Fairlib title

Tesi di laurea in:
INTELLIGENT SYSTEMS ENGINEERING

Relatore
Prof. Giovanni Ciatto

Candidato
Valerio Di Zio

Correlatori
Prof. Andrea Omicini
Prof.ssa Roberta Calegari

Abstract

Automated decision-making systems are rapidly permeating socially sensitive domains such as finance, healthcare, justice, and autonomous mobility. While these data-driven solutions can increase efficiency, they can also perpetuate or amplify existing inequities whenever the underlying algorithms exhibit unfair behaviour. This thesis provides a systematic investigation of algorithmic fairness, clarifying the multiple, often competing, formal definitions of fairness adopted in the literature and mapping them to the practical risks of bias and discrimination that arise throughout the machine-learning pipeline.

After surveying the principal sources of bias—data imbalance, historical prejudice, model opacity, and feedback loops—the work reviews state-of-the-art mitigation strategies grouped into three families: pre-processing (data-repair and re-sampling techniques), in-processing (fairness-aware losses, constraints, and regularisers), and post-processing (prediction-adjustment and explanation tools). Building upon these foundations, the thesis introduces FairLib: a modular, open-source library that unifies bias-diagnosis metrics, visual analytics, and mitigation algorithms behind a consistent API. FairLib is designed to be model-agnostic, seamlessly integrating with popular ML frameworks, and to facilitate reproducible experimentation through configurable pipelines.

A preliminary evaluation on the canonical Adult income dataset indicates that selected FairLib pipelines can measurably reduce unfairness while leaving overall predictive accuracy broadly unchanged. Although limited to one benchmark and modest in scope, these findings suggest that systematic fairness interventions are achievable without prohibitive performance trade-offs.

By coupling a critical analysis of fairness concepts with a practical, extensible toolkit, this thesis aims to foster greater transparency and accountability in artificial-intelligence systems and to assist practitioners in deploying models that respect fundamental principles of equity.

A tutte le persone che hanno reso questi anni magici e indimenticabili.

Contents

25

Chapter 1

Introduction

Artificial-intelligence (AI) systems increasingly mediate decisions about credit, employment, healthcare, justice, and mobility. Their scale and opacity mean that even small statistical biases can translate into systematic, large-scale harm. The goal of *algorithmic fairness* is therefore to ensure that automated decisions do not introduce—nor amplify—unjustified discrimination against protected groups [MMS⁺21]. Achieving that goal, however, remains challenging because unfairness can emerge at many points in the machine-learning (ML) pipeline, from data collection to model deployment, and because technical, legal, and ethical notions of "fair" often conflict in practice [SG21].

1.1 From Bias to Mitigation: Scope of This Report

1. **Diagnose the problem:** We survey the landscape of algorithmic unfairness, definitions of fairness, sources of bias (representation, measurement, aggregation, and historical), and the feedback loops that can entrench disparities. We also review state-of-the-art mitigation strategies that operate at three key stages of the ML lifecycle: *pre-processing* (data repair and reweighting), *in-processing* (fairness-constrained learning), and *post-processing* (output adjustment).

2. **Offer a practical solution:** In this study we present **FairLib**, a modular Python library that consolidates a broad spectrum of fairness metrics and bias-mitigation algorithms within a single, production-oriented interface. By elevating bias detection, remediation, and continuous auditing to integral components of the machine-learning development lifecycle, FairLib significantly reduces the practical barriers to deploying equitable intelligent systems.

1.2 Why Another Library? Limitations of the Status Quo

Although IBM's *AI Fairness 360* (AIF360) [BDH+18] and Microsoft's *Fairlearn* [BDE+20] have stimulated important research, they remain ill-suited for day-to-day engineering because they exhibit four persistent shortcomings:

- **Fragmented APIs and terminology:** Conceptually identical operations are exposed through incompatible class names, argument conventions, and data structures, forcing practitioners to maintain fragile "glue code" whenever they switch between metrics or mitigation algorithms.

- **Verbosity and operational complexity:** Even elementary tasks—such as computing a single fairness metric or launching a standard debiasing pass—require boiler-plate scripts and nested wrapper objects, raising the learning curve and discouraging rapid experimentation.

- **Binary-only focus:** Most built-in metrics and mitigation routines assume *binary* sensitive attributes and *binary* classification targets; multi-class settings are either unsupported or require significant user re-implementation.

- **Integration barriers and rigid internal models:** Neither toolkit plugs cleanly into existing *PyTorch* or *scikit-learn* pipelines: mitigation methods are hard-wired to internally defined estimators that can be adjusted only through superficial hyper-parameters. A practitioner cannot, for example,

inject a domain-specific neural network as the base learner for a fairness algorithm without rewriting large portions of the codebase.

1.3 FairLib Objectives

FairLib is explicitly engineered so that each objective eliminates a corresponding limitation identified above:

- **Unifying fragmented APIs:** A coherent, scikit-learn–style interface (`fit`, `transform`, `predict`) and consistent terminology replace the incompatible conventions that plague existing libraries, removing the need for brittle integration code.

- **Reducing verbosity and complexity:** Group- and individual-level metrics share a vectorised NumPy core, so computing a metric or launching a debiasing routine becomes a one-liner, dramatically accelerating prototyping.

- **Extending beyond binary settings:** All fairness metrics and evaluation utilities natively support multi-class targets and both binary and categorical sensitive attributes, while debiasing algorithms offer varying levels of support for multi-class settings together addressing key gaps in existing fairness libraries.

- **Providing end-to-end debiasing:** Harmonised *pre-* and *in-processing* modules (with post-processing on the roadmap) let practitioners inject fairness at the point most compatible with data-governance and privacy constraints.

- **Enabling customisable trade-offs and models:** Every mitigation algorithm surfaces its fairness penalty as an hyperparameter and accepts a user-supplied estimator, which can range from linear models to bespoke neural networks, thereby supporting automatic exploration of the fairness–utility frontier without sacrificing architectural flexibility.

- **Ensuring seamless workflow integration:** Native PyTorch wrappers and drop-in scikit-learn compatibility allow FairLib components to slot directly

into production notebooks, model registries, and CI/CD pipelines, overcoming the rigid, "model-locked" designs of prior toolkits.

Chapter 2

Background

2.1 Introduction to AI Fariness

The use of machine learning algorithms has permeated every aspect of our lives: from movie recommendations to hiring decisions and even to high-stakes areas such as loan approvals or risk assessments in the judicial system. Although these systems offer significant advantages—such as processing vast amounts of information without fatigue—they are also susceptible to biases that can lead to unfair decisions. As machine learning systems continue to be deployed in increasingly sensitive domains, ensuring fairness becomes a critical concern for developers, policymakers, and society at large [MMS+21].

The concept of fairness in AI is multifaceted and involves technical, legal, and ethical considerations. While technical solutions can address some aspects of bias, a comprehensive approach requires understanding the social contexts in which these systems operate and the historical inequalities they might perpetuate or amplify. This report provides an overview of fairness in AI, including its definitions, origins of bias, mitigation strategies, and applications across various domains.

2.1.1 What is Fairness?

Fairness refers to the principle that every individual or group should be treated justly and impartially. In decision-making, this means that processes, whether human or automated, should not introduce unjustified discrimination based on

inherent or acquired characteristics such as gender, ethnicity, or socioeconomic status [MNDWLS24].

The notion of fairness is deeply rooted in philosophical and legal traditions, with concepts such as distributive justice, procedural fairness, and equal opportunity providing frameworks for understanding what constitutes fair treatment [KKBK21]. In the context of automated decision systems, fairness extends these principles to ensure that algorithms do not perpetuate or amplify existing social biases [BCG21].

For example, in a recruitment system, fairness requires that candidates be evaluated solely on their skills and qualifications, free from biases related to gender or ethnicity [Che23]. Similarly, in credit scoring, a fair system would assess creditworthiness based on relevant financial factors rather than characteristics protected by anti-discrimination laws [GGR24].

It's worth noting that fairness is context-dependent, and what constitutes fair treatment may vary across different domains and cultures. Nevertheless, certain fundamental principles, such as non-discrimination and equal treatment of similar cases, remain consistent across contexts [KKBK21].

2.1.2 The Importance of Fairness in AI Systems

Ensuring fairness in artificial intelligence is crucial, especially when these systems make decisions that directly affect people's lives [CDG23, YW24]. Unfair AI systems can perpetuate and even exacerbate existing social inequalities, affecting individuals' access to opportunities, resources, and services.

A canonical example is the COMPAS system, used in U.S. courts to assess the risk of recidivism. Studies have shown that COMPAS tends to produce higher false-positive rates for African-American defendants compared to Caucasian defendants under similar conditions. This means that Black defendants were more likely to be incorrectly labeled as high-risk, potentially leading to harsher sentencing or denial of parole [YW24].

Beyond the justice system, fairness concerns arise in numerous domains:

- **Employment:** Resume screening algorithms might inadvertently favor candidates from certain demographic groups or educational backgrounds [KBA24].

- **Healthcare:** Diagnostic tools calibrated on data from predominantly one demographic group may be less accurate for others[NHS+24].

- **Financial Services:** Credit scoring algorithms might systematically disadvantage certain communities, reinforcing historical patterns of exclusion [Fer23].

- **Education:** Automated evaluation systems might penalize students whose learning styles or language patterns differ from the norm [KBA24].

The consequences of unfair AI systems extend beyond individual harms to societal impacts. When automated systems systematically disadvantage certain groups, they can erode trust in technology, reinforce stereotypes, and contribute to social polarization. Instead, fair AI systems have the potential to promote inclusiveness, reduce discrimination and create more equitable outcomes [Fer23].

2.1.3 Origins of Bias and the Feedback Loop

Bias in AI can stem from multiple sources, each requiring different approaches for mitigation:

- **Bias in the Data:** If the training data is skewed or unrepresentative of the true population, the algorithm will learn and reproduce these biases, sometimes even amplifying them. For instance, facial recognition systems trained predominantly on light-skinned faces perform worse on darker-skinned individuals [BG18].

- **Bias in the Problem Formulation:** How we define the problem and choose variables can introduce bias. For example, using proxies like ZIP codes instead of more direct measures can inadvertently incorporate demographic biases if neighborhoods are segregated.

- **Bias in the Algorithms:** Even with unbiased data, design choices—such as optimization functions, regularization techniques, or modeling assumptions—can introduce unfairness. For instance, prioritizing overall accuracy might lead to models that perform well on majority groups but poorly on minorities [NHS+24].

- **Bias in Evaluation:** If we evaluate models using metrics that don't account for fairness considerations, we might deploy systems that appear successful but discriminate in practice.

A particularly concerning phenomenon is the **feedback loop**, where biased predictions influence future data collection, which in turn reinforces the original bias [EFN+17]. For instance, predictive policing algorithms might direct more officers to neighborhoods already experiencing higher policing, leading to more arrests, which then reinforce the prediction that these areas have higher crime rates. This creates a self-fulfilling prophecy that can entrench and amplify existing disparities.

Historical biases present in society can also manifest in AI systems through the data they're trained on. Language models trained on internet text may inherit sexist or racist associations present in their training corpus. Word embeddings trained on standard text corpora reflect gender and racial stereotypes present in society [CBN17].

2.1.4 Types of Bias and Sources of Discrimination

Understanding the various types of bias that can affect AI systems is essential for developing effective mitigation strategies. A number of frameworks have been proposed to categorize these biases, including those affecting data, model design, and deployment contexts [MMS+21, SG21].

- **Representation Bias** arises when the dataset does not adequately reflect the diversity of the population. For instance, Buolamwini and Gebru found that commercial facial recognition systems perform worst on darker-skinned women, illustrating severe underrepresentation in benchmark datasets [BG18, SG21].

- **Measurement Bias** occurs when proxies are used in place of direct measurements. For example, arrest records used in predictive policing introduce racial disparities due to historically biased law enforcement practices [EFN+17, SG21].

- **Aggregation Bias**, **Evaluation Bias**, **User Interaction Bias**, and **Temporal Bias** are discussed as distinct sources of harm that emerge when models are designed, validated, or deployed without accounting for subgroup-specific differences or time-based shifts. These were formally characterized in lifecycle frameworks proposed by Suresh and Guttag [SG21].

- **Historical Bias** reflects the reproduction of long-standing social inequalities. Language models, for instance, have been shown to encode and replicate gender and racial stereotypes learned from text corpora [CBN17].

- **Algorithmic Bias** stems from model design decisions, including feature selection, optimization objectives, or regularization strategies, which may produce disparate outcomes even when the training data is balanced [MMS+21].

Discrimination in AI can be **direct**, involving explicit use of protected attributes, or **indirect**, where neutral criteria disproportionately impact certain groups. These forms can be explainable, meaning they can be justified by relevant non-protected attributes, or unexplainable, which is considered unfair or illegal. Frameworks for quantifying and mitigating such discrimination have been proposed to distinguish between the two [KZC12, CDG23].

Understanding these distinctions is crucial for ethical and legal assessments of AI. While certain disparities may reflect legitimate factors (e.g., qualifications), others are manifestations of unjustified bias that require targeted safeguards in both design and deployment phases.

2.1.5 Definitions of Fairness

One of the central challenges in developing fair AI systems is that there is no single, universally agreed-upon definition of fairness. Different notions of fairness capture different intuitions and may be more appropriate in different contexts [VR18] [MPB+21].

Some of the most commonly used formal definitions include:

- **Demographic Parity:** Requires that the probability of a positive outcome is the same across all groups, regardless of protected attributes. Mathemat-

ically, for a decision $\hat{Y}$ and a protected attribute A:

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b) \text{ for all values } a, b \text{ of } A$$

This definition aims for equal representation but may not be appropriate when the base rates genuinely differ across groups.

- **Equalized Odds:** Demands that, conditioned on the true outcome, both true positive and false positive rates are equal across groups [HPS16]:

$$P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = b)$$
$$\text{for } y \in \{0, 1\} \text{ and all values } a, b \text{ of } A$$

This ensures that the algorithm's errors are distributed fairly across groups.

- **Equal Opportunity:** A relaxation of equalized odds that focuses solely on ensuring that individuals who truly belong to the positive class have an equal chance of being correctly classified:

$$P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b) \text{ for all values } a, b \text{ of } A$$

This is particularly relevant in contexts like hiring, where we want qualified candidates to have equal chances regardless of their demographic group.

- **Predictive Parity:** Requires that the precision of the classifier is the same across all groups:

$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b) \text{ for all values } a, b \text{ of } A$$

This ensures that a positive prediction has the same meaning regardless of group membership.

Importantly, these definitions can conflict with each other [KMR16]. This creates inherent trade-offs that must be navigated based on the specific context and ethical priorities.

The choice of fairness definition should be guided by the specific application domain, the nature of the task, the potential harms of different types of errors,

and the ethical values being prioritized. For instance, in medical diagnosis, we might prioritize equality of false negative rates to ensure that serious conditions aren't missed for any group, while in criminal justice, we might prioritize equality of false positive rates to avoid wrongful punishment.

2.1.6 Trade-offs Between Fairness and Other Objectives

An important consideration in implementing fair AI systems is the potential trade-off between fairness and other objectives such as accuracy, efficiency, or interpretability.

The most widely discussed trade-off is between fairness and accuracy. Imposing fairness constraints typically reduces a model's overall predictive performance, as measured by conventional metrics like accuracy or F1 score. This creates a Pareto frontier where improvements in fairness come at the cost of reduced accuracy, and vice versa. The extent of this trade-off varies depending on the specific fairness definition, the dataset characteristics, and the learning algorithm used.

Beyond accuracy, fairness may also trade off against:

- **Interpretability:** More complex models that incorporate fairness constraints may be harder to interpret, reducing transparency and making it difficult to identify sources of bias;

- **Computational Efficiency:** Many fairness-aware algorithms require additional computational resources for training and inference, potentially limiting their applicability in resource-constrained environments;

- **Privacy:** Ensuring fairness often requires collecting and analyzing sensitive demographic data, which may conflict with privacy objectives or regulations like GDPR;

- **Individual Utility:** Group fairness measures may sometimes reduce utility for specific individuals who would have received positive outcomes under an unfair but more accurate model;

- **Short-term vs. Long-term Fairness:** Interventions that promote fairness in the short term may have unintended consequences that reduce fairness in the long term, or vice versa.

Achieving fairness in AI systems isn't just a technical challenge — it involves making value-based decisions about which goals matter most in a given context. These decisions should include input from people with different backgrounds, especially those who might be affected by the system, and should be guided by ethical principles, legal standards, and domain-specific knowledge.

It's also important to remember that these trade-offs aren't fixed. They depend on things like data quality, model complexity, and how the problem is defined. Better data collection, more advanced algorithms, and clearer definitions of fairness can help ease — though not completely solve — these tensions [KMR16, CR20].

2.2 Methods for Mitigating Bias

Researchers have developed numerous approaches to address bias in AI systems, which can be categorized based on the stage of the machine learning pipeline where they're applied:

2.2.1 Pre-processing techniques

These methods focus on transforming the input data to remove discriminatory patterns before model training. Examples include:

- Reweighting or resampling data to balance representation across groups;

- Transforming features to remove correlations with protected attributes [KZC12];

- Learning fair representations that preserve task-relevant information while obscuring protected attributes [MMS$^+$21];

- Data augmentation to generate synthetic samples for underrepresented groups [MMS$^+$21].

2.2.2 In-processing techniques

These approaches incorporate fairness considerations directly into the learning algorithm, often by modifying the objective function or adding constraints. Examples include:

- Adversarial debiasing, which uses an adversarial approach to remove information about protected attributes from the learned representations [MMS+21];

- Fair classification with constraints, which explicitly incorporates fairness criteria as constraints in the optimization problem [CDG23];

- Fair reinforcement learning, which modifies reward functions to account for fairness considerations [MMS+21].

2.2.3 Post-processing techniques

These methods adjust the output of already trained models to ensure fairness. Examples include:

- Threshold optimization, which applies different decision thresholds for different groups to equalize error rates;

- Calibration techniques that ensure predictions have the same meaning across groups;

- Reject option classification, which identifies and manually handles cases in the critical region between positive and negative classifications [KZC12].

2.2.4 Comparative Analysis of Mitigation Approaches

Each of these methods presents trade-offs. **Pre-processing** techniques offer flexibility and compatibility with any model but may not fully eliminate bias. **In-processing** techniques often yield better fairness-accuracy trade-offs but require access to model internals. **Post-processing** techniques are typically easier to implement but can be less robust [MMS+21, SG21].

The choice of mitigation strategy should be guided by practical considerations such as:

- Access to protected attributes during training and deployment;

- Computational resources available;

- Regulatory requirements regarding the use of protected attributes;

- The specific fairness definition being targeted.

Importantly, bias mitigation is not a one-time fix. Effective fairness in AI requires ongoing auditing, evaluation, and adjustment as societal values, population characteristics, and contexts evolve [SG21].

Recent work emphasizes the importance of domain-specific fairness frameworks. For instance, fairness in healthcare may prioritize equitable health outcomes over statistical parity, whereas fairness in education or finance may involve different risk-benefit trade-offs [MMS^{+}21].

2.3 Fairness in Different Application Domains

The implementation of fairness principles varies significantly across different domains, reflecting the diverse nature of AI applications:

- **Criminal Justice:** Beyond the COMPAS example, many risk assessment tools are used to make decisions about bail, sentencing, and parole. Fairness is especially important here because these decisions have serious consequences for people's lives and happen within a system that has a history of discrimination. Research shows that even factors that seem neutral, like past arrests, can reflect racial bias caused by unequal policing [BHJ^{+}17, EFN^{+}17]. Fairness in this area means finding the right balance between keeping the public safe and protecting individual rights, while also being careful not to repeat injustices from the past.

- **Healthcare:** AI is playing a growing role in healthcare, from diagnosing illnesses to suggesting treatments and managing how resources are distributed. One study [OPVM19] found that a commonly used algorithm gave lower risk

scores to Black patients than to White patients with similar health conditions, which led to fewer referrals for extra care. Ensuring fairness in healthcare requires more than just accounting for demographics — it also means recognizing variations in how diseases appear, how different groups respond to treatment, and the unequal access to medical services.

- **Finance:** Credit scoring, loan approval, insurance pricing, and other financial algorithms can significantly impact individual economic opportunities. Studies have found evidence of disparate impacts in mortgage lending, with certain demographic groups receiving higher interest rates or being denied loans more frequently, even after controlling for relevant financial factors [BMSW22]. Fairness in financial contexts must balance risk assessment with ensuring equal access to financial services and preventing the perpetuation of historical economic disparities.

- **Education:** AI systems are used for admissions decisions, student assessment, personalized learning, and resource allocation in educational settings. Automated grading systems, for instance, may penalize certain writing styles or language patterns associated with particular cultural backgrounds. Fairness in education requires considering diverse learning styles, cultural contexts, and educational backgrounds while ensuring that AI systems support rather than hinder educational equity [MMS+21].

- **Employment:** AI-powered resume screening and hiring tools can reflect and reproduce workplace discrimination if trained on biased historical hiring data. For instance, algorithms may deprioritize resumes with minority-associated names or nontraditional educational paths [Che23]. Fair employment AI demands careful review of what constitutes "merit" and ensures equal opportunity.

- **Public Services:** Government systems increasingly use AI for eligibility screening, benefits administration, and fraud detection. These services impact vulnerable populations most directly. Fairness in public services is crucial to prevent administrative exclusion or harm, and must be designed with high accountability standards [SG21].

These diverse applications highlight the need for domain-specific approaches to fairness that consider the unique ethical challenges, stakeholder perspectives, and regulatory frameworks relevant to each context. Moreover, they emphasize that technical solutions alone are insufficient—organizational practices, institutional policies, and legal frameworks must also evolve to support fair AI deployment.

2.4 Fairness Algorithms and Metrics

The state-of-the-art in algorithmic fairness is built upon several foundational pillars that have evolved over time. These pillars form a comprehensive framework for addressing bias in machine learning systems through a systematic approach. First, researchers have developed various metrics to identify and quantify bias, enabling precise measurement of different fairness notions across demographic groups. Second, a rich ecosystem of mitigation algorithms has emerged, categorized by their position in the machine learning pipeline: pre-processing techniques that transform training data before model development; in-processing methods that incorporate fairness directly into the learning algorithms; and post-processing approaches that adjust model outputs to ensure fair predictions. Together, these components provide practitioners with a diverse toolkit to address fairness concerns across different contexts and applications, balancing the often competing goals of accuracy and equity. The following sections examine each of these pillars in greater detail, highlighting key approaches and their theoretical foundations.

2.4.1 Fairness Metrics

Fairness metrics are essential for quantifying bias in machine learning models. They provide a means to evaluate how well a model adheres to various fairness definitions, enabling practitioners to identify and address potential disparities. Commonly used metrics include:

Statistical Parity Difference (SPD)

Measures whether positive prediction rates are balanced across different demographic groups. Formally:

$$\text{SPD} = P(\hat{Y} = 1 \mid A = a) - P(\hat{Y} = 1 \mid A = b) \tag{2.1}$$

A value close to 0 indicates that the model assigns positive outcomes at equal rates to both groups. SPD is useful for detecting allocative bias—whether one group is favored in outcome distribution [DHP+11].

Disparate Impact (DI)

Quantifies the ratio of positive prediction rates between different demographic groups.

$$\text{DI} = \frac{P(\hat{Y} = 1 \mid A = a)}{P(\hat{Y} = 1 \mid A = b)} \tag{2.2}$$

Values below 0.8 (or above 1.25) are typically seen as indicative of potential disparate impact, according to the 80% rule. DI is widely used in legal and regulatory contexts as a clear and interpretable indicator [FFM+15].

Equal Opportunity Difference

Evaluates whether true positive rates are equal across protected groups.

$$\text{EOD} = P(\hat{Y} = 1 \mid Y = 1, A = a) - P(\hat{Y} = 1 \mid Y = 1, A = b) \tag{2.3}$$

A value near 0 implies that among the individuals who should receive a positive outcome ($Y = 1$), each group is equally likely to be correctly identified. EOD is crucial for ensuring equal quality of service [HPS16].

2.4.2 Pre-processing methods

A Pre-processing algorithm transform the training data to remove bias before model training. These techniques act directly on the data—by transforming features, learning new representations, or rebalancing weights—to reduce the influ-

ence of protected attributes and improve fairness regardless of the downstream model. Commonly used pre-processing methods includes:

Learning Fair Representations (LFR)

This technique learns a compressed data representation that is informative for prediction, reconstructs the input, and obscures group membership. A simplified view of the objective is:

$$L_{\text{fair}} = \text{prediction loss} + \text{reconstruction loss} + \text{fairness penalty} \qquad (2.4)$$

More formally, the training minimizes a combined loss:

$$L_{\text{LFR}} = \lambda_y \, \ell(\hat{y}(z), y) + \lambda_x \, \|x - \hat{x}(z)\|^2 + \lambda_a \, D(z, A) \qquad (2.5)$$

where:

- $z = f(x)$ is the latent representation,

- $\hat{y}(z)$ is the prediction output,

- $\hat{x}(z)$ is the reconstructed input,

- $D(z, A)$ penalizes statistical dependence between the representation and the protected attribute.

This enables training fair models on top of a transformed space where sensitive information is less present [ZWS$^+$13].

Disparate Impact Remover

This method transforms features to reduce their dependency on the protected attribute. At a high level, the transformation pushes each feature toward a "fair" version that is equally distributed across groups:

$$x' = (1 - \alpha) \cdot x + \alpha \cdot x_{\text{fair}} \qquad (2.6)$$

The formal transformation applies quantile mapping using **Cumulative Distribution Functions**:

$$x' = (1 - \alpha)\,x + \alpha\,F^{-1}(F_{x|A}(x)) \tag{2.7}$$

where $F_{x|A}$ is the **CDF** of the feature within group A, and F^{-1} is the global inverse **CDF**. The parameter $\alpha \in [0, 1]$controls how strongly features are adjusted. This technique reduces disparate impact by making feature distributions more similar across groups [FFM$^+$15].

Reweighing

Modifies the importance of each training sample based on how frequent its label-group combination is. The goal is to simulate a balanced dataset where the label distribution is independent of group membership. Each sample with label y and group a receives weight:

$$w(y, a) = \frac{P(Y = y)}{P(Y = y \mid A = a)} \tag{2.8}$$

This rebalancing ensures that minority or underrepresented combinations receive more influence during training, encouraging models to make fairer decisions [KC12].

2.4.3 In-processing methods

An In-processing algorithm modifies the learning algorithm to incorporate fairness constraints during model training. These methods can be more effective than pre-processing, as they directly influence the model's decision-making process. Common in-processing techniques include:

FaUCI

Augments any stochastic-gradient–based learner with a generic penalised loss that can host any group-fairness metric (statistical parity difference, equalised odds, disparate impact, . . .) and works with binary, categorical, or continuous sensitive attributes.

$$\min_{\boldsymbol{\theta}} \ \mathcal{L}\text{task}(\boldsymbol{\theta}) \ + \ \lambda\,\Phi\big(\widehat{f\boldsymbol{\theta}},\, A\big), \tag{2.9}$$

where

- $\mathcal{L}_{\text{task}}$ is the ordinary prediction loss (cross-entropy, MSE, . . .),

- A is the protected attribute,

- Φ is the user-chosen fairness metric expressed as a differentiable surrogate,

- $\lambda > 0$ trades off accuracy and fairness.

Because Φ is a plug-in term, the same implementation can enforce several metrics simply by swapping Φ [MCCO24].

Adversarial Debiasing

Train a predictor and an adversary simultaneously: the predictor tries to forecast the label Y; the adversary tries to recover the protected attribute A from the predictor's internal representation (or logits). If the adversary fails, the representation is (approximately) independent of A.

$$\min_{\boldsymbol{\theta}} \left[\mathcal{L}\text{task}(\boldsymbol{\theta}) \;-\; \lambda \max \boldsymbol{\phi} \mathcal{L}_{\text{adv}}\big(\boldsymbol{\theta}, \boldsymbol{\phi}\big) \right], \tag{2.10}$$

where

- $\boldsymbol{\phi}$ are the adversary parameters;

- $\mathcal{L}\text{adv}$ is a cross-entropy loss on predicting A.

At equilibrium the classifier is both accurate (small $\mathcal{L}\text{task}$) and uninformative about A (large adversary loss) [ZLM18].

Prejudice Remover

Make the prediction $\widehat{Y}$ statistically independent of the protected attribute by adding a mutual-information regulariser to the learner's objective.

$$\min_{\boldsymbol{\theta}} \; \mathcal{L}_{\text{task}}(\boldsymbol{\theta}) \;+\; \eta \, I\big(\widehat{Y}; A\big), \tag{2.11}$$

where

- $I(\widehat{Y}; A)$ is the mutual information between the model's output and the sensitive attribute (estimated with differentiable approximations);

- η controls the strength of the fairness penalty.

Driving $I(\widehat{Y}; A)$ toward zero removes indirect prejudice that might leak through correlated features [KAAS12].

Meta-Fair Classifier:

Solve classification and fairness as a single constrained optimisation problem. The user specifies (i) a target fairness metric $C(\boldsymbol{\theta})$ (e.g. statistical-rate gap, false-discovery-rate ratio) and (ii) an acceptable tolerance τ. **Primal form.**

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{task}}(\boldsymbol{\theta})$$
$$\text{s.t. } |C(\boldsymbol{\theta})| \leq \tau. \tag{2.12}$$

Lagrangian reduction. In practice the constraint is dualised,

$$\min_{\boldsymbol{\theta}} \left[\mathcal{L}_{\text{task}}(\boldsymbol{\theta}) + \lambda \, |C(\boldsymbol{\theta})| \right], \tag{2.13}$$

and solved with an exponentiated-gradient meta-algorithm that adapts λ during learning, offering provable convergence guarantees for a large family of convex surrogates [CHKV20].

2.4.4 Post-processing methods

After a model has been fully trained, post-processing acts only on its outputs (scores or hard labels). This means you can retrofit fairness guarantees without touching the training data or the learning algorithm—very handy when you must comply with new regulations but cannot retrain. Below are three classic methods, each with a short plain-language description followed by the minimal mathematics that drives it.

Equalized Odds

Modifies predictions to equalize error rates across protected groups by applying group-specific thresholding or probabilistic flipping to model outputs. It ensures that both the **false positive rate (FPR)** and **false negative rate (FNR)** are equal across groups:

$$\Pr(\hat{Y} = 1 \mid Y = 0, A = a) = \Pr(\hat{Y} = 1 \mid Y = 0, A = a') \quad \text{(FPR equality)} \quad (2.14)$$

$$\Pr(\hat{Y} = 0 \mid Y = 1, A = a) = \Pr(\hat{Y} = 0 \mid Y = 1, A = a') \quad \text{(FNR equality)} \quad (2.15)$$

This is often achieved by solving a linear program that adjusts predictions post hoc while minimising accuracy loss [HPS16].

Calibrated Equalized Odds

Adjusts predictions while maintaining calibration—that is, the predicted probability $\hat{P}(Y = 1|X)$ must match the true frequency:

$$\Pr(Y = 1 \mid \hat{P} = p, A = a) = p \quad \text{for all groups } a. \quad (2.16)$$

It seeks a transformation of scores that retains this calibration while bringing **FPR** and **FNR** differences between groups below a small threshold ε. This allows a trade-off between fairness and probabilistic interpretability of outputs [PRW$^+$17].

Reject Option Classification

Introduces a rejection option for uncertain predictions that might reflect bias. Specifically, in a band around the decision boundary (typically $p \in [0.5-\theta, 0.5+\theta]$), predictions are flipped in favour of the disadvantaged group:

$$\widetilde{Y} = \begin{cases} 1 & \text{if } A \text{ is disadvantaged and } |p - 0.5| \leq \theta, \\ 0 & \text{if } A \text{ is advantaged and } |p - 0.5| \leq \theta, \\ \text{original prediction} & \text{otherwise.} \end{cases} \quad (2.17)$$

This improves statistical parity by reducing unfair decisions near the boundary where the model is least confident [KKZ12].

2.5 Fairness Toolkits and Libraries Overview

IBM's *AI Fairness 360* (AIF360) and Microsoft's *Fairlearn* are two prominent open-source toolkits designed to operationalise fairness techniques in machine learning. Both provide extensive libraries of fairness metrics and bias-mitigation algorithms: AIF360 offers several detection metrics and nine pre-, in-, and post-processing mitigations [BDH+18], whereas Fairlearn complements a dashboard for diagnostic visualisation with reduction- and post-processing-based mitigations for both classification and regression [BDE+20]. Despite this breadth, recent empirical studies reveal substantial usability barriers [DNL+22]. Further hurdles arise from API and documentation design: inconsistent terminology (e.g., `CorrelationRemover` in Fairlearn versus `DisparateImpactRemover` in AIF360) and limited explanatory material oblige many users to copy code from toy examples rather than adapt algorithms systematically [DNL+22]. Finally, the libraries' research-centric architecture complicates integration into production pipelines, especially where access to protected attributes is constrained by privacy or compliance requirements [HWVD+19]. In short, while AIF360 and Fairlearn lower the entry barrier to technical fairness assessments, their current incarnations remain far from "plug-and-play." Addressing gaps in task coverage, guidance, and workflow integration is therefore essential if such tools are to fulfil their promise in real-world machine-learning practice.

2.6 Current Challenges in Operationalizing Fairness

In operationalizing algorithmic fairness, practitioners face multiple interrelated challenges. **First**, numerous formal definitions of fairness exist (e.g., demographic parity, equalized odds), and many of these criteria conflict or cannot be satisfied simultaneously [KMR16, XZ24]. This forces stakeholders to decide which defini-

tion best aligns with their context and values. **Second**, pursuing fairness often involves trade-offs with model accuracy or other performance measures, since fairness constraints can degrade predictive accuracy [CDPF+17]. **Third**, bias in data presents a major obstacle: historical datasets may reflect societal prejudices, and models can inadvertently use proxy features correlated with protected attributes, leading to residual unfairness even after interventions [BS16]. **Fourth**, fairness solutions must be tailored to each domain's legal and social context; for example, anti-discrimination laws in some sectors restrict the use of sensitive attributes and mandate specific fairness criteria, necessitating domain-specific approaches [VB17]. **Finally**, even when technical solutions are available, deploying them in real-world systems is difficult. Open-source fairness toolkits (e.g., AI Fairness 360 [BDH+18]) provide many metrics and mitigation algorithms, but practitioners often struggle with their usability and integration into existing workflows [HWVD+19, LS21]. These challenges highlight that achieving fairness in practice requires not only better algorithms, but also careful attention to context, data quality, and human-centered tool design.

Bibliography

[BCG21]   Teresa Bono, Karen Croxson, and Adam Giles. Algorithmic fairness
           in credit scoring. *Oxford Review of Economic Policy*, 37(3):585–
           617, 09 2021.

[BDE+20]   Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman
           Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and
           Kathleen Walker. Fairlearn: A toolkit for assessing and improving
           fairness in ai. Technical Report MSR-TR-2020-32, Microsoft, May
           2020.

[BDH+18]   Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoff-
           man, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacque-
           lyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar,
           Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan
           Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and
           Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting,
           understanding, and mitigating unwanted algorithmic bias, 2018.

[BG18]   Joy Buolamwini and Timnit Gebru. Gender shades: Intersec-
           tional accuracy disparities in commercial gender classification. In
           Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the
           1st Conference on Fairness, Accountability and Transparency*, vol-
           ume 81 of *Proceedings of Machine Learning Research*, pages 77–91.
           PMLR, February 2018.

[BHJ⁺17]   Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and
             Aaron Roth. Fairness in criminal justice risk assessments: The
             state of the art, 2017.

[BMSW22]   Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wal-
             lace. Consumer-lending discrimination in the fintech era. *Journal
             of Financial Economics*, 143(1):30–56, 2022.

[BS16]   Solon Barocas and Andrew D. Selbst. Big data's disparate impact.
             California Law Review, 104:671, 2016.

[CBN17]   Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Seman-
             tics derived automatically from language corpora contain human-
             like biases. *Science*, 356(6334):183–186, April 2017.

[CDG23]   Sam Corbett-Davies and Sharad Goel. The measure and mismea-
             sure of fairness. *Journal of Machine Learning Research*, 24:1–117,
             2023.

[CDPF⁺17]   Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and
             Aziz Huq. Algorithmic decision making and the cost of fairness.
             In *Proceedings of the 23rd ACM SIGKDD International Conference
             on Knowledge Discovery and Data Mining*, KDD '17, page 797–806,
             New York, NY, USA, 2017. Association for Computing Machinery.

[CGCMO23]   Roberta Calegari, Gabriel G. Castañé, Michela Milano, and Barry
             O'Sullivan. Assessing and enforcing fairness in the ai lifecycle.
             In Edith Elkind, editor, *Proceedings of the Thirty-Second Interna-
             tional Joint Conference on Artificial Intelligence, IJCAI-23*, pages
             6554–6562. International Joint Conferences on Artificial Intelli-
             gence Organization, 8 2023. Survey Track.

[Che23]   Zhisheng Chen. Ethics and discrimination in artificial intelligence-
             enabled recruitment practices. *Humanities and Social Sciences
             Communications*, 10(1):567, 2023.

[CHKV20]   L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees, 2020.

[CR20]   Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, April 2020.

[DHP⁺11]   Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011.

[DNL⁺22]   Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. In *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT '22, page 473–484. ACM, June 2022.

[EFN⁺17]   Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing, 2017.

[Fer23]   Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, December 2023.

[FFM⁺15]   Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2015.

[GGR24]   Ana Cristina Bicharra Garcia, Marcio Gomes Pinto Garcia, and Roberto Rigobon. Algorithmic discrimination in the credit domain: what do we know about it? *AI & SOCIETY*, 39(4):2059–2098, 2024.

[HPS16]   Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[HWVD+19]   Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro
            Dudik, and Hanna Wallach. Improving fairness in machine learn-
            ing systems: What do industry practitioners need? In *Proceedings
            of the 2019 CHI Conference on Human Factors in Computing Sys-
            tems*, CHI '19, page 1–16. ACM, May 2019.

[KAAS12]   Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun
            Sakuma. Fairness-aware classifier with prejudice remover regular-
            izer. pages 35–50, 09 2012.

[KBA24]   Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil
            Aryal. The pursuit of fairness in artificial intelligence models: A
            survey, 2024.

[KC12]   F. Kamiran and T.G.K. Calders. Data preprocessing techniques for
            classification without discrimination. *Knowledge and Information
            Systems*, 33(1):1–33, 2012.

[KKBK21]   Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke
            Kreuter. Distributive justice and fairness metrics in automated
            decision-making: How much overlap is there?, 2021.

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

[KMR16]   Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. In-
            herent trade-offs in the fair determination of risk scores, 2016.

[KNP+25]   Dewant Katare, David Solans Noguero, Souneil Park, Nicolas
            Kourtellis, Marijn Janssen, and Aaron Yi Ding. Analyzing and
            mitigating bias for vulnerable road users by addressing class imbal-
            ance in datasets. *IEEE Open Journal of Intelligent Transportation
            Systems*, 6:590–604, 2025.

[KZC12]   Faisal Kamiran, Indrė Zliobaitė, and Toon Calders. Quantifying
            explainable discrimination and removing illegal discrimination in

automated decision making. *Knowledge and Information Systems*, 1:in press, 06 2012.

[LCZ⁺24]   Xinyue Li, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. Bias behind the wheel: Fairness testing of autonomous driving systems, 2024.

[LS21]   Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[MCCO24]   Matteo Magnini, Giovanni Ciatto, Roberta Calegari, and Andrea Omicini. Enforcing fairness via constraint injection with FaUCI. In Roberta Calegari, Virginia Dignum, and Barry O'Sullivan, editors, *AEQUITAS 2024: Fairness and Bias in AI*, volume 3808 of *CEUR Workshop Proceedings*, pages 8:1–8:13. CEUR-WS, October 2024. Proceedings of the 2nd Workshop on Fairness and Bias in AI co-located with 27th European Conference on Artificial Intelligence (ECAI 2024).

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

[MNDWLS24]   Marie Mirsch Née Decker, Laila Wegner, and Carmen Leicht-Scholten. Procedural fairness in algorithmic decision-making: the role of public engagement. *Ethics and Information Technology*, 27, 11 2024.

[MPB⁺21]   Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

[MWK23]   Aboli Marathe, Rahee Walambe, and Ketan Kotecha. In rain or shine: Understanding and overcoming dataset bias for improving

robustness against weather corruptions for autonomous vehicles, 2023.

[NHS⁺24] Khalida Walid Nathim, Nada Abdulkareem Hameed, Saja Abdul-fattah Salih, Nada Adnan Taher, Hayder Mahmood Salman, and Dmytro Chornomordenko. Ethical ai with balancing bias mitigation and fairness in machine learning models. In *2024 36th Conference of Open Innovations Association (FRUCT)*, pages 797–807, 2024.

[OPVM19] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[PRW⁺17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[SG21] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, page 1–9. ACM, October 2021.

[SGA19] Ismail Abiodun Sulaimon, Ahmed Ghoneim, and Mubarak Al-rashoud. A new reinforcement learning-based framework for unbiased autonomous software systems. In *2019 8th International Conference on Modeling Simulation and Applied Optimization (ICM-SAO)*, pages 1–6, 2019.

[VB17] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

[VR18] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*,

FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[XZ24]   Ruicheng Xian and Han Zhao. A unified post-processing framework for group fairness in classification, 2024.

[YSE+23]   Jenny Yang, Andrew Soltan, David Eyre, Yang Yang, and David Clifton. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ digital medicine*, 6:55, 03 2023.

[YW24]   Chih-Cheng Rex Yuan and Bow-Yaw Wang. Ensuring fairness with transparent auditing of quantitative bias in ai systems, 2024.

[ZHL+23]   Jie Zhu, Mengsha Hu, Xueyao Liang, Amy Zhang, Ruoming Jin, and Rui Liu. Fairness-sensitive policy-gradient reinforcement learning for reducing bias in robotic assistance, 2023.

[ZLM18]   Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning, 2018.

[ZWS+13]   Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, June 2013. PMLR.

Acknowledgements

Optional. Max 1 page.