP_u}{TP_u + FP_u}\right|$$

Equation 2-21: Positive Predictive Parity (PPP)

where TP_p and TP_u represent the number of true positives for the privileged and unprivileged groups, respectively, and FP_p and FP_u represent the number of false positives for the privileged and unprivileged groups, respectively.

Example use case: In a job application screening process, the Positive Predictive Parity can be used to ensure that the model's predictions do not disproportionately favor applicants from different demographic groups (e.g., gender or race) in terms of true positive predictions, which could lead to unfair hiring decisions or a biased workforce.

Similar metrics such as the Negative Predictive Parity can be easily derived from the previous explanations and are not listed here.

G. Metrics based on overall model performance:

These metrics focus on the parity in overall model performance between demographic groups:

G.1 Overall Accuracy Parity (OAP): The Overall Accuracy Parity is a fairness metric that measures the difference in overall accuracy rates between two demographic groups. It falls into the separation category, as it compares the overall performance across different groups without considering the relationship between the protected attribute and other input features.

$$\text{OAP} = \left| \frac{TP_p + TN_p}{TP_p + TN_p + FP_p + FN_p} - \frac{TP_u + TN_u}{TP_u + TN_u + FP_u + FN_u} \right|$$

Equation 2-22: Overall Accuracy Parity (OAP)

where TP_p and TP_u represent the number of true positives for the privileged and unprivileged groups, respectively; TN_p and TN_u represent the number of true negatives for the privileged and unprivileged groups, respectively; FP_p and FP_u represent the number of false positives for the privileged and unprivileged groups, respectively; and FN_p and FN_u represent the number of false negatives for the privileged and unprivileged groups, respectively.

Context use cases: The Overall Accuracy Parity is relevant when it is crucial to monitor and minimize the differences in overall accuracy rates between different demographic groups. It is useful in situations where the accuracy of the model is important across various contexts, such as recommendation systems, content moderation, or customer support. This metric helps ensure that the model's predictions do not disproportionately impact certain demographic groups, promoting fairness and reducing the potential for discrimination.

Table 2-2 summarizes the fairness metrics discussed in this section. A few clarifying comments should be made: Firstly, the categories of Treatment Equality Difference (TED) and Treatment Equality Ratio (TER) appear to be hardly used within the established literature on fairness metrics and require further elucidation; therefore, they do not appear in the table. Secondly, the characterization of Fairness Through Unawareness (FTU) as a metric could be misleading; it is more appropriately categorized as an approach that disregards sensitive attributes in the pursuit of fairness. However, it's important to note that this method does not inherently provide a measure of fairness, as it often fails to consider proxy variables correlated with sensitive attributes.

Thirdly, Counterfactual Fairness (CF) and Indirect Bias, while vital conceptual tools in the study of algorithmic fairness, are not typically represented as quantifiable metrics. Similarly, the fourth observation pertains to Fairness Constraints as proposed by Hardt et al., 2016. This is not a metric but rather an algorithmic approach that incorporates fairness constraints into the optimization process during model training. Thus, it may seem inappropriate to present it as a quantifiable metric, however, in combination with other metrics it is a widely used approach and has therefore been included in this section.

Table 2-2: Fairness Metrics Comparison

Metric Name	Category	Prob.	Range	Meaning of Average
Statistical/Demographic Parity Difference (SPD)	Ratio/ Difference	No	-1 to 1, 0 is best	Average difference in outcomes between groups
Demographic Parity Ratio (DPR)	Ratio/ Difference	No	0 to ∞, 1 is best	Average ratio of outcomes between groups
Equalized Odds Difference (EOD)	Ratio/ Difference	No	-1 to 1, 0 is best	Average difference in true positive rates and false positive rates between groups
Equalized Odds Ratio (EOR)	Ratio/ Difference	No	0 to ∞, 1 is best	Average ratio of true positive rates and false positive rates between groups
Predictive Equality Difference (PED)	Ratio/ Difference	No	-1 to 1, 0 is best	Difference in false positive rates between groups
Predictive Equality Ratio (PER)	Ratio/ Difference	No	0 to ∞, 1 is best	Ratio of false positive rates between groups
Average Odds Difference (AOD)	Aggregates	No	-1 to 1, 0 is best	Average of differences in false positive rates and true positive rates between groups
Conditional Demographic Disparity (CDD)	Aggregates	Yes	0 to ∞, 1 is best	Measure of disparity in outcomes given the condition
Calibration	Prob. Pred.	Yes	0 to 1, 1 is best	Measure of the agreement between predicted probabilities and actual outcomes
Probability Integral Transform (PIT)	Prob. Pred.	Yes	0 to 1, .5 is best	Measure of how much a distribution deviates from uniform distribution
Normalized Mutual Information (NMI)	Prob. Pred.	Yes	0 to 1, 1 is best	Measure of the mutual dependency between the predicted and actual outcomes
Generalized Entropy Index (GEI)	Distance	Yes	0 to ∞, 0 is best	Measure of inequality in outcomes
Kullback-Leibler Divergence (KLD)	Distance	Yes	0 to ∞, 0 is best	Measure of the information lost when the predicted probability distribution is used to approximate the true probability distribution
Counterfactual Fairness (CF)	Concept	N/A	N/A	Measure of fairness based on the idea that the outcome should be the same in the actual world and a counterfactual world
Fairness Through Unawareness (FTU)	Approach	N/A	N/A	Approach that ignores sensitive attributes to achieve fairness
Indirect Bias (Dwork et al, 2012)	Concept	N/A	N/A	Measure of fairness based on the correlation between sensitive attributes and outcomes
Fairness Constraints (Hardt et al., 2016)	Approach	N/A	N/A	Method of adding constraints to the optimization problem to enforce fairness
False Negative Rate Difference (FNRD)	Error Rates	No	0 to 1, 0 is best	Difference in false negative rates between groups
False Positive Rate Difference (FPRD)	Error Rates	No	0 to 1, 0 is best	Difference in false positive rates between groups
Positive Predictive Parity (PPP) or Predictive Parity	Error Rates	No	0 to 1, 0 is best	Difference in positive predictive values between groups
Overall Accuracy Parity (OAP)	Perform.	No	0 to 1, 0 is best	Difference in overall accuracy rates between groups

2.2.4 Explainable AI (XAI)

As ML models and algorithms become increasingly sophisticated, their decision-making processes often grow more complex and less transparent. This phenomenon, sometimes referred to as the "black box" problem, presents challenges to trustworthiness and fairness, particularly in contexts where understanding the basis for decisions is paramount. This issue is tackled head-on by the field of explainable artificial intelligence (XAI), a rapidly growing area within the broader AI discipline.

XAI is premised on the belief that to be trusted and effectively used, an AI system's decisions should be interpretable and explainable to its users. It seeks to create systems whose actions and decisions can be easily understood by humans. This is particularly important in high-stakes domains such as healthcare, finance, and criminal justice, where opaque decision-making can lead to disastrous outcomes and exacerbate existing biases.

Explainable AI or XAI has been regarded as an increasingly important part of AI/ML both in- and outside the AI community, as it helps explain the reasons why a certain prediction has been made and besides and probably most importantly, it bridges part of the gap to the outmost important debate around bias and fairness in automated decision-making processes as can be observed in recent legal foundations such as the GDPR (e.g., article 22 on automated decisions) or the new EU Act on AI (yet to be passed), to name a few.

XAI is usually organized into the categories of so-called white-box (WBM) and black-box models (BBM). As white-box models are transparent and relatively easily interpretable, the focus lies on the black-box models, however, a short introduction to the most common white-box models is provided first and then a deep-dive into different explainability techniques for black-box models based on a specific HR recruitment tool example serving as a red thread throughout this section.

Figure 2-3 (Thampi, A., 2022) illustrates the roadmap of this section:

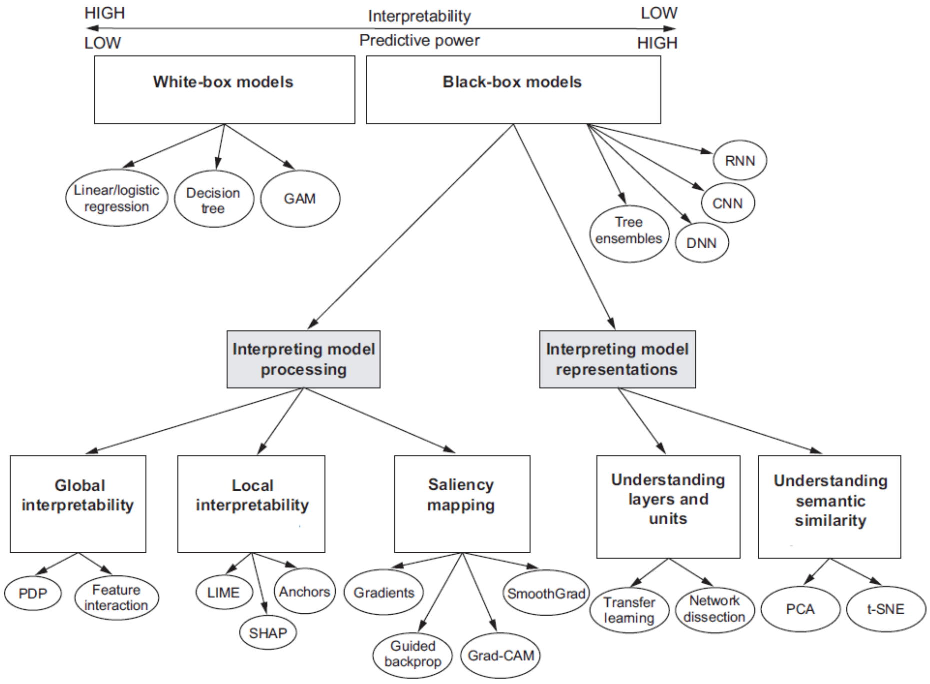


Figure 2-3: Map of Interpretability Techniques

As black-box models do not offer an intuitive explanation of their outcome and are inherently opaque and less interpretable, yet they usually outperform white-box models and have higher predictive power, the focus lies on the various interpretability techniques usually applied in the XAI area for black-box models. Figure 2-4 (Thampi, A., 2022) shows where BBM and WBM are placed on the interpretability (x-)axis and predictive power (y-)axis:

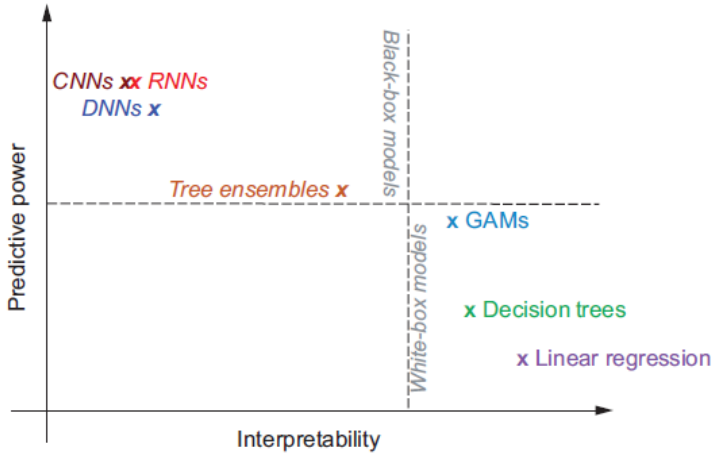


Figure 2-4: Models and Predictive Power

1) White-box models

In the realm of Explainable AI (XAI), white-box models refer to ML algorithms that provide inherently interpretable outcomes. Due to their transparency, the predictions made by these models can be clearly traced back to the input features. Common examples of white-box models include linear and logistic regression, decision trees, and generalized additive models (GAMs). Some of them are only briefly explained as they usually do not pose any real challenges when it comes to interpreting their outcomes.

Linear and Logistic Regression: Linear regression models assume a linear relationship between the input features and the output variable. The coefficient of each feature in the regression represents the change in the output variable for a one-unit change in that feature, while holding all other features constant. Logistic regression is used for binary classification problems, and its output can be interpreted as the logarithm of the odds of the positive class. The coefficients have similar interpretations to those in linear regression.

Decision Trees: Decision trees are graphical models that make decisions based on a set of binary rules. Each internal node of the tree tests a particular feature of the input, each branch corresponds to a result of the test, and each leaf node assigns a prediction to the input. The interpretation of a decision tree follows the path from the root to a leaf.

Generalized Additive Models (GAMs): GAMs are an extension of linear models that allow the response variable to depend on smooth transformations of the predictors, offering more flexibility. GAMs maintain the property of additivity, making them interpretable while also capable of capturing more complex patterns in the data.

2) Black-box models

Black-box models are ML models that are typically more complex and not as easily interpretable as their white-box counterparts. These models, which include tree ensembles, deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs), can generate highly accurate predictions but do not readily disclose how they reach their decisions, making them the primary focus for XAI techniques.

Tree Ensembles: Tree ensembles, such as Random Forests and Gradient Boosting Machines, are powerful ML models that combine predictions from multiple decision trees to make a final prediction. While individual decision trees are interpretable, ensembles of trees lose this interpretability because their predictions are derived from many trees, each potentially having a different structure (e.g., Friedman, J., 2001).

Deep Neural Networks (DNNs): DNNs are a type of artificial neural network with multiple hidden layers between the input and output layers. The internal workings of these networks are not easily interpretable due to the large number of parameters and complex transformations they use (LeCun, Y. et al., 2015).

Convolutional Neural Networks (CNNs): CNNs are a type of deep learning model primarily used for image processing. The complex layers of convolutions and transformations make it challenging to interpret their decision-making process (Krizhevsky, A., et al., 2012).

Recurrent Neural Networks (RNNs): RNNs are deep learning models used for sequential data. They incorporate loops to allow information to persist across timesteps, adding another layer of complexity that makes them harder to interpret (Elman, J., 1990).

For illustration purposes, the HR recruitment tool is used as a red thread and an example, which in turn also enhances the discussion of the case study in section 5.2 HR Recruitment Process.

Interpreting Model Processing

Black-box models require additional interpretability techniques to understand their decisions. These techniques are usually divided into two categories: Global and local interpretability.

Global interpretability refers to an overall understanding of the model's behavior across the feature space. This includes identifying which features are most important for the model's predictions on average, as well as how features interact with each other to affect predictions. Two commonly used techniques for global interpretability are Partial Dependence Plots (PDPs) and feature interaction.

- **Partial Dependence Plots (PDPs):** PDPs are graphical visualizations that show the marginal effect one or two features have on the predicted outcome of a ML model (Friedman, J., 2001). PDPs are a powerful tool for the exploration and interpretation of complex model behaviors, allowing users to visualize the average model prediction behavior for different values of the chosen feature(s), while marginalizing over the distribution of the other features. For a global perspective, PDPs can summarize the influence of selected features over the entire dataset by integrating out the other features.
  In the HR recruitment scenario, PDPs could be used to understand how specific applicant attributes such as "years of experience" or "level of

education" broadly influence the model's recommendation for interview selection. For instance, a PDP plot showing the effect of "years of experience" might reveal a positive trend, indicating that as the years of experience increase, the model's propensity to recommend the candidate for an interview also increases.

- **Feature Interaction:** Feature interactions provide insight into how combinations of feature values can affect the output of the model. They can reveal complex dependencies that are not visible from the inspection of individual features. When two features interact, the effect of one feature on the output changes depending on the value of the other feature (Greenwell, B. et al., 2018).
  For the HR recruitment tool, we consider the features "years of experience" and "level of education". An interaction effect could be observed if the effect of "years of experience" on the model's interview recommendation changes based on the "level of education". For instance, the model could place a higher weight on experience for applicants with a bachelor's degree compared to those with a master's degree, indicating an interaction between these features.

Local interpretability focuses on understanding the model's predictions at the individual instance level. This includes understanding why the model made a specific prediction for a specific data point. LIME, SHAP, and Anchors are commonly used techniques for local interpretability.

- **Local Interpretable Model-agnostic Explanations (LIME)** aims at explaining individual predictions. The fundamental premise of LIME is to locally approximate complex models with simpler, more interpretable models (Ribeiro, M.T. et al., 2016). To generate a LIME explanation, the following steps are typically taken:
  a) Choose an instance requiring explanation.
  b) Perturb this instance, producing a dataset of similar instances, and make predictions for these instances using the original model.
  c) Assign each instance in the perturbed dataset a weight based on its similarity to the original instance.
  d) Fit an interpretable model (e.g., a linear regression or decision tree) to the weighted perturbed dataset, with the original model's predictions serving as the target.
  e) Extract feature importance from the interpretable model to serve as the explanation for the original instance.

  Consider the HR recruitment scenario where an AI model is used to predict which candidates should be shortlisted. If a particular candidate is not recommended by the model, LIME can be used to understand the reasons behind this decision. The candidate's data is selected as the instance requiring explanation. Perturbed versions of this data are generated by slightly modifying feature values (e.g., experience, skills, qualifications). The model's predictions for these perturbed instances are obtained and used to fit a simpler model, from which feature importance is extracted to explain the model's original prediction. This might reveal, for instance, that the candidate's lack of specific skills or experience contributed most to the model's decision not to recommend them.

- **SHAP (SHapley Additive exPlanations)** values are grounded in Shapley values from cooperative game theory and serve to distribute the prediction value fairly among the features (Lundberg, S. et al. 2017). SHAP values offer the average contribution of each feature to the prediction for a specific instance, considering all possible combinations of features. The computation of SHAP values involves:

  a) Identifying all possible subsets of features.

  b) For each feature, calculating the contribution it makes to the prediction when added to different subsets of features.

  c) Calculating the Shapley value for each feature by averaging its contributions across all possible feature subsets.

  Using the same HR recruitment scenario, if a candidate is recommended by the model, SHAP can be used to identify which features contributed most to this decision. All possible subsets of features (e.g., combinations of experience, skills, qualifications) are considered. The contribution of each feature to the prediction is calculated when it is included in these different subsets. The average of these contributions gives the SHAP value for each feature, providing a measure of how much each feature contributed to the decision to recommend the candidate. This could reveal that the candidate’s particular skillset and years of experience were the main contributors to the model's decision.

- **Anchors** are model-agnostic rules that explain individual predictions with high precision (Ribeiro, M.T. et al., 2018). Anchors are derived from the concept of anchor points in geometry, which remain fixed and provide a reference point for other measures. In terms of interpretability, anchors are features that provide a "sufficient condition" for a certain prediction, meaning that as long as the anchors hold true, the prediction will remain the same.
  In the context of the HR recruitment tool, let's assume that an applicant is predicted by the model to be highly suitable for an interview. To generate an anchor for this prediction, the tool would generate multiple similar profiles by slightly modifying the applicant's features. The tool would then identify which features, when unchanged, result in the same prediction of interview suitability. These identified features form the anchor for the prediction and provide a basis for interpreting the model's recommendation.

Saliency mapping techniques are used to highlight regions in the input that are most relevant for a model's prediction. While saliency mapping is most often used in image-based models, the techniques can be conceptually applied to other data types.

- **Gradients:** The gradient of the output with respect to the input, often known as a gradient saliency map, gives an idea of how the model output changes with small changes to the input. For example, a deep learning model used in an HR recruitment context may take a variety of factors into account, like qualifications, skills, experience, etc. By computing the gradient of the output with respect to each of these factors, we can understand which factors have the largest impact on the hiring decision (Simonyan, K. et al., 2013).

- **Guided Backpropagation:** This method modifies the standard backpropagation algorithm to focus only on positive contributions to the output. It effectively provides a 'guided' tour of the factors that led to the final prediction, by discarding any factors that had a negative or neutral influence on the decision. In our HR scenario, if a candidate is rejected, guided backpropagation can help us identify the features that were most instrumental in this decision (Springenberg, J.T. et al., 2014).
- **SmoothGrad:** This technique reduces noise in gradient saliency maps by taking an average over multiple noisy versions of the input. For an HR tool, this could help in understanding whether small changes in a candidate's profile significantly alter the hiring recommendation. If the tool's recommendation changes dramatically with small alterations to a candidate's profile, it may indicate that the tool is overly sensitive to certain features (Smilkov, D. et al., 2017).

  In the HR recruitment example, these techniques could be used to interpret a model that uses textual data from resumes. For instance, a deep learning model could be trained to classify applicants based on the textual content of their resumes, and a saliency map could reveal which words or phrases are most influential in the model's decision-making process. For instance, the phrase "project management certification" might be highlighted as being particularly influential in positive recommendations.

Interpreting Model Representations

Apart from understanding the process by which a model makes a prediction, it is also crucial to understand the internal representations that a model learns.

Understanding layers and units

Deep learning models, such as neural networks, learn representations of data in a hierarchical manner, where each layer of the network captures different levels of abstraction in the data. Interpreting these representations can provide insight into what the model has learned.

- **Transfer Learning** involves using a pre-trained model (typically on a large benchmark dataset) and adapting it for a new, similar task (Yosinski, J. et al., 2014). When the layers of the pre-trained model are examined, the initial layers often capture universal features like edges and curves, while the deeper layers capture more task-specific features.
- **Network Dissection** is a technique for quantifying the interpretability of latent representations of CNNs by evaluating the alignment between individual units and a set of semantic concepts (Bau, D. et al., 2017). The technique uses a broad set of semantic segmentation labels to annotate the units with human-interpretable concepts.

In the HR recruitment tool, imagine a model pre-trained on job descriptions to recommend candidates for an interview is used. By using transfer learning, we can understand which aspects of the job descriptions (e.g., required skills or qualifications) are most important for the model. With network dissection, we could identify which units in the network align with key concepts, such as "leadership experience" or "Python programming", providing further insight into the model's decision-making process.

Understanding Semantic Similarity (PCA, t-SNE)

Dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) can be used to visualize high-dimensional data, such as the representations learned by a model, in two or three dimensions (Van Der Maaten, L. et al., 2008).

In the context of the HR recruitment tool, imagine the tool learns a high-dimensional representation of each candidate based on their resume. Applying PCA or t-SNE could allow us to visualize these representations in 2D or 3D, potentially revealing clusters of similar candidates. This could provide valuable insights, such as the grouping of candidates with similar skills or experience levels, and could help in understanding the model's decision-making process.

2.2.5 Fairness vs. Accuracy and other Trade-offs

The discussion of fairness and accuracy trade-offs forms an essential component of understanding the ethical implications of AI algorithms, particularly those used in high-stakes decision-making scenarios. These trade-offs primarily involve balancing societal norms of 'fairness' and the potential social costs that come with prioritizing one over the other.

Corbett-Davies, S. et al. (2017) addressed this tension in their study of the COMPAS algorithm (cf. section 5.4), demonstrating that optimizing for public safety results in decisions that negatively impact defendants of color, thus revealing a conflict between the minimization of violent crime and satisfying common notions of fairness. Their conclusion highlighted a significant implication: satisfying legal and societal fairness definitions may result in a higher release rate of high-risk defendants, which could detrimentally affect public safety.

Moreover, this negative impact on public safety might disproportionately affect different demographic groups, subsequently engendering another form of 'fairness cost.' This highlights that fairness is not a one-dimensional construct but rather a multifaceted concept with far-reaching implications.

As such, the challenge faced by developers and operators of these algorithms is determining how to mitigate potential biases without reinforcing existing societal inequalities. One possible strategy is to explore avenues for reducing disparities between groups without compromising the overall performance of the AI model. In particular, this becomes critical in scenarios where there appears to be a trade-off between fairness and accuracy.

Notably, some scholars and practitioners posit that there are opportunities for improving both fairness and accuracy in algorithms simultaneously. One approach is through the thorough investigation and resolution of bugs in the software, which might impede the model from maximizing overall accuracy. Another strategy involves addressing under-representation in the training data sets. Including more diverse data could improve accuracy in decision-making and reduce unfair results, as demonstrated in Buolamwini's facial detection experiments (Buolamwini, J., et al., 2018).

Additionally, as highlighted by Sara Holland from Google (Barton, N., et al. 2019), understanding and managing the risk tolerance associated with these types of trade-offs is critical . Decisions about whether the social costs of the trade-offs are justifiable, whether stakeholders are open to algorithm-based solutions, or if human intervention is necessary for framing the solution, need careful deliberation.

Hence, understanding and navigating the trade-offs between fairness, accuracy, and other societal objectives is integral to the development and deployment of AI algorithms, especially those with high societal impact.

3 Tools and Approaches for Improving Trustworthiness

Chapter 3 provides a comprehensive examination of the tools and approaches designed to enhance the trustworthiness of AI systems. Trustworthiness in this context encompasses various dimensions including data quality, algorithmic fairness, transparency, and system robustness.

In the first three sections, some techniques of bias mitigation lay the groundwork. This approach comprises several intricate and closely intertwined stages—namely pre-processing, in-processing, and post-processing—which correspond respectively to data collection and pre-processing, model selection and training, evaluation and validation, and quantifying bias and fairness as shown in figure 3-1:

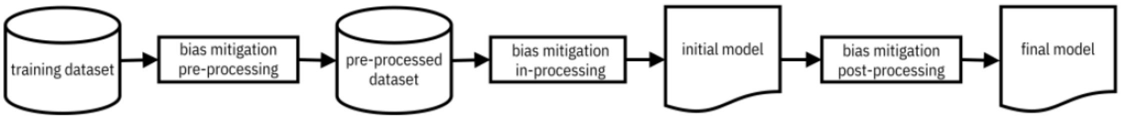

training dataset
bias mitigation pre-processing
pre-processed dataset
bias mitigation in-processing
initial model
bias mitigation post-processing
final model

Figure 3-1: Bias Mitigation in Different Stages of the ML Pipeline

In the **preprocessing** phase, techniques for data collection and preprocessing are paramount. These include methods to ensure representative data sampling and robust preprocessing techniques to mitigate initial data bias, such as disparate impact analysis, reweighing, and optimized preprocessing.

The **in-processing** stage, on the other hand, concentrates on model selection and training. The choice of algorithm, fairness constraints integrated into the optimization function, adversarial de-biasing, and regularization techniques are examples of the critical choices to be made at this stage. Each choice can have a profound impact on the model's capacity to offer fair predictions.

The **postprocessing** phase is concerned with model evaluation, validation, and adjustment based on evaluation outcomes. Techniques to evaluate and validate model outcomes include multiple bias detection methods, fairness metrics (such as equality of opportunity, demographic parity), and adjustment techniques like threshold adjusting and equalized odds postprocessing.

Section 3.4 discusses the role of Explainable AI (XAI) in fortifying system robustness and transparency, while Section 3.5 investigates causality-based fairness methods and their potential in bias mitigation.

Section 3.6 delivers a comparative overview of selected tool-based bias mitigation solutions, such as IBM's AI Fairness 360, Google's What-if Tool, Meta's Fairness Flow, Carnegie Mellon's Aequitas, and Themis AI. A summary comparison in Section 3.6.6 concludes the chapter by providing a succinct appraisal of these tools.

Through a comprehensive exploration of these aspects, chapter 3 contributes to a deeper understanding of strategies and tools intended for improving trustworthiness in AI systems.

3.1 Pre-Processing: Data Collection and Cleaning

During preprocessing, techniques such as data cleaning, feature extraction, and sampling are employed to structure the data for the ML model. The fundamental goal in this stage is to ensure a robust, unbiased dataset that serves as the basis for a fair and accurate ML model.

Table 3-1 provides an overview of the most commonly applied techniques to mitigate unfairness at this stage:

Category	Technique	Description
Bias Mitigation Techniques	Reweighing	Assigning instance weights to data points to ensure balanced representation across different groups
	Reject Option-Based Classification	Changing predictions to make them fairer by providing favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty
	Disparate Impact Remover	Modifying feature values to increase group fairness while preserving rank-ordering within groups
Data Augmentation Techniques	Counterfactual Data Augmentation	Augmenting the dataset by generating synthetic instances through flipping protected attribute values
Sampling-Based Techniques	Oversampling	Artificially boosting the number of instances in the less-represented class in an imbalanced dataset
	Undersampling	Balance the class distribution by reducing the number of instances in the majority class
Data Transformation Techniques	Optimized Preprocessing	Learning a probabilistic transformation to modify the features and labels in the training data to reduce bias
	Learning Fair Representations	Learning a transformed representation of the data that minimizes the ability to predict protected attributes
	Correlation or Proxy Remover	Removing features that are highly correlated with protected attributes
Data Editing Techniques	Relabeling	Modifying labels in the dataset to reduce bias

	Massaging the Dataset	Changing class labels for a chosen subset of instances to enforce demographic parity
Cluster-Based Techniques	Fair k-Means Clustering	Implementing k-means clustering algorithm to ensure equal cluster distributions across protected and unprotected groups

Table 3-1: Pre-Processing Mitigation Techniques

The most common of these techniques are described in detail and illustrated by an example which will bring forward some of the discussions on credit scoring and loan approval of section 5.3 Automated Credit Scoring where we consider a scenario where an organization is building a ML model to predict whether a person will default on a loan. The available data includes various features such as age, income, employment status, credit history, and a protected attribute like gender. The goal is to build a model that makes accurate predictions while being fair and not discriminating against any gender.

Bias Mitigation Techniques

Reweighing is a pre-processing technique employed to ensure fairness in ML. It involves the assignment of weights to instances in the training dataset in such a way that fairness constraints are satisfied in expectation across different demographic groups. It is a mechanism employed to mitigate potential biases, specifically by providing equalized odds across distinct demographic groups.

In the loan approval scenario, gender can be considered a protected attribute. If an original dataset is imbalanced and demonstrates differences in the proportion of loan approvals across gender groups, it may result in the learning model acquiring a gender bias. This is where reweighing becomes essential.

The process of reweighing involves assigning different weights to the instances of the dataset such that the statistical parity difference between the two gender groups is minimized. In simple terms, this ensures that the odds of loan approval are similar for individuals across gender lines. This technique, thus, alters the importance of instances from different demographic groups in a way that contributes towards reducing bias in the decision-making model. It is important to note that reweighing does not modify the features or the labels, but it impacts the way the ML algorithm learns from the instances.

The method of reweighing was introduced by Kamiran, F. et al. (2012) in their paper "Data Preprocessing Techniques for Classification without Discrimination". They proposed this technique to alter the distribution of training data such that discrimination and bias were minimized while maintaining the rank order of instances within each group.

Reject Option-based Classification (ROC) is a fair ML pre-processing technique introduced by Kamishima et al. (2012) in their work "Fairness-Aware Classifier with Prejudice Remover Regularizer". This technique operates on the principle of avoiding potentially unfair classification decisions by abstaining from making a prediction when uncertainty is high.

In the context of loan approval, where gender serves as the protected attribute, ROC can be instrumental in promoting fairness. The classifier, equipped with ROC, may abstain from making a loan approval or rejection decision if the predicted probability is within a predefined threshold of uncertainty.

This specific area of uncertainty, often referred to as the "rejection region", is situated around the decision boundary. It is typically set where the classifier's confidence in making a prediction is low. This abstention from potentially unfair decisions is especially beneficial in ensuring fairness when prediction errors could disproportionately affect the less privileged gender group.

Although ROC is a post-processing method, it can be implemented at the pre-processing stage if abstained instances are removed from the dataset or relabeled before training the classifier.

The Disparate Impact Remover (DIR) is a pre-processing technique used in fair ML pipelines to address discrimination and bias. The methodology behind the DIR is to adjust feature values to enhance group fairness while keeping within-group rank ordering unaltered. It addresses a specific type of discrimination called disparate impact, where a decision disproportionately disadvantages individuals from a certain group.

Considering a loan approval scenario, where gender is the protected attribute, DIR can be implemented by adjusting the feature values for both gender groups to mitigate potential disparities. For example, if a feature like "credit score" disproportionately impacts loan approval decisions across genders, the DIR can be used to modify credit scores such that their distribution becomes similar for both genders.

The DIR works by first ranking the individuals within each group (e.g., male and female) based on the value of the feature that may cause disparate impact (e.g., credit score). Then it adjusts the feature values such that the distributions across the gender groups become similar while maintaining the relative order of the individuals within each group. This adjustment ensures that the disparate impact caused by this feature is minimized.

The Disparate Impact Remover was first introduced by Feldman, M. et al. (2015) in their paper "Certifying and Removing Disparate Impact". This method is particularly effective for tackling indirect discrimination that arises due to certain seemingly innocuous features being correlated with both the protected attribute and the outcome.

Sampling-Based Techniques

Synthetic Minority Over-sampling Technique (SMOTE) is a robust method used for addressing class imbalance in datasets, which can contribute to bias in ML models. This approach was introduced by Chawla, N. et al. (2002) in their paper "SMOTE: Synthetic Minority Over-sampling Technique".

Unlike conventional oversampling methods, which replicate minority class instances, SMOTE generates synthetic examples that enhance the feature space of the minority class. It works by selecting instances that are close in the feature space, drawing a line between these instances, and creating new instances along this line.

In the context of a loan approval scenario where gender is a protected attribute, SMOTE can be applied to the class representing the minority gender, which may be underrepresented in the loan approval instances. If, for example, fewer females are approved for loans in the original dataset, SMOTE can help to generate synthetic female profiles who have been approved for loans. These synthetic instances will be a blend of features from female borrowers who have been approved for loans, thus increasing the representation of this group in the approval class.

Similar undersampling techniques exist for the majority class and are not further explained here.

Data Transformation Techniques

Optimized Preprocessing is a fair ML pre-processing technique that involves learning a probabilistic transformation, which modifies the features and labels in the training data to achieve a fairer model. This technique was presented by Calmon, F. et al. (2017) in the paper "Optimized Preprocessing for Discrimination Prevention".

The primary goal of Optimized Preprocessing is to learn a transformation that maps instances in the original dataset to instances in a new dataset, where this transformation is chosen to optimize for a number of objectives, such as preserving as much information as possible from the original data, improving the accuracy of a predictive model, and satisfying fairness constraints.

In a loan approval scenario, where gender is the protected attribute, Optimized Preprocessing can be utilized to transform the dataset in a way that reduces gender-based disparities. This could involve modifying features like credit history, employment status, or income levels in a way that reduces their correlation with gender, while still retaining the overall patterns and relationships within the data necessary for accurate predictions.

The transformations generated through this technique ensure that the decision-making process is less influenced by the protected attribute, helping to prevent potential discrimination. However, it is crucial to ensure that the transformed data still maintains sufficient utility for accurate prediction.

Learning Fair Representations (LFR) is a pre-processing technique that aims to create a fair representation of the input data by minimizing the ability to predict the protected attribute from the transformed data. The technique was introduced by Zemel, R. et al. (2013) in their paper "Learning Fair Representations".

The primary objective of LFR is to find a transformation of the original dataset into a new representation that preserves as much useful information as possible for prediction purposes while simultaneously reducing any bias related to the protected attribute. It employs a variant of adversarial training to achieve this goal.

In the context of a loan approval scenario, where gender serves as the protected attribute, LFR would strive to find a new representation of the dataset where it becomes challenging to predict the gender based on the transformed features. This could involve transforming features like credit score, income, and employment status in such a way that the correlation between these features and the gender attribute is minimized.

By creating a representation in which gender cannot be easily inferred, LFR aims to ensure that the subsequent classifier built on this representation is less likely to rely on gender when making loan approval decisions, thereby reducing potential discrimination.

However, it's essential to ensure that the transformed data maintains enough utility for accurate prediction. This balance between fairness and accuracy is a key challenge in implementing LFR and similar techniques.

The **Correlation or Proxy Remover** is a preprocessing technique employed in fair ML to reduce or remove the correlation between the protected attribute and the predictor variables in the dataset. This approach seeks to mitigate discrimination by ensuring that the predictor variables used by a ML model do not serve as proxies for the protected attribute.

In a loan approval scenario, where gender is the protected attribute, there may exist certain predictor variables like occupation or educational background, which might indirectly correlate with gender. These variables, if not treated, can potentially serve as proxies, leading the model to indirectly discriminate based on gender.

The Correlation or Proxy Remover methodology involves identifying these proxies and either removing them or adjusting their values such that their correlation with the protected attribute is minimized. This can involve statistical techniques such as residualization or decorrelation methods.

The concept of removing correlation or proxies has been an integral part of fairness-related research, but it's not necessarily associated with a single paper. However, it is significantly featured in works like "Fairness through Awareness" by Dwork, C. et al. (2012) and "A Survey on Bias and Fairness in Machine Learning" by Mehrabi, N. et al. (2019).

It is crucial to mention that this technique, while effective, should be used cautiously as removing or altering variables could lead to loss of critical information required for the prediction task.

Data Editing Techniques

Relabeling is a pre-processing technique used in fair ML to adjust labels of instances in a dataset to mitigate potential bias and ensure fairness. It involves changing the labels of certain instances based on a specified fairness criterion or objective.

Considering the loan approval scenario where gender is the protected attribute: If the original dataset exhibits bias, such as an uneven approval rate between genders, relabeling can be applied to adjust the labels (approved or not approved) of some instances to reduce this disparity.

The instances to be relabeled can be chosen in various ways. For example, borderline instances, whose classification confidence is low, can be relabeled to promote fairness. It is also possible to perform relabeling in a way that balances the number of positive and negative instances for each group, thereby enforcing demographic parity.

Relabeling was introduced and utilized in several works to mitigate bias. One such study is "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness" by Kusner, M. et al. (2017), which introduces a technique for relabeling based on counterfactual fairness.

"**Massaging the Dataset**" is a pre-processing technique employed in fair ML to amend biases in datasets, and it involves the modification of class labels to improve fairness measures. This technique is related to relabeling, but with more specific adjustments targeting a defined fairness criterion.

In the loan approval scenario, massaging would involve adjusting the labels (approved or not approved) of some instances to decrease any discrepancies in approval rates between genders. The instances for which the labels are modified are typically those that are closest to the decision boundary of a classifier trained on the original dataset.

The term "Massaging the Dataset" was coined by Kamiran, F. et al. (2012) in their work "Data preprocessing techniques for classification without discrimination". While it may sound informal, it is a widely recognized term in the academic literature on fairness in ML.

Cluster-Based Techniques

Fair k-Means Clustering is an adaptation of the standard k-means clustering algorithm, aimed at ensuring that the generated clusters respect fairness criteria with respect to a protected attribute. The method was proposed by Chierichetti et al. in their work "Fair Clustering Through Fairlets".

Fair k-Means Clustering first partitions the dataset into smaller sets, referred to as "fairlets", in such a way that each set satisfies a specified fairness condition. For example, each fairlet may be required to contain an equal number of instances from each category of a binary protected attribute. Then, the standard k-means algorithm is run on these fairlets to form the final clusters.

Taking the loan approval scenario where customer segmentation needs to be performed fairly with respect to gender: Using Fair k-Means Clustering, the customer data can be divided into fairlets in a way that each fairlet contains an equal number of male and female applicants. The k-means algorithm is then applied to these fairlets to generate the final customer segments, ensuring that the clusters are fair with respect to gender.

It is not claimed that these preprocessing steps are a complete list of all existing bias mitigation techniques at this stage of the data pipeline, however, it provides an overview of the most widely used techniques, some of which are also applied in the following case studies.

3.2 In-Processing: Model Selection and Training

In-Processing techniques apply bias mitigation during the model training phase. Here, fairness constraints are directly incorporated into the optimization process of a learning algorithm, or fairness regularization terms are added to the objective function that the algorithm seeks to minimize. The following list of in-processing techniques is not exhaustive, but it provides some of the mostly applied ones. An in-depth description of these techniques is beyond the scope of this project.

- The **Prejudice Remover** method, introduced by Kamiran et al. (2012), is an example of an in-processing algorithm where a regularization term is included in the logistic regression's objective function. This regularizer specifically penalizes discriminatory predictions on the basis of a protected attribute, such as gender.
- The **Adversarial Debiasing** technique, introduced by Zhang et al. (2018), involves a two-player game in which a classifier is trained to make predictions while an adversary is trained to predict the protected attribute from the classifier's predictions. The aim is to train a classifier that not only performs well on the task at hand but also confuses the adversary, thereby reducing the ability to predict the protected attribute, hence reducing the discrimination.
- The **Exponentiated Gradient Reduction** approach, introduced by Agarwal et al. (2018), aims at reducing discriminatory behavior by learning a probabilistic classifier as a mixture of classifiers that minimizes both the empirical risk and a chosen notion of disparity.
- The **GerryFair Classifier** introduced by Kearns et al. (2017), ensures fairness by adding a fairness regularizer to the objective function that corresponds to the fairness violation in terms of a chosen disparity measure.
- **Grid Search Reduction**, as presented by Agarwal et al. (2018), uses a grid search to select hyperparameters that generate fair classifiers. It combines multiple disparity metrics with traditional error metrics to arrive at a satisfactory trade-off.
- The **Meta-Fair Classifier** proposed by Celis et al. (2019) is an approach that treats fairness constraints as penalties in a Lagrangian framework. The classifier works by optimizing both fairness and accuracy measures.

Example: Loan approval

The loan approval decision system uses logistic regression as a base algorithm, and gender is a protected attribute. One way to ensure fairness during the model training phase is to use the Prejudice Remover method. A regularization term that penalizes discriminatory predictions based on gender is added to the objective function. The trained model would then not only aim for accurate predictions but also consider the fairness constraints related to gender discrimination during the prediction.

In the Adversarial Debiasing scenario, an adversarial network would be trained in parallel to the loan decision classifier. The classifier would predict loan approval, and the adversarial network would attempt to predict the applicant's gender from the classifier's loan approval predictions. The classifier's goal is to make accurate loan approval decisions while ensuring that the adversarial network cannot accurately predict the applicant's gender, thereby reducing bias.

For other techniques like Exponentiated Gradient Reduction, GerryFair Classifier, Grid Search Reduction, and Meta-Fair Classifier, the key idea would be similar: modifying the classifier's training process to enforce fairness by reducing disparity on the protected attribute (gender), while maintaining performance on the loan approval prediction task.

These techniques provide an opportunity to integrate fairness considerations directly into the model training process. However, as with all fairness interventions, it is important to ensure that these measures do not inadvertently introduce new forms of bias or unfairness, and therefore ongoing evaluation and validation is necessary.

3.3 Post-Processing: Evaluation Metrics & Outcome Manipulation

There are a series of possible postprocessing techniques in the data pipeline, however, the first step is to quantify bias and fairness to be able to apply bias mitigation techniques. In section 2.2.3 Fairness Metrics a wealth of metrics was already introduced according to different purely technical categorization criteria. Here, on the other hand, the idea is to combine three of the independently applied criteria to provide a holistic understanding of these categories, i.e., group versus individual, observational versus causality-based, and independence versus separation versus sufficiency fairness criteria. Many academic papers refer to some of the categories, however, they do not combine all of them. Table 3-2 on page 63 is based on a series of different papers, among others, Dwork, C. et al. (2012), Hardt, M. et al. (2016), Castelnovo, A. et al. (2022).

Group fairness and individual fairness are two contrasting conceptual frameworks used to characterize and evaluate fairness in ML algorithms.

- **Group fairness**, also known as statistical or demographic fairness, operates on the principle of demographic parity. It advocates that an algorithm should yield similar outcomes, on average, for different demographic groups defined by a protected attribute. Metrics falling under this category, such as Demographic Parity Difference or Equalized Odds Ratio, often measure disparities in aggregate outcomes or error rates between these groups. They provide a high-level perspective on fairness, focusing on the distributional aspects of decisions across different groups, but do not account for differences among individuals within those groups.
- **Individual fairness**, also known as similarity-based or instance-based fairness, focuses on treating similar individuals similarly, regardless of their group membership. It operates on the principle of treating similar cases alike. This principle posits that two individuals who are similar in terms of attributes relevant to a decision should receive similar outcomes. Individual fairness metrics such as Counterfactual Fairness or Fairness Through Unawareness tend to focus on ensuring fairness at an individual level, taking into account the heterogeneity within demographic groups.

While both these frameworks aim to ensure fairness, they sometimes lead to conflicting outcomes due to their different underlying principles. Group fairness may lead to situations where similar individuals from different demographic groups are treated differently, while individual fairness might overlook systemic biases that affect entire demographic groups. Balancing the trade-off between these two fairness perspectives is an ongoing challenge in the field of ML fairness.

Observational and causal fairness represent two distinct paradigms for fairness evaluation in ML, each focusing on different aspects of data and decision-making processes.

- **Observational fairness** is based on the statistical relationships found in the observed data. In this approach, fairness metrics are computed from the joint distribution of predictions, outcomes, and protected attributes, without assuming any causal relationships among them. The key focus here is to ensure equal treatment across groups defined by the protected attributes purely based on the observed statistics. Metrics like Demographic Parity Difference, Equalized Odds Ratio, or Calibration fall under this category. These observational metrics do not, however, capture the potential causes of unfairness, and therefore might fail to address underlying structural biases present in the data-generating process.

- **Causal fairness**, on the other hand, takes into consideration the causal relationships among variables. It goes beyond observational data and seeks to capture the underlying causal structure that generates the data. The premise here is that the causes of unfairness often trace back to societal structures and systemic biases. Hence, fairness should be defined based on these causal structures. This perspective gives rise to fairness metrics such as Counterfactual Fairness or Fairness Through Unawareness. These causal metrics aim to ensure that the model's predictions would remain the same if the protected attribute were counterfactually changed, addressing biases that might not be apparent from observational statistics alone.

While causal fairness provides a more principled way to handle fairness by considering the underlying causes of biases, it also demands a higher level of data requirement and methodological rigor. It often requires knowledge about the causal structure of the data, which might not be easily available or accurately discernible.

Independence, separation, and sufficiency represent three key criteria for fairness evaluation in ML, each focusing on a specific relationship between the model's predictions, the protected attribute, and the true outcomes.

- **Independence**: This criterion, also known as demographic parity or statistical independence, requires the model's predictions to be independent of the protected attributes. The idea here is to ensure that decisions do not depend on protected attributes, which could lead to disparate treatment of different demographic groups. A fairness metric adhering to the independence criterion, such as Demographic Parity Difference, would measure the disparity in outcomes across groups defined by the protected attribute.
- **Separation**: Also known as conditional procedure accuracy equality, the separation criterion requires the model's errors to be independent of the protected attributes. This means the model should have similar performance, such as similar rates of false positives and false negatives, across different demographic groups. Metrics adhering to the separation criterion, such as Equalized Odds Difference or Treatment Equality Difference, measure the disparity in error rates across different demographic groups.
- **Sufficiency**: Sufficiency, also known as conditional use accuracy equality, requires the predicted outcomes to capture all the information necessary for the decision, given the true outcome and the protected attribute. This means, for a given outcome, the decision should be the same across different demographic groups. Metrics adhering to the sufficiency criterion, such as Predictive Equality Difference, measure disparities in how the predictions relate to the true outcomes across different demographic groups.

These three criteria provide distinct perspectives on fairness, focusing on different aspects of the decision-making process. However, it is important to note that these criteria cannot generally be satisfied simultaneously unless the outcome is independent of the protected attribute. This tension among the criteria illustrates the inherent complexity and trade-offs involved in achieving fairness in ML. Table 3-2 categorizes the metrics introduced in section 2.2.3:

ID	Metric Name	Group/ Individual Fairness	Observational / Causal Fairness	Independence / Separation/ Sufficiency
A.1	Statistical/Demographic Parity Difference	Group	Observational	Independence
A.2	Demographic Parity Ratio	Group	Observational	Independence
A.3	Equalized Odds Difference	Group	Observational	Separation
A.4	Equalized Odds Ratio	Group	Observational	Separation
A.5	Treatment Equality Difference	Group	Observational	Sufficiency
A.6	Treatment Equality Ratio	Group	Observational	Sufficiency
A.7	Predictive Equality Difference	Group	Observational	Sufficiency
A.8	Predictive Equality Ratio	Group	Observational	Sufficiency
B.1	Average Odds Difference	Group	Observational	Separation
B.2	Conditional Demographic Disparity	Group	Observational	Separation
C.1	Calibration	Group	Observational	Sufficiency
C.2	Probability Integral Transform	Group	Observational	Sufficiency
C.3	Normalized Mutual Information	Group	Observational	Sufficiency
D.1	Generalized Entropy Index	Group	Observational	Sufficiency
D.2	Kullback-Leibler Divergence	Group	Observational	Independence
E.1	Counterfactual Fairness	Individual	Causal	Independence
E.2	Fairness Through Unawareness	Individual	Causal	Independence
E.3	Indirect Bias	Individual	Causal	Independence
E.4	Fairness Constraints	Individual	Causal	Independence

Table 3-2: Categorization of Fairness Metrics

Once the fairness metrics are evaluated, different postprocessing techniques can be applied.

Postprocessing techniques are crucial when dealing with ML models that have already been trained and where the training data or process cannot be altered. This scenario might arise when using pre-trained models from a third party or when organizational or legal restrictions exist on modifying the training data. In these situations, bias mitigation is carried out by manipulating the output predictions to ensure fairness criteria are met. This essentially involves altering the predicted outcomes (e.g., loan approval or denial) to align with desired group fairness metrics.

The primary postprocessing strategies are generally based on thresholding, calibration, or equalized odds. Postprocessing can be carried out with or without the availability of validation data. When validation data is available, the postprocessing can follow the observational fairness perspective, i.e., "what you see is what you get." In the absence of validation data, the causal fairness perspective, i.e., "we're all equal," can be employed.

Thresholding techniques involve manipulating the decision threshold for binary predictions. Predicted labels can be 'flipped' according to certain criteria. This might involve adjusting the decision threshold for a certain demographic group or flipping the predicted outcome for individuals whose predicted scores lie close to the decision boundary, similar to the massaging technique in preprocessing. However, a random selection of individuals for flipping might be seen as procedurally unfair.

Calibration techniques, such as Platt Scaling, adjust the predicted probabilities to better match the observed probabilities in each group. These techniques are often used in conjunction with thresholding to ensure the model's predictions are not only fair but also reliable.

The **Equalized Odds Postprocessing** method, proposed by Hardt et al. (2016), provides a principled approach to achieve equalized odds (similar true positive and false positive rates) across groups. This technique alters the predictions in a way that ensures equal opportunity for all demographic groups in terms of loan approval, given the true label.

The **Fairness-aware Classifier Calibration** technique, introduced by Menon, A. et al. (2018), aims to calibrate the classifier's output scores to achieve fairness. This technique involves finding an optimal transformation of the predicted scores into a new score, which is then thresholded to a binary decision. It is a potent postprocessing tool and works well empirically without being computationally intensive.

However, all of these approaches necessitate the availability of the protected attribute in the deployment data. For situations where the protected attribute is not available, techniques such as the Fair Score Transformer can be employed. The Fair Score Transformer works on the continuous scores output by the base classifier, making it a desirable choice in the category of postprocessing bias mitigation.

Example: Loan approval system

The system uses a ML model to predict whether a loan application should be approved or denied. Suppose that the model has already been trained and deployed, and it is discovered that it is resulting in unfair outcomes, particularly discriminating against female applicants. In such cases, various postprocessing techniques can be used to address this bias:

1. Thresholding Techniques:

Suppose that female applicants consistently receive lower scores from the model than they should, and as a result, they are denied loans more often. To address this, a unique decision threshold could be set for female applicants. The threshold for loan approval for this group could be lowered, such that applicants who would have previously been denied a loan (based on the original threshold) are now approved. This would result in more balanced loan approval rates between male and female applicants. However, careful tuning is required to ensure the threshold does not result in overcorrection and a different kind of unfairness.

2. Calibration Techniques:

Calibration techniques like Platt Scaling can be employed to adjust the predicted probabilities of loan approval for female applicants. If it is found that the model's predicted probabilities do not align with actual loan repayment rates, Platt Scaling can be used to recalibrate these predictions. The objective here would be to ensure that, across the range of predicted probabilities, the model's confidence in its predictions is justified by the actual loan repayment rates.

3. Equalized Odds Postprocessing:

Equalized Odds Postprocessing, as proposed by Hardt et al. (2016), can be utilized to ensure fairness in both true positive and false positive rates between male and female applicants. This technique would adjust the predictions of the model in a manner that guarantees that, given the true label (loan repaid or not), the chance of being approved is the same across both demographic groups.

4. Fairness-aware Classifier Calibration:

The fairness-aware classifier calibration, proposed by Menon, A. et al. (2018), could also be implemented. Here, the continuous score output from the model would be transformed optimally into a new score. This new score would then be thresholded to make a binary loan approval decision, ensuring that the loan approval rates are fair across both male and female demographic groups. This technique is especially effective when the protected attribute is not available in the deployment data.

In conclusion, postprocessing techniques offer the ability to manipulate the predictions of an existing model to achieve a fairer outcome. However, it is critical to continuously evaluate and validate these techniques to ensure that they're not introducing new forms of bias or unfairness.

3.4 Robust AI Systems with XAI

The typical data mining process as described in the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, comprises the phases of (1) problem specification, (2) data understanding, (3) data preparation, (4) modeling, (5) evaluation, (6) deployment and monitoring, which are arranged in a continuous circle. Building on top of this well-known methodology and focusing on the interpretable and explainable part of the standard process, understanding or interpretation and explanation phases are embedded in the process as shown in figure 3-2 (Thampi, J., 2022) below.

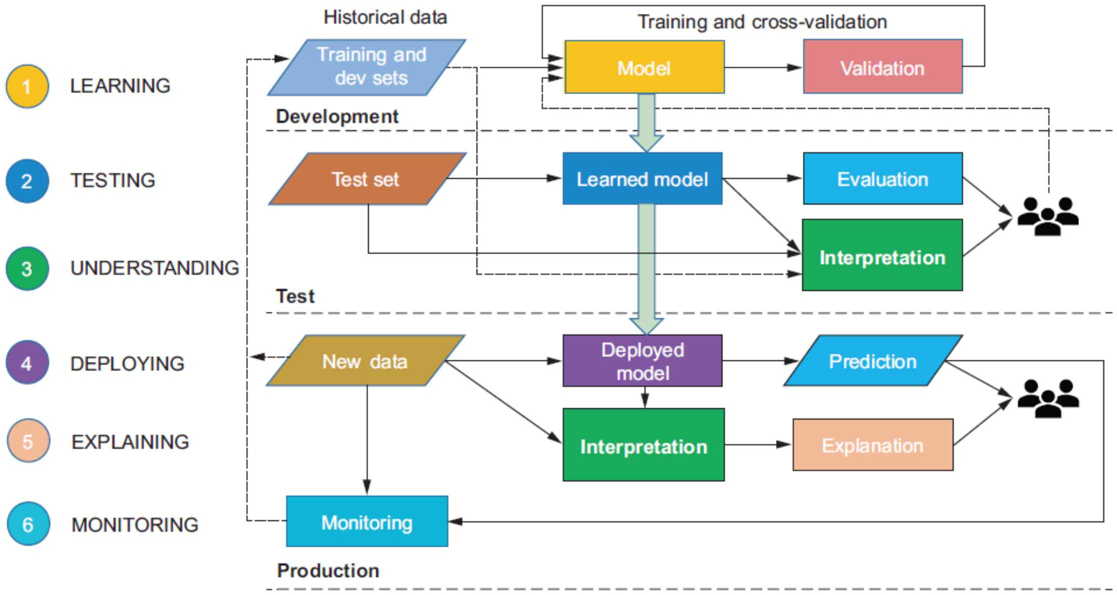


Figure 3-2: Embedded Interpretation and Explanation Phases in XAI

1) Within the robust AI system, the **learning phase** unfolds within a development environment, utilizing two key subsets of data: the training set and the development set. The **training set** plays an instrumental role in enabling the ML model to learn the mapping function 'f' from the provided input features 'X'. Subsequent to this training process, the **dev set** is employed for validation purposes, serving as the basis for tuning the model. Tuning the model constitutes a vital iterative process that seeks to determine the optimum model parameters, also known as hyperparameters, which yield the highest performance. Such optimization revolves around the model's performance on the development set. The iterative nature of this process entails repeated cycles of adjustment and validation, which persist until the model demonstrates an acceptable level of performance. The learning phase, therefore, encapsulates the initial training and fine-tuning of the ML model, setting the foundation for the subsequent stages of robust AI system deployment.
2) Upon completion of the learning phase, the focus transitions to the **testing phase**, carried out within the test environment. This stage introduces a novel subset of data known as the **test set**, distinct from the data used during training. The central aim of the testing phase is to garner an unbiased appraisal of the model's accuracy. At this juncture, stakeholders and experts are brought in to assess the system's functionality and the model's performance when applied to the test set. This additional layer of testing, often referred to as user acceptance testing (UAT), signifies the final stage in the software system's development process. UAT is intended to ensure that

the model meets the defined specifications and is ready for deployment, thereby promoting the integrity of the AI system.

3) Preceding deployment and following the testing phase is the **model understanding phase**. This phase is dedicated to unraveling the underlying mechanisms of the model's decision-making process. The primary inquiry in this phase revolves around explicating how the model formulates its predictions based on the input features. This process includes the interpretation of significant features and their interactions within the model, comprehension of the patterns learned by the model, identification of any blind spots, and the assessment of potential biases in the data. This stage ensures that such biases are not carried forward by the model into its predictive functionality. The understanding phase plays a critical role in safeguarding the AI system from potential pitfalls such as data leakage and bias issues. By providing a deeper insight into the model's operations, this phase contributes to enhancing the trustworthiness and robustness of the AI system prior to its full deployment.
4) The **deploying phase** signifies the transition of the learned model from the development environment into the production system. This critical juncture exposes the model to a new, previously unseen dataset, marking the commencement of the model's active function in real-world scenarios. Within this stage, the model is expected to perform its predictive duties effectively on the fresh data, delivering its predictions with a quantified confidence measure. These generated insights are then consumed and interpreted by system experts, who relay actionable information to the end users. Therefore, the deploying phase represents a significant milestone where the AI system transitions from a theoretical and development context to a practical and operational one.
5) Following the deployment, the **explaining phase** takes precedence, the primary goal of which is to elucidate how the model arrives at predictions when exposed to novel data within the production environment. Interpreting the model's decisions on this new data allows the insights generated by the model to be exposed, if necessary, to expert users. These users may challenge or require further information regarding the decisions made by the deployed model. Additionally, the explaining phase seeks to construct a human-readable explanation that can be disseminated to a wider audience of end users within the AI system. Such an interpretation step is crucial for addressing potential regulatory noncompliance and enhancing the transparency and trustworthiness of the AI system. By ensuring that model predictions can be understood by various stakeholders, the explaining phase fosters greater accountability and reliability within the deployed AI system.
6) The final phase within the robust AI system is the **monitoring phase**, executed within the production environment. The primary objective of this phase is to continually observe the data distribution as well as the performance of the deployed model. This tracking process is instrumental in identifying any changes in the data distribution or potential dips in the model's performance. Should such deviations occur, it necessitates a return to the learning phase. During this cycle, the new data gleaned from the production environment is incorporated into the existing datasets to retrain the models. The monitoring phase, therefore, provides a feedback loop within the AI system, ensuring that the model continues to learn, adapt, and maintain its performance and relevance in the face of evolving data trends. This iterative process contributes to the robustness and longevity of the AI system in the production environment.

3.5 Causality-Based Fairness Methods

This section provides a short introduction to structural causal models (SCM) (Pearl, J., 2009) and how they can help explain fairness issues in a ML context. In order to avoid abstract explanations throughout this subsection, a reduced toy set of variables from the credit scoring and loan approval case study from section 5.3 Automated Credit Scoring is used to illustrate causal modelling, where a series of variables such as current income, credit history, current employment, loan amount repayment period, and gender as protected attribute, predict the binary target variable if the loan is approved or not by the banking institution. The objective is not to give a comprehensive overview of structural causal models built on Bayesian networks (Koller, D., 2009), but to set the context for the bias and fairness discussions from a different angle.

SCMs offer a powerful approach to encapsulating and interrogating the cause-and-effect relationships among a set of variables (Pearl, J., 2009). An SCM comprises two elements: structural equations and a graph. Structural equations dictate how each variable is generated as a function of its direct causes (parents) and an independent error term. The graph, expressed as a Directed Acyclic Graph (DAG), symbolizes the variables as nodes and causal relationships as directed edges.

In the *credit scoring example* with the binary loan approval target, the nodes and edges could be expressed as follows in figure 3-3:

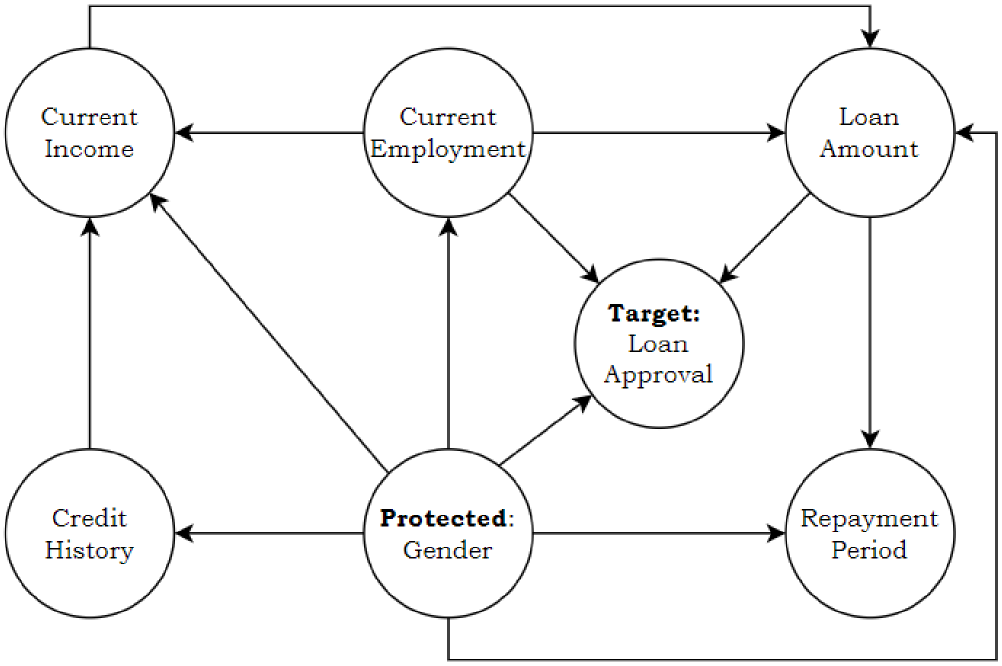


Figure 3-3: Structural Causal Model

In this SCM, variables are connected by functional relationships which represent potential causal relationships, and the noise terms represent unobserved factors that also influence these variables.

Gender is not a function of other variables, it is an intrinsic attribute of the individual, so $gender = f_0(U_0)$.

Credit history might be dependent on gender, due to societal biases and some unobserved factors like the individual's financial habits or unexpected expenses. Thus, it can be modeled as $credit\_history = f_1(gender, U_1)$.

Current income could be dependent on credit history and gender as people with good credit history may have stable income sources, and there might be a gender income gap due to societal biases. It is also dependent on some unobserved factors such as their education, skills, etc. Therefore, $current\_income = f_2(credit\_history, gender, U_2)$.

Current employment is dependent on gender, due to potential hiring biases, and unobserved factors like the local economy, the individual's professional skills, and their field of work, so it could be modeled as $current\_employment = f_3(gender, U_3)$.

Loan amount could be influenced by current income, current employment and gender as individuals with higher income and stable jobs might apply for larger loans, and there could be biases in how much loan men and women apply for. There could also be other unobserved factors like the purpose of the loan, so $loan\_amount = f_4(current\_income, current\_employment, gender, U_4)$.

Repayment period could depend on loan amount and gender as larger loans typically have longer repayment periods, and women may choose longer repayment periods due to income instability. There could also be other unobserved factors like the individual's personal choice or the bank's policies, thus $repayment\_period = f_5(loan\_amount, gender, U_5)$.

Finally, loan approval could depend on current employment, loan amount, gender and unobserved factors like the bank's internal policies or the credit officer's judgement, and hence, it can be modeled as $loan\_approval = f_6(current\_employment, loan\_amount, gender, U_6)$.

In the language of SCMs, when we say we "do" or "intervene on" a variable, we mean we actively manipulate or set its value, which is represented by the $do(variable = x)$ operator in the causal model. For instance, the conditional distribution of loan approval given a specific action or intervention on current employment and loan amount can be denoted as $P\big(\, loan\_approval \mid do(current\_employment = x, loan\_amount = y) \big)$.

This formulation provides a way to express the idea of making an intervention in the system, as opposed to just passively observing it. With SCMs, it is possible to model complex phenomena, examine potential causal relationships, make interventions, and predict the outcomes of those interventions. In the context of fairness in AI, SCMs provide a way to make explicit the potential influence of a protected attribute (like gender in this example) on the target variable, offering a basis for mitigating bias.

3.5.1 Basic Causal Structures

In the context of causal analysis, the understanding of basic causal structures: **chains**, **common cause** (also known as forks), and **common effect** (also known as colliders) is pivotal. These structures describe different configurations of causal relationships among three variables and serve as the elementary building blocks for more complex causal diagrams (Spirtes, P., et al., 2001).

- **Causal Chains** are sequential relationships where one variable influences another, which in turn affects a third variable. In the credit scoring model described above, an example of a causal chain is credit_history -> current_income -> loan_amount. Here, the credit history of an individual may impact their current income, which subsequently influences the loan amount they are approved for.
- **Common Cause** structures (or **forks**) occur when a single variable influences two other variables. For instance, gender acts as a common cause affecting both credit_history and current_employment in the credit scoring model. In such a scenario, it is important to consider that any observed correlation between credit_history and current_employment might be due to their common cause gender and not a direct causal relationship between them.
- **Common Effect** structures (or **colliders**) involve a situation where two variables exert an influence on a third one. A common effect in the credit scoring model can be seen in the current_employment and loan_amount variables that together affect loan_approval. In this situation, loan_approval is the common effect of current_employment and loan_amount.

These basic causal structures provide a preliminary understanding of how variables may causally relate to each other. By identifying these structures in a causal diagram, it is possible to better understand the complex interplay of factors that lead to the outcome of interest.

3.5.2 Relationship Structures

When delving deeper into the realm of causal modelling, complex relationships are encountered that involve more than three variables. Concepts like **d-separation**, **backdoor paths**, and **confounders** become indispensable for identifying and addressing bias (Pearl, J., 2009).

D-Separation is a criterion used to determine whether a set of variables are conditionally independent given a particular conditioning set in a given causal diagram. To put it simply, if two nodes in a graph are d-separated, then changes in one node do not influence the other. In our credit scoring example, given the nodes current income and loan amount, when we condition (or observe) on current employment, these two nodes are d-separated, implying that current income does not influence loan amount when current employment is observed.

Figure 3-4 illustrates how the different structures are connected or separated depending on:

- The structure itself (causal chain, common cause, common effect), and
- The observed variable.

Therefore, if the loan amount is not observed in a causal chain the structure is connected but becomes separated as soon as the loan amount is known. Similarly, in a common cause, if the current employment is not known the structure is connected and becomes separated as soon as the current employment is observed. In the common effect structure, however, the logic is reversed, i.e., if the variable is observed the structure is connected and separated if not observed ((gray-shaded when observed).

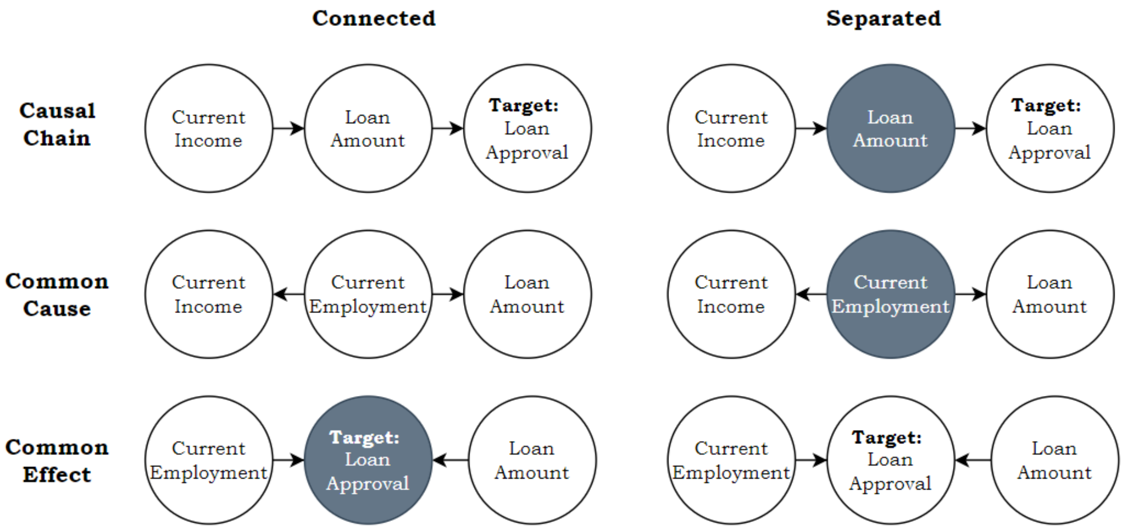


Figure 3-4: Connected and Separated Causal Structures

A **backdoor path** is a sequence of steps along the edges in a causal graph, originating from the treatment (T) and ending at the outcome (Y), which starts with an edge pointing towards T (hence "backdoor"). These paths can introduce spurious associations between T and Y if they're not correctly accounted for, causing bias in estimating the treatment effect.

In a causal diagram, a path is blocked if it includes any of the following:

a) A causal chain (A->B->C) or common cause (A<-B->C) structure with the middle node being observed or conditioned upon.
b) A common effect (A->B<-C, often called a "collider") structure with the middle node being unobserved or unconditioned upon.

Backdoor paths can contain unblocked causal chains and unblocked common cause motifs, i.e., motifs where the middle node is not observed. The inequality between the interventional distribution $P(Y|do(t))$ and the associational distribution $P(Y|T)$ due to unblocked backdoor paths is known as confounding bias. The variables that form the middle node of common cause structures along a backdoor path are referred to as confounding variables or confounders.

This can be illustrated with the credit scoring example in figure 3-5 with unobserved common cause or observed common effect:

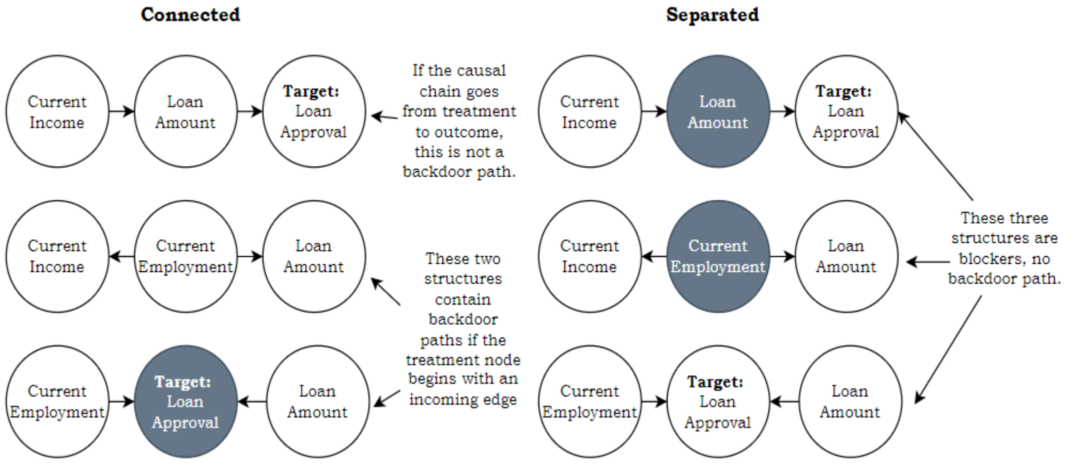


Figure 3-5: Backdoor Paths via Common Cause or Common Effect

For illustrative purposes, the original SCM is slightly reduced to show the two backdoor paths when conditioning (or observing) on current income as illustrated in figure 3-6:

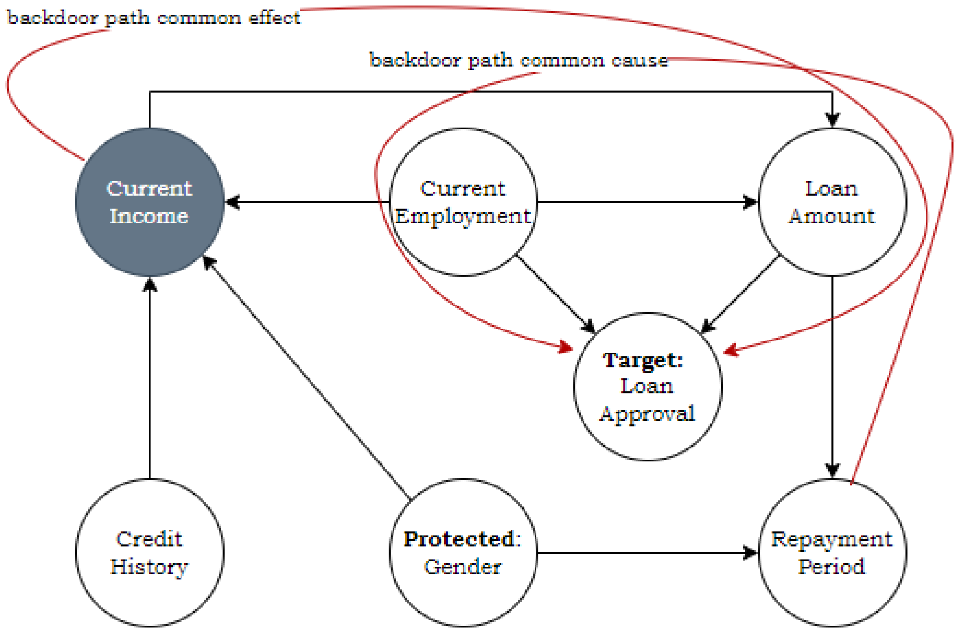


Figure 3-6: The Two Backdoor Paths in the Loan Approval Example

It is now assumed that gender only influences the current income which in turn holds for almost all industrialized economies[8]. If the current income is observed, the common effect structure gets connected and both credit history and gender are confounders along the path towards the target loan approval. Likewise, if current employment is not observed as a common cause, a backdoor is opened through this structure.

3.5.3 Types of Data in Causal Inference

In causal inference, it is important to distinguish between two types of data: **observational and interventional data** (e.g., Hernán, M., et al., 2020). The fundamental difference between these two types of data lies in the way they are generated.

Observational data are the kind of data collected in natural circumstances, without any interference in the system from which the data are drawn. For instance, when a bank collects data about customers' gender, age, credit history, current income, loan amount, and approval decisions in their everyday operations, that's observational data. This data reflects the current policies and patterns of the bank.

[8] World Economic Forum: Global Gender Gap Report 2022, https://www.weforum.org/reports/global-gender-gap-report-2022

Interventional data, on the other hand, come from situations where some variable in the system has been manipulated, typically in an experimental setting. If the bank decided to test a new policy and actively adjusted customers' loan amounts based on a specific criterion, the data collected under this new policy would be considered interventional data.

The importance of this distinction comes to light when we consider the Structural Causal Models (SCMs). An SCM describes the data generating process under all possible interventions, not just the ones that have occurred naturally. The relationship between variables in an SCM is given by a set of equations that represent how each variable is influenced by its parents. When we intervene on a variable in an SCM, we effectively replace its equation with a new one. This is represented in our SCM by the $do(variable = x)$ operator. For example, $P\big(Y\big|do(T)\big)$ represents the distribution of Y when we set the value of T by intervention.

In our *loan approval example*, an intervention might be something like "what would happen to the loan approval decisions if we were to change the loan amounts while keeping other factors the same?" To answer this question, we would use interventional data or apply methods that can emulate interventional data from observational data, such as adjusting for confounders.

The crucial point is that while observational data allow us to estimate correlations, they often fall short of providing causal relationships because of potential confounding. Interventional data, on the other hand, gives us the ability to uncover causal effects directly, but they are more challenging to obtain due to the practical and ethical considerations involved in intervening in real-world systems.

3.5.4 Causal Discovery

Causal discovery is a statistical process that involves deducing the causal relationships among different variables in a system. The key idea is to identify the structural connections between variables that are capable of explaining the observed correlations within the data, with the ultimate aim of building a causal graph. Such a graph can represent the data-generating process and aid in making predictions and interventions. Causal discovery is broadly divided into two primary branches: methods based on conditional independence testing and methods that rely on explicit assumptions about the form of the underlying causal mechanisms.

Conditional Independence Testing

The first branch is based on the principle of conditional independence testing (Spirtes, P., et al., 2001, PC algorithm), relying on a concept known as the "faithfulness" or "stability" assumption. This concept posits that the observed dependence and independence relationships among the variables reflect the underlying causal structure.

Under this assumption, the goal is to find a graph (or, more precisely, an equivalence class of graphs known as a "Markov equivalence class") such that the graph's implied conditional independence relationships match those observed in the data. This approach is often referred to as constraint-based causal discovery. However, the fundamental limitation of this approach is that it typically yields a class of equivalent models that cannot be distinguished based on the observed data.

Causal Discovery Based on Explicit Assumptions

The second branch is based on making explicit assumptions about the form of the causal relationships. For example, some methods assume that causal relationships are linear or involve additive noise. Other methods may use the entropy of the distribution of the residuals to distinguish between alternative models. This approach is often referred to as score-based causal discovery, as the aim is to find a model that best fits the data according to some scoring criterion. Unlike the conditional independence testing approach, these methods can yield a unique solution. However, they require stronger assumptions.

In our loan approval example, one might start with a list of potential variables - including gender, credit history, current income, loan amount - and use a causal discovery algorithm to infer a plausible causal graph. The resulting graph can provide valuable insights into the causal mechanisms that underlie loan approval decisions. It is important to note, however, that any discovered graph should be interpreted cautiously. It provides a hypothesis about the causal structure, but this hypothesis needs to be validated using additional information, preferably from domain knowledge or intervention experiments.

3.5.5 Causal Inference

Causal inference involves estimating the effect of interventions from observational data. While the discovery process provides a hypothesis about the causal structure, it is often silent about the size or strength of the causal effects. Causal inference techniques fill this gap, allowing us to make quantitative predictions about the effect of potential interventions (Hernán, M., et al., 2020). One of the most common methods used in causal inference is adjustment for confounders. The idea here is to control for variables that can induce spurious correlations between the treatment and the outcome, as illustrated earlier with the loan approval backdoor path examples.

Another approach is to use **propensity score** methods, which estimate the probability of receiving the treatment given the observed covariates. The propensity score can then be used to balance the treatment and control groups on the observed covariates, thereby reducing bias due to confounding. Detailed descriptions of causal inference go beyond the scope of this project, as the focus lies on fairness measures.

3.5.6 Techniques to Improve Fairness

In the field of fairness in ML, it is crucial to consider the underlying causal structure, since it directly influences the patterns of disparity and discrimination in predictions. The methods of imposing fairness constraints have been proposed in the context of causal models to ensure fair outcomes (Kusner, M. et al., 2017) and are Causal Fairness Constraints and Counterfactual Fair Loss Function

Causal fairness constraints leverage the causal structure of the data to provide fairness guarantees. The central idea is to restrict the learning of predictive models such that they only leverage the causal effects of the sensitive attributes that are considered fair according to some fairness criterion.

Using the loan approval process as an example, suppose we want our model to be fair with respect to gender. A causal fairness constraint could ensure that the effect of gender on the loan approval decision, through paths other than the legitimate ones (e.g., through credit history), is eliminated. Thus, any disparity in loan approval rates between men and women can only be explained by these legitimate factors. The advantage of using causal fairness constraints is that they directly target the mechanisms that lead to unfair outcomes. However, the application of these methods requires knowledge about the causal structure, which is often not fully known in practice.

Counterfactual fairness is a fairness criterion based on counterfactual reasoning. A predictor is said to be counterfactually fair if the prediction for an individual would have been the same in a counterfactual world where the individual's sensitive attribute was different.

Counterfactual fairness can be achieved by constructing predictors that are functions of the "fair" variables. In the reduced loan approval example, a fair variable might be the credit history, which is not directly influenced by gender. A counterfactually fair predictor would only use information from these fair variables to make its prediction.

To apply counterfactual fairness, a causal model needs to be specified that includes counterfactual variables. These are variables that capture what would have happened under different interventions. In the loan approval example, a counterfactual variable might be the credit history that a person would have had if they had been of a different gender. In that regard, it is not enough to simply change the gender which is commonly called “attribute flipping”.

Both causal fairness constraints and counterfactual fairness provide a principled way of ensuring fairness in predictions. They leverage the power of causal reasoning to go beyond mere statistical parity, allowing us to capture more nuanced notions of fairness (Kilbertus, N. et al., 2017).

Causality-based methods for fairness have become increasingly critical due to their ability to understand and act upon the causal structures that produce unfair outcomes. They move beyond correlation-based approaches, that treat observed variables as independent entities, and delve into the underlying mechanisms that drive relationships between these variables. By doing so, they enable a more nuanced understanding of fairness and discrimination, addressing the sources rather than the symptoms of unfairness.

When using ML models for decision-making processes, such as in the loan approval example, fairness is a central concern. Traditional methods for enforcing fairness primarily focus on ensuring statistical parity, which entails that the decision outcomes are balanced across different sensitive groups. However, such approaches are limited in that they do not account for the inherent complexities and causal structures present within real-world data. Causality-based methods for fairness take into account these underlying structures and their impacts on outcomes. They provide a framework to analyze and mitigate direct and indirect forms of discrimination, allowing for a more comprehensive fairness analysis.

3.6 Tool-based Bias Mitigation

In the realm of AI systems, the challenge of mitigating bias and ensuring fairness is substantial. Given the increasing integration of AI in decision-making processes, from loan approval to job recruitment, this challenge necessitates effective and efficient tools for bias detection and mitigation. Consequently, several state-of-the-art toolkits have been developed by leading tech companies and academic institutions to address these issues. Toolkits such as IBM's AI-Fairness 360, Google's What-if Tool, Facebook's Fairness Flow, Carnegie Mellon's Aequitas, and Themis AI offer a variety of features and capabilities designed to improve the fairness of AI systems. This section will delve into the specifics of these toolkits, exploring their methodologies, functionalities, and potential application areas, thus providing a comprehensive understanding of the current landscape of tool-based bias mitigation in AI systems.

3.6.1 IBM: AI-Fairness 360

IBM's AI-Fairness 360 (AIF360) serves as an open-source library designed to facilitate the detection and mitigation of biases within ML models.[9] As a Python package, it furnishes an extensive array of fairness metrics alongside bias detection and mitigation algorithms. Its applicability primarily rests with allocation or risk assessment problems where protected attributes are clearly demarcated. It therefore acts as a springboard for dialogue among stakeholders concerning decision-making workflows.

Selecting suitable fairness metrics and bias mitigation algorithms from the extensive offerings of AIF360 depends on the specific application under consideration. Fairness metrics can be broadly categorized into two forms: individual and group fairness metrics. Furthermore, the latter can be segmented into metrics applicable to training data and those relevant to models. The selection of metrics largely hinges on the user's worldview, falling broadly under "we're all equal" (WAE) or "what you see is what you get" (WYSIWYG). Both versions of metrics—difference and ratio—are accessible, subject to user preference. The toolkit encompasses the following metrics and classes:

1. Metrics:
   1.1 Individual Fairness: Use SampleDistortionMetric class.
   1.2 Group Fairness:
   1.3 Training data: Use DatasetMetric class (and its children classes, such as BinaryLabelDatasetMetric).
   1.4 Models: Use ClassificationMetric class.

2. Group Fairness Metrics:
   2.1 We're All Equal (WAE) worldview:

[9] IBM used to provide an open web access to the AI Fairness 360 toolkit at http://aif360.mybluemix.net/, however, in June 2023 it was removed without any clear explanations. The source code of the library can still be accessed at https://github.com/Trusted-AI/AIF360 and https://pypi.org/project/aif360/.

- 2.1.1 Demographic parity metrics: disparate_impact and statistical_parity_difference.
- 2.2 What You See Is What You Get (WYSIWYG) worldview:
  - 2.2.1 Equality of odds metrics: average_odds_difference and average_abs_odds_difference
- 2.3 In-between worldviews:
  - 2.3.1 Other group fairness metrics (including some labeled as equality of opportunity): false_negative_rate_ratio, false_negative_rate_difference, false_positive_rate_ratio, false_positive_rate_difference, false_discovery_rate_ratio, false_discovery_rate_difference, false_omission_rate_ratio, false_omission_rate_difference, error_rate_ratio, and error_rate_difference.

Bias mitigation algorithms within AIF360 are stratified into pre-processing, in-processing, or post-processing, contingent upon the user's capacity to intervene within the ML pipeline. The earliest possible intervention category is recommended by AIF360. The selection of algorithms is influenced by a multitude of factors, such as dataset attributes, transparency prerequisites, and the specific fairness metric to be optimized. The consequence of ameliorating one fairness metric on others can be intricate.

Figure 3-7 ((Bellamy, R. et al., 2018) shows the typical so-called fairness pipeline where fairness measures or bias mitigation techniques can be applied during pre-, in-, and post-processing:

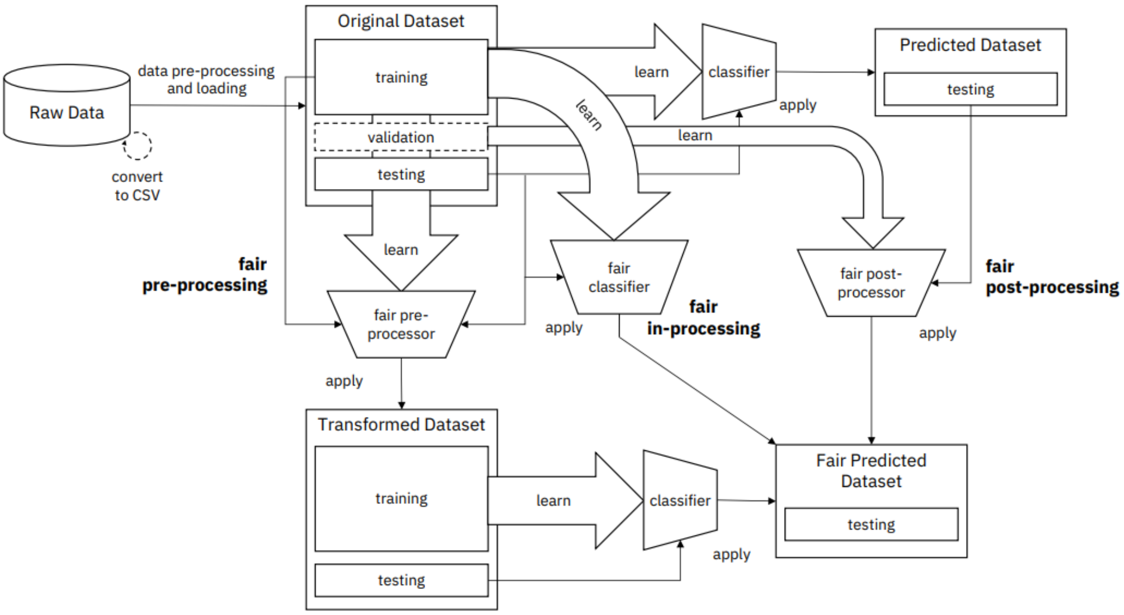


Figure 3-7: AI 360 Fairness Pipeline

The available fairness measures are listed below per phase:

1. **Pre-processing:**
   1.1 Reweighing: Changes weights applied to training samples without altering feature or label values.
   1.2 Disparate Impact Remover: Yields modified datasets in the same space as the input training data.
   1.3 Optimized Pre-processing: Addresses both group fairness and individual fairness, while providing transparency on the transformation.
   1.4 Learning Fair Representations (LFR): Transforms dataset into a latent space.
2. **In-processing:**
   2.1 Prejudice Remover: Limited to learning algorithms that allow for regularization terms.
   2.2 Adversarial Debiasing: Allows for a more general set of learning algorithms.
3. **Post-processing:**
   3.1 Equalized Odds Post-processing: Two algorithms with a randomized component.
   3.2 Reject Option: A deterministic algorithm.

Some of the AIF360 algorithms (for instance, optimized pre-processing and reject option) require the specification of which fairness metric to optimize, whereas others (for instance, disparate impact remover and equalized odds post-processing) do not have such a requirement. An enhancement in one fairness metric might have convoluted repercussions on other fairness metrics.

It is noteworthy that the selection of a mitigation algorithm is not necessarily a straightforward task due to its dependence on numerous variables, such as the characteristics of the dataset, the requirements for transparency, and the specific fairness metric that is being optimized. As such, the careful evaluation of these factors is recommended in the selection process.

For instance, the pre-processing algorithm 'Reweighing' changes weights attached to training samples without altering feature or label values, making it a good choice for scenarios where there's a need to maintain the original data integrity. On the other hand, 'Disparate Impact Remover', another pre-processing algorithm, yields modified datasets in the same space as the input training data, making it a viable choice when it is acceptable to modify the data to achieve fairness.

In the in-processing category, the 'Prejudice Remover' is limited to learning algorithms that allow for regularization terms. It is therefore most suitable for datasets and models where adding a regularization term is feasible. In contrast, 'Adversarial Debiasing' allows for a more general set of learning algorithms, thus providing a more versatile option for a wide range of datasets and models.

In the post-processing category, 'Equalized Odds Post-processing' includes two algorithms with a randomized component. It is best suited for applications that can tolerate a degree of randomness in the output. The 'Reject Option', in contrast, is a deterministic algorithm that is suitable for use cases where determinism in the outcome is of prime importance.

The aforementioned algorithms within the AIF360 toolkit, however, may not always lead to the perfect balance in fairness metrics. This is due to the fact that improving one fairness metric may have complicated, and often unintended, effects on other fairness metrics. This trade-off between different fairness metrics is an important factor to consider when choosing and applying these algorithms, and in many cases, it might require a careful and systematic evaluation to achieve the desired balance in fairness.

3.6.2 Google: What-if Tool

The Google What-If Tool[10] is an integral part of the open-source TensorBoard web application and it aims to facilitate the task of analyzing ML models without the necessity for code. By offering an interactive visual interface for scrutinizing the input and output of ML models and their results, this tool empowers users to evaluate model performance across various data segments, dissect individual data points, and assess the effect of theoretical changes to input features.

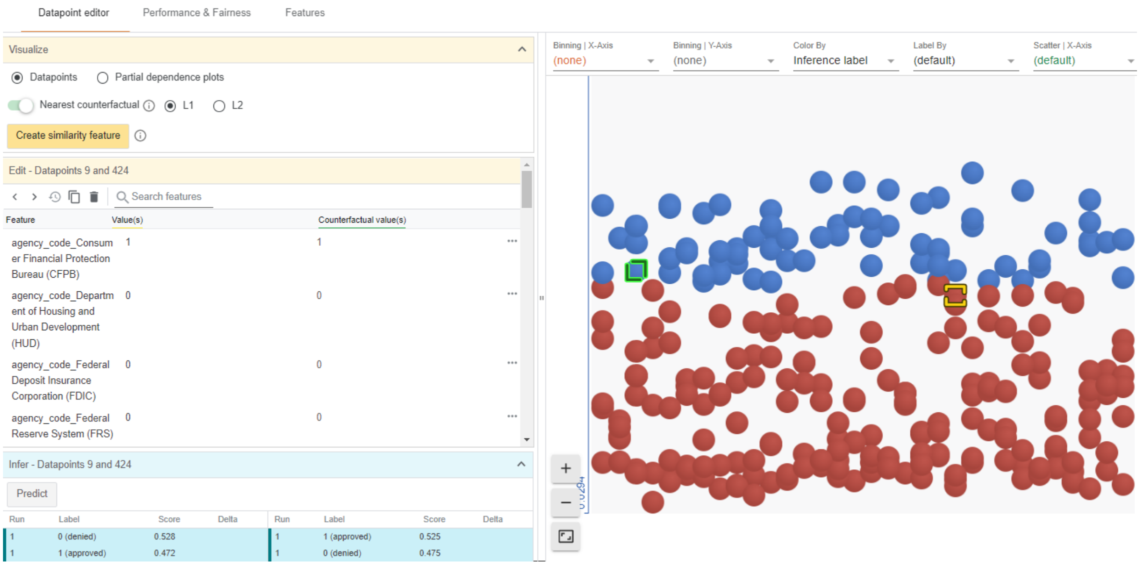
Inherent to the What-If Tool are numerous features such as the Facets-driven automatic visualization of datasets, the capability to manually edit examples within datasets to observe the consequential effects, and the automatic generation of partial dependence plots that illustrate the alteration of model predictions corresponding to changes in single features.

Moreover, the tool provides support for counterfactual analysis, which enables users to juxtapose a datapoint to the most analogous point with a divergent model prediction. This assists in the understanding of the model's decision boundaries. In the context of the What-If Tool, counterfactuals represent the identification of the closest datapoint to a specific instance where the model offers a different prediction. Through the examination of counterfactuals, insights into the decision boundaries of the model can be attained, along with an understanding of the factors impacting its predictions.

For instance, users can manually modify a datapoint to observe how these alterations influence the model's prediction. This practice assists in the identification of critical features that are instrumental in a model's decision-

[10] The main documentation can be found at https://cloud.google.com/ai-platform/prediction/docs/using-what-if-tool and on Github https://pair-code.github.io/what-if-tool/.

making process, consequently leading to an improved understanding of the model's behavior. Figure 3-8 illustrates this with a mortgage example.



Edit - Datapoints 9 and 424

Feature	Value(s)	Counterfactual value(s)
agency_code_Consumer Financial Protection Bureau (CFPB)	1	1
agency_code_Department of Housing and Urban Development (HUD)	0	0
agency_code_Federal Deposit Insurance Corporation (FDIC)	0	0
agency_code_Federal Reserve System (FRS)	0	0

Infer - Datapoints 9 and 424

Run	Label	Score	Delta	Run	Label	Score	Delta
1	0 (denied)	0.528		1	1 (approved)	0.525	
1	1 (approved)	0.472		1	0 (denied)	0.475	

Binning | X-Axis (none)
Binning | Y-Axis (none)
Color By Inference label
Label By (default)
Scatter | X-Axis (default)

Figure 3-8: Google What-if: Mortgage Example for Counterfactuals

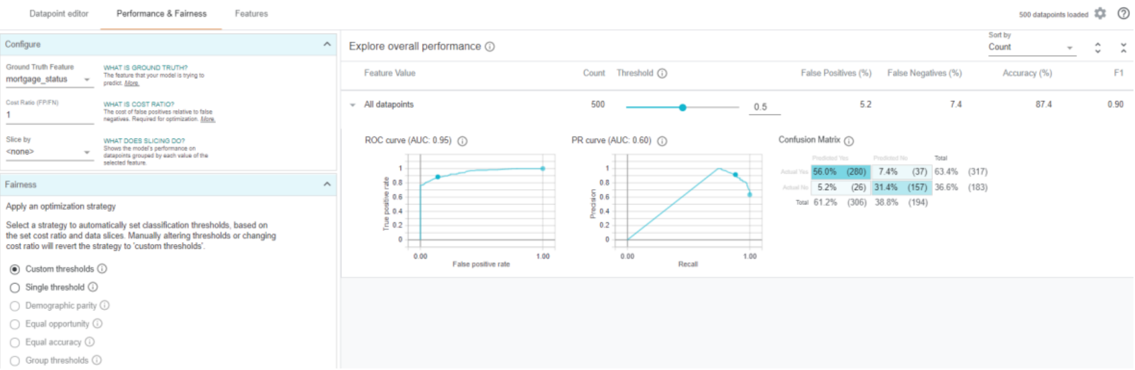
Suppose a user is examining a ML model built to determine the likelihood of mortgage approval for an applicant based on various input features such as income, credit score, agency, employment history, and the size of the loan. A specific instance - say, an applicant who was denied a mortgage - can be selected. The user can then employ the counterfactual analysis feature of the What-If Tool to find the most similar applicant who was approved a mortgage by the model.

This helps to identify the specific features that might be influencing the model's decisions. For instance, it might reveal that a small increase in income or a marginal improvement in credit score could have resulted in mortgage approval for the original applicant. This analysis can be further deepened by manually altering datapoints, such as income or credit score, to observe how these changes affect the model's prediction.

The manual exploration of counterfactuals, as facilitated by the What-If Tool, uncovers critical insights into the feature importance and decision-making process of the model. In our mortgage example, such insights could be crucial for understanding whether the model is making fair or biased decisions and could inform necessary improvements or alterations to the model to mitigate any identified bias.

In addition to counterfactual analysis, the 'Analysis of Performance and Algorithmic Fairness' feature of the What-If Tool allows users to examine the impact of different classification thresholds on diverse numerical fairness criteria as shown in figure 3-9. For instance, users might want to compare the model's performance when trained on two different slices of data - say, applicants from urban areas versus rural areas - and adjust the classification thresholds to satisfy certain fairness constraints, such as equal opportunity.

This could reveal if the model is consistently fair across different demographic groups or if it is biased towards one group over another.



Ground Truth Feature: mortgage_status

WHAT IS GROUND TRUTH? The feature that your model is trying to predict. More.

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. More.

Slice by: <none>

WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

- ● Custom thresholds ⓘ
- ○ Single threshold ⓘ
- ○ Demographic parity ⓘ
- ○ Equal opportunity ⓘ
- ○ Equal accuracy ⓘ
- ○ Group thresholds ⓘ

Sort by: Count

Feature Value	Count	Threshold ⓘ	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	500	0.5	5.2	7.4	87.4	0.90

	Predicted Yes		Predicted No		Total	
Actual Yes	56.0%	(280)	7.4%	(37)	63.4%	(317)
Actual No	5.2%	(26)	31.4%	(157)	36.6%	(183)
Total	61.2%	(306)	38.8%	(194)		

Figure 3-9: Google What-if: Mortgage Example for Metrics and Performance

The Google What-If Tool's interactive and visually driven approach not only simplifies the process of evaluating and debugging ML models but also provides users with an intuitive understanding of how their models operate, respond to changes, and make decisions. Therefore, it serves as an instrumental resource in developing transparent, fair, and reliable ML systems.

3.6.3 Facebook / Meta: Fairness Flow

Fairness Flow[11] is a proprietary tool developed by Facebook, presently known as Meta, that functions as an aide to data scientists and engineers in their pursuit of ensuring their ML models' fairness. This instrument offers a selection of metrics that allow users to ascertain fairness and identify potential biases that may have inadvertently crept into their models. The tool is part of Meta's broader commitment to responsible AI technologies, which underscores the development of inclusive models that offer effective solutions and just treatment to all individuals and communities.

The primary utility of Fairness Flow lies in its capability to detect statistical biases in artificial intelligence models and labels within Meta's internal applications. It attempts to highlight instances of model and label biases, which can transpire when a model systematically misestimates outcomes for different groups, or if human labelers impose inconsistent standards.

Embedded within the Python library, Fairness Flow provides an application programming interface (API) for analyzing performance and fairness metrics. This tool facilitates an examination of the performance of models or human-labeled training data across different groups, enabling engineers to discern if enhancements are necessary to ensure an equitable performance. The process may involve adjustments to the training or test dataset, an examination of the

[11] In-depth documentation can be found at https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/ and https://ai.facebook.com/resources/frameworks-and-tools/

prominence of specific features, or exploration of varying degrees of model complexity.

The operational mechanism of Fairness Flow involves segmenting the data into pertinent groups and subsequently calculating the model's performance for each group. The tool then scrutinizes several fairness metrics such as the representation of each group in the dataset, the model's proficiency in content classification or ranking, and whether the model over- or under-predicts for specific groups. Variations in performance across these groups might indicate fairness concerns that necessitate further investigation.

Additionally, Fairness Flow assesses potential bias in labels by comparing labels assigned by annotators with labels produced by experts, assuming the latter to be the ground truth. This comparison aids in determining the accuracy of the labeling process and any biases that may have been introduced.

Despite its strengths, Fairness Flow does have limitations. It cannot analyze all types of models, and the approach to fairness can differ depending on the goals of the AI system. The choice of the appropriate metric relies heavily on the specific product, its context, and the potential impacts of incorrect predictions on users and vulnerable groups.

It is crucial to note that Fairness Flow is an internal tool, and the external evaluation of its effectiveness depends solely on the communications published by Meta. Despite Meta's assurances, there is some skepticism regarding the tool's ability to completely address bias and fairness issues (e.g. Greene, T., 2021).

3.6.4 Carnegie Mellon: Aequitas

Aequitas[12] was created by the Center for Data Science and Public Policy at the University of Chicago, and is maintained by Carnegie Mellon, to promote the use of data science in policy research and practice. Their work includes education, data science projects with various partners, and developing new methods and open-source tools for data-driven public policy and social impact in a fair and equitable manner.

Aequitas is an open-source bias audit toolkit that automates the process of assessing fairness in binary classification models. It provides a flexible and extensible way to measure fairness and helps data scientists and policymakers understand, communicate, and act on algorithmic bias.

It is a tool designed to audit risk assessment systems for two types of biases: biased actions or interventions and biased outcomes. To conduct these audits, data is needed about the overall population, protected attributes (e.g., race, gender, age, income), the set of individuals recommended for intervention or action, and actual outcomes for selected and non-selected individuals.

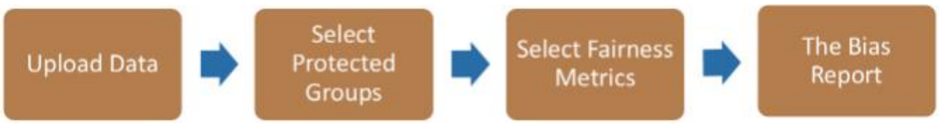
Aequitas can be utilized through three methods: the Web Audit Tool, which generates a bias report; the Python Library, which allows for the generation of

[12] A general introduction can be found at http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/, the online web tool at http://aequitas.dssg.io/ and the Python libraries on Github https://github.com/dssg/aequitas.

bias and fairness metrics on data and predictions; and the Command Line Tool, which produces a report using one's own data and predictions.

The tool provides a bias report, detailed fairness and bias statistics, and an interactive bias dashboard.

The **Web Audit Tool** provides a first and reduced approach to bias and fairness via a series of metrics. Anyone can carry out the four basic steps at the link http://aequitas.dssg.io/, i.e.:



1) **Upload Data:**
The web tool only allows for binary target (label_value) and predicted (score) variables. All other features need to be categorical or are discretized into bins if continuous values are provided.
2) **Select Protected Attributes:**
One or several protected attributes (race, sex, age, etc.) can be selected manually, by majority group for every attribute, or automatically by choosing the group with the lowest bias metric.
3) **Select Fairness Metrics:**
A few fairness metrics can be selected. The complete list comprises Equal Parity, Proportional Parity, False Positive Rate Parity, False Discovery Rate Parity, False Negative Rate Parity, False Omission Rate Parity. Additionally, a fairness threshold (disparity intolerance) can be set, being defaulted to 80% as this is the threshold usually applied from a legal perspective.
4) **The Bias Report:**

All output fairness metrics are explained per protected attribute chosen in step 2, explaining what the metric means, why it matters, and which groups failed the test. The following so-called fairness tree in figure 3-10 provides an overview of the possible metrics of the bias report:

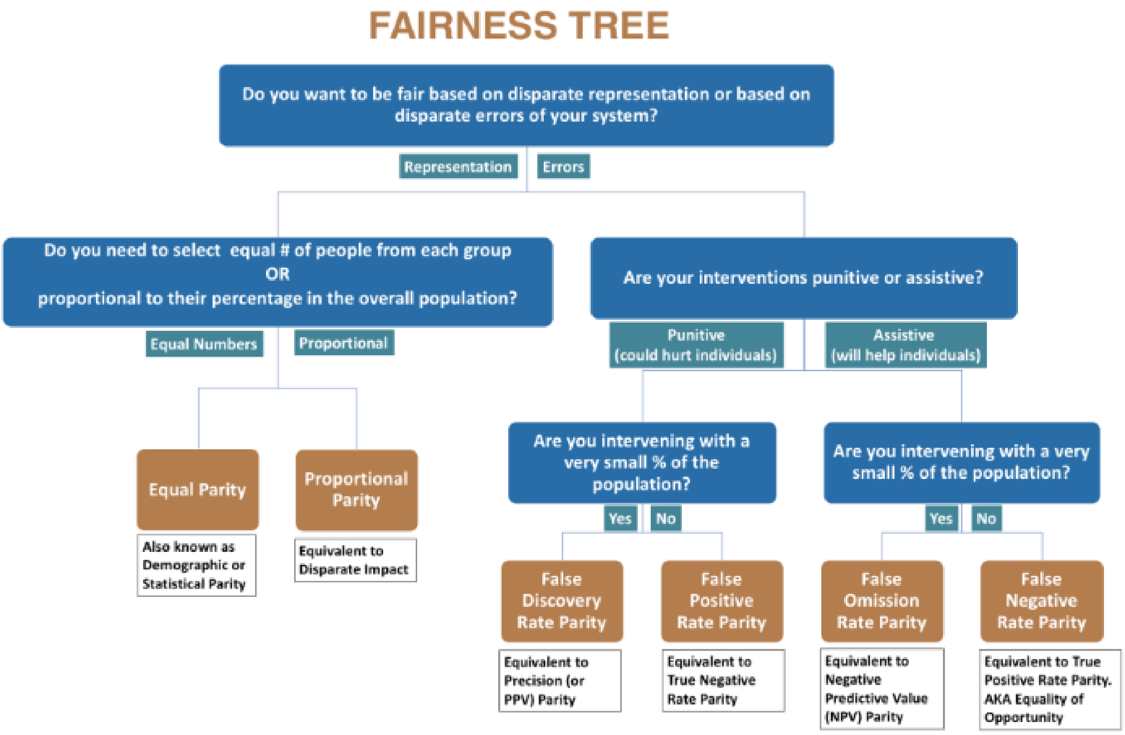

FAIRNESS TREE
Do you want to be fair based on disparate representation or based on disparate errors of your system?
Representation
Errors
Do you need to select equal # of people from each group OR proportional to their percentage in the overall population?
Are your interventions punitive or assistive?
Equal Numbers
Proportional
Punitive (could hurt individuals)
Assistive (will help individuals)
Equal Parity
Also known as Demographic or Statistical Parity
Proportional Parity
Equivalent to Disparate Impact
Are you intervening with a very small % of the population?
Are you intervening with a very small % of the population?
Yes
No
Yes
No
False Discovery Rate Parity
Equivalent to Precision (or PPV) Parity
False Positive Rate Parity
Equivalent to True Negative Rate Parity
False Omission Rate Parity
Equivalent to Negative Predictive Value (NPV) Parity
False Negative Rate Parity
Equivalent to True Positive Rate Parity. AKA Equality of Opportunity

Figure 3-10: Fairness Tree for the Aequitas Bias Report[13]

This is a simple and straightforward way to check for several basic fairness metrics for binary classifications.

In a similar way, the **Python library** (https://pypi.org/project/aequitas/) also focuses on binary classification, however, 13 instead of only 6 metrics and an increased flexibility for initial data handling are provided. Besides, the data visualization can be enhanced via several out-of-the box graphs such bar charts and treemaps. Nonetheless, the underlying basic models have the same restrictions as the web audit tool, i.e., a simple binary classifier with categorical or discretized continuous variables.

The command-line interface is simply an extension of the previous Python library, i.e., the tool can be accessed interactively through the CLI for the data upload, but the functionality and restrictions remain the same as for the other two utilization methods. Further details can be found at Github (https://github.com/dssg/aequitas).

In summary, Aequitas is an easy-to-use tool and offers some interesting options for a first bias and fairness approach for binary classifiers. Nonetheless, further in-depth analysis is needed for a more comprehensive understanding of fairness issues in AI.

3.6.5 Themis AI

Themis ML is an AI fairness tool which was originally developed by the MIT Computer Science & Artificial Intelligence Laboratory (CSAIL) and allows users

[13] e.g. http://aequitas.dssg.io/audit/7cwgs8fa/compas_for_aequitas/

to evaluate the fairness of their ML models by comparing the model's performance on different subsets of data. It provides a set of easy-to-understand fairness metrics and visualization tools for understanding and improving model fairness (cf. Bantilan, N., 2017).

In the context of Themis ML, discrimination is defined as biased preferences towards or against certain social groups, resulting in unfair treatment regarding specific outcomes. Fairness, on the other hand, is the opposite of discrimination. A ML algorithm is considered fair if its predictions do not favor one social group over another for outcomes with socioeconomic, political, or legal significance, such as loan approvals.

Although Themis ML is still an open-source library and publicly accessible[14] for anyone who wishes to make use of the ML fairness options, it has also emerged as a spin-off company (Themis AI, https://themisai.io/) with a more holistic approach throughout the entire AI cycle, including out-of-the-box industry-focused solutions as the so-called AI Guardian and AI certification programs.

The focus in this research lies on the publicly available open-source Python library, which in turn has been built on top of the Sci-kit Learn, Numpy and Pandas libraries and the corresponding data pipeline interfaces such as:

- Transformer: Preprocess raw data for model training
- Estimator: Train models to perform a classification task
- Scorer: Evaluate performance of different models
- Predictor: Predict outcomes for new data

Therefore, applying Themis-ML can be performed in a similar way to Sci-Kit Learn. The Readthedocs and Github documentation present a rather reduced scope of functionalities (the items marked with an asterisk are listed on the documentation but currently not implemented):

1. Measuring Discrimination:
   1.1 Mean difference: Measures the disparity in outcomes between two social groups.
   1.2 Normalized mean difference: Similar to mean difference but scales values based on the maximum possible discrimination in a dataset.
   1.3 *Consistency:** Compares an observation's target label with those of its neighbors, with lower scores indicating less individual-level discrimination.
   1.4 *Situation Test Score:** Assesses discrimination only for disadvantaged individuals by computing a score between 0 and 1, where 0 indicates no discrimination and 1 indicates maximum discrimination.

2. Mitigating Discrimination:
   2.1 Preprocessing:
      2.1.1 Relabeling (Massaging): Modifies the original dataset by changing some instance labels to achieve a more balanced distribution.

[14] https://pypi.org/project/themis-ml/ and https://themis-ml.readthedocs.io/en/latest/

2.1.2 *Reweighing:** Assigns different weights to the instances in the dataset, emphasizing or de-emphasizing their influence on the learning algorithm.

2.1.3 *Sampling:** Involves oversampling the minority class, undersampling the majority class, or a combination of both to create a balanced dataset.

2.2 Model Estimation:

2.2.1 Additive Counterfactually Fair Estimator: A method that models the relationship between protected attributes and outcomes while accounting for confounding variables.

2.2.2 *Prejudice Remover Regularized Estimator:** A learning algorithm that minimizes discriminatory behavior by adding a penalty term based on the prejudicial impact of the model.

2.3 Postprocessing:

2.3.1 Reject Option Classification: Introduces a decision threshold in the classification process, allowing instances within the threshold to be reconsidered for fairer treatment.

2.3.2 *Discrimination-aware Ensemble Classification:** Combines multiple classifiers in an ensemble, taking into account their individual discriminatory behavior to improve fairness.

The latest library updates date back to February 2018, and as mentioned above, most of the envisioned enhancements have never been implemented, which in turn seems to underpin the fact that further developments have been moved to the spin-off company Themis AI.

3.6.6 Tool Summary Comparison

The previous sections have provided a detailed exploration of various fairness tools developed by diverse entities. To distill the salient aspects of each tool, a comparative analysis is presented in table 3-3, highlighting the specific advantages and drawbacks of each tool. This approach aids in identifying the distinctive features and utility of each tool in different contexts.

A common advantage across many of the tools is the emphasis on ease of use. This aspect is primarily facilitated by interactive interfaces, detailed documentation, and an emphasis on user-friendly metrics. For instance, Google's What-if Tool and MIT's Themis AI prioritize visualization and simple metrics, making them accessible to non-experts. Such features enhance the usability of the tools and encourage broader adoption by data scientists, engineers, and even stakeholders with limited technical expertise.

Another key strength is the comprehensive range of metrics and algorithms that some tools offer. IBM's AI Fairness 360, for example, provides a broad selection of fairness metrics and bias mitigation algorithms. This feature gives users the flexibility to choose the best fit for their specific use-case, thus enhancing the tool's versatility.

Tool	Pros	Cons
IBM: AI-Fairness 360	Comprehensive set of metrics and algorithms	Requires some understanding of fairness metrics and mitigation techniques
	Open-source and well-documented	May require more effort to integrate with existing model pipelines
Google: What-if Tool	Interactive and visual interface	Limited to specific ML frameworks
	Easy to use for non-experts	Primarily focused on understanding model behavior rather than mitigating bias
Facebook: Fairness Flow	Focuses on model evaluation and understanding biases	Limited to specific ML frameworks
	Provides actionable insights	Only for internal Facebook use, no open source
Carnegie Mellon: Aequitas	Automation of bias auditing	Focused on binary classification models only
	Flexible and extensible	Requires deep understanding of fairness metrics
MIT: Themis AI	Easy-to-understand metrics	May not provide as much depth or customizability as other tools
	Visualization tools	Open source but not updated anymore, now also a spin-off company

Table 3-3: AI Fairness Toolkit Comparison

However, these tools are not without their limitations. One common shortcoming is the restrictive applicability of some tools, often limited to specific ML frameworks or types of models. For example, Google's What-if Tool and Facebook's Fairness Flow are only compatible with certain ML frameworks. Furthermore, Aequitas by Carnegie Mellon University primarily focuses on binary classification models, limiting its scope of application.

Another notable constraint is the lack of support for intersectional protected attributes in many tools. This absence can pose a significant challenge as biases often operate at the intersection of multiple attributes, and tools lacking this feature may not fully detect or mitigate such biases.

Lastly, the accessibility of some tools, such as Facebook's Fairness Flow, is restricted as they are for internal use only. This limitation curtails their application beyond the organization that developed them.

In summary, while these tools offer valuable features for fairness evaluation and bias mitigation, they also come with certain constraints that need to be considered. Selecting the right tool would therefore depend on the specific requirements of the project, the technical expertise of the users, and the data and models being used.

4 Ethical, Legal, and Social Implications

Chapter 4 of this analysis focuses on the Ethical, Legal, and Social Implications of artificial intelligence (AI) and AD-MS. This multifaceted exploration is organized into two significant sections.

Section 4.1, "AI Bias and Fairness from a Legal Perspective," delves into a critical analysis of the legal frameworks that address issues of bias, fairness, and transparency in AI, with a particular focus on the EU, and to a lesser extent on the US and China. This includes examining various laws and directives such as the General Data Protection Regulation (GDPR)[15] and the future Artificial Intelligence Act (AIA)[16] in the EU, the Algorithmic Accountability Act (AAA)[17] in the United States, and the Personal Information Protection Law (PIPL)[18] in China. The importance of accountability and transparency in AD-MS is highlighted in this subsection, emphasizing how biases in AI can impact legal aspects and the possible legal consequences therein.

Section 4.2, "Ethical and Social Implications of AI," moves the discussion to the broader societal and ethical aspects of AI implementation and use. This includes a detailed exploration of the potential unintended consequences of AI, such as the amplification of existing societal biases and the potential for discrimination. This discussion will unpack how the current deployment of AI in society could, if not checked, inadvertently exacerbate social inequities.

This section concludes with an overview of the need to strike a balance between accuracy, fairness, and privacy in AD-MS. This delicate interplay involves careful consideration of the trade-offs between these values and emphasizes the necessity for ethical decision-making in the design, development, and deployment of AI. By recognizing and addressing these issues, AI can be better harnessed to benefit society as a whole while minimizing potential harm and injustices.

4.1 AI Bias and Fairness from a Legal Perspective

This section outlines the major efforts of the lawmakers of the main players in the AI space, i.e., USA, China, and the EU. Although the first two have a clear head start, the focus lies on the EU for being the most active region in implementing new laws. Even the major companies in the ascending generative AI market urge to implement new laws and regulations such as Sam Altman, CEO of OpenAI, the creator of ChatGPT, although the reasons might deviate from the fairness discussion in this project.[19]

[15] Regulation (EU) 2016/679 - https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

[16] Proposal for Artificial Intelligence Act - https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

[17] H.R.6580 - Algorithmic Accountability Act of 2022 - https://www.congress.gov/bill/117th-congress/house-bill/6580/text

[18] Personal Information Protection Law - https://www.china-briefing.com/news/the-prc-personal-information-protection-law-final-a-full-translation/

[19] The reasons for these companies might be rather related to create a so-called “moat” as their technology cannot be easily defended due to transfer learning.

Table 4-1 illustrates the recent legislators' efforts to implement a legal AI framework. Only the most prominent laws and initiatives are mentioned, and it is not aimed at providing a comprehensive list.

Law or Directive	Year in Force	Most Relevant Articles	Region	Short Description
General Data Protection Regulation (GDPR)	2018	Article 5(1)(a), 5(1)(c), 22, Article 12, Article 13, Article 14, Recital 71	EU	Ensuring transparency and responsibility, preventing unintended consequences and bias, and balancing accuracy, fairness, and privacy in the processing of personal data.
Digital Services Act (DSA)	2022	Article 11, 12, 13, 14, 15, 24, 27, 30, 42 (Transparency) , Article 16, 17, 20, 21, 25, 28, 31 (Preventing Bias), Article 26 (paragraph 3), 54 (Balancing Accuracy, Fairness, and Privacy)	EU	A regulation aimed at creating a safer digital space with European values and rules at its core.
Artificial Intelligence Act (AIA)	Not in force as of June 2023	Article 10, Article 15, Article 20, Article 41, Article 52	EU	A proposal for regulation addressing transparency, responsibility, unintended consequences, bias, and balancing accuracy, fairness, and privacy for high-risk AI systems.

Sam Altman on regulations in the NY Times: "OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing", https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html

Guidelines from the High-level Expert Group on Artificial Intelligence[20]	2018 2019 2020 2020	The recommendations from the group have acted as reference points for legislative actions undertaken by the Commission and its member states.	EU	- Ethics Guidelines for Trustworthy AI - Policy and Investment Recommendations for Trustworthy AI - The final Assessment List for Trustworthy AI (ALTAI) - Sectoral Considerations on the Policy and Investment Recommendations
Algorithmic Accountability Act (AAA)	2022	N/A	United States	Legislation requiring companies to assess the impacts of automated systems they use and sell, providing new transparency about how these systems are used, and empowering consumers to make informed choices about the automation of critical decisions.
Civil Rights Act, Title II and III	1964	Title II - Section 202, Title III	United States	Legislation used to prevent discrimination in public spaces and facilities, extended to digital spaces to prevent bias in AI systems.
Personal Information Protection Law (PIPL)	2021	Article 24	China	A law equivalent to GDPR in China, requiring the explicit consent of individuals for automated decision-making processes, including those involving AI.
New Generation Artificial Intelligence Governance Principles	2019	N/A	China	A set of principles emphasizing ethical aspects that AI development should adhere to, such as fairness, justice, and respect for human rights.

Table 4-1: Overview of the Most Important AI Laws and Regulations

[20] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

The latest initiative, ratified by the European Parliament in June 2023[21], is the AI Act, which is not supposed to come into force before 2026, however, it represents one of the most ambitious law initiatives in this context and is depicted more thoroughly.

AI Act:

The EU AI Act is a legislative proposal introduced by the European Commission in April 2021 and ratified by the European Parliament on June 14th, 2023. It aims to create a legal framework for artificial intelligence in the EU to ensure AI systems are developed and used safely and responsibly. The proposed AI Act addresses various aspects of AI, including transparency, accountability, and human oversight.

The main objectives of the AI Act are to:

1) Establish a single market for AI to facilitate the scaling up of AI solutions across the EU.
2) Ensure that AI applications respect fundamental rights, EU values, and follow ethical guidelines.
3) Promote investment in AI research and development to strengthen the EU's position in the global market.

The AI Act categorizes AI systems into three risk categories: minimal, limited, and high risk. AI systems considered high-risk must comply with strict regulatory requirements, such as transparency, data quality, documentation, and human oversight. Limited-risk AI systems, like chatbots, should be transparent about their AI-driven nature. Minimal-risk AI systems will have limited regulatory intervention.

An overview of some of the most important provisions and key concepts is provided:

1) Scope and definitions (Articles 1-4): These articles define key terms and concepts used in the AI Act, such as "AI system," "provider," "user," and "high-risk AI system." They also set the scope of the regulation, which covers both AI systems developed in the EU and those imported into the EU.
2) High-risk AI systems (Article 6): Article 6 outlines the criteria for identifying high-risk AI systems. Such systems may have a significant impact on people's rights, safety, or other important aspects of their lives. Examples include biometric identification, critical infrastructure management, and AI applications in employment and education.
3) Conformity assessment (Articles 7-43): These articles establish the requirements for high-risk AI systems to ensure they meet legal compliance before being placed on the market. These requirements include transparency, data quality, documentation, and human oversight. High-risk AI systems must undergo a conformity assessment to verify that they meet the necessary requirements.

[21] e.g. Zakrzewski, C. et al.: “Europe moves ahead on AI regulation, challenging tech giants’ power”, Washington Post, June14th, 2023, https://www.washingtonpost.com/technology/2023/06/14/eu-parliament-approves-ai-act/

4) Transparency obligations (Article 52): This article requires that AI systems intended to interact with humans, like chatbots, must be designed so that users are aware they are interacting with an AI system and not a human. This ensures that users can make informed decisions about whether to engage with the AI system.
5) National competent authorities (Articles 58-60): These articles describe the role of national authorities in monitoring and enforcing the AI Act. Each EU member state must designate one or more national competent authorities to oversee the application of the AI Act.
6) European Artificial Intelligence Board (Article 61): The AI Act proposes the establishment of the European Artificial Intelligence Board, an independent body that will advise and assist the European Commission and national competent authorities on AI-related matters.
7) Fines and penalties (Articles 71-72): The AI Act sets out penalties for non-compliance, including administrative fines that can be as high as 6% of a company's annual global turnover or €30 million, whichever is higher, for the most severe infringements.

Focusing on the EU legislation, the following three principles seem of utmost importance from an AI fairness perspective:

- Ensuring transparency and responsibility
- Preventing unintended consequences and bias
- Balancing accuracy, fairness, and privacy

All of which are briefly summarized as to how they are reflected in the most prominent EU regulations, the GDPR and the upcoming AI Act.

Ensuring Transparency and Responsibility

- **GDPR:**

Article 12: Transparent information, communication, and modalities for the exercise of the rights of the data subject.

Article 13: Information to be provided where personal data is collected from the data subject.

Article 14: Information to be provided where personal data has not been obtained from the data subject.

- **AI Act:**

Article 52: Transparency obligations for AI systems intended to interact with natural persons, ensuring that users are aware they are interacting with an AI system.

Article 41: Record-keeping, requiring providers of high-risk AI systems to maintain documentation that demonstrates the system's compliance with the AI Act.

Preventing Unintended Consequences and Bias

- **GDPR:**

Recital 71: Emphasizes the importance of preventing discriminatory effects on individuals due to automated decision-making, including profiling.

- **AI Act:**

Article 10: Data and data governance, requiring high-risk AI systems to be trained, validated, and tested using high-quality datasets that are representative and respect privacy.

Article 15: Testing and validating high-risk AI systems, ensuring their performance is consistent, accurate, and does not produce undesired effects.

Balancing Accuracy, Fairness, and Privacy

- **GDPR:**

Article 5(1)(c): Data minimization principle, requiring personal data to be adequate, relevant, and limited to what is necessary for the purpose of processing.

Article 22: Addresses automated decision-making, including profiling, that has legal or similarly significant effects on individuals, requiring organizations to provide meaningful information about the logic involved, the significance of the decision, and the consequences for the data subject.

- **AI Act:**

Article 20: Human oversight of high-risk AI systems, ensuring there is an appropriate level of human control to reduce the risk of errors and unintended consequences.

Article 10: Data and data governance, requiring high-quality datasets for high-risk AI systems that respect privacy, and helps balance accuracy and fairness while adhering to privacy requirements.

4.1.1 AI Act Assessment by Different Key Stakeholders

The proposed Artificial Intelligence Act (AIA) by the European Union has stimulated diverse responses from several key stakeholders. These stakeholders span academia, nonprofit organizations, legal and data experts, small and medium enterprises (SMEs), and think tanks. The following summarizes their viewpoints based on the EU Commission’s feedback initiative which received a total of 133 documents, however, the feedback period was already closed in August 2021. Some additional sources are added in the following brief overview:

- The **Future of Life Institute**, an independent nonprofit, stresses the importance of AI providers considering the societal impact of their applications, beyond just individual-level implications (Future of Life Institute, 2021)[22].
- The **Leverhulme Centre for the Future of Intelligence** and the Centre for the Study of Existential Risk at the University of Cambridge underline the potential of the AIA to set global standards for reducing AI-related risks and

[22] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665546_en

enabling benefits. They propose more flexibility in changing the list of restricted and high-risk systems (University of Cambridge institutions, 2021)[23].

- **Access Now Europe**, a digital rights organization, expresses concerns about the sufficiency of the AIA in protecting fundamental rights, particularly in relation to biometric applications like emotion recognition and AI polygraphs. They advocate for more robust measures such as outright bans (Access Now Europe, 2021)[24].
- **Michael Veale and Frederik Zuiderveen Borgesius**, legal and digital rights experts, reveal that the AIA heavily relies on self-assessment for compliance, raising questions about enforcement and the effectiveness of third-party verification (Veale, M. et al., 2021).
- **The Future Society**, a nonprofit advocating for responsible AI adoption, recommends improvements in information flow between national and European institutions and emphasizes the importance of analyzing incident reports from member states (The Future Society, 2021)[25].
- **Nathalie A. Smuha, Emma Ahmed-Rengers**, and colleagues criticize the AIA's ability to accurately recognize wrongs and harms associated with AI systems and allocate responsibility. They also contend that it lacks an effective framework for enforcing legal rights and duties (Smuha, N. et al., 2021).
- **The European DIGITAL SME Alliance**, a network of ICT SMEs, calls for improvements to avoid overburdening SMEs and emphasizes the need for SMEs' active participation in the development of standards for conformity assessments (European DIGITAL SME Alliance, 2021)[26].
- **The Center for Data Innovation** has estimated that the AIA will cost €31 billion over the next five years and reduce AI investments by nearly 20%. However, Meeri Haataja and Joanna Bryson argue that the Act will likely be much cheaper as it primarily covers a small proportion of high-risk AI applications (Center for Data Innovation[27], 2021 / Haataja, M. et al., 2021).
- Lastly, **Nathalie Smuha** distinguishes between societal harm and individual harm in the context of the AI Act. She argues that the Act's proposal focuses almost exclusively on individual harm, overlooking the need for protection against societal harms posed by AI (Smuha, N., 2021).

These viewpoints highlight the importance of comprehensive and effective AI regulations that balance the protection of individual and societal rights, consider the potential cost implications, and ensure a flexible and practical approach for all stakeholders involved, but also show that the new challenges of AI and especially generative AI are not fully covered by the new AI Act.

[23] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665626_en

[24] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665462_en

[25] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665611_en

[26] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665574_en

[27] https://www2.datainnovation.org/2021-aia-costs.pdf

4.1.2 Assessment of Legal Frameworks in the EU, USA, and China

One can claim that the legal landscapes for AI regulation and anti-discrimination vary considerably across the European Union, United States, and China. There are several dimensions for comparison, including the scope of protections, enforcement mechanisms, and the emphasis on transparency, accountability, and privacy.

In the European Union, a comprehensive regulatory approach is evident, particularly with the proposed Artificial Intelligence Act and the Digital Services Act, which cover a broad range of issues associated with AI, including transparency, bias, and accountability. Additionally, the EU's General Data Protection Regulation has provisions relevant to AI, such as data minimization, the right to explanation, and rules against solely automated decision-making. In the field of anti-discrimination law, the EU has established a series of directives such as the Race Equality Directive and Employment Equality Directive. While these directives primarily cover employment, they provide a legal basis for addressing algorithmic bias and discrimination.

In contrast, the United States has a more sector-specific approach to AI regulation (cf. Barocas, S., et al., 2016). The Algorithmic Accountability Act of 2022 mandates impact assessments for automated decision systems but does not cover all aspects of AI systems. Anti-discrimination laws, such as the Civil Rights Act, Fair Housing Act, and Equal Credit Opportunity Act, provide a legal framework to address discrimination and bias in AI systems, but these laws were not designed with AI systems in mind and may not cover all forms of discrimination that can arise from these systems.

In China, the regulatory approach to AI is still developing, with draft provisions for AI security management released for public comment. The country's approach is characterized by a focus on security and state control, with less emphasis on transparency and individual rights (cf. Creemers, R., 2018). In terms of anti-discrimination law, China has regulations against discrimination in employment, but a comprehensive anti-discrimination law is still under proposal.

In conclusion, while the EU, US, and China have legal mechanisms to address bias and discrimination in AI, there are significant differences in the approach and scope of these mechanisms. The EU stands out for its comprehensive regulatory approach, while the US relies more on sector-specific laws and China's legal framework is still in development. Nevertheless, in all three regions, the existing legal frameworks may not fully cover the unique challenges posed by AI, indicating a need for further legal developments in this field.

4.2 Ethical and Social Implications of AI Systems

The ethical and social Implications of AI systems serve as a critical facet of study in the field of AI. As AI systems become pervasive in various spheres of society, it is paramount to understand and address the broader societal implications, design, and usability issues, as well as privacy concerns these technologies may engender.

The societal impact of AI is multifaceted, encompassing issues such as prejudice bias, economic inequality, and environmental harm. These systems, though designed to optimize certain tasks, may inadvertently contribute to societal disparities, propagate harmful stereotypes, or further burden the environment.

Recalling the overview table of section 2.1, some of the issues and risks are treated in this section as marked in bold in table 4-2, whereas the others are briefly explained at the end of the section:

Category	Specific Issues
Societal Impact	Exacerbation of Historic Human Bias, Exclusion, Unfair Punishment, Economic Inequality, **Environmental Impact**, **Labor Market Distortions**
Design and Usability Issues	Lack of Transparency, Accessibility and Usability
Privacy Concerns	Privacy Violations, **Data Collection**

Table 4-2: Specific AI Limitations

Moreover, AI systems' design and usability can significantly affect their efficacy and fairness. A lack of transparency within AI models, often referred to as the 'black box' problem, can lead to mistrust and confusion, hindering the system's adoption and undermining its potential benefits. Likewise, poor accessibility and usability can limit the reach and effectiveness of AI technologies, potentially excluding certain groups from their benefits.

Privacy concerns arise due to the data-intensive nature of these AI systems. They require vast amounts of data for training and operation, raising concerns about data misuse, consent, and individuals' right to privacy. These concerns are compounded by the increasing use of AI in sensitive domains like healthcare, finance, and public services.

This section aims to delve into a selection of thee issues, providing some additional notions of the ethical and social implications of AI systems, however, it is only meant to briefly describe some societal problems which arise due to the implementation of AI systems and cannot be understood as a complete view because of the project's scope.

- **Environmental Impact**

The underpinnings of AI are deeply embedded in environmental concerns that stretch beyond mere considerations of carbon footprints. As the lifeblood of AI, electrical energy requires substantial carbon resources, and this is often overlooked amidst the tech sector's efforts to portray an image of sustainability and carbon neutrality. Despite attempts to diminish their environmental impact, massive digital infrastructures such as Amazon Web Services or Microsoft's Azure are voracious consumers of energy, contributing to a perpetually increasing carbon footprint (Crawford, K., 2021).

In parallel, the rapid growth of AI brings about the expansion of computational needs, which in turn intensifies these environmental concerns. AI model training requires significant energy, with initial investigations into this domain revealing startling figures. For example, one research found that running a single natural language processing (NLP) model generates more than 660,000 pounds of CO2 emissions or the equivalent of 125 round-trip flights from New York to Beijing (Strubell, E. et al., 2019). Moreover, it is not just model training that is energy-consuming. Data centers operating these models also have high electricity demands, with projections indicating that the power requirements of these centers could see a fifteenfold increase by 2030 (Belkhir, L. et al., 2019).

Efforts to counteract these environmental implications by corporations have been diverse. While companies like Apple, Google, and Microsoft have made commitments to carbon neutrality or even negativity, their involvement with fossil fuel companies offers a contradictory narrative. In essence, while they pledge to reduce their carbon footprints, they simultaneously enable the operations of the industries most responsible for climate change.

Moreover, the environmental implications of AI are not constrained solely to energy consumption and carbon emissions. The high water demand of data centers presents another critical environmental challenge, and likewise, the rare mineral extraction with sometimes disastrous consequences across the world to build the numerous devices which are necessary to support the AI systems (Abrahams, D., 2017).

Overall, the increasing computational requirements of AI contribute to a variety of environmental problems, contradicting the environmentally friendly image often promoted by the tech industry. The expanding footprint of AI systems underscores the urgent need for comprehensive environmental considerations within the tech industry's growth strategies.

- **Labor Market Impact**

The implications of Artificial Intelligence (AI) on the labor market are both transformative and profound, with observable consequences on job conditions and potential future disruptions across various industries.

AI's influence on work conditions is palpable, with Amazon's logistic fulfillment centers serving as a telling example. These facilities have implemented AI-driven technologies, such as automation and robotics, to streamline their operations. As a result, the role of human labor has undergone significant changes, with employees needing to adapt to the increasingly mechanized environment. This adaptation has often manifested in strenuous physical labor and stringent productivity targets, leading to reported declines in job satisfaction and increases in workplace injuries (Crawford, K., 2021). The shift towards AI-driven workplaces raises essential questions about workers' rights and welfare, the quality of work, and the impact of AI on physical and psychological health.

The long-term impact of AI on the labor market extends beyond altering current job conditions. A looming concern is the potential displacement of human workers as AI capabilities advance. For example, according to Kai-Fu Lee (2019), automation could dramatically reshape the labor landscape. In particular, repetitive, routine jobs and those that require less creativity and social interaction are at a higher risk of being replaced by AI technologies as shown in the figure 4-1 (Lee, K.-F., 2019) below:

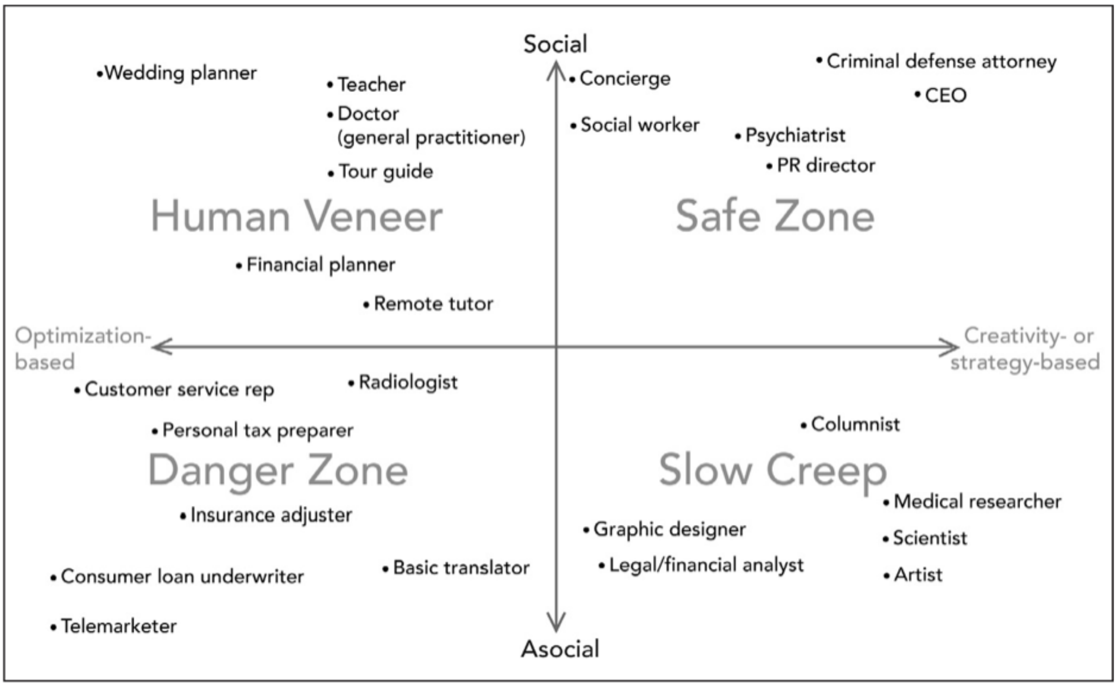


Figure 4-1: Different Professions Being Threatened by AI Replacement

This threat of automation spans across industries, from transportation, where self-driving technology might eliminate trucking jobs, to the service industry, where AI could automate tasks currently performed by cashiers, wait staff, and customer service representatives. In the manufacturing sector, where automation has already led to significant job displacement, AI could further amplify this trend.

However, it is important to note that while AI has the potential to displace certain jobs, it can also create new ones and transform existing roles, requiring a shift in the skills workers need. The challenge is in managing this transition and ensuring that workers displaced by AI can be retrained and reskilled to take on new roles in the evolving job market, which is most likely not possible in the short run (European Commission and the Council of Economic Advisers in the US, 2021).

In conclusion, the impact of AI on the labor market is multifaceted, with observable effects on work conditions and looming prospects of job displacement. A balanced and responsible approach to AI implementation is needed to mitigate adverse outcomes and leverage its potential benefits for labor. Such an approach might involve regulatory oversight, a commitment to workers' rights, and initiatives to support workforce transition through retraining and education programs.

- **Data Collection**

The practice of data collection in AI poses profound ethical, societal, and political challenges that require close examination (Seaver, N., 2018). An uncritical acceptance of data as an objective representation of reality is widespread, despite the underlying biases embedded in the training data that often reflect dominant cultural norms (Bolukbasi, T. et al., 2016). As a result, AI systems, though seemingly neutral, have the potential to inadvertently perpetuate stereotypes and discriminatory behaviors (Barocas, S. et al., 2016). This poses considerable ethical dilemmas that necessitate immediate attention.

The perspective of AI as a mathematical tool rather than a potential subject of research has historically minimized ethical scrutiny, a perspective that requires reconsideration in light of AI's integration into sensitive domains (Holstein et al., 2019). Ethical issues are often downplayed by AI professionals and researchers, indicating a broader trend of overlooking the responsibility for potential harm arising from the deployment of these systems (Selbst, A. et al., 2019). It is vital for AI professionals to contemplate the repercussions of their work and realize that it could contribute to cultural harm if not conducted mindfully.

The widespread culture of data extraction has led to the subtle privatization of what was formerly public data (Zuboff, S., 2019). Data collected, often without informed consent, is used to build AI systems with profound influence over various life aspects, leading to a power imbalance favoring private entities with large data pipelines (Pasquale, F., 2015). Far from being a neutral technical process, the collection and classification of data is essentially a social and political intervention with substantial implications (Boyd, D. et al., 2012). The power dynamics underlying these practices warrant closer scrutiny.

In summary, these considerations highlight the need for an extensive overhaul in AI and data ethics, encompassing a rigorous examination of AI systems, a deep understanding of ethical implications, and a reassessment of the prevailing data extraction culture. It is crucial to uphold transparency, accountability, and fairness in these fast-evolving fields (Crawford, K. et al., 2019).

Due to the focus if this master’s project the remaining topics are only briefly mentioned, but in-depth literature on all topics is added.

- **Exacerbation of Historic Human Bias:** AI systems, built and trained on historic data, can inadvertently carry forward and amplify existing societal biases. This happens when the data the models learn from reflects ingrained human prejudices, leading to outcomes that might perpetuate these biases further, thereby influencing decision-making in areas such as hiring, law enforcement, and lending (O'Neil, C., 2016).

- **Exclusion:** AI systems, if not designed inclusively, can result in exclusionary practices. This can occur when certain groups are not represented in the data on which the AI models are trained. For instance, voice recognition systems may fail to recognize certain accents if the training data is not diverse, leading to certain populations being denied the benefits of the technology (Buolamwini, J. et al, 2018).

- **Unfair Punishment:** AI systems, specifically those deployed in the criminal justice system, have raised concerns of unfair punishment. These systems can impact parole, sentencing, and bail decisions. But, if not carefully regulated, they can reinforce existing systemic biases, leading to unjust outcomes (cf. section 5.4).

- **Economic Inequality:** The broad deployment of AI can potentially amplify economic inequality. This can occur through job automation where AI and robotics replace certain job roles, affecting workers with lower levels of education disproportionately and thus exacerbating income disparity (Brynjolfsson, E. et al., 2014).

- **Lack of Transparency:** A significant challenge within AI systems is the lack of transparency. The complexity of these systems can make it difficult to understand their decision-making process, leading to what is often referred to as 'black box' AI. This opacity can hinder trust in these systems, making it crucial to develop techniques for better interpretability (Castelvecchi, D. 2016).

- **Accessibility and Usability:** The effective adoption of AI also depends on its accessibility and usability. However, if the design of these systems is not intuitive or does not consider the diverse needs of its users, it may limit their potential benefits (Giaccardi, E. et al., 2016).

- **Privacy Violations:** AI systems often require vast amounts of data, raising significant privacy concerns. These systems can collect and analyze sensitive user data, posing potential risks to privacy if the data is misused or inadequately protected (Zuboff, S. 2019).

5 Case Studies

The pursuit of understanding trustworthy ML and bias mitigation involves substantial theoretical discussions and the development of abstract principles. However, the application of these ideas and methods within a practical context provides the most meaningful evaluation of their strengths and weaknesses. To this end, this chapter introduces a series of case studies, serving as the bridge between the theoretical framework discussed in the previous chapters and its practical implications.

5.1 Case Studies Overview

The primary objective of these case studies is to facilitate the application of the explored concepts, tools, and approaches to real-world scenarios of ML. This process aims to foster a more comprehensive understanding of the various aspects of trustworthy ML and the mitigation of bias. The case studies endeavor to demonstrate how abstract principles of trustworthiness, fairness, and explainability are implemented in tangible situations, thus allowing the effectiveness and potential limitations of these principles to be evaluated.

Diverse areas, including HR recruitment process, automated credit scoring, and predictive policing and recidivism profiling, are addressed in these case studies, with each one bringing unique challenges and insights. For each case, the steps of data collection and preprocessing, model selection and training, evaluation and validation metrics, and quantifying bias and fairness are examined, although each of the case studies focuses on certain aspects:

1) **HR Recruitment Processes:** Multimodal input data and detection of protected attributes via unstructured data
2) **Automated Credit Scoring:** Black-box models and interpretability (XAI)
3) **Recidivism Profiling:** Review of ProPublica's investigation on recidivism scores in US courts via different fairness metrics

Through these detailed analyses, a deeper understanding of the complexities of implementing trustworthy ML algorithms and the multidimensional nature of bias, as well as how it can be effectively mitigated, can be achieved. These case studies are expected to provide compelling evidence of the theories and tools discussed and offer insights into areas where further research and development are needed.[28]

5.2 HR Recruitment Process

The examination of the Human Resources (HR) recruitment process represents a highly significant case study in the context of fairness and bias in ML. Recruitment serves as a critical gateway to opportunities, making fairness a fundamental consideration in the process. The potential for biases, often unconsciously introduced, can have severe implications for equality of

[28] The coding examples can be retrieved at https://github.com/sw-upm/trustworthy-ai

opportunity, particularly when these biases are associated with protected attributes such as race and gender.

A wealth of research underscores these concerns. For instance, studies have identified how personal characteristics can unconsciously influence decision-making processes. Researchers have demonstrated that identical resumes bearing names associated with different racial groups can receive different rates of interview invitations, reflecting an implicit bias in the recruitment process (Abrams, D. et al., 2012). Moreover, the misuse of information is also prevalent. Prospective employees' credit histories are often reviewed by employers, potentially disadvantaging minority groups, despite the lack of a proven correlation between credit history and job performance (Board of Governors of the Federal Reserve System, 2020).

Bias can also be encoded indirectly, even when algorithms are restricted from considering protected characteristics explicitly. For example, hiring algorithms can favor words more frequently used in applications from men, such as "executed" or "captured", indicating a gender bias (Datta, A., et al., 2015).

Further, a study on Amazon illustrates the risk of AI systems learning and perpetuating existing biases. In this case, the company's AI recruiting tool was found to be biased against women, as the system had learned from a historical data set primarily composed of male candidates' resumes (Dastin, J., 2018).

Given these examples, the case study of the HR recruitment process provides crucial insights into how AI can both help reduce bias but also risk perpetuating and scaling bias if not carefully managed. Therefore, the scrutiny of this process is essential in understanding how ML models can be designed and utilized to ensure fairness in decision-making.

5.2.1 Restrictions on HR Data and Algorithms

In investigating the complications of employing specific datasets for the analysis of AI in HR recruitment, several salient factors need to be elucidated. These comprise of the confidential character of the data, the corporate culture of non-disclosure, and the proprietorial nature of algorithms utilized by third-party recruitment tool providers.

Primarily, the sensitive nature of candidate data acts as a significant deterrent for examination. Candidate data inherently includes private and confidential information such as personal identifiers, education history, and career information. Given the requirements of regulations like General Data Protection Regulation (GDPR), it becomes challenging to utilize such data, even for research purposes. GDPR, along with similar data protection policies, stipulates strict regulations on the disclosure, processing, and transfer of personal data, especially without explicit consent from the individuals concerned.

Furthermore, a culture of non-disclosure is prevalent within companies. Typically, companies are wary about the release of datasets, even in anonymized forms, due to a myriad of reasons, ranging from concerns about data misuse, competitive advantages, to potential legal implications. This guarded approach constrains the ability of researchers to perform comprehensive and detailed analyses.

Moreover, third-party recruitment tool providers maintain a tight hold on their proprietary algorithms, making the analysis further complicated. Tools like

LinkedIn Talent Solutions, Jobvite, and Workable[29] are some of the dominant providers in the market and are noted for their non-disclosure of underlying algorithms. Without transparency into the operating mechanisms of these tools, comprehensive analysis and verification of their biases or fairness are practically impossible.

Aside from these factors, other barriers in evaluating specific datasets may include the lack of representativeness in the collected data, the dynamic nature of the data due to continually evolving job markets, and the difficulty in comparing outcomes due to the variance in recruitment criteria across different companies and positions (e.g., Foster, I. (ed.) et al., 2016). Furthermore, the unavailability of detailed metadata about the dataset collection process might lead to inherent biases, thereby restricting the reliability of the dataset for empirical analyses.

Finally, the black box problem associated with AI systems can also pose a substantial challenge. Many AI models, such as deep learning algorithms, can be opaque and complex, making it difficult to understand the decision-making process and hence, complicating the analysis.

In sum, these complexities and constraints inherent in the HR recruitment process and the associated data highlight the intricacies involved in conducting a fair and objective analysis.

Therefore, the case study focuses on a synthetic dataset which incorporates bias in the HR recruitment process as described in the following sections.

5.2.2 Automatic Recruitment Bias in Multimodal Datasets

The following case study describes the findings of the academic paper “Bias in Multimodal AI: Testbed for Fair Automatic Recruitment” by Peña, A. et al. (2020), presents the result, and proposes a series of future directions for investigation.

The researchers propose an automated recruitment testbed, FairCVtest, to study how the HR recruitment algorithms are affected by sensitive elements and biases within data. FairCVtest uses a set of multimodal synthetic profiles that have been consciously scored with gender and racial biases. This helps illustrate how the AI behind recruitment tools can extract sensitive information from unstructured data, potentially leading to unfair decision-making. The difficulty in detecting and preventing biases in multimodal models, which utilize multiple heterogeneous sources of information, including structured and unstructured data, is highlighted. Hence, the aim is to study how multimodal ML is influenced by biases in training datasets, evaluate the ability of neural networks to learn biased target functions from multimodal sources of information, and develop a discrimination-aware learning method that eliminates sensitive information from the learning process.

A model was built based on this multimodal input data, and a scoring system was established to rank candidates according to their merits. However, bias can sneak in at different stages of the learning process, including the data collection, preprocessing, and even in defining the target function and learning strategy.

[29] https://business.linkedin.com/es-es/talent-solutions, https://www.jobvite.com/, https://www.workable.com/

Dataset

To test these biases, a research dataset named FairCVdb was created, containing 24,000 synthetic resume profiles, 80% being used for the training and 20% for the validation sets. These profiles consist of 12 features from 5 information blocks, as well as 2 demographic attributes (gender and ethnicity), and a face photograph. The five blocks are education attainment, availability, previous experience, presence of a recommendation letter, and proficiency in 8 different languages. Figure 5-1 shows schematically which variables are used as Information blocks in a resume and personal attributes that can be derived from each one. The number of crosses represent the level of sensitive information (+++ = high, ++ = medium, + = low).[30]

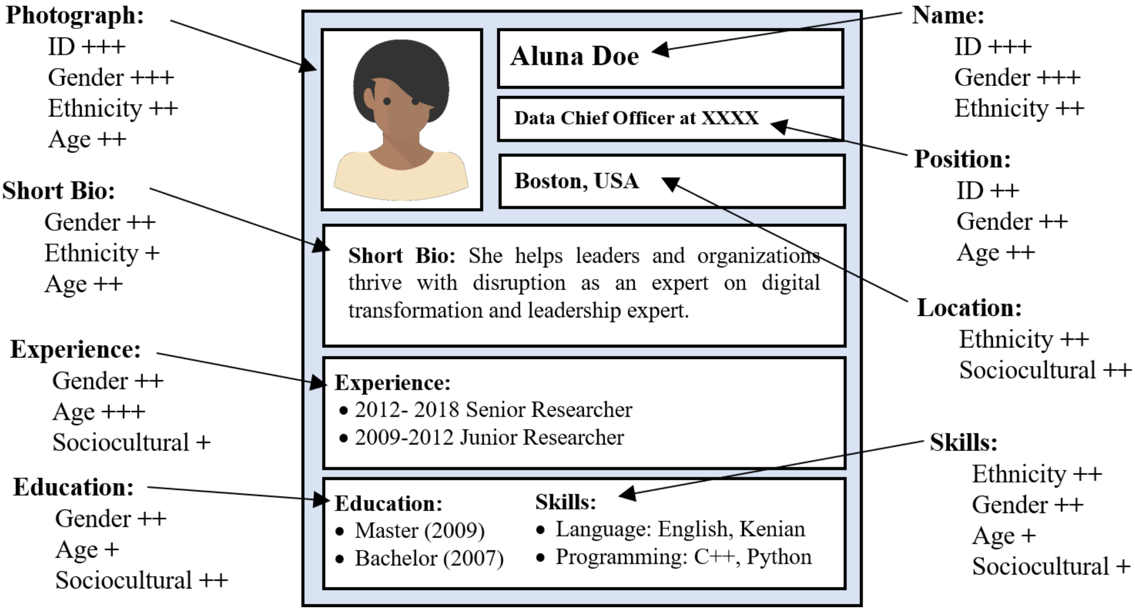


Figure 5-1: Multimodal Information in Resume

Problem Formulation

Each profile was assigned a score based on a linear combination of these competencies. Importantly, the scores were calculated without taking gender or ethnicity into account, creating an unbiased set of scores. However, two additional sets of scores were generated that incorporated gender and ethnicity biases, simulating real-world scenarios where the recruitment process might be influenced by such biases.

Hence, the problem formulation is simply based on minimizing a loss function where the target function is a resume score function.

Loss function: $\min_{w} \sum_{x^j \in S} \mathcal{L}(O(x^j \mid w), T^j)$

Where:

w is the model parameter vector.

x is an individual input sample obtained from the resume.

[30] The full dataset and code can be viewed in Github: https://github.com/BiDAlab/FairCVtest

T is the target value representing the score within the interval [0, 1] ranking the candidates according to their merits.

$\mathcal{S}$ is the set of training samples.

$O(x|w)$ represents the output of the model with parameters w given the input x

$\mathcal{L}$ is the loss function used to calculate the discrepancy between the model output and the target value.

And the target score function: $T^j = \beta^j + \sum_{i=1}^{n} \alpha_i x_i^j$

Where:

$n = 12$ is the number of features (competencies).

α_i are the weighting factors for each competency x_i^j

ϵ^j is some Gaussian noise to include a small degree of variability (i.e., to cater for slightly different scores for two profiles with the same competencies).

These scores T^j serve as ground truth in the experiments, but they are generated without considering gender or ethnicity information to ensure they are unbiased and equally distributed among different demographic groups.

The unbiased scores, referred to as T_U, are used as a baseline. However, two additional target functions are introduced to simulate biased scenarios: gender bias T_G and ethnicity bias T_E.

Biased scores are generated by applying a penalty factor δ to certain individuals in specific demographic groups. This introduces a simulated cognitive bias, where individuals from certain groups may receive lower scores compared to others with the same competencies.

By comparing the performance of models trained on unbiased scores T_U with those trained on biased scores T_G or T_E , the impact of cognitive biases introduced by humans, protocols, or automatic systems can be analyzed.

Experiments and First Results

The experiments consist of four scenarios, differing in the use of gender attribute and target function (unbiased or biased).

The four scenarios are:

Scenario 1: Training with unbiased scores, including the gender attribute.

Scenario 2: Training with gender-biased scores, including the gender attribute.

Scenario 3: Training with gender-biased scores, without the gender attribute.

Scenario 4: Training with gender-biased scores, including feature embedding from the face photograph.

In all scenarios, the models were constructed as feedforward neural networks. In Scenario 4, a pretrained ResNet-50 model was used to obtain feature embeddings from face photographs as shown below in figure 5-2:

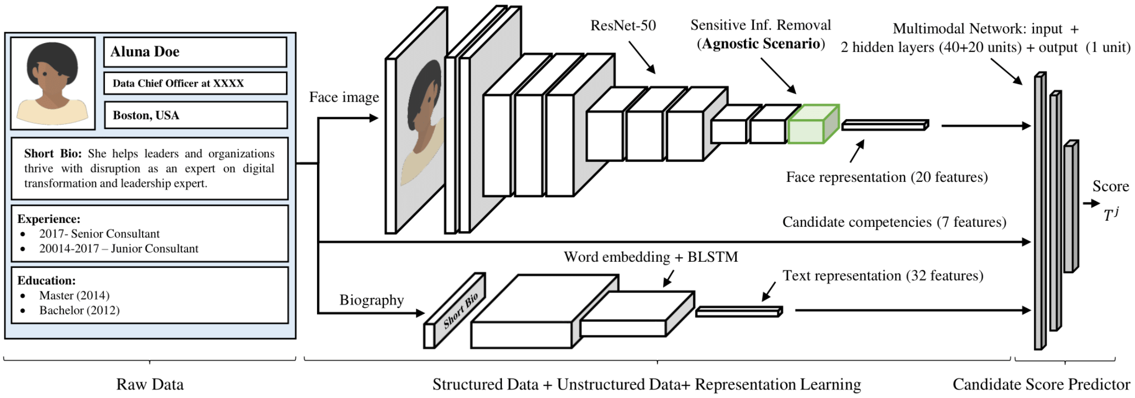

Aluna Doe
Data Chief Officer at XXXX
Boston, USA
Short Bio: She helps leaders and organizations thrive with disruption as an expert on digital transformation and leadership expert.
Experience:
• 2017- Senior Consultant
• 20014-2017 – Junior Consultant
Education:
• Master (2014)
• Bachelor (2012)
Face image
ResNet-50
Sensitive Inf. Removal (Agnostic Scenario)
Multimodal Network: input + 2 hidden layers (40+20 units) + output (1 unit)
Face representation (20 features)
Candidate competencies (7 features)
Word embedding + BLSTM
Text representation (32 features)
Biography
Short Bio
Score
T^j
Raw Data
Structured Data + Unstructured Data+ Representation Learning
Candidate Score Predictor

Figure 5-2: Multimodal Architecture - ResNet-50 & Fully-connected Layer

In the experiment, the recruitment tool was trained using 80% of the synthetic profiles and a 20% validation set. Performance evaluation based on validation loss and KLD showed that adding gender and ethnicity information could enhance model accuracy but might also introduce significant bias into the recruitment process.

The study further revealed that even if the gender attribute was not explicitly provided, the model could still detect gender from facial features. This suggests that AI recruitment tools might inadvertently learn to discriminate based on latent features if trained on biased datasets.

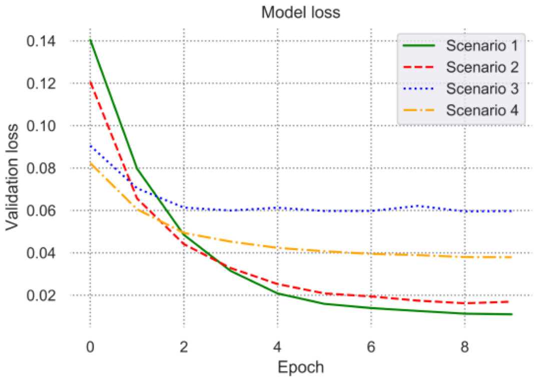


Figure 5-3: Validation Loss per Scenario

Figure 5-3 presents the validation loss during the training process for different scenarios, serving as an indicator of each network's performance on scoring applicant resumes. In Scenarios 1 and 2, where the network has access to all influential features, the models perform more precisely. The presence of Gaussian noise prevents the loss from converging to zero. Scenario 3 performs poorly due to the absence of a correlation between the bias in scores and the network inputs. Scenario 4 falls in between the others, with the network discovering gender features in the face embeddings, despite not being trained for gender recognition. It is noted that the validation loss is lower when biased scores and sensitive features are available (Scenario 2) compared to when the network is blind to sensitive features (Scenarios 3 and 4).

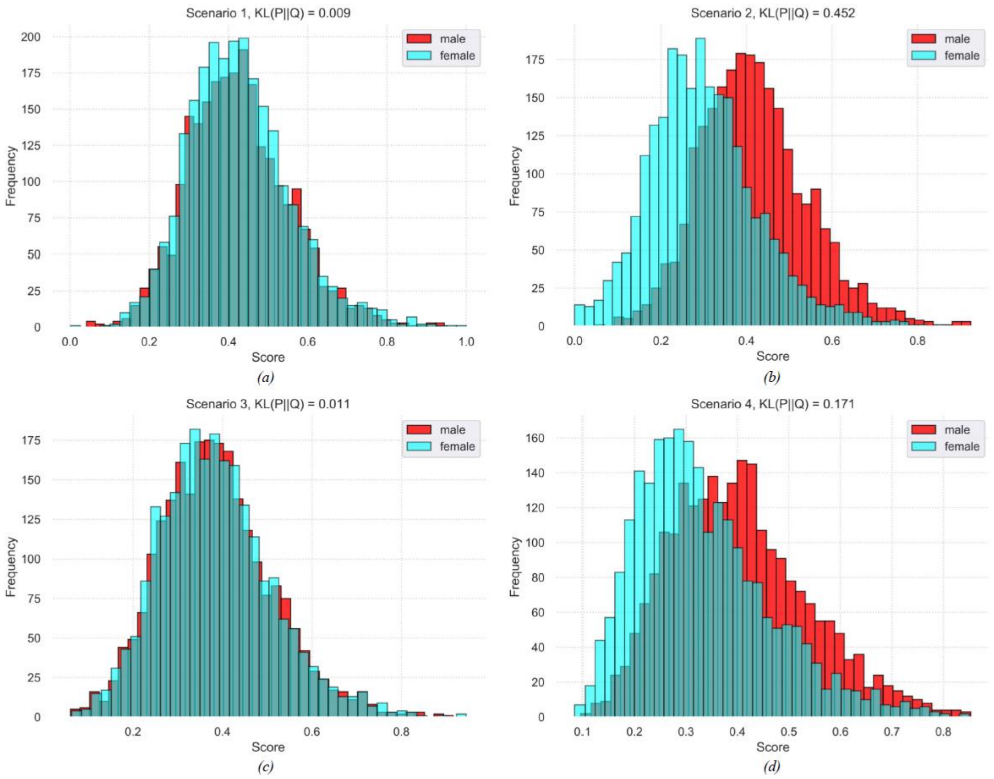


Figure 5-4: Kullback-Leibler Divergence with Gender Bias in Distribution

Figure 5-4 illustrates the score distributions predicted in each scenario by gender, revealing the presence of bias in certain scenarios. The KLD is used to measure the bias's impact on the classifier output. In Scenarios 1 and 3, there's no gender difference in the scores, evidenced by a KL divergence close to zero. Scenario 1 achieves this due to the use of unbiased scores during training, rendering gender information in the input irrelevant, while Scenario 3 achieves it by ensuring no gender information in the training data, with balanced classes. Despite a drop in performance, the absence of this information results in a fairer model. Scenario 2 displays the most pronounced difference between male-female classes due to the explicit provision of gender information. In Scenario 4, the network is capable of detecting gender information from face embeddings and correlating them with the injected bias. This reveals the presence of gender bias, even when gender was not explicitly available during training, indicating the gender is inferred from latent features in the face image. In this case, the KL divergence is lower than in Scenario 2 but significantly higher than in the unbiased scenarios.

Experiment Enhancement with Unbiased Model FairCVTest and Results

The model is enhanced with an additional regularization to cater for unbiased learning as proposed by Morales, A. et al. (2019), which they call agnostic representation through SensitiveNets.

The optimization formula is enhanced as follows:

$$\min_{w} \sum\nolimits_{x^j \in \mathcal{S}} \left(\mathcal{L}\left(O\left(x^j \mid w\right), T^j\right) + \Delta^j\right)$$

This method, originally developed for privacy enhancement in face biometrics, incorporates an adversarial regularizer to eliminate sensitive information from the learned representations. In this context, the term Δ^j is generated using a sensitiveness detector and quantifies the amount of sensitive information present in the learned model represented by the parameter vector *w*. The face representation used in Scenario 4 is trained using this method, which is referred to as the "Agnostic scenario" in subsequent experiments.

The simulated recruitment experiment, assuming a tool to perform initial screening, used all 24,000 resumes as input. In each scenario, the top 100 scores were selected. Table 5-1 shows the gender distribution for these selections. Scenarios 1 and 3, where the classifier exhibits no gender bias, had almost no discrepancy in the percentage of candidates chosen from different gender groups. However, in Scenarios 2 and 4, the gender bias was significant, more pronounced in Scenario 2 with a 74% difference. The difference was 54% in Scenario 4. The application of a sensitive features removal technique dramatically reduced this discrepancy from 54% to 0%, effectively rectifying the gender bias. This finding underlines the potential dangers of these recruitment tools in terms of fairness and highlights potential solutions.

Scenario	**Bias**	**Input Features**			**Gender**		**Delta**
		Merits	Dem	Face	Male	Female	
1	no		yes	no	51%	49%	2%
2	yes	yes	yes	no	87%	13%	74%
3	yes	yes	no	no	50%	50%	0%
4	yes	yes	no	yes	77%	23%	54%
Agnostic	yes	yes	no	yes	50%	50%	0%

Table 5-1: Distribution of the Top 100 Candidates

Conclusions and Discussions

The case study introduces FairCVtest, an open-source framework developed to understand how biases in data affect AI recruitment tools. It uses deep learning to analyze the ability of AI to expose and use sensitive data. The framework uses 24,000 synthetic job applicant profiles. Biases in gender and ethnicity were incorporated into the scoring of these profiles, resulting in discrimination in the generated candidate scores. This discrimination was apparent both when these attributes were given explicitly and when only a face image was provided. The findings show that biases can arise from unstructured data combined with historical biases, particularly with datasets gathered from historical sources that lack diversity representation. The study also explores ways to mitigate these biases, specifically using a method called SensitiveNets (Morales, A. et al., 2019), which removes sensitive information during the learning process, improving fairness.

This HR recruitment model approach based on a synthetic toy set already provides some interesting insights by demonstrating that unintended bias can emerge from unstructured data such as face images or resume texts if the model is trained with historically biased data.

The paper presents a thorough study on the biases inherent in AI-based automated recruitment systems and proposes an experimental framework, FairCVtest, to explore this critical issue. However, there could be some potential limitations and avenues for future work:

Limitation in Synthetic Data: While FairCVdb is a valuable tool, it uses synthetic profiles instead of real-world data. The dynamics of real-world biases could be much more complex and multifaceted than the biases artificially introduced in the synthetic data. Future studies should involve real-world recruitment data to more accurately understand how biases manifest in AI recruitment systems. As mentioned before, this seems to be an extremely challenging task for the time being.

Bias Scope: This study mainly focuses on gender and ethnicity biases. While these are important, there are other types of biases that could be present in recruitment processes, such as age, disability, socio-economic status, or educational background biases. Although the toy set can be easily adapted to gender and ethnicity, the focus is always on a single sensitive attribute and never on a combination of various, e.g., how would an elderly black woman be treated in such a recruitment process?

Fairness Metric: The paper utilizes SensitiveNets to train fair models, but it does not discuss in depth how fairness is quantified or what metrics are used to evaluate the fairness of these models except KLD. Future work could delve into the development of robust fairness metrics for AI recruitment systems.

Preventive Measures: Although the paper experiments with SensitiveNets to reduce biases in AI, it could explore additional methods or preventative measures to reduce or eliminate these biases. There is a wide range of de-biasing techniques that could be employed, including pre-processing, in-processing, and post-processing methods.

User-Specific Bias: The paper notes the recent shift from group-based bias analysis to user-specific bias analysis, and it suggests that FairCVtest will be updated to incorporate such user-specific biases. Exploring user-specific biases could be a promising direction for future work, as this approach accounts for the individuality and uniqueness of each candidate.

Legal and Ethical Implications: While the study is technical, it could also delve into the legal and ethical implications of using AI in recruitment. This would provide a more holistic understanding of the problem and could lead to the formulation of guidelines or best practices for AI recruitment.

Explainability and Transparency: AI models, especially deep learning models, are often seen as black boxes, where the decision-making processes are not transparent. Future work could explore methods to increase the transparency and explainability of AI recruitment systems, allowing users to understand why certain decisions are made, although as mentioned in the introductory part of this section, this also proves very challenging due to the proprietary nature of the algorithms applied by automated HR recruitment tool providers.

5.3 Automated Credit Scoring

The evaluation of automated credit scoring serves as another critical case study in the context of fairness and bias in ML. Credit scoring significantly impacts an individual's access to various financial products and services, and as such, biases within this system can have profound implications on economic opportunity and financial inclusion.

Historically, conventional credit scoring methods have relied on a set of factors such as credit history, current debt levels, and income, among others. While these factors can be effective indicators of creditworthiness, they may also introduce biases, either directly or indirectly, against certain demographic groups. For instance, lower-income individuals or recent immigrants may lack substantial credit history, thus negatively affecting their credit scores despite their potential creditworthiness.

ML algorithms, while promising for their ability to analyze complex relationships and large datasets, can perpetuate or even exacerbate these biases if not properly managed. For example, a study by the National Bureau of Economic Research found that algorithms used for FinTech lending discriminate against Latinx and African American borrowers (Bartlett et al., 2019). This discrimination is not because these algorithms use race explicitly but because they pick up proxies for race in the data.

In another instance, an investigation by the Federal Trade Commission found that an AI system used for credit decisions was discriminating against certain customers, even though the model was not explicitly using any protected characteristics (Federal Trade Commission, 2021). This highlights the subtlety with which biases can be encoded within these systems and the importance of ensuring fairness in their design and implementation.

Given these examples, the scrutiny of automated credit scoring is crucial to understanding how AI systems can be developed and used to uphold fairness and eliminate bias in this important area of financial decision-making.

However, the first step is to get a clear understanding of how certain black-box models should be interpreted. The idea of this short case study is to leverage and illustrate two of the presented XAI techniques, namely SHAP and LIME, via a well-known benchmark dataset.

5.3.1 German Credit Dataset

The German Credit dataset is a renowned collection of data, widely used in risk analysis and ML research, obtained from the Statlog project at the University of California, Irvine's Machine Learning Repository[31]. This dataset comprises information from 1,000 loan applicants who were customers of a German bank from the 1990s. The primary purpose of this dataset is to aid in the assessment of credit risk, that is, to determine whether a loan applicant should be categorized as 'Good' or 'Bad' credit risk based on a collection of attributes.

Each instance in this dataset is described by a set of 20 diverse features, both categorical and numerical. The categorical features provide information about the applicant's credit history, employment status, personal status and sex, property ownership, and other socio-economic indicators. On the other hand, the numerical features include credit duration, credit amount, installment rate, residence duration, age, and other related variables. From a fairness perspective age and gender (sex) are usually chosen as protected attributes.

The target variable, 'classification', differentiates between 'Good' and 'Bad' credit risk. These classes are balanced to prevent any bias towards a specific category, with approximately 70% of the instances labeled as 'Good' and 30% labeled as 'Bad'.

[31] http://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

The German Credit dataset's nature and multi-aspect representation of an individual's financial status make it a regular resource for developing and evaluating credit risk assessment models. The dataset's size and diversity, including both categorical and numerical features, provides opportunities for various preprocessing, feature engineering, and modeling techniques, making it a commonly used benchmark in the field of ML and risk prediction. Annex 2 lists each of the features with a short description.

5.3.2 XAI Analysis

In this short case study, a ML analysis is performed on the German Credit Data dataset to predict customer credit risk.

Initially, the dataset is imported from the UCI Machine Learning Repository, and the target class labels are replaced for binary classification purposes. The dataset consists of both numerical and categorical features. For preprocessing, numerical features are standardized, which rescales them to have a mean of 0 and a standard deviation of 1, making the values more suitable for the applied ML algorithms. The categorical features are converted into numerical form through label encoding, followed by one-hot encoding to ensure they are properly interpreted by the ML algorithms.

The data is then split into a training set and a test set, with a test size of 20% of the total data. The Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training set to address the issue of class imbalance, generating synthetic samples of the minority class to balance the class distribution.

Subsequently, five different black-box ML models are trained on the balanced training data: XGBoost, LGBM, AdaBoost, CatBoost, and HistGradientBoost. Each model's performance is evaluated based on four different metrics: accuracy, F1 score, recall, and precision. These metrics provide insight into the model's ability to correctly classify instances, balance precision and recall (F1), correctly identify positive instances (recall), and the proportion of true positive predictions among all positive predictions (precision).

After training and evaluation, the performance of the models is compared and ranked based on the four metrics. Table 5-2 shows the result, HistGradientBoost being the one with the best performance, also illustrated in figure 5-5:

Model	Accuracy	F1	Recall	Precision
HistGradientBoost	0.82	0.88	0.91	0.84
CatBoost	0.82	0.87	0.91	0.82
AdaBoost	0.80	0.86	0.87	0.84
XGBoost	0.79	0.86	0.91	0.81
LGBM	0.78	0.85	0.91	0.80

Table 5-2: Performance on Boosting Models

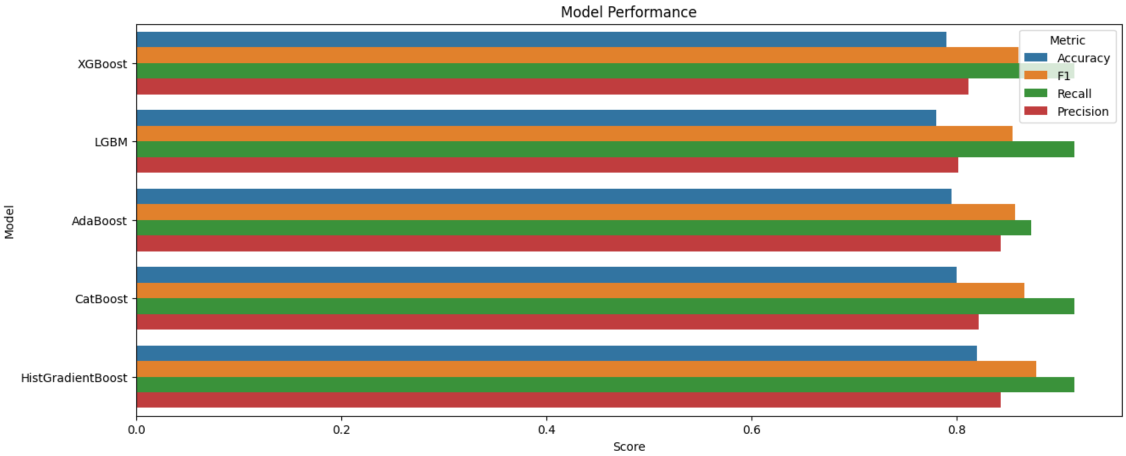


Figure 5-5: Boosting Model Performance Comparison

To provide a deeper understanding of model, the SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) techniques are used for explainability. The best performing model, HistGradientBoost, is chosen for this analysis.

SHAP provides a **summary plot** displaying the most important features and their negative (blue) and positive (red) impacts on the model's output as illustrated in figure 5-6:

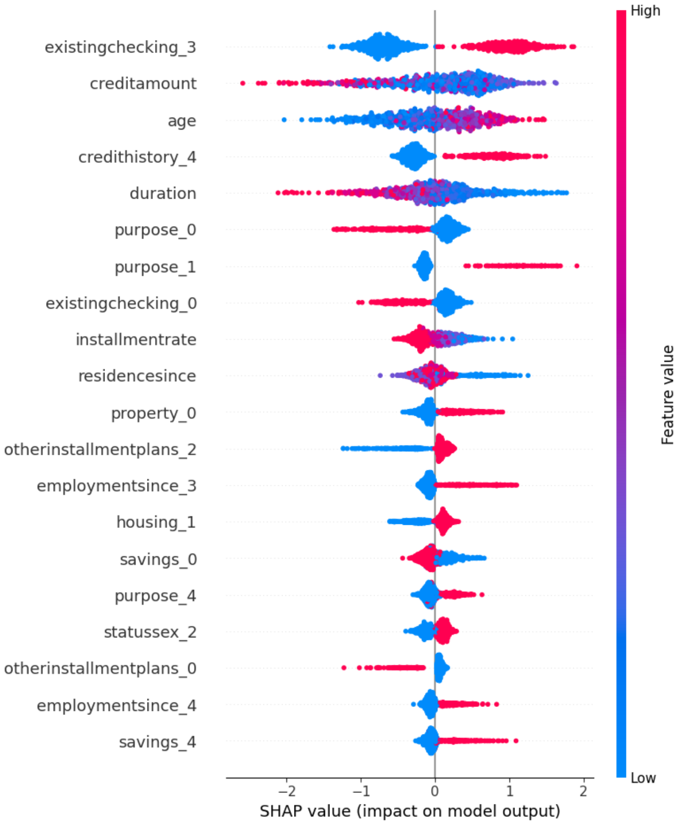


Figure 5-6: SHAP Summary Plot on Credit Data

The SHAP summary plot provides a holistic view of the feature importance and their effects on the prediction model. Here's an interpretation of the first three features as presented in the summary plot:

- **existingchecking_3:** This feature seems to have the highest impact on the model's output. It demonstrates a clear bifurcation in its SHAP values, with a separate cluster of negative (blue) and positive (red) SHAP values. The blue points on the left show that lower values of existingchecking_3 decrease the model's output, while the red points on the right show that higher values of this feature increase the model's output. Hence, existingchecking_3 has a significant and varying impact on the model's predictions.
- **creditamount:** The SHAP values for creditamount are mostly blue and overlap between positive and negative effects, indicating this feature has a generally negative effect on the model's output. However, the presence of both red and blue points throughout its range implies that creditamount can increase or decrease the model's output depending on its value. The relationship between creditamount and the output is likely complex, potentially non-monotonic, and warrants further investigation.
- **age:** The feature age shows a similar pattern as creditamount, with both red and blue points scattered across its range. This suggests age has a mixed impact on the model's predictions, and its effect is not simply positive or negative but varies based on its value. The slightly more blue values indicate that higher age values might more frequently decrease the model's output, though there is considerable variance.

A **SHAP dependence plot** is also produced to further explore the relationship between age and the prediction as shown in figure 5-7:

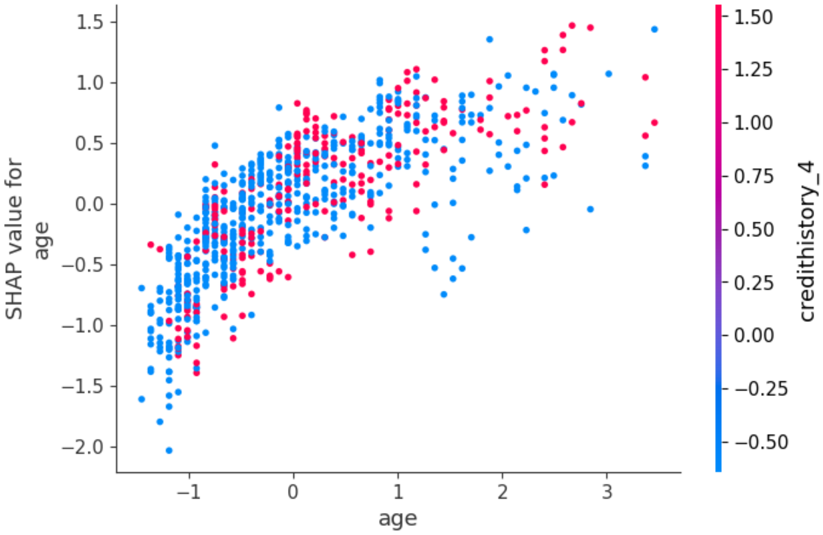


Figure 5-7: SHAP Dependence Plot for Age on Credit Data

The SHAP dependence plot provides a detailed view of the relationship between a specific feature and its corresponding SHAP values. This allows us to visualize and understand the complex interplay between this feature and the model's predictions.

The feature age is depicted along the x-axis, while the SHAP values are represented along the y-axis. Each point represents a specific instance in the data.

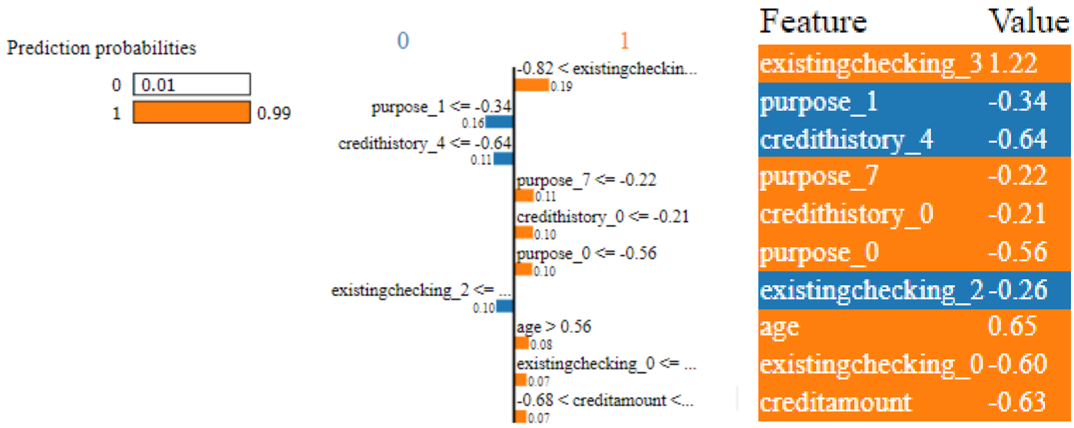
The predominantly blue points imply that lower age values frequently lower the model's output (since blue is often associated with a negative SHAP value), while the few red points suggest that certain higher age values increase the output (since red typically corresponds to positive SHAP values). This suggests that the relationship between age and the model's output is complex and potentially non-linear, with both higher and lower age values being associated with both higher and lower model outputs.

The SHAP values on the y-axis range from -2 to 1.5. A SHAP value essentially quantifies the contribution of a feature to the prediction of each instance relative to the prediction's baseline value. A negative SHAP value implies that the presence of a feature pushes the model's prediction lower than the baseline, while a positive SHAP value suggests that the feature increases the prediction. The magnitude of these values signifies the strength of the effect. In this case, age values that yield SHAP values around -2 have a strong negative impact on the model's prediction, while those around 1.5 have a strong positive impact.

The standardization of age makes the feature have a mean of 0 and a standard deviation of 1. This range implies that the age values in this dataset are mainly within 3 standard deviations from the mean.

On the right side of the dependence plot, the feature credithistory_4 (=delay in paying off in the past) is displayed with values ranging from -0.5 to 1.5. The color of each dot in the plot corresponds to the value of credithistory_4 for that particular instance, suggesting that credithistory_4 may interact with age in affecting the model's predictions. The values are encoded representations of the categorical feature "delay in paying off in the past". The fact that these values are represented by colors in the main plot signifies an interaction effect. Specifically, the SHAP values (model's output) might not just depend on age alone, but also on the interaction between age and credithistory_4. For instance, a particular age might have a positive impact on the model's output for one value of credithistory_4 but a negative impact for a different value.

Additionally, an instance from the test set is explained with **LIME**, giving insight into how each feature contributes to the final prediction for that specific instance.



Feature	Value
existingchecking_3	1.22
purpose_1	-0.34
credithistory_4	-0.64
purpose_7	-0.22
credithistory_0	-0.21
purpose_0	-0.56
existingchecking_2	-0.26
age	0.65
existingchecking_0	-0.60
creditamount	-0.63

Figure 5-8: LIME Explanation for a Specific Instance

The LIME output for instance 190 provides the following interpretation (the exact meaning of the one-hot encoded variables can be checked in annex 2):

- **existingchecking_3:** The feature existingchecking_3 (=checking account with highest amount) continues to have a positive impact on the prediction with a weight of 1.22. An increase in the value of this feature would make the model more likely to predict the positive class for this instance.
- **purpose_1:** With a weight of -0.34, this feature has a negative influence on the prediction. An increase in the value of purpose_1 (=new car) for this instance would push the model's prediction towards the negative class.
- **credithistory_4:** This feature negatively affects the prediction with a weight of -0.64. This means when credithistory_4 (=delay in paying off in the past) is high, it leads the model to predict the negative class.
- **purpose_7 and credithistory_0:** Both features have a negative impact on the prediction with weights of -0.22 and -0.21 respectively, pushing the prediction towards the negative class when their values are high.
- **age:** This feature has a weight of 0.65, suggesting that a higher age pushes the prediction towards the positive class for this instance.
- **existingchecking_0 and creditamount:** These features negatively impact the prediction with weights of -0.60 and -0.63 respectively. This implies that an increase in these features would lead the model's prediction towards the negative class for this instance.

In conclusion, this exercise served as a brief illustration of the interpretability techniques applied to black-box models using the German Credit dataset. By leveraging SHAP and LIME, it was to gain insights into the feature importance and understand the model's decision-making process on specific instances. However, it is important to note that this exercise was conducted solely for illustrative purposes, and no generalizable conclusions or inferences should be drawn from this specific example. Interpretability methods should be used as tools to aid understanding and provide transparency in complex models, but further investigation and evaluation are necessary to draw robust conclusions in real-world scenarios.

5.4 Predictive Policing and Recidivism Profiling - COMPAS

ML's role in predictive policing and recidivism profiling, particularly the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), serves as a pertinent case study in the analysis of fairness and bias within AD-MS. These systems carry substantial influence over life-altering decisions in the context of justice administration, underscoring the necessity for transparency, fairness, and accountability.

COMPAS is a predictive algorithm applied within the criminal justice system to estimate the likelihood of a defendant's recidivism. Its deployment, however, has been controversial due to issues related to potential racial bias. Specifically, the tool was analyzed based on its application in Broward County, Florida, and it was found that Black defendants were more likely to be falsely labeled as high-risk reoffenders compared to their White counterparts (Angwin, J., et al., 2016).

However, this depiction of COMPAS is complex and multifaceted. Some researchers argue that COMPAS can help reduce bias. For instance, Kleinberg, J. et al. (2018) suggest that such predictive algorithms, when correctly applied, can improve decision-making processes and reduce human bias. On the other hand, oversampling issues may arise due to overpolicing in certain neighborhoods, potentially biasing the algorithm towards these regions.

The controversy surrounding COMPAS underscores an essential tension in designing fair ML systems—the trade-off between different fairness metrics. For instance, balancing predictive parity (equal predictive accuracy across groups) and false positive rate balance (equal false positive rates across groups) is a challenging yet crucial aspect of ensuring algorithmic fairness.

In summary, this examination of COMPAS illuminates the nuanced challenges in constructing and applying fair and trustworthy AI systems within the context of criminal justice. The lessons derived from this case study are integral to guiding future efforts in designing fair AI systems for high-stakes decision-making.

Other prominent predictive policing cases comprise the following ones and are only meant to illustrate the growing concern about the need to apply a bias and fairness framework as depicted in the previous sections.

The (CPD) in 2016 faced controversy over its use of a predictive policing algorithm known as the Strategic Subject List (SSL). The algorithm was designed to identify individuals who were predicted to be involved in violent crime. Critics contended that the algorithm disproportionately targeted communities of color, exacerbating existing racial biases and contributing to over-policing in these communities. This case has been thoroughly discussed in the paper titled, " The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement" (Ferguson, A.G., 2017).

In 2018, the Los Angeles Police Department (LAPD) encountered criticism for its use of predictive policing algorithms such as PredPol and Operation LASER. Critiques highlighted that the algorithm disproportionately targeted communities of color, reinforcing existing racial biases, and led to over-policing. The analysis and criticism of these methods can be found in the paper, " Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data" (Richardson, R. et al., 2019).

The Predictive Policing (PredPol) algorithm, widely used across various police departments in the U.S., has attracted criticism for its potential to reinforce racial biases and contribute to over-policing in communities of color. PredPol uses historical crime data to predict where crimes are likely to occur, but critics argue that this approach can reinforce existing patterns of bias and discrimination. The discussion surrounding the implications of PredPol can be found in the academic paper, "Predictive Policing and the Politics of Patterns" (Kaufmann, M. et al., 2018).

However, the focus in the rest of this section lies on the COMPAS algorithm developed by Northpointe[32] first in the late 1990s and then applied across the US in courts applying pretrial release risk, general recidivism, and violent recidivism scales. The following analysis comes with a series of caveats which need to be mentioned as they limit the scope and depth of the analysis:

1) **Proprietary software:** A pervasive critique of proprietary software, such as COMPAS, revolves around the issue of transparency and due process.

[32] Courtview Justice Solutions Inc., Constellation Justice Systems Inc., and Northpointe Inc. were merged as Equivant in January 2017.

Since the underpinning algorithms and computational procedures that these systems employ are often safeguarded as trade secrets, they are typically inaccessible for public scrutiny or examination.

2) **Dataset:** The issue of dataset availability is a key factor that complicates the analysis of the efficacy and fairness of proprietary algorithms like COMPAS. With the Broward County, Florida dataset being the only one publicly available for scrutiny, there are inherent limitations in the analysis that can be performed on COMPAS's effectiveness and potential biases.

Notwithstanding these limitations, an analysis can be performed based on the reduced dataset and the respective models' predictions. The starting point is the ProPublica[33] investigation from 2016 (Angwin, J. et. al., 2016)[34], which provoked civil disturbance and controversy.

Furthermore, some of the fairness tools presented in section 3.5 Tool-based Bias Mitigation are applied to the dataset, statistical fairness measures analyzed, and conclusions drawn from these standard tools.

Finally, additional analysis is carried out based on own Python notebooks, making use of different libraries as described below.

5.4.1 The COMPAS Dataset and Initial Exploratory Data Analysis

ProPublica's investigation into the COMPAS algorithm focused on its application in Broward County, Florida, a large jurisdiction that extensively employs the COMPAS tool in pretrial decision-making. The choice of Broward County was also influenced by Florida's robust open-records laws.

Following a public records request, ProPublica secured COMPAS scores for two consecutive years (2013 and 2014) from the Broward County Sheriff's Office. This dataset encompassed 18,610 individuals who were scored during that period. To align the study with the county's primary usage of the COMPAS tool, only scores associated with pretrial decisions were retained, resulting in a refined sample of 11,757 individuals. Each of these defendants received three COMPAS scores: 'Risk of Recidivism', 'Risk of Violence', and 'Risk of Failure to Appear'. These scores, on a scale from 1 to 10, were categorized by COMPAS as 'Low' (1-4), 'Medium' (5-7), or 'High' (8-10).

To construct a comprehensive criminal history for each individual, public criminal records from the Broward County Clerk's Office were gathered until April 1, 2016. The mean time defendants were not incarcerated was calculated as approximately 623 days, with a standard deviation of approximately 329 days. These records were matched with the COMPAS scores using individuals' first and last names and date of birth. This process led to the downloading of approximately 80,000 criminal records.

[33] ProPublica describe themselves as "an independent, nonprofit newsroom that produces investigative journalism with moral force." https://www.propublica.org/about/

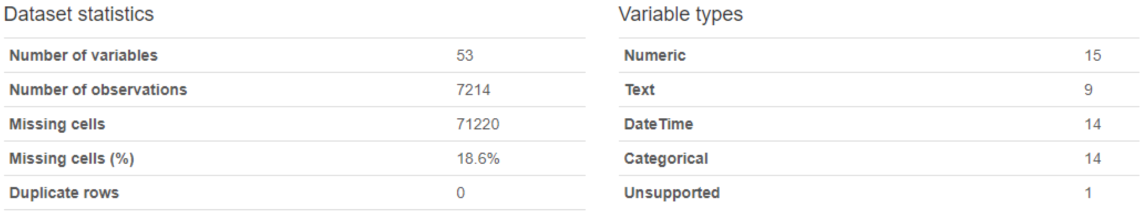
[34] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Race classification followed the Broward County Sheriff's Office's system, which categorizes defendants as Black, White, Hispanic, Asian, and Native American. In 343 cases, the race was marked as 'Other'. Additionally, incarceration records were compiled from jail records provided by the Broward County Sheriff's Office (January 2013 to April 2016) and from public records on the Florida Department of Corrections website.

Finally, the dataset used in most of the analyses (including the one from ProPublica) and academic papers, is the one for a two-year period (2013-2014) recidivism score with 52 features, one binary target variable (recidivist or not) and 7,214 instances, although most of the published analyses use between 14 and 29 features, including that of ProPublica.

The dataset table in annex 2 provides a brief overview of all 53 variables used in the 2-year recidivism dataset.

The raw dataset basic statistics already show that quite a few entries are missing as shown in figure 5-9:



Dataset statistics

Number of variables	53
Number of observations	7214
Missing cells	71220
Missing cells (%)	18.6%
Duplicate rows	0

Variable types

Numeric	15
Text	9
DateTime	14
Categorical	14
Unsupported	1

Figure 5-9: COMPAS Dataset Statistics

A closer look in figure 5-10 reveals that most of the missing values refer to recidivism (r_*) and violent recidivism events (vr_*), which in turn can be easily explained as only those individuals who recidivate can generate the corresponding datapoints. Violent recidivism is the only column without any values and can be skipped as the following one is_violent_recid already conveys the same information.

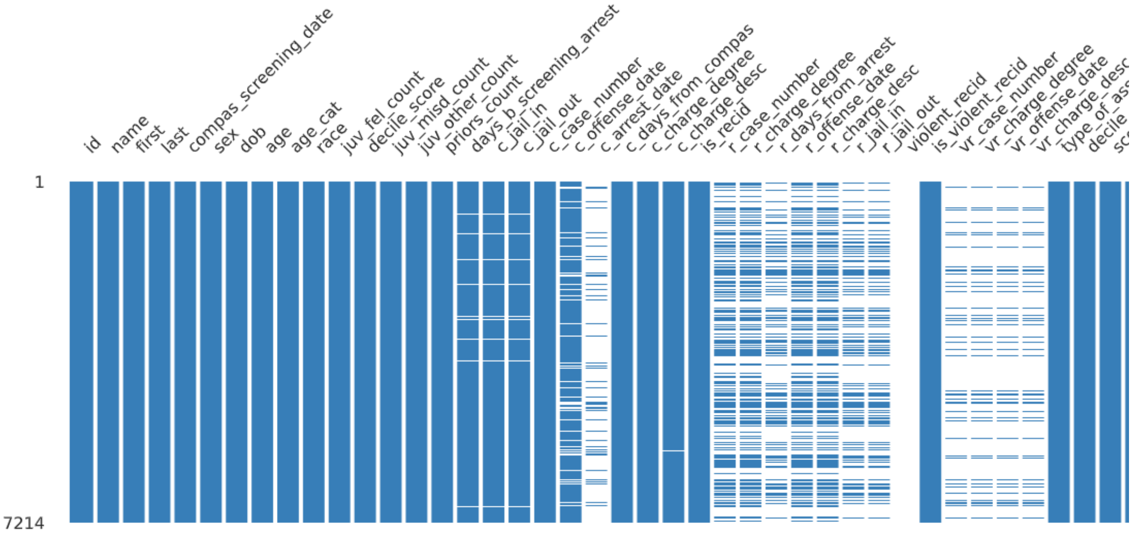
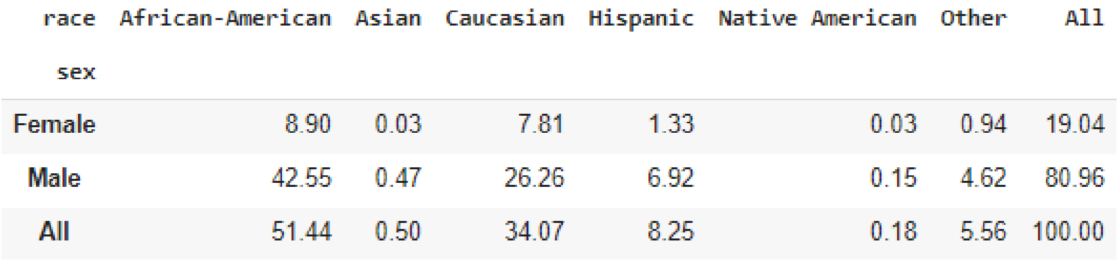


Figure 5-10: Missing Values in the COMPAS Dataset

A series of protected attributes are directly included in the dataset, above all race and gender, hence, fairness through awareness plays an important role in this analysis. The distribution of the main protected variables of sex and race are very skewed as can be observed in the following contingency (or crosstab) table 5-3 analysis:

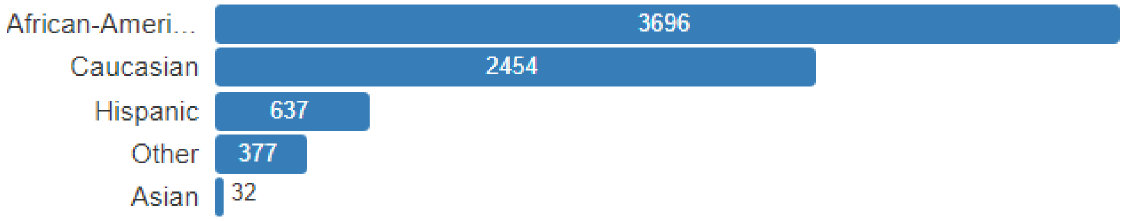


race	African-American	Asian	Caucasian	Hispanic	Native American	Other	All
sex							
Female	8.90	0.03	7.81	1.33	0.03	0.94	19.04
Male	42.55	0.47	26.26	6.92	0.15	4.62	80.96
All	51.44	0.50	34.07	8.25	0.18	5.56	100.00

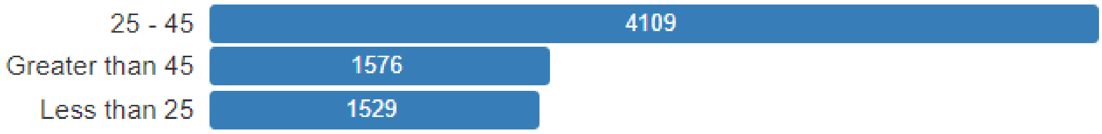
Table 5-3: Contingency Table of Main Protected Variables

Afro-Americans account for 51.44% and Caucasian for 34.07% (total of 85.51%) and men for 80.96% compared to 19.04% women of the entire dataset. The skewed ground truth for race, sex and age groups is important to keep in mind for the analysis in the following sections as illustrated in figure 5-11:

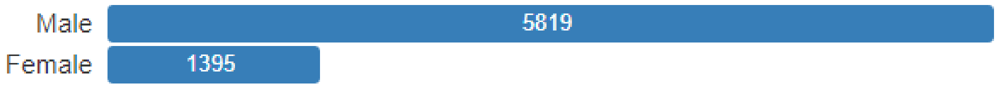
Race:



Age:



Gender:



However, the binary target variable of recidivism is a lot more balanced with 55% representing no recidivism compared to 45% of recidivism:

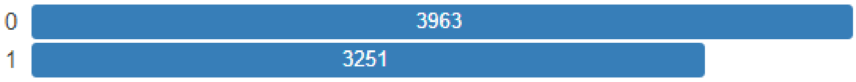


Figure 5-11: Protected Attributes and Target Distributions

It is worth noting that Northpointe claimed to have included 137 variables in its COMPAS model[35] in 2015, however, the exact process and weighting of these 137 features are proprietary and have not been publicly disclosed in detail, which is part of the ongoing debate about the transparency and accountability

[35] Originally published by Northpointe, but now only available at : https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf

of such tools. These features include criminal history, rehabilitation behavior, and responses to a detailed questionnaire. The specific set of features used, and their role in the overall predictive algorithm, is not publicly available due to the proprietary nature of the software.

Figure 5-12 depicts the even decile distribution for blacks and the diminishing distribution for whites, as observed in the initial analysis of COMPAS decile scores. The analysis considered a cohort of 6,172 defendants who had either not been arrested for a new offense or had recidivated within a two-year period. The histograms indicate that scores for white defendants exhibited a skew towards lower-risk categories, whereas black defendants displayed a more uniform distribution across scores.

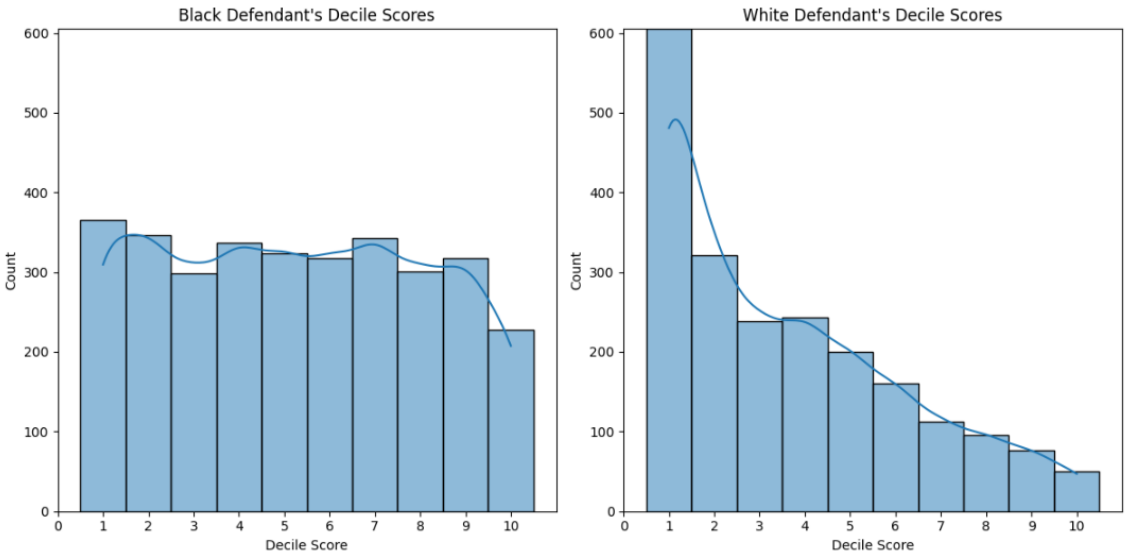


Figure 5-12: Even Decile Distribution for Blacks, Diminishing for Whites

5.4.2 Standard Fairness Tools

The COMPAS dataset with a two-year recidivism score is used in numerous academic papers and a series of standard fairness tools. One can easily claim that this dataset has evolved as one of the main benchmark datasets in the AI bias and fairness academic community.

Aequitas from Carnegie Mellon University provides a concise and brief introduction to the topics.

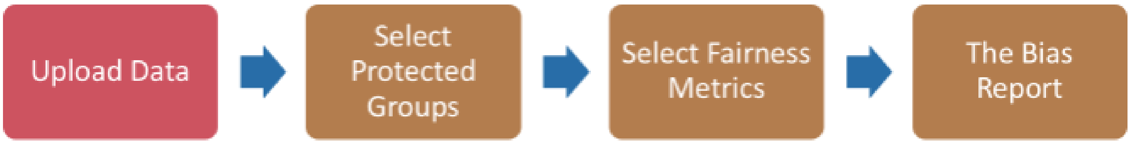
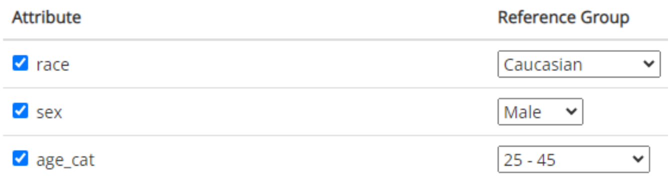

Upload Data → Select Protected Groups → Select Fairness Metrics → The Bias Report

Figure 5-13: Aequitas Online Fairness Process

2. Select protected attributes that need to be audited for bias as in figure 5-14:



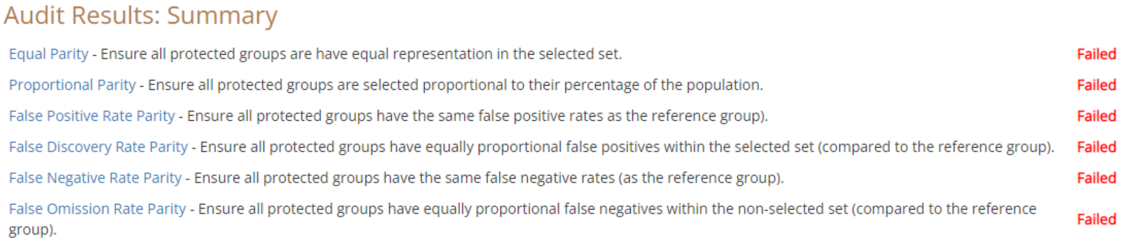
Attribute	Reference Group
☑ race	Caucasian
☑ sex	Male
☑ age_cat	25 - 45

Figure 5-14: Reference Protected Attribute

3. Select Fairness Metrics to Compute:

Equal Parity
Proportional Parity
False Positive Rate Parity
False Discovery Rate Parity
False Negative Rate Parity
False Omission Rate Parity

4. Enter your Disparity Intolerance (in %):



Audit Results: Summary

Equal Parity - Ensure all protected groups are have equal representation in the selected set. Failed

Proportional Parity - Ensure all protected groups are selected proportional to their percentage of the population. Failed

False Positive Rate Parity - Ensure all protected groups have the same false positive rates as the reference group). Failed

False Discovery Rate Parity - Ensure all protected groups have equally proportional false positives within the selected set (compared to the reference group). Failed

False Negative Rate Parity - Ensure all protected groups have the same false negative rates (as the reference group). Failed

False Omission Rate Parity - Ensure all protected groups have equally proportional false negatives within the non-selected set (compared to the reference group). Failed

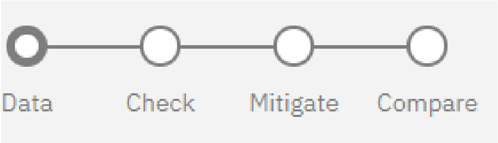
Figure 5-15: Aequitas Audit Results Summary on COMPAS

Figure 5-15 shows that all fairness metrics fail comparing the privileged group of white (Caucasian) males between 25-45 compared to the other groups. Further details and explanations can be retrieved in the Aequitas tool, e.g., as ProPublica (cf. section 5.4.4) claims that the false positive rate is different between Whites and Blacks (African-American), this can be easily retrieved with the following results showing for which groups the audit failed: For race (with reference group as Caucasian)

Other with 0.63X Disparity
African-American with 1.91X Disparity
Asian with 0.37X Disparity
Native American with 1.60X Disparity

According to the Aequitas audit, Blacks are 1.91 more likely to give a false positive, meaning that a black offender is 1.91 times more likely to be falsely predicted as a reoffender than a white offender, supporting ProPublica's view. Further metrics are available, however, it is only a reduced set as can be observed in figure 5-10.

AI Fairness 360 from IBM offers a more holistic view on statistical fairness metrics and mitigation measures. The process goes beyond a simple metrics output and includes a series of additional tools as depicted below:



Predict a criminal defendant’s likelihood of reoffending.

Protected Attributes:

- Sex, privileged: Female, unprivileged: Male

- Race, privileged: Caucasian, unprivileged: Not Caucasian

Protected Attribute: Race

Privileged Group: Caucasian, Unprivileged Group: Not Caucasian

Accuracy with no mitigation applied is 66%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics

Figures 5-16 and 5-17 show how different fairness metrics are applied across different groups compared to the privileged (Caucasian) group, and the strength of the corresponding bias:

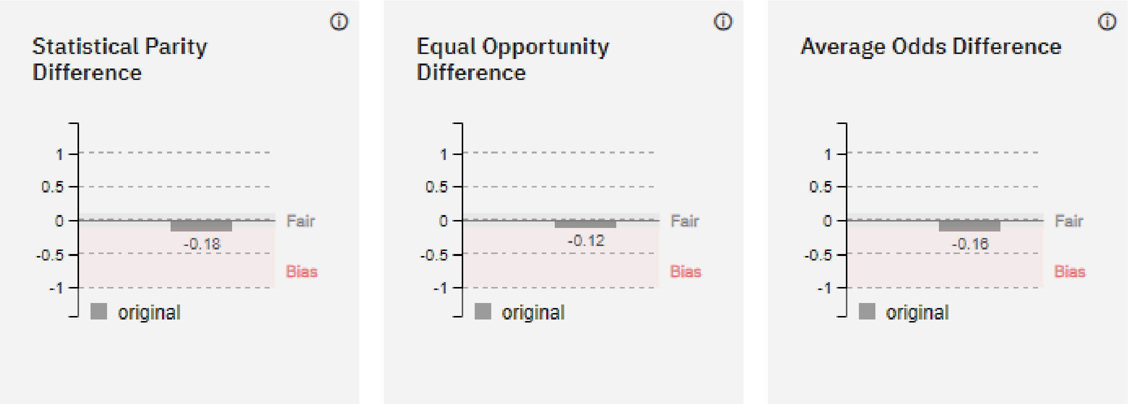


Figure 5-16: AIF360 Fairness Measures – 1

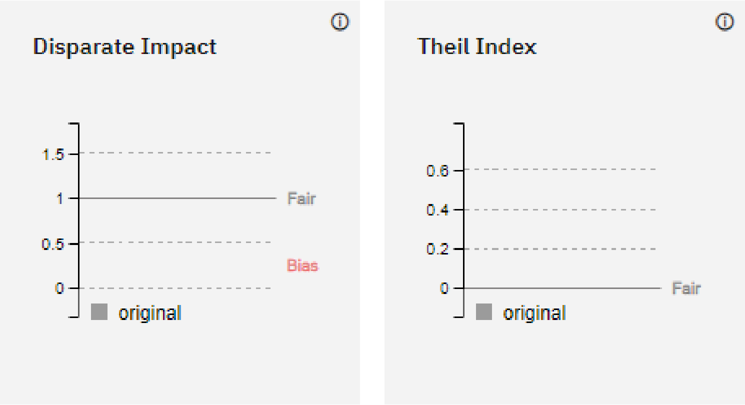


Figure 5-17: AIF360 Fairness Measures – 2

The first two following available mitigation measures focus on the data (pre-processing), whereas the third on the classifier (in-processing) and the last one on the predictions (post-processing):

Reweighing: Weights the examples in each (group, label) combination differently to ensure fairness before classification.

Optimized Pre-processing: Learns a probabilistic transformation that can modify the features and the labels in the training data.

Adversarial Debiasing: Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.

Reject Object Based Classification: Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

For this specific COMPAS dataset the in-processing adversarial debiasing does not yield any clear improvement on bias, however, the pre- and post-processing eliminate the original bias, e.g., with reweighing, bias against unprivileged group was reduced to acceptable levels for 4 of 4 previously biased metrics (0 of 5 metrics still indicate bias for unprivileged group) as can be observed in figures 5-18 and 5-19:

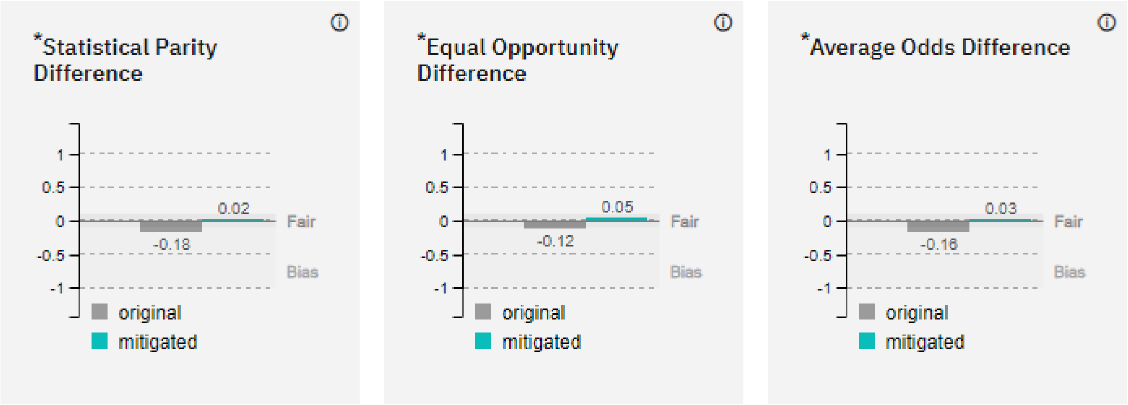


Figure 5-18: Bias Reduction via Reweighing – 1

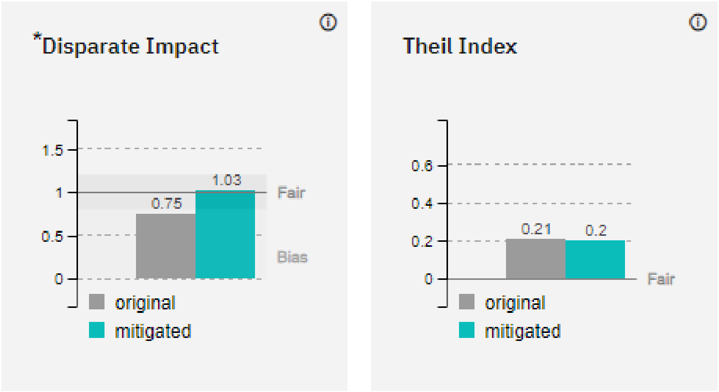


Figure 5-19: Bias Reduction via Reweighing – 2

5.4.3 ProPublica Analysis

ProPublica published the aforementioned article (Angwin, J. et al., 2016) and supporting material on their website[36]. Additionally, the corresponding R code

[36] The article can be reviewed at this link: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-

can be accessed on Github[37], which in turn has been partially reproduced in Python for this master's project to check the validity of the statements made. Own comments and interpretations have been added where it seemed suitable.

Firstly, data cleaning involved removing rows from the dataset under the following circumstances reducing the set to 6172 instances:

- When the date of the crime for which a defendant received a COMPAS score is not within a month of their arrest, the accuracy of the offense data is questioned and the row is consequently removed.
- The 'is_recid' marker is set to -1 when there is no corresponding COMPAS case found, indicating the potential absence of recidivism data.
- Cases pertaining to minor traffic violations, denoted by a 'c_charge_degree' of 'O', are discarded as they don't usually lead to imprisonment.
- The dataset has been narrowed down to data from Broward County, excluding cases that do not represent individuals who either relapsed within a two-year period or spent at least two years away from a correctional facility.

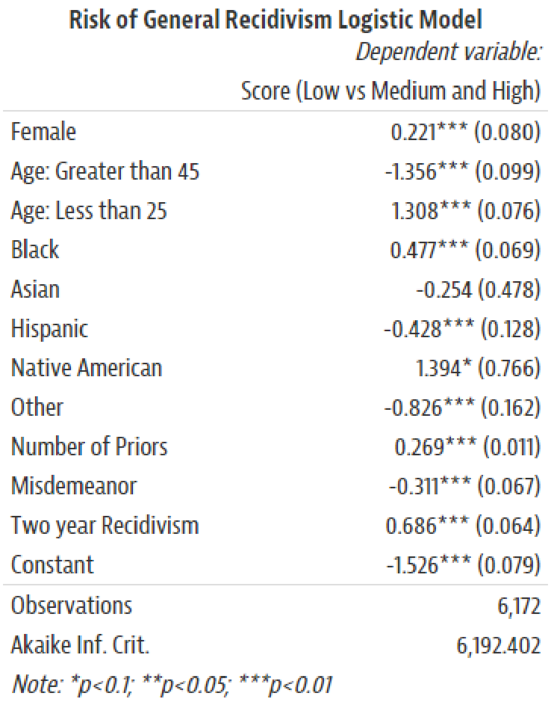
Some first dataset statistics and exploratory data analysis resemble the ones already depicted in the previous section and are omitted here.

Racial Bias in COMPAS via Logistic Regression:

Section 5.4.3 explores the potential racial bias in COMPAS scores. Once erroneous rows are eliminated from the dataset, the initial inquiry centers around possible significant disparities in COMPAS scores among different racial groups. To assess this, certain variables are transformed into factors (such as score, gender, age, race, prior count, crime type, and two-year recidivism), and a logistic regression (LR) is executed comparing low scores to high scores as shown in table 5-4:

sentencing/, meanwhile the methodology has been published here: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm/

[37] Github repository: https://github.com/propublica/compas-analysis/



Risk of General Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	0.221*** (0.080)
Age: Greater than 45	-1.356*** (0.099)
Age: Less than 25	1.308*** (0.076)
Black	0.477*** (0.069)
Asian	-0.254 (0.478)
Hispanic	-0.428*** (0.128)
Native American	1.394* (0.766)
Other	-0.826*** (0.162)
Number of Priors	0.269*** (0.011)
Misdemeanor	-0.311*** (0.067)
Two year Recidivism	0.686*** (0.064)
Constant	-1.526*** (0.079)
Observations	6,172
Akaike Inf. Crit.	6,192.402

*Note: *p<0.1; **p<0.05; ***p<0.01*

Table 5-4: Logistic Regression Coefficients

The findings indicate that, when adjusting for variables like crime severity, prior arrests, and future criminal activity, Black defendants are 45% more likely to receive a higher COMPAS score than their white counterparts. Gender bias is also evident as women are found to be 19.4% more likely to receive a higher score than men. A notable revelation is the pronounced age bias, with individuals under 25 being 2.5 times more likely to receive a higher score compared to middle-aged defendants.

The COMPAS system also provides a score designed to assess an individual's risk of violent recidivism. The accuracy of this score is akin to that of the Recidivism score. A similar methodology, logistic regression, can be utilized to investigate possible racial bias within these scores. The results indicate a pronounced discrepancy, with the violent score overestimating the recidivism rate for Black defendants by 77.3% in comparison to White defendants. Furthermore, an age-related bias is identified, with defendants under the age of 25 being 7.4 times more likely to receive a higher score than middle-aged defendants.

Predictive Accuracy of COMPAS:

The predictive accuracy of the COMPAS scoring system, in terms of determining whether an offender is classified as Low, Medium, or High risk, was evaluated using a Cox Proportional Hazards model. This model, also utilized by Northpointe, the company behind COMPAS, in their validation study[38], is a statistical technique for exploring the relationship between the survival of a subject and several explanatory variables. It operates on the assumption that

[38] Northpointe Validation Study: https://journals.sagepub.com/doi/abs/10.1177/0093854808326545

the effects of the predictors are multiplicative with respect to the hazard rate and are constant over time.

Findings suggest that individuals categorized as High risk are 3.5 times more likely to recidivate. However, the concordance of the COMPAS system, or its predictive power, stands at 63.6%, which is notably lower than the 68% accuracy reported in the Northpointe study. The accuracy increases slightly to 66% when using the COMPAS decile scores.

The presence of racial disparities in the functioning of the algorithm was examined by introducing a race interaction term in the Cox model. The outcome revealed a similar disparity to that observed in the logistic regression analysis. Specifically, High risk White defendants were found to be 3.61 times more likely to recidivate than their Low-risk counterparts, while High-risk Black defendants were 2.99 times more likely to recidivate than their Low-risk peers. This indicates that the model may not behave consistently across different racial groups as illustrated in figure 5-20:

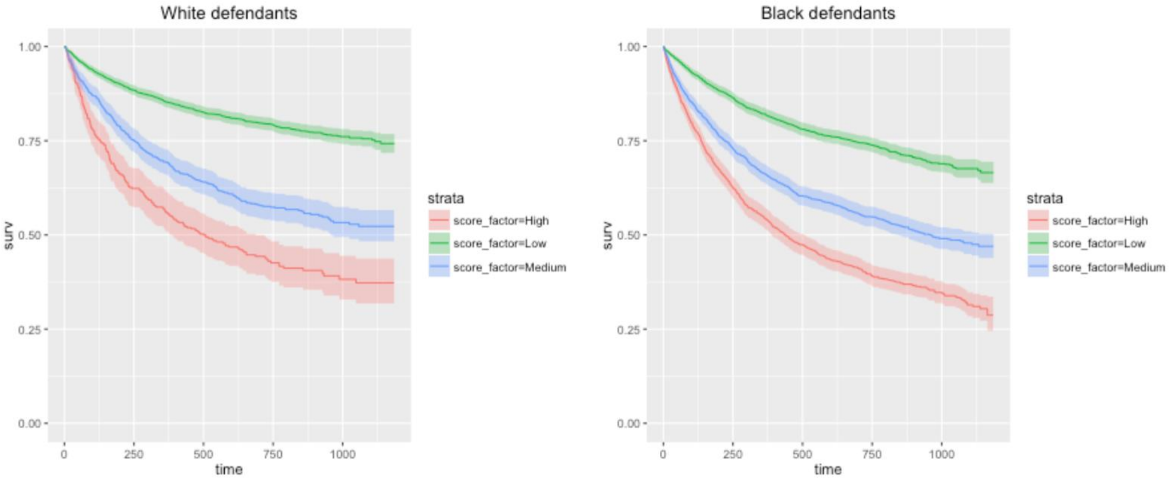


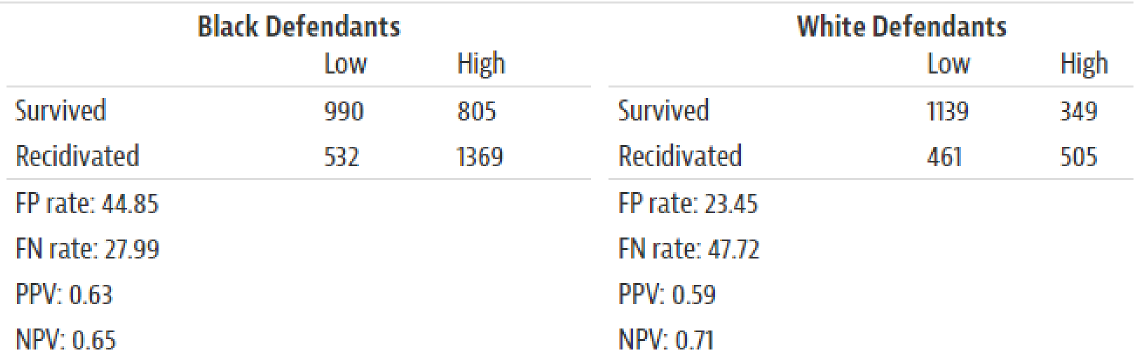
Figure 5-20: Survival Analysis Based on Cox Proportional Hazard Model

Directions of the Racial Bias:

The preceding analysis revealed overprediction of future recidivism for African-American defendants by the COMPAS algorithm, yet the directional bias remains unexplored. A more nuanced understanding of overprediction and underprediction can be achieved by comparing COMPAS scores across racial demographics.

In aggregate, the false positive rate — or the rate at which defendants are incorrectly predicted to recidivate — is 32.35%. However, this rate exhibits racial disparity, reaching 44.85% for African-American defendants, compared to a lower 23.45% for White defendants. Consequently, African-American defendants are 91% more likely than White defendants to receive higher COMPAS scores without committing more crimes over a two-year period.

The COMPAS scoring system also exhibits a higher misclassification rate for White reoffenders as low-risk, doing so 70.4% more frequently than for African-American reoffenders as shown in table 5-5. This underscores a dual-directional bias in the COMPAS scoring system.



Black Defendants			White Defendants		
	Low	High		Low	High
Survived	990	805	Survived	1139	349
Recidivated	532	1369	Recidivated	461	505
FP rate: 44.85			FP rate: 23.45		
FN rate: 27.99			FN rate: 47.72		
PPV: 0.63			PPV: 0.59		
NPV: 0.65			NPV: 0.71		

Table 5-5: False Positive and False Negative Rates for Blacks and Whites

In this line, similar results are confirmed for violent risk scores between Black (African-American) and White (Caucasian) defendants.

In summary, the critique suggests that despite its widespread use, the COMPAS system has inherent biases and inaccuracies that raise serious questions about its fairness and reliability, emphasizing the following points:

- **Racial Bias:** The COMPAS algorithm overestimates the likelihood of recidivism for African-American defendants compared to White defendants, indicating a racial bias, especially via False Positives and False Negatives rates. The COMPAS algorithm exhibits a high false positive rate, especially among African-American defendants, as well as a high false negative rate, particularly among White reoffenders.
- **Predictive Accuracy:** The accuracy of the COMPAS system's predictive power, as measured by concordance, is lower than reported by Northpointe in their validation study.

As a matter of fact, the criticism goes beyond racial bias and also suggests a gender and age bias. The applied models and techniques, however, are identical to the ones illustrated here, hence, no additional insights are gained in this respect.

The issue of **contextual understanding** arises when considering that algorithms, such as the one underlying COMPAS, are devoid of any real-world understanding. They operate purely on mathematical principles, identifying patterns within training data. Consequently, these algorithms may attribute significance to patterns where none exist in a real-world context, leading to spurious associations or overfitting. In the case of COMPAS, the algorithm might, for example, attribute a higher risk of recidivism to certain demographics based on spurious patterns in the training data, leading to potential racial bias and unfair outcomes. Furthermore, the concept of omitted variable bias comes into play. If a variable that influences the outcome is not included in the model, the model may incorrectly assign its predictive power to a correlated variable. Suppose that a significant factor correlated with race was omitted from the COMPAS algorithm; this could lead to the algorithm attributing higher predictive power to race, resulting in racial bias, as was noted in the ProPublica analysis. The issue of algorithmic fragility becomes evident when considering the performance of the COMPAS algorithm. While it might perform well under certain conditions, it can also perform poorly or exhibit bias when the data it is trained on is unbalanced, incomplete, or contains biased patterns. As such, the ProPublica analysis demonstrated that the COMPAS algorithm may overpredict recidivism for African-American defendants, underlining the potential fragility of this predictive model.

6 Results and Conclusions

The project undertakes an exploration of the complex landscape of AI systems, their shortcomings, biases, and the path towards their trustworthy implementation. In this regard, the project comprises the four building blocks of:

1) A theoretical framework to **define, measure and explain the biases and shortcomings** of AI systems,
2) The approaches to **mitigating these biases and improve the trustworthiness** in the AI systems,
3) Enhancing the previous technical analysis with an **ethical, legal, and social perspective**,
4) And finally, combining all three building blocks into **three case studies** with their respective coding implementation.

All initial research questions were tackled and answered as shortly summarized in this chapter.

Research Question 1: How can methods and techniques be applied to develop AD-MS that **enhance transparency and interpretability**, thereby improving their explainability?

> Transparency in AI systems refers to their ability to provide clear, understandable explanations of their decision-making process. However, due to the black-box nature of some advanced AI models such as ensemble models, DNN, CNN or RNN, achieving complete transparency can be challenging. As explained in section 2.2.4, a myriad of techniques in interpreting both the model processing (globally via PDP and feature interaction or locally via LIME, SHAP or anchors) and the model representation (transfer learning, network dissection, PCA, and t-SNE) have been explained and under which circumstances they can be applied to achieve transparency.

Research Question 2: In what ways do **data-driven approaches unintentionally encode human biases** and introduce new ones, and what are the implications of these biases for fairness in AD-MS?

> Fairness, another critical aspect, necessitates that AI systems do not discriminate against any particular group or individual. This principle often faces hurdles due to biases ingrained within the data on which the AI is trained. AI, although extremely potent in many domains, is not devoid of limitations as analyzed in section 2.1. One key challenge is that of bias, a systemic inequality that could arise from multiple sources. These sources often revolve around the data used to train AI systems, the subjective decisions made during model development, or a lack of representation of certain groups in the data. Bias, when present in AI systems, can lead to harmful consequences, including discrimination against specific demographics, reinforcement of societal prejudices, or even misinformed decision-making.
>
> The analysis of the most prevalent and substantial biases within AI systems illustrated a wide variety. These ranged from gender and racial biases in facial recognition systems to socio-economic biases in credit scoring models. The potential harms of such biases include unjust

societal repercussions, unfair allocation of resources, and exacerbation of existing societal inequalities. Further investigation revealed that these biases are often unintentionally encoded within AI systems through data-driven approaches. Data-driven AI systems are only as unbiased as the data they are trained on. If the data carries an inherent bias, the AI system is likely to replicate and amplify that bias in its outcomes.

Research Question 3: How can **fairness in AD-MS be effectively measured**, particularly considering the complex relationships between input features, protected attributes, and target variables?

The analysis on fairness metrics in section 2.2.3 in AD-MS revealed both their indispensability and limitations. Fairness metrics provide quantitative ways to measure and mitigate bias in AI systems. However, an exclusive reliance on them is too narrow and restrictive. The inherent complexity of fairness cannot be entirely captured by a single metric or a set of metrics. This is perfectly illustrated by COMPAS case study in section 5.4 where ProPublica's evaluation primarily used the 'false positive parity' as a measure of fairness, concluding that the COMPAS system was biased against black defendants, whereas Northpointe's argument was based on the predictive parity which is achieved when, for every group, the proportion of positive predictions that are true (known as positive predictive value) is the same. More often than not, there is a necessity for subjective human judgment to assess and ensure fairness as circumstances vary greatly from one case to another.

Research Question 4: What are the **key trade-offs between performance and fairness** in machine learning models, and how can these trade-offs be navigated in practice to balance optimal outcomes with fairness considerations?

A deep-dive into the trade-offs between different fairness metrics in section 2.2.3 brought to light a significant challenge – it is impossible to simultaneously achieve all fairness metrics. For instance, ensuring equal false positive rates across different groups (demographic parity) may not necessarily lead to similar positive predictive values for these groups (predictive parity).

Moreover, the evaluation of key trade-offs discovered between performance and fairness in ML models revealed a key conundrum as explained in section 2.2.5. In many cases, maximizing the performance of an AI system might require accepting a certain degree of unfairness, and vice versa. This trade-off scenario necessitates careful and considered decisions based on the context of the application, societal norms, and legal requirements.

Research Question 5: What methods can be investigated and applied to **minimize the potential for AI systems to introduce and perpetuate discriminatory practices**, reproduce, reinforce, and exacerbate existing biases, and create feedback loops from deployed systems?

The investigation into methods for reducing potential discrimination and bias in AI systems identified effective strategies across pre-, in-, and post-processing stages as described in sections 3.1, 3.2., and 3.3. Pre-processing techniques like 'reweighing' and 'Disparate Impact Remover' helped minimize initial data bias by ensuring balanced representation and increasing group fairness respectively. During in-processing, the 'Reject Option-Based Classification' technique adjusted the fairness of predictions by changing the distribution of favorable outcomes, thereby reducing bias during model training. Post-processing techniques also played a crucial role in bias mitigation after predictions had been made. The case study on the COMPAS dataset in section 5.4 showed the effects of a series of these techniques. Moreover, standard fairness tools presented in section 3.6 played a pivotal role. Tools such as IBM's AI-Fairness 360, Google's What-if Tool, Carnegie Mellon's Aequitas, and Themis AI were evaluated for their capabilities in identifying and mitigating bias. These tools offered unique yet complementary capabilities ranging from data preprocessing to model analysis, allowing for effective bias mitigation. IBM's AI-Fairness 360, for instance, proved valuable for checking and enhancing model fairness, while Google's What-if Tool was helpful in visually analyzing model behavior. These tools, along with others examined, demonstrated their effectiveness as a preliminary point of analysis, offering quantifiable bias measures in AI systems. In conclusion, the integrated application of these techniques and tools could minimize the potential for AI systems to introduce and perpetuate discriminatory practices.

Research Question 6: How can effective methods for **incorporating causality into fairness-aware AD-MS** be applied to mitigate bias and discrimination in decision-making processes?

Applying causality into fairness-aware AD-MS is instrumental to mitigating bias and discrimination. A causal understanding helps in identifying not just the 'what' but also the 'why' behind potential bias in decision-making processes. As demonstrated in the loan approval example in section 3.5, understanding causal structures can expose the real drivers behind apparent discrimination. For instance, if loan approvals are less frequent for a particular demographic group, it is not enough to adjust decisions to balance approvals artificially. It is vital to understand why this disparity exists in the first place. A causal examination can reveal indirect paths of bias, like socio-economic factors, that ultimately affect the loan approval rates. Through techniques such as Causal Discovery and Causal Inference, the sources of bias that might be hidden in correlation-based analyses can be detected, quantified, and addressed. Specific fairness techniques like Counterfactual Fairness can then be applied to reduce these biases effectively. Incorporating causality into fairness-aware AD-MS facilitates the detection, understanding, and mitigation of bias and discrimination in decision-making processes by identifying the root causes and enabling targeted interventions.

Research Question 7: How can **multimodal input features be handled in fairness-aware models**, and what strategies can be employed to mitigate non-apparent bias?

In the HR Recruitment case study in section 5.2, the FairCVtest was leveraged, a framework used to understand biases within AI recruitment tools. The testbed used multimodal synthetic profiles, consciously scored with gender and racial biases, simulating how AI recruitment tools may inadvertently extract sensitive information from unstructured data, leading to unfair decision-making. Findings demonstrated that bias can infiltrate various stages of the learning process, including data collection, preprocessing, and in defining the target function. Through the use of a method called SensitiveNets, it was possible to remove sensitive

> information during the learning process, demonstrating an effective strategy for mitigating non-apparent bias. This approach illustrates the ability to improve fairness, even in the complex landscape of multimodal inputs where biases are often difficult to identify and address. However, understanding the cause of biases in AI systems, including their roots in data collection, preprocessing, and target function definition, remains a critical task. It is essential to continue exploring and implementing methods, such as SensitiveNets, to effectively handle multimodal input features and mitigate non-apparent bias in fairness-aware models. Simultaneously, the importance of robust fairness metrics and transparency in AI recruitment systems cannot be underestimated, as they are key components to ensure the successful mitigation of bias in the decision-making process.

In conclusion, chapter 6 provided a summary of the current state of AI systems, the inherent biases within them, and the paths one can take to mitigate these biases and improve fairness and trustworthiness. Through an exploration of these aspects across theoretical, technical, ethical, legal, and social dimensions, along with in-depth case studies, light has been shed on several pertinent aspects related to AI systems' shortcomings and biases. Despite these extensive investigations, the vast expanse of AI fairness research and its multidisciplinary character inherently leaves a myriad of open research questions. It is crucial to acknowledge that the current understanding, although expansive, is still developing. As AI systems continue to evolve, so will their associated biases, ethical implications, and societal impact. Therefore, this project does not represent a comprehensive exploration but rather a step in the ongoing journey towards more responsible, trustworthy AI systems. In the forthcoming chapter 7. Future Directions, some of the unexplored areas and open questions are explained, signaling possible paths for future research in the quest for more ethical and trustworthy AI.

7 Future Directions

In light of the topics and issues discussed thus far, several open research questions emerge as potential focal points for future exploration in the field of AI fairness, trustworthiness, and ethics. These research questions include, but are not limited to:

1) **Fairness in Multi-class Classification**: The adaption of fairness metrics and approaches to more complex settings, such as multi-class classification problems and environments modeled as Markov decision processes (MDPs), presents a rich vein for future research exploration (e.g., Corbett-Davies, S. et al., 2018). Fairness in this context can be particularly challenging because the interactions between different classes can lead to intricate bias patterns.

2) **Interpretability vs. Performance:** How can we create sophisticated AI systems that do not compromise on interpretability or transparency for the sake of improved performance? What novel methods or approaches could be developed to maintain a balance between complexity, performance, and interpretability?

3) **Robust Fairness Metrics:** What would a more comprehensive set of fairness metrics that captures the intricate nuances of fairness in a diverse set of contexts look like? Can there be universal standards for fairness or do they always need to be context-specific?

4) **Bias in Reinforcement Learning:** How can we understand and mitigate biases in reinforcement learning, particularly when they emerge from the interaction between the agent and the environment, resulting in biased policies or biased reward functions?

5) **Bias in Non-Traditional Data:** As AI begins to use more non-traditional types of data such as images, videos, and voice, how can we detect and mitigate biases present in these forms of data? How do we ensure fairness in multimodal learning systems beyond the simple techniques presented?

6) **Causality in AI:** Can we further incorporate causal reasoning into AI models to better understand and counteract the roots of bias and discrimination? How can the principles of causality be practically applied to various use cases in different industries?

7) **Ethical and Societal Impact Assessment:** How can we best measure and evaluate the ethical and societal impacts of AI systems, and how can this be integrated into the development process of these systems?

8) **Regulatory Frameworks and AI:** What should effective regulation of AI systems look like, ensuring they are used responsibly and fairly, without hindering innovation?

9) **AI in the Global Context:** How can we ensure fairness and inclusivity in AI systems globally, considering the diversity and variations in societal, cultural, and ethical norms around the world?

10) **Bias Mitigation over Time:** As AI systems continue to learn and evolve over time, how can we ensure that bias mitigation techniques are still effective? What new techniques might be necessary to account for this ongoing learning process?

11) **Trustworthy AI in Critical Fields:** How can we ensure trustworthiness and fairness in AI systems that operate in critical sectors like healthcare, law enforcement, or autonomous driving where errors can have dire consequences?

These research questions point to some of the many avenues of exploration that lie ahead. They represent not just theoretical curiosities, but pressing real-world concerns that demand the attention of researchers, practitioners, policymakers, and society at large. Indeed, as we forge ahead on our journey to build more ethical, fair, and trustworthy AI systems, these questions - and many others that may arise - must be at the forefront of our explorations.

One research question seems to be of paramount importance as it impacts especially vulnerable minorities of our societies which is the **intersectionality** of different protected attributes (e.g., black, elderly women). Only recently focus has shifted from a rather narrow perspective on a single protected attribute where possibly no bias exists to a more inclusive one (e.g., Dixon-Fyle, S. et al., 2023). The following discussion shows the importance of the necessary research in this area and illustrates how this project could be enhanced in the future.

12) **Intersectionality**: Intersectionality, a term first coined by Kimberlé Crenshaw in the late 1980s (Crenshaw, K., 1989), is a concept that examines how various forms of oppression, such as racism and sexism, can intersect and compound to create unique experiences of disadvantage for individuals. In the context of AI fairness, intersectionality considers how biases in AI systems may differentially impact individuals based on the intersection of their various identities, such as race, gender, and socio-economic status.
In AD-MS, intersectional biases are prevalent and impactful. They arise when the biases present in the training data, resulting from systemic societal issues, are replicated and even amplified in the outcomes generated by AI systems. A classic example is seen in facial recognition systems (Buolamwini, J. et al., 2018), which have been found to have higher error rates for individuals who are female or have darker skin tones. These biases are further exacerbated for individuals who belong to both these groups – i.e., dark-skinned women.
Fairness metrics, the quantitative measures used to assess bias in AI systems, have historically struggled to adequately account for intersectional bias. Most fairness metrics focus on individual protected attributes, such as race or gender, and do not consider the compound impact of these attributes. This is a significant limitation because bias does not exist in isolation – the intersection of various protected attributes often results in unique forms of discrimination that are not captured by traditional fairness metrics.
The AI fairness tools currently in use, such as IBM's AI-Fairness 360 and Google's What-if Tool, have made significant strides in identifying and mitigating bias in AI systems. However, these tools still face challenges when it comes to addressing intersectional biases. While these tools

provide mechanisms for addressing bias on individual protected attributes, they often do not adequately account for the compounded effect of multiple intersecting attributes.

Real-world case studies provide stark evidence of the impact of intersectional bias. For instance, AI systems used in recruitment processes have been found to disadvantage women of color disproportionately (Kazim, E. et al., 2021). These systems tend to favor resumes that resemble those of people already successful in the field, who are often white and male. As such, the impact of the AI's decision-making process is most detrimental to those at the intersection of multiple marginalized groups.

When handling intersectional biases, there can be significant trade-offs between performance and fairness. An AI model optimized for performance may inadvertently amplify intersectional biases present in the training data, resulting in unfair outcomes. On the other hand, models optimized for fairness might sometimes suffer reduced predictive performance. Thus, striking a balance between performance and fairness when addressing intersectional biases is a complex and delicate task.

The presence of intersectional bias in AI systems raises several ethical and legal considerations. From an ethical standpoint, it is paramount to ensure that AI systems do not perpetuate societal inequities but instead contribute towards their reduction. Legally, intersectional bias can lead to discriminatory outcomes, which may violate anti-discrimination laws in many jurisdictions.

Therefore, addressing intersectional bias is a critical challenge in AI fairness. As the field advances, it is essential to develop more sophisticated fairness metrics, tools, and methodologies that can account for the unique experiences and disadvantages arising from the intersection of various protected attributes. Future research in this area is not only crucial for advancing the technical field but is also a moral imperative to ensure AI systems are fair and just for all users.

In conclusion, it is crucial to underscore the indispensable value of fostering further interdisciplinary research between computer science, law, social sciences, and other relevant fields. The complex and multifaceted nature of fairness in AI necessitates the expertise and perspectives from diverse academic disciplines. Technological solutions alone may not be sufficient to tackle the deeply embedded biases in AI systems, as these biases often stem from systemic social issues. Collaboration between computer scientists, legal scholars, and social scientists could facilitate a more comprehensive understanding of AI fairness issues and contribute to the development of more holistic and effective solutions.

Moreover, this collaboration should extend beyond academia. AI developers, legislators, and social scientists must also engage in active dialogue to ensure the development of fairer AI systems. AI developers can provide insights into the technological possibilities and limitations, legislators can clarify the legal boundaries and requirements, and social scientists can elucidate the societal impacts and norms. This collaboration would foster a reciprocal understanding and create a synergy that could significantly accelerate the progress towards fairer AI systems.

Furthermore, considering the findings regarding ethical, legal, and social implications, it is evident that there is a pressing need for more comprehensive and widely applicable AI ethics education. The development of AI systems is not merely a technical endeavor; it is also a deeply ethical one. AI developers should be equipped with a robust understanding of ethical considerations, legal constraints, and social implications. Moreover, this education should not be limited to those directly involved in AI development. Given the pervasive impact of AI on society, it is crucial for all stakeholders, from policymakers to the general public, to have a basic understanding of AI ethics. This will facilitate informed decision-making and encourage a culture of responsibility and critical engagement with AI technology.

As we move forward into an increasingly AI-driven world, these avenues of exploration present not only as academic pursuits but as necessary steps towards creating a fairer and more inclusive future. As we continue to shape this future, it is paramount that we do so with a commitment to fairness, inclusivity, and respect for all individuals. The path towards trustworthy AI systems is undeniably challenging and complex, yet it is a journey we must undertake, guided by the shared principles of justice and equality.

8 Bibliography

[1] Abraham, D. S., The Elements of Power: Gadgets, Guns, and the Struggle for a Sustainable Future in the Rare Metal Age. Yale University Press, 2017.

[2] Abrams, D. S. et al., “Are Emily and Greg More Employable Than Lakisha and Jamal?,” 2012. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/ARE-EMILY-AND-GREG-MORE-EMPLOYABLE-THAN-LAKISHA-AND-Abrams-Bede/87c46ef0ccd1b6ae214a454da33321fd3333e924

[3] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H., “A Reductions Approach to Fair Classification,” ArXiv, Mar. 2018, Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/A-Reductions-Approach-to-Fair-Classification-Agarwal-Beygelzimer/19cb02117084b023c28da2fb356679806a299890

[4] Alan Nelson, “Unequal treatment: confronting racial and ethnic disparities in health care,” Choice Reviews Online, vol. 40, no. 10, pp. 40-5843-40–5843, Jun. 2003, doi: 10.5860/CHOICE.40-5843.

[5] Altman, A., “Discrimination,” in The Stanford Encyclopedia of Philosophy, E. N. Zalta, Ed., Winter 2020.Metaphysics Research Lab, Stanford University, 2020. Accessed: Jun. 16, 2023. [Online]. Available: https://plato.stanford.edu/archives/win2020/entries/discrimination/

[6] Alvarez, J. M. and Ruggieri, S., “Counterfactual Situation Testing: Uncovering Discrimination under Fairness given the Difference,” 2023, doi: 10.48550/ARXIV.2302.11944.

[7] Andrew Smith, “Using Artificial Intelligence and Algorithms.” Federal Trade Commission, 2020. [Online]. Available: https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial-intelligence-and-algorithms

[8] Angwin, J. L., Jeff, “Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks,” ProPublica, May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[9] Arnold, M. et al., “FactSheets: Increasing trust in AI services through supplier’s declarations of conformity,” IBM J. Res. & Dev., vol. 63, no. 4/5, p. 6:1-6:13, Jul. 2019, doi: 10.1147/JRD.2019.2942288.

[10] Bagdasaryan, E. and Shmatikov, V., “Differential Privacy Has Disparate Impact on Model Accuracy,” presented at the Neural Information Processing Systems, May 2019. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Differential-Privacy-Has-Disparate-Impact-on-Model-Bagdasaryan-Shmatikov/3b1941105317edaef6ac5995089d6d916e5fb483

[11] Bantilan, N., “Themis-ml: A Fairness-aware Machine Learning Interface for End-to-end Discrimination Discovery and Mitigation.” arXiv, Oct. 18, 2017. doi: 10.48550/arXiv.1710.06921.

[12] Barocas, S., Hardt, M., and Narayanan, A., “Limitations and Opportunities,” 2018, [Online]. Available: http://fairmlbook.org

[13] Barocas, S. and Selbst, A. D., “Big Data’s Disparate Impact,” SSRN Journal, 2016, doi: 10.2139/ssrn.2477899.

[14] Bartlett, K. and Crenshaw, K., “Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics [1989],” K. T. Bartlett and R. Kennedy, Eds., Routledge, Feb. 2018, pp. 57–80. doi: 10.4324/9780429500480-5.

[15] Bartlett, R. P., Morse, A., Stanton, R., and Wallace, N., “Consumer Lending Discrimination in the FinTech Era,” SSRN Journal, 2017, doi: 10.2139/ssrn.3063448.

[16] Barton, N. T. L., Paul Resnick, and Genie, “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms,” Brookings, May 22, 2019. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ (accessed Jun. 16, 2023).

[17] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A., “Network Dissection: Quantifying Interpretability of Deep Visual Representations,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3319–3327, Jul. 2017, doi: 10.1109/CVPR.2017.354.

[18] Belkhir, L. and Elmeligi, A., “Carbon footprint of the global pharmaceutical industry and relative impact of its major players,” Journal of Cleaner Production, vol. 214, pp. 185–194, Mar. 2019, doi: 10.1016/j.jclepro.2018.11.204.

[19] Bellamy, R. et al., “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” ArXiv, Oct. 2018, Accessed: Jun. 27, 2023. [Online]. Available: https://www.semanticscholar.org/paper/AI-Fairness-360%3A-An-Extensible-Toolkit-for-and-Bias-Bellamy-Dey/c8541b1dc813f3a638d7acc79e5f972e77f3c5a7

[20] Benjamin Müller, “How Much Will the Artificial Intelligence Act Cost Europe?” Center for Data Innovation, 2021. [Online]. Available: https://www2.datainnovation.org/2021-aia-costs.pdf.

[21] Bertsimas, D., Orfanoudaki, A., and Wiberg, H., Interpretable clustering: an optimization approach, vol. 110, no. 1. Springer US, 2021, p. 138. doi: 10.1007/s10994-020-05896-2.

[22] Binns, R., “Fairness in Machine Learning: Lessons from Political Philosophy,” Decision-Making in Computational Design & Technology eJournal, Dec. 2017, Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Fairness-in-Machine-Learning%3A-Lessons-from-Binns/2a944564c2466883ec14a6f6ef461f0e34d21b38

[23] Board of Governors of the Federal Reserve System., “Report on the Economic Well-Being of U.S. Households in 2019.” 2020. [Online]. Available: https://www.federalreserve.gov/publications/files/2019-report-economic-well-being-us-households-202005.pdf

[24] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A., “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” presented at the NIPS, Jul. 2016. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Man-is-to-Computer-Programmer-as-Woman-is-to-Word-Bolukbasi-Chang/ccf6a69a7f33bcf052aa7def176d3b9de495beb7

[25] Boyd, D. and Crawford, K., “CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon,” Information, Communication & Society, vol. 15, no. 5, pp. 662–679, Jun. 2012, doi: 10.1080/1369118X.2012.678878.

[26] Brand, D. J., “Algorithmic decision-making and the law,” eJournal of eDemocracy and Open Government, vol. 12, no. 1, pp. 115–132, 2020, doi: 10.29379/jedem.v12i1.576.

[27] Brynjolfsson, E. and McAfee, A. P., “The second machine age: work, progress, and prosperity in a time of brilliant technologies, 1st Edition,” 2014. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/The-second-machine-age%3A-work%2C-progress%2C-and-in-a-of-Brynjolfsson-McAfee/a3ee6c2ee186160306cdcc9ebc03865dff2f754d

[28] Buolamwini, J. and Gebru, T., “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” presented at the FAT, Jan. 2018. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Gender-Shades%3A-Intersectional-Accuracy-Disparities-Buolamwini-Gebru/18858cc936947fc96b5c06bbe3c6c2faa5614540

[29] Caliskan, A., Bryson, J. J., and Narayanan, A., “Semantics derived automatically from language corpora contain human-like biases,” Science, vol. 356, no. 6334, pp. 183–186, Apr. 2017, doi: 10.1126/science.aal4230.

[30] Calmon, F., Wei, D., Ramamurthy, K., and Varshney, K., “Optimized Data Pre-Processing for Discrimination Prevention,” ArXiv, Apr. 2017, Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Optimized-Data-Pre-Processing-for-Discrimination-Calmon-Wei/b6aea5d9b79f2a49a25453b23e73f5e0ac58f1e0

[31] Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C., “A clarification of the nuances in the fairness metrics landscape,” Sci Rep, vol. 12, no. 1, p. 4209, Mar. 2022, doi: 10.1038/s41598-022-07939-1.

[32] Castelvecchi, D., “Can we open the black box of AI?,” Nature, vol. 538, no. 7623, pp. 20–23, Oct. 2016, doi: 10.1038/538020a.

[33] Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K., “Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees,” Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 319–328, Jan. 2019, doi: 10.1145/3287560.3287586.

[34] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., “SMOTE: Synthetic Minority Over-sampling Technique,” jair, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[35] Chen-Oster v. Goldman Sachs, Goldman Gender Case. 2018. [Online]. Available: https://goldmangendercase.com/

[36] Chohlas-Wood, A., Coots, M., Goel, S., and Nyarko, J., “Designing Equitable Algorithms,” 2023, doi: 10.48550/ARXIV.2302.09157.

[37] Chouldechova, A., “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments,” Big Data, vol. 5, no. 2, pp. 153–163, Jun. 2017, doi: 10.1089/big.2016.0047.

[38] Chouldechova, A. and Roth, A., “A snapshot of the frontiers of fairness in machine learning,” Commun. ACM, vol. 63, no. 5, pp. 82–89, Apr. 2020, doi: 10.1145/3376898.

[39] Corbett-Davies, S. and Goel, S., “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning,” ArXiv, Jul. 2018, Accessed: Jul. 14, 2023. [Online]. Available: https://www.semanticscholar.org/paper/The-Measure-and-Mismeasure-of-Fairness%3A-A-Critical-Corbett-Davies-Goel/acd6de3ac2a3d9449aae51b87fbb03f6f0020954

[40] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A., “Algorithmic decision making and the cost of fairness,” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. Part F1296, pp. 797–806, 2017, doi: 10.1145/3097983.3098095.

[41] Crawford, K., Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven London, 2022.

[42] Crawford, K., Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas and Amba Kak, V. M., Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker., “AI Now 2019 Report,” New York: AI Now Institute, 2019, [Online]. Available: https://ainowinstitute.org/AI_Now_2019_Report.html

[43] Crawford, Kate, Roel Dobbe, Theodora Dryer, and Genevieve Fried, “AI Now Report 2019.” : AI Now Institute, 2019. [Online]. Available: https://ainowinstitute.org/wp-content/uploads/2023/04/AI_Now_2019_Report.pdf

[44] Creedon, T. B., Zuvekas, S. H., Hill, S. C., Ali, M. M., McClellan, C., and Dey, J. G., “Effects of Medicaid expansion on insurance coverage and health services use among adults with disabilities newly eligible for Medicaid,” Health Services Research, vol. 57, no. S2, pp. 183–194, Dec. 2022, doi: 10.1111/1475-6773.14034.

[45] Creemers, R., “China’s Social Credit System: An Evolving Practice of Control,” SSRN Journal, 2018, doi: 10.2139/ssrn.3175792.

[46] d’Alessandro, B., O’Neil, C., and LaGatta, T., “A Data Scientist’s Guide to Discrimination-Aware Classification Authors:,” 2019. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/A-Data-Scientist%E2%80%99s-Guide-to-Discrimination-Aware-d'Alessandro-O'Neil/2592aa0de9955a5b5bfc0039387dacb5874a1107

[47] d'Alessandro, B., O'Neil, C., and LaGatta, T., "Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification," Big Data, vol. 5, no. 2, pp. 120–134, Jun. 2017, doi: 10.1089/big.2016.0048.

[48] Dastin, J., "Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women," Reuters, pp. 296–299, Oct. 2022. doi: 10.1201/9781003278290-44.

[49] Datta, A., Tschantz, M. C., and Datta, A., "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination," ArXiv, Aug. 2014, Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Automated-Experiments-on-Ad-Privacy-Settings%3A-A-of-Datta-Tschantz/770d59a2cc0dbbbc964cac1c7e2d51897222e507

[50] Dietz, G. and Den Hartog, D. N., "Measuring trust inside organisations," Personnel Review, vol. 35, no. 5, pp. 557–588, Sep. 2006, doi: 10.1108/00483480610682299.

[51] Doshi-Velez, F. and Kim, B., "Towards A Rigorous Science of Interpretable Machine Learning," arXiv: Machine Learning, Feb. 2017, Accessed: Jun. 27, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Towards-A-Rigorous-Science-of-Interpretable-Machine-Doshi-Velez-Kim/5c39e37022661f81f79e481240ed9b175dec6513

[52] Dougherty, C., "Google Photos Mistakenly Labels Black People 'Gorillas,'" New York Times, Jul. 2015. [Online]. Available: https://archive.nytimes.com/bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/

[53] Duckworth, A. L. and Yeager, D. S., "Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes," Educational Researcher, vol. 44, no. 4, pp. 237–251, May 2015, doi: 10.3102/0013189X15584327.

[54] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R., "Fairness through awareness," Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226, Jan. 2012, doi: 10.1145/2090236.2090255.

[55] Elman, J. L., "Finding Structure in Time," Cognitive Science, vol. 14, no. 2, pp. 179–211, Mar. 1990, doi: 10.1207/s15516709cog1402_1.

[56] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S., "Runaway Feedback Loops in Predictive Policing," presented at the FAT, Jun. 2017. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Runaway-Feedback-Loops-in-Predictive-Policing-Ensign-Friedler/2a33a1e161d6e3bc91e1f33ad7172d29d3ce0b73

[57] EU High-Level Expert Group on AI, "Ethics guidelines for trustworthy AI | Shaping Europe's digital future," Apr. 08, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed Jun. 27, 2023).

[58] European Commission and the Council of Economic Advisers (CEA), “THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE FUTURE OF WORKFORCES IN THE EUROPEAN UNION AND THE UNITED STATES OF AMERICA.” US-EU Trade and Technology Council Inaugural Statement, 2021.

[59] European DIGITAL SME Alliance, “DIGITAL SME reply to the AI Act consultation.” 2021. [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665574_en.

[60] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S., “Certifying and Removing Disparate Impact,” Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268, Aug. 2015, doi: 10.1145/2783258.2783311.

[61] Ferguson, A. G., “The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement,” Dec. 2020, doi: 10.18574/nyu/9781479854608.001.0001.

[62] Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J., Eds., Big Data and Social Science: A Practical Guide to Methods and Tools. Boca Raton, FL, 2016.

[63] Frank A. Pasquale, “The black box society: the secret algorithms that control money and information,” in Choice Reviews Online, Jun. 2015, pp. 52-5426-52–5426. doi: 10.5860/CHOICE.190706.

[64] Frey, C. B. and Osborne, M. A., “The future of employment: How susceptible are jobs to computerisation?,” Technological Forecasting and Social Change, vol. 114, pp. 254–280, Jan. 2017, doi: 10.1016/j.techfore.2016.08.019.

[65] Friedman, J. H., “Greedy function approximation: A gradient boosting machine.,” Ann. Statist., vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.

[66] Future of Life Institute (FLI), “Future of Life Institute (FLI) reply to the AI Act consultation.” 2021. [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665546_en

[67] Genovesi, S. et al., “Standardizing fairness-evaluation procedures: interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans,” AI Ethics, May 2023, doi: 10.1007/s43681-023-00291-8.

[68] Giaccardi, E., Cila, N., Speed, C., and Caldwell, M., “Thing Ethnography: Doing Design Research with Non-Humans,” Proceedings of the 2016 ACM Conference on Designing Interactive Systems, pp. 377–387, Jun. 2016, doi: 10.1145/2901790.2901905.

[69] Gohar, U. and Cheng, L., “A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges,” 2023, doi: 10.48550/ARXIV.2305.06969.

[70] Goodman, B. and Flaxman, S., “European union regulations on algorithmic decision making and a ‘right to explanation,’” AI Magazine, vol. 38, no. 3, pp. 50–57, 2017, doi: 10.1609/aimag.v38i3.2741.

[71] Green, C. R. et al., “The Unequal Burden of Pain: Confronting Racial and Ethnic Disparities in Pain,” Pain Med, vol. 4, no. 3, pp. 277–294, Sep. 2003, doi: 10.1046/j.1526-4637.2003.03034.x.

[72] Greene, T., “Facebook’s feckless ‘Fairness Flow’ won’t fix its broken AI,” TNW | Deep-Tech, Mar. 31, 2021. https://thenextweb.com/news/facebooks-feckless-fairness-flow-wont-fix-its-broken-ai (accessed Jul. 03, 2023).

[73] Greenwell, B., M., “pdp: An R Package for Constructing Partial Dependence Plots,” The R Journal, vol. 9, no. 1, p. 421, 2017, doi: 10.32614/RJ-2017-016.

[74] Grgic-Hlaca, N., Zafar, M. B., Gummadi, K., and Weller, A., “The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making,” 2016. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/The-Case-for-Process-Fairness-in-Learning%3A-Feature-Grgic-Hlaca-Zafar/fdb6a159cb65f4d1147224998d56e67f0398948b

[75] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q., “On Calibration of Modern Neural Networks,” presented at the International Conference on Machine Learning, Jun. 2017. Accessed: Jun. 27, 2023. [Online]. Available: https://www.semanticscholar.org/paper/On-Calibration-of-Modern-Neural-Networks-Guo-Pleiss/d65ce2b8300541414bfe51d03906fca72e93523c

[76] Haataja, M. and Bryson, J. J., “What costs should we expect from the EU’s AI Act?” Aug. 27, 2021. doi: 10.31235/osf.io/8nzb4.

[77] Hardin, R., “Conceptions and explanations of trust.,” 2001. Accessed: Jun. 27, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Conceptions-and-explanations-of-trust.-Hardin/9c4361728123cdc47c3a88414bf7b11a7427b38f

[78] Hardt, M., Price, E., and Srebro, N., “Equality of Opportunity in Supervised Learning,” ArXiv, Oct. 2016, Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Equality-of-Opportunity-in-Supervised-Learning-Hardt-Price/d42b11ce90c9c69a20ed015b73dc33e0e4100a7b

[79] Hardt, M. and Recht, B., “Patterns, predictions, and actions: A story about machine learning,” 2021, [Online]. Available: http://arxiv.org/abs/2102.05242

[80] Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., and Wallach, H., “Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?,” Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–16, May 2019, doi: 10.1145/3290605.3300830.

[81] Hooker, S., “The hardware lottery,” Commun. ACM, vol. 64, no. 12, pp. 58–65, Dec. 2021, doi: 10.1145/3467017.

[82] ICML, “Conference on Machine Learning 2019 - Trustworthy Machine Learning,” Accessed: Jun. 27, 2023. [Online]. Available: https://icml.cc/Conferences/2019

[83] Jang, T., Zheng, F., and Wang, X., “Constructing a Fair Classifier with the Generated Fair Data,” 35th AAAI Conference on Artificial Intelligence, AAAI 2021, vol. 9A, pp. 7908–7916, 2021, doi: 10.1609/aaai.v35i9.16965.

[84] Jeklin, A., “The AI Atlas,” no. July, pp. 1–23, 2016.

[85] Jia, R. and Liang, P., “Adversarial Examples for Evaluating Reading Comprehension Systems,” Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2021–2031, 2017, doi: 10.18653/v1/D17-1215.

[86] Johndrow, J. E. and Lum, K., “An algorithm for removing sensitive information: Application to race-independent recidivism prediction,” Ann. Appl. Stat., vol. 13, no. 1, Mar. 2019, doi: 10.1214/18-AOAS1201.

[87] Kaltheuner, F. and Bietti, E., “Data is power: Towards additional guidance on profiling and automated decision-making in the GDPR,” Journal of Information Rights, Policy and Practice, vol. 2, no. 2, pp. 1–17, 2018, doi: 10.21039/irpandp.v2i2.45.

[88] Kamiran, F. and Calders, T., “Data preprocessing techniques for classification without discrimination,” Knowl Inf Syst, vol. 33, no. 1, pp. 1–33, Oct. 2012, doi: 10.1007/s10115-011-0463-8.

[89] Kamiran, F., Calders, T., and Pechenizkiy, M., “Discrimination Aware Decision Tree Learning,” 2010 IEEE International Conference on Data Mining, pp. 869–874, Dec. 2010, doi: 10.1109/ICDM.2010.50.

[90] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J., “Fairness-Aware Classifier with Prejudice Remover Regularizer,” P. A. Flach, T. De Bie, and N. Cristianini, Eds., in Lecture Notes in Computer Science, vol. 7524. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 35–50. doi: 10.1007/978-3-642-33486-3_3.

[91] Kaufmann, M., Egbert, S., and Leese, M., “Predictive Policing and the Politics of Patterns,” The British Journal of Criminology, vol. 59, no. 3, pp. 674–692, Apr. 2019, doi: 10.1093/bjc/azy060.

[92] Kazim, E., Koshiyama, A. S., Hilliard, A., and Polle, R., “Systematizing Audit in Algorithmic Recruitment,” J. Intell., vol. 9, no. 3, p. 46, Sep. 2021, doi: 10.3390/jintelligence9030046.

[93] Kearns, M., Neel, S., Roth, A., and Wu, Z. S., “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness,” presented at the International Conference on Machine Learning, Nov. 2017. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Preventing-Fairness-Gerrymandering%3A-Auditing-and-Kearns-Neel/19930147204c97be4d0964e166e8fe72ac1d6c3d

[94] Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B., "Avoiding Discrimination through Causal Reasoning," ArXiv, Jun. 2017, Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Avoiding-Discrimination-through-Causal-Reasoning-Kilbertus-Rojas-Carulla/ee2c3b095ad226a211fd674ec1d4c960c07af107

[95] Kirkpatrick, K., "It's not the algorithm, it's the data," Commun. ACM, vol. 60, no. 2, pp. 21–23, Jan. 2017, doi: 10.1145/3022181.

[96] Kleinberg, J., Mullainathan, S., and Raghavan, M., "Inherent Trade-Offs in the Fair Determination of Risk Scores," p. 23 pages, 2017, doi: 10.4230/LIPICS.ITCS.2017.43.

[97] Kozodoi, N., Jacob, J., and Lessmann, S., "Fairness in credit scoring: Assessment, implementation and profit implications," European Journal of Operational Research, vol. 297, no. 3, pp. 1083–1094, Mar. 2022, doi: 10.1016/j.ejor.2021.06.023.

[98] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[99] Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R., "Counterfactual Fairness," presented at the NIPS, Mar. 2017. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Counterfactual-Fairness-Kusner-Loftus/043f084e379a44608c470059c2aa174a323e9774

[100] Lambrecht, A. and Tucker, C., "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads," Management Science, vol. 65, no. 7, pp. 2966–2981, Jul. 2019, doi: 10.1287/mnsc.2018.3093.

[101] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[102] Lee, K.-F., Ai Superpowers: China, Silicon Valley, and the New World Order. Boston, 2019.

[103] Leverhulme Centre for the Future of Intelligence and Centre for the Study of Existential Risk, "Leverhulme Centre reply to the AI Act consultation." University of Cambridge, 2021. [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665626_en

[104] Li, M., Mickel, A., and Taylor, S., "'Should This Loan be Approved or Denied?': A Large Dataset with Class Assignment Guidelines," Journal of Statistics Education, vol. 26, no. 1, pp. 55–66, 2018, doi: 10.1080/10691898.2018.1434342.

[105] Li, S., Yu, J., Du, X., Lu, Y., and Qiu, R., "Fair Outlier Detection Based on Adversarial Representation Learning," Symmetry, vol. 14, no. 2, Art. no. 2, Feb. 2022, doi: 10.3390/sym14020347.

[106] Lipton, Z. C., McAuley, J., and Chouldechova, A., “Does mitigating ML’s impact disparity require treatment disparity?,” presented at the Neural Information Processing Systems, Nov. 2017. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Does-mitigating-ML's-impact-disparity-require-Lipton-McAuley/932404745d960291925b3f27b71734dff5b23633

[107] Llongueras, M. D. ; M., A., “Consorcios y valor de Shapley,” Congreso Galego de Estadística e Investigación de Operacións. “IX Congreso Galego de Estadística e Investigación de Operacións,” no. June, pp. 173–179, 2010.

[108] Long, R., “Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness,” Journal of Moral Philosophy, vol. 19, no. 1, pp. 49–78, Nov. 2021, doi: 10.1163/17455243-20213439.

[109] Lum, K. and Isaac, W., “To Predict and Serve?,” Significance, vol. 13, no. 5, pp. 14–19, Oct. 2016, doi: 10.1111/j.1740-9713.2016.00960.x.

[110] Lundberg, S. M. and Lee, S.-I., “A Unified Approach to Interpreting Model Predictions,” ArXiv, May 2017, Accessed: Jul. 02, 2023. [Online]. Available: https://www.semanticscholar.org/paper/A-Unified-Approach-to-Interpreting-Model-Lundberg-Lee/442e10a3c6640ded9408622005e3c2a8906ce4c2

[111] Lütz, F., “Gender equality and artificial intelligence in Europe. Addressing direct and indirect impacts of algorithms on gender-based discrimination,” ERA Forum, vol. 23, no. 1, pp. 33–52, 2022, doi: 10.1007/s12027-022-00709-6.

[112] Maaten, L. and Hinton, G. E., “Visualizing Data using t-SNE,” Journal of Machine Learning Research, 2008, Accessed: Jul. 02, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Visualizing-Data-using-t-SNE-Maaten-Hinton/1c46943103bd7b7a2c7be86859995a4144d1938b

[113] Maitland, C., “Trust in cyberspace,” Telecommunications Policy, vol. 24, no. 6–7, pp. 628–631, Aug. 2000, doi: 10.1016/S0308-5961(00)00046-X.

[114] Mayer, R., Davis, J. H., and Schoorman, F., “An integrative model of organizational trust, Academy of Management Review, : .,” 1995. Accessed: Jul. 08, 2023. [Online]. Available: https://www.semanticscholar.org/paper/An-integrative-model-of-organizational-trust%2C-of-%3A-Mayer-Davis/3c7098f78e6446131e90095cebb781dee9904b26

[115] McCarthy, J. M., Marvin; Rochester, Nathan; Shannon, Claude, “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence,” Dartmouth Summer Research Project 1956, 1955, [Online]. Available: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

[116] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A., “A Survey on Bias and Fairness in Machine Learning,” ACM Comput. Surv., vol. 54, no. 6, pp. 1–35, Jul. 2022, doi: 10.1145/3457607.

[117] Menon, A. and Williamson, R. C., “The cost of fairness in binary classification,” presented at the FAT, Jan. 2018. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/The-cost-of-

fairness-in-binary-classification-Menon-Williamson/9d55b5fc3f3b92324f4a9c46d5b66d11895ba565

[118] Mitchell, B., “An Examination of Murder,” Murder and Penal Policy, no. December, pp. 40–73, 1990, doi: 10.1007/978-1-349-20745-9_2.

[119] Mitchell, S., Potash, E., Barocas, S., D’Amour, A., and Lum, K., “Algorithmic fairness: Choices, assumptions, and definitions,” Annual Review of Statistics and Its Application, vol. 8, pp. 141–163, 2021, doi: 10.1146/annurev-statistics-042720-125902.

[120] Morales, A., Fierrez, J., and Vera-Rodríguez, R., “SensitiveNets: Learning Agnostic Representations with Application to Face Recognition,” ArXiv, Feb. 2019, Accessed: Jul. 08, 2023. [Online]. Available: https://www.semanticscholar.org/paper/SensitiveNets%3A-Learning-Agnostic-Representations-to-Morales-Fierrez/7d0b7a42368d7fb78ade5e21cad713b5c5611ee4

[121] Mueller, B., “How Much Will the Artificial Intelligence Act Cost Europe?,” Center for Data Innovation, 2021.

[122] Narayanan, A., “21 Fairness Definitions and Their Politics,” 2019, [Online]. Available: https://fairmlbook.org/tutorial2.html

[123] Nedlund, E., “Apple Card is accused of gender bias. Here’s how that can happen | CNN Business,” CNN, Nov. 12, 2019. https://www.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html (accessed Jun. 16, 2023).

[124] Nizami, M. and Bogliolo, A., “Investigation and Mitigation of Bias in Explainable AI (short paper),” presented at the BEWARE@AI*IA, 2022. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Investigation-and-Mitigation-of-Bias-in-Explainable-Nizami-Bogliolo/7e13939030e15bec52da99c6daf0aa2f44332aa0

[125] O’Neil, C., Weapons Of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown/Archetype, 2017.

[126] Orwat, C., “Diskriminierungsrisiken durch Verwendung von Algorithmen,” Antidiskriminierungsstelle des Bundes, pp. 1–400, 2019.

[127] Pagano, T. P. et al., “Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods,” Big Data and Cognitive Computing, vol. 7, no. 1, pp. 15–15, 2023, doi: 10.3390/bdcc7010015.

[128] Patton, D. U., Brunton, D.-W., Dixon, A., Miller, R. J., Leonard, P., and Hackman, R., “Stop and Frisk Online: Theorizing Everyday Racism in Digital Policing in the Use of Social Media for Identification of Criminal Conduct and Associations,” Social Media + Society, vol. 3, no. 3, p. 205630511773334, Jul. 2017, doi: 10.1177/2056305117733344.

[129] Pearl, J., “Causality: Models, Reasoning, and Inference,” Cambridge University Press, Sep. 2009. doi: 10.1017/CBO9780511803161.

[130] Pena, A., Serna, I., Morales, A., and Fierrez, J., “Bias in Multimodal AI: Testbed for Fair Automatic Recruitment,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 129–137, Jun. 2020, doi: 10.1109/CVPRW50498.2020.00022.

[131] Pessach, D. and Shmueli, E., “Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings,” Expert Systems with Applications, vol. 185, p. 115667, Dec. 2021, doi: 10.1016/j.eswa.2021.115667.

[132] Phillips-Brown, M., “Algorithmic neutrality,” 2023, doi: 10.48550/ARXIV.2303.05103.

[133] Rawls, J., A Theory of Justice: Original Edition. Harvard University Press, 1971. doi: 10.2307/j.ctvjf9z6v.

[134] Rempel, J. K., Holmes, J. G., and Zanna, M. P., “Trust in close relationships.,” Journal of Personality and Social Psychology, vol. 49, no. 1, pp. 95–112, Jul. 1985, doi: 10.1037/0022-3514.49.1.95.

[135] Ribeiro, M. T., Singh, S., and Guestrin, C., “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, Aug. 2016, doi: 10.1145/2939672.2939778.

[136] Richardson, R., Schultz, J., and Crawford, K., “Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice,” Feb. 2019. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Dirty-Data%2C-Bad-Predictions%3A-How-Civil-Rights-Data%2C-Richardson-Schultz/9a43ab4a3d1aab2095bfbba60a1ddb8396d5c084

[137] Russell, C., Kusner, M. J., Loftus, J. R., and Silva, R., “When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness,” presented at the NIPS, Dec. 2017. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/When-Worlds-Collide%3A-Integrating-Different-in-Russell-Kusner/c076c741790674226109178 86c6566b2504e52e2

[138] Schneier, B., Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World. New York: W. W. Norton & Company, 2016.

[139] Seaver, N., “What Should an Anthropology of Algorithms Do?,” Cult. Anthropol., vol. 33, no. 3, pp. 375–385, Aug. 2018, doi: 10.14506/ca33.3.04.

[140] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J., “Fairness and Abstraction in Sociotechnical Systems,” Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59–68, Jan. 2019, doi: 10.1145/3287560.3287598.

[141] Silberg, J. and Manyika, J., “Notes from the AI frontier : Tackling bias in AI (and in humans) Article,” McKinsey Global Institute, pp. 1–8, 2019.

[142] Silberg, J. and Manyika, J., “Notes from the AI frontier: Tackling bias in AI (and in humans),” McKinsey Glob. Inst., 2019, pp. 1–8. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Notes-

from-the-AI-frontier%3A-Tackling-bias-in-AI-in/945fc1294764084b60aa12810cc6e45e6eb67116

[143] Simonyan, K., Vedaldi, A., and Zisserman, A., “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,” CoRR, Dec. 2013, Accessed: Jul. 02, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Deep-Inside-Convolutional-Networks%3A-Visualising-and-Simonyan-Vedaldi/dc6ac3437f0a6e64e4404b1b9d188394f8a3bf71

[144] Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M., “SmoothGrad: removing noise by adding noise,” ArXiv, Jun. 2017, Accessed: Jul. 02, 2023. [Online]. Available: https://www.semanticscholar.org/paper/SmoothGrad%3A-removing-noise-by-adding-noise-Smilkov-Thorat/f538dca4def5167a32fbc12107b69a05f0c9d832

[145] Smith, J. J. and Beattie, L., “RecSys Fairness Metrics: Many to Use But Which One To Choose?,” 2022, doi: 10.48550/ARXIV.2209.04011.

[146] Smuha, N. A., “Beyond the individual: governing AI’s societal harm,” Internet Policy Review, vol. 10, no. 3, Sep. 2021, doi: 10.14763/2021.3.1574.

[147] Smuha, N. A. et al., “How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission’s Proposal for an Artificial Intelligence Act,” SSRN Journal, 2021, doi: 10.2139/ssrn.3899991.

[148] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. A., “Striving for Simplicity: The All Convolutional Net,” CoRR, Dec. 2014, Accessed: Jul. 02, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Striving-for-Simplicity%3A-The-All-Convolutional-Net-Springenberg-Dosovitskiy/0f84a81f431b18a78bd97f59ed4b9d8eda390970

[149] Strubell, E., Ganesh, A., and McCallum, A., “Energy and Policy Considerations for Deep Learning in NLP,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650, 2019, doi: 10.18653/v1/P19-1355.

[150] Sundiatu Dixon-Fyle, Klaudia Gegotek, Nyasha Tsimba, Tania Holt, and Tunde Olanrewaju, “Race in the UK workplace: The intersectional experience | McKinsey.” https://www.mckinsey.com/bem/our-insights/race-in-the-uk-workplace-the-intersectional-experience?stcr=981732D320D049E1ABEC5AA0A772574C&cid=other-eml-dre-mip-mck&hlkid=a38f23a33ce246c2a7b0f8b1cb08e793&hctky=14863471&hdpid=c73d766a-9eef-4f74-bcea-d7b7295192e0#/ (accessed Jul. 08, 2023).

[151] Suresh, H. and Guttag, J., “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle,” Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–9, Oct. 2021, doi: 10.1145/3465416.3483305.

[152] Thampi, A., Interperetable AI: Building Explainable Machine Learning Systems. Manning, 2022, p. 275.

[153] The Future Society, “The Future Society reply to the AI Act consultation.” 2021. [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665611_en

[154] Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A., “Robustness May Be at Odds with Accuracy,” arXiv: Machine Learning, May 2018, Accessed: Jun. 27, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Robustness-May-Be-at-Odds-with-Accuracy-Tsipras-Santurkar/1b9c6022598085dd892f360122c0fa4c630b3f18

[155] Varshney, K. R., Trustworthy Machine Learning. Independently published, 2022.

[156] Veale, M. and Zuiderveen Borgesius, F., “Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach,” Computer Law Review International, vol. 22, no. 4, pp. 97–112, Aug. 2021, doi: 10.9785/cri-2021-220402.

[157] Wachter, S., Mittelstadt, B., and Russell, C., “Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law,” SSRN Electronic Journal, pp. 1–51, 2021, doi: 10.2139/ssrn.3792772.

[158] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S., “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding,” Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 353–355, 2018, doi: 10.18653/v1/W18-5446.

[159] Xenidis, R., “Tuning EU equality law to algorithmic discrimination: Three pathways to resilience,” Maastricht Journal of European and Comparative Law, vol. 27, no. 6, pp. 736–758, Dec. 2020, doi: 10.1177/1023263X20982173.

[160] Xiang, A., “Reconciling Legal and Technical Approaches to Algorithmic Bias,” Tennessee Law Review, vol. 88, no. 3, pp. 649–649, 2021.

[161] Xu, J. et al., “Algorithmic fairness in computational medicine,” eBioMedicine, vol. 84, pp. 104250–104250, 2022, doi: 10.1016/j.ebiom.2022.104250.

[162] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H., “How transferable are features in deep neural networks?,” presented at the NIPS, Nov. 2014. Accessed: Jul. 02, 2023. [Online]. Available: https://www.semanticscholar.org/paper/How-transferable-are-features-in-deep-neural-Yosinski-Clune/081651b38ff7533550a3adfc1c00da333a8fe86c

[163] Yu, P. K., “The Algorithmic Divide and Equality in the Age of Artificial Intelligence,” Florida Law Review, vol. 72, no. 19, pp. 19–44, 2020.

[164] Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P., “Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment,” 26th International World Wide Web Conference, WWW 2017, pp. 1171–1180, 2017, doi: 10.1145/3038912.3052660.

[165] Zemel, R., Wu, L. Y., Swersky, K., Pitassi, T., and Dwork, C., “Learning Fair Representations,” presented at the International Conference on Machine Learning, Jun. 2013. Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Learning-Fair-Representations-Zemel-Wu/37c3303d173c055592ef923235837e1cbc6bd986

[166] Zhang, B. H., Lemoine, B., and Mitchell, M., “Mitigating Unwanted Biases with Adversarial Learning,” Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 335–340, Dec. 2018, doi: 10.1145/3278721.3278779.

[167] Zhang, J. and Bareinboim, E., “Fairness in Decision-Making — The Causal Explanation Formula,” in Proceedings of the AAAI Conference on Artificial Intelligence, Apr. 2018. doi: 10.1609/aaai.v32i1.11564.

[168] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W., “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 2979–2989. doi: 10.18653/v1/D17-1323.

[169] Žliobaitė, I., “A survey on measuring indirect discrimination in machine learning,” ArXiv, Oct. 2015, Accessed: Jun. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/A-survey-on-measuring-indirect-discrimination-in-%C5%BDliobait%C4%97/bd0cb6f62619649616316ca3c1348f568c63a852

[170] Zuboff, P. S., The Age of Surveillance Capitalism The Fight for a Human Future at the New Frontier of Power. 2018.

9 Annexes

Annex 1: Anti-Discrimination Laws in the EU, USA, and China

Anti-discrimination laws in the EU, USA, and China

Law or Directive	Year in Force	Most Relevant Articles	Region	Short Description
Civil Rights Act, Title II and VII	1964	Title II - Section 202, Title VII	United States	Legislation used to prevent discrimination in public spaces, facilities and employment, potentially extended to digital spaces to prevent bias in AI systems.
Fair Housing Act	1968	Title VIII	United States	Protects people from discrimination when they are renting or buying a home, getting a mortgage, seeking housing assistance, or engaging in other housing-related activities, relevant in the context of AI systems used in housing and real estate.
Education Amendments Act, Title IX	1972	Title IX	United States	Prohibits discrimination on the basis of sex in any federally funded education program or activity, relevant for AI systems used in educational settings.

Equal Credit Opportunity Act	1974	N/A	United States	Prohibits credit discrimination on the basis of race, color, religion, national origin, sex, marital status, age, or because someone receives public assistance, relevant for AI systems used in credit scoring.
Americans with Disabilities Act	1990	Title I, II, III	United States	Prohibits discrimination against people with disabilities in several areas, including employment, transportation, public accommodations, communications, and access to state and local government programs and services, could be extended to AI systems.
Race Equality Directive (2000/43/EC)	2000	Article 2, 3	EU	Prohibits all forms of racial or ethnic discrimination in various areas, including employment, education, social protection, and access to goods and services, relevant for AI systems used in these sectors.

Employment Equality Directive (2000/78/EC)	2000	Article 2, 3	EU	Prohibits discrimination in employment on the grounds of religion or belief, disability, age, or sexual orientation, relevant for AI systems used in hiring and employment.
Equal Treatment Directive (proposed)[39]	Proposed, not in force as of 2023	N/A	EU	A proposed directive that would extend EU anti-discrimination protections beyond employment to include areas like social protection, education, and access to goods and services.
Regulations on Employment Services	2007	Article 3	China	Prohibits discrimination in employment on grounds of ethnicity, race, gender, religious belief, etc., relevant for AI systems used in hiring and employment.
Women's Protection Law	2023	N/A	China	Aiming to give women stronger protection against sexual harassment and gender discrimination

Table 9-1: Anti-Discrimination Laws in the EU, USA, and China

[39] https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vktj9botz0zd

Annex 2: German Credit Dataset

All variables of the German Credit dataset are encoded as depicted in the "value range" column. This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0): http://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

No.	Description	Type	Value Range
1	Status of existing checking account	qualitative	A11: ... < 0 DM A12: 0 <= ... < 200 DM A13: ... >= 200 DM / salary assignments for at least 1 year A14: no checking account
2	Duration in month	numerical	
3	Credit history	qualitative	A30: no credits taken/ all credits paid back duly A31: all credits at this bank paid back duly A32: existing credits paid back duly till now A33: delay in paying off in the past A34: critical account/ other credits existing (not at this bank)
4	Purpose	qualitative	A40: car (new) A41: car (used) A42: furniture/equipment A43: radio/television A44: domestic appliances A45: repairs A46: education A47: (vacation - does not exist?) A48: retraining A49: business A410: others
5	Credit amount	numerical	
6	Savings account/bonds	qualitative	A61: ... < 100 DM A62: 100 <= ... < 500 DM A63: 500 <= ... < 1000 DM A64: .. >= 1000 DM A65: unknown/ no savings account
7	Present employment since	qualitative	A71: unemployed A72: ... < 1 year A73: 1 <= ... < 4 years A74: 4 <= ... < 7 years A75: .. >= 7 years

8	Installment rate in percentage of disposable	numerical	income
9	Personal status and sex	qualitative	A91: male: divorced/separated A92: female: divorced/separated/married A93: male: single A94: male: married/widowed A95: female: single
10	Other debtors / guarantors	qualitative	A101: none A102: co-applicant A103: guarantor
11	Present residence since	numerical	
12	Property	qualitative	A121 : real estate A122 : if not A121 : building society savings agreement/ life insurance A123 : if not A121/A122 : car or other, not in attribute A124 : unknown / no property
13	Age in years	numerical	
14	Other installment plans	qualitative	A141: bank A142: stores A143: none
15	Housing	qualitative	A151: rent A152: own A153: for free
16	Number of existing credits at this bank	numerical	
17	Job	qualitative	A171: unemployed/ unskilled - non-resident A172: unskilled - resident A173: skilled employee/ official A174: management/ self-employed/ highly qualified employee/ officer
18	Number of people being liable to provide	numerical	
	maintenance for		
19	Telephone	qualitative	A191: none A192: yes, registered under the customer's name
20	Foreign worker	qualitative	A201: yes A202: no

Table 9-2: German Credit Datset

Annex 3: COMPAS Dataset

The COMPAS dataset as provided by ProPublica does not contain a description of the 53 variables. Although it is a widely used benchmark dataset and some of the variables seem self-explanatory, none of the researchers provides a clear overview of all variables. The following table is meant to close this gap and show all 53 variables used in the 2-year recidivism dataset published by ProPublica in 2016.

This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0): https://github.com/propublica/compas-analysis

No.	Column name	Description
1	id	unique ID that identifies each suspect
2	name	name of the suspect
3	first	first name of the suspect
4	last	last name of the suspect
5	compas_screening_date	COMPAS screening date of the suspect
6	sex	sex of the suspect
7	dob	date of birth of the suspect
8	age	age of the suspect at the time of the survey
9	age_cat	age category of the suspect
10	race	race of the suspect
11	juv_fel_count	the number of felony charges as a juvenile
12	decile_score	recidivism score from 1 to 10
13	juv_misd_count	the number of misdemenor charges as a juvenile
14	juv_other_count	the number of other charges as a juvenile
15	priors_count	the number of prior conviction for the suspect
16	days_b_screening_arrest	the count of days between screening date and (original) arrest date. If they are too far apart, that may indicate an error. If the value is negative, that indicate the screening date happened before the arrest date.
17	c_jail_in	start timestamp of incarceration
18	c_jail_out	end timestamp of incarceration
19	c_case_number	charge case number of the suspect
20	c_offense_date	charge offense date of the suspect
21	c_arrest_date	charge arrest date of the suspect
22	c_days_from_compas	the number of days between committing an offense and going to jail
23	c_charge_degree	charge degree of the suspect
24	c_charge_desc	charge description of the suspect

25	is_recid	whether the suspect recidivates
26	r_case_number	recidivism case number of the suspect
27	r_charge_degree	recidivism charge degree of the suspect
28	r_days_from_arrest	number of days between the person get re-arrested from the re-offense date
29	r_offense_date	recivism offense date of the suspect
30	r_charge_desc	recidivism charge description of the suspect
31	r_jail_in	time and date when the suspect goes in the jail for recidivism
32	r_jail_out	time and date when the suspect gets released from the jail for recidivism
33	violent_recid	violent recidivism, all missing values (can be omitted)
34	is_violent_recid	violent recidivism crime indicator of the suspect
35	vr_case_number	violent_case_number of the suspect
36	vr_charge_degree	violent_charge_degree of the suspect
37	vr_offense_date	violent_offense_date of the suspect
38	vr_charge_desc	violent_charge_description of the suspect
39	type_of_assessment	constant 'Risk of Recidivism' for all rows, can be omitted
40	decile_score	repition of column 12
41	score_text	decile score text: low, medium, high
42	screening_date	COMPAS screening date of the suspect
43	v_type_of_assessment	constant 'Risk of Violence' for all rows, can be omitted
44	v_decile_score	violent recidivism score from 1 to 10
45	v_score_text	violent recidivism score text: low, medium, high
46	v_screening_date	COMPAS screening date of the suspect for violent crimes
47	in_custody	custody start date
48	out_custody	custody end date
49	priors_count	the number of prior conviction for the suspect
50	start	survival analysis: start point of the suspect entering the survival analysis
51	end	survival analysis: end point of the suspect entering the survival analysis
52	event	binary indicator that denotes whether the event of recidivism has occurred or not
53	two_year_recid	target: two year recidivism (binary 0 / 1)

Table 9-3: COMPAS Dataset as Compiled by ProPublica in 2016