ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

Corso di Laurea Magistrale in Ingegneria e Scienze Informatiche

Addressing Fairness in AI Systems: Design and Development of a Pragmatic (Meta-)Methodology

Tesi di laurea in:
INTELLIGENT SYSTEMS ENGINEERING

Relatore
Prof. Giovanni Ciatto

Candidato
Mattia Matteini

Correlatori
Prof. Roberta Calegari
Prof. Andrea Omicini

IV Sessione di Laurea
Anno Accademico 2023–2024

Abstract

Max 2000 characters, strict.

Optional. Max a few lines.

Contents

List of Figures

List of Listings

Chapter 1

Introduction

What is Fairness? Fairness, from an ethical and social perspective, is the principle of treating individuals and groups equitably, ensuring that no one is unjustly advantaged or disadvantaged due to biases, discrimination, or arbitrary distinctions. It is deeply rooted in moral philosophy, legal systems, and societal norms, aiming to promote justice, equality, and inclusion. A just society requires reducing social inequalities, and ensuring that opportunities and resources are distributed in a way that acknowledges both individual merit and systemic disadvantages. The concept of fairness evolves based on cultural, historical, and contextual factors, reflecting a society's commitment to ethical treatment and social cohesion.

Fairness in AI. AI From a technical point of view, fairness in Artificial Intelligence (AI) refers to the development and deployment of AI systems that minimize biases and prevent discriminatory outcomes. It involves designing systems that ensure equitable treatment across different demographic groups, particularly those historically marginalized or disadvantaged. The main challenges in this field are building `fair-by-design` systems, namely such systems in which the fairness problem is addressed since the very beginning of the process, and detecting biases in already existing systems, mitigating them if possible.

Mattia Matteini: troppo lungo? forse meglio spezzare in più paragrafi

Before the advent of fairness, AI systems were developed with the primary goal of optimizing performance metrics, such as accuracy and efficiency. Nowadays, that fairness is becoming a crucial aspect to consider, accuracy is no more the only metric to optimize, it is necessary to find a balance between accuracy and fairness. Besides that, fairness can also be in contrast with the performance of the model, making difficult to find a good trade-off between these two aspects.

Fairness is becoming crucial because AI systems increasingly influence decision-making processes in various sectors of society, including hiring, lending and health-care. If AI models are biased, they can perpetuate and even amplify existing societal inequalities, leading to unjust outcomes and tremendous effects on individuals and communities. Ensuring fairness in AI enhances trust, transparency, and accountability, making AI systems more ethical, reliable, and beneficial for society.

AI has undergone significant advancements over the past few decades, causing an enormous increase in its adoption across various domains, until becoming pervasive in the daily life of people. This also caused a growing of biases in AI systems, as discriminations are intrinsically part of the human history, and consequently of the data that AI systems are trained on.

AI is now widely used in critical domains such as healthcare, finance, education, and criminal justice, where biased decisions can have life-altering consequences. For instance, in healthcare, biased algorithms may lead to misdiagnosis or unequal treatment recommendations for different demographic groups; in finance, AI-driven credit scoring models can reinforce discriminatory lending practices, limiting access to financial resources; in the criminal justice system, biased predictive policing and risk assessment tools can disproportionately target marginalized communities. Given these risks, ensuring fairness in AI is essential to preventing discrimination, maintaining ethical standards, and safeguarding individuals.

On Multidisciplinarity. Achieving fairness in AI requires a multidisciplinary approach that integrates insights from computer science, ethics, law and social sciences. Technical methods alone cannot fully address fairness, as fairness is deeply tied to societal values, human rights, and legal frameworks. Socio-legal experts help define fairness principles, ensure compliance with anti-discrimination laws and analyze the societal impact. The intersection of these fields highlights that AI fairness is not merely a technical challenge but a complex, multidimensional issue requiring collective effort and interdisciplinary research and collaboration.

An impactful example regarding the work of legal experts in the field of Artificial Intelligence is the *AI Act*. The AI Act, proposed by the European Union, is a comprehensive regulatory framework designed to ensure that AI systems are safe, transparent, and aligned with fundamental rights. It categorizes AI applications into different risk levels—unacceptable risk, high risk, limited risk, and minimal risk—imposing stricter requirements on higher-risk systems, such as those used in hiring, law enforcement, and healthcare. These requirements include transparency, human oversight, and bias mitigation. However, translating these legal constraints into practical technical steps is not trivial.

Mattia Matteini: add some citation

Concepts like fairness, accountability, and explainability are difficult to quantify, and AI models often operate as black boxes, making compliance complex. While the AI Act sets an important precedent for AI governance, its effective implementation requires further collaboration between policymakers, legal experts, and computer scientist to bridge the gap between regulation and technical feasibility.

Measuring Fairness. At one point, in order to assess the fairness of an AI system, is important to have a way to "measure" how much the system is fair and in what terms. Remarking what said before, fairness is very context-dependent, and there is not a single way to measure it.

The need to cover multiple aspects of fairness has led to the introduction of various *fairness metrics*—statistical formulas that quantify fairness in different ways, each capturing a slightly different aspect of fairness. These fairness metrics, are a set of indexes that can be used to detect biases in AI systems, and they can be used indeed to evaluate the fairness of a model.

In the following, are listed some of the most common fairness metrics used in the literature ([IML23]):

- *Statistical Parity Difference* (SPD) measures the difference between the probability of the privileged and unprivileged classes receiving a favorable outcome. This measure should be equal to 0 to be fair.

  Formally it is defined as:

  $$SPD = P(\hat{Y} = 1 | A = a) - P(\hat{Y} = 1 | A = b)$$

where A is the sensitive attribute, $\hat{Y}$ is the predicted outcome, and a and b are the privileged and unprivileged groups, respectively.

- *Disparate Impact* (DI) compares the proportion of individuals that receive a favorable outcome for two groups, a privileged group and an unprivileged group. This measure should be equal to 1 to be fair.

  Formally it is defined as:

  $$DI = P(\hat{Y} = 1|A = a)/P(\hat{Y} = 1|A = b)$$

  where A is the sensitive attribute, $\hat{Y}$ is the predicted outcome, and a and b are the privileged and unprivileged groups, respectively.

Mattia Matteini: add more metrics?

Structure of the Thesis

Mattia Matteini: At the end, describe the structure of the paper

Chapter 2

Background

Mattia Matteini: manca una sezione Lack of engineering methodology?

2.1 AI Lifecycle

Since the very beginning of the AI era, the standard lifecycle consists of the following "traditional" steps: (i) data collection and processing, (ii) model training, (iii) system evaluation. Obviously, this workflow in the latest years have increased in complexity and now, with the newer innovations and powerful models and architectures, it may appear even almost minimalistic, but it still represents the core of all AI systems.

However, when fairness is taken into account, each step needs to be revisited in order to obtain an equitable, impartial, and fair AI system.

To achieve this goal, the technical perspective is not enough. Fairness is a multidisciplinary concept that involves social, legal, and ethical aspects. Therefore, the AI lifecycle needs to be constrained by socio-legal requirements that engineers must consider during the development process. This includes understanding the societal impact of AI systems, ensuring compliance with legal standards, and adhering to ethical guidelines.

There are also many differences between the socio-legal and technical perspectives. Regarding the AI lifecycle, engineers tend to focus on technical aspects and few development phases, in fact the major part of the literature speaks only about

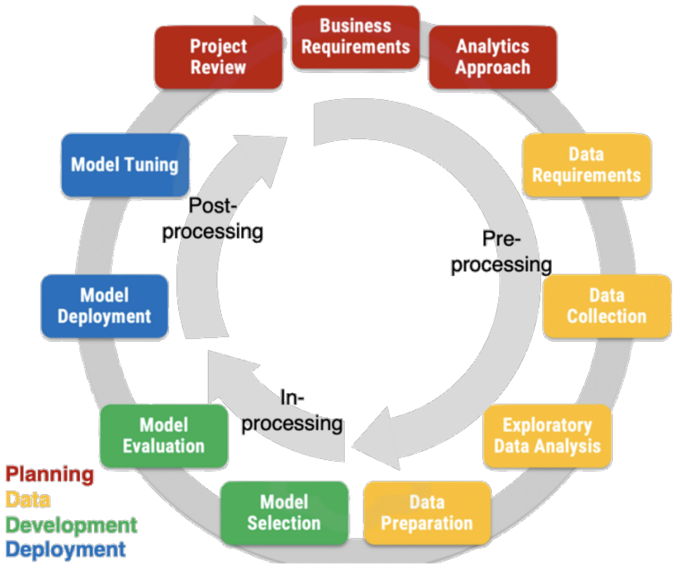

Figure 2.1: Fair AI lifecycle

Mattia Matteini: la citazione va bene anche a fianco dell'immagine?

pre-processing, *in-processing* and *post-processing* (Figure 2.1 [CCMO23]). Respectively, *pre-processing* involves data collection and preparation, *in-processing* refers to the model training phase, and *post-processing* deals with the fair evaluation of the AI system.

Often engineers adopt reductionist approaches addressing a field that is not their own, discarding the big picture of social, economic, and institutional constraints. On the other hand, socio-legal experts consider a broader range of activities and phases. They focus on "building blocks" for fair AI such as risk assessment, stakeholder identification, regulatory analysis, and fundamental human rights impact assessment. In particular, with respect to fundamental rights impact assessments, these will be legally required for some AI systems, yet no standard for implementing them has emerged so far.

2.2 Practical Issues

2.2.1 What is (un)fair?

Fairness in AI (and beyond) is inherently subjective, shaped by cultural values, ethical theories, and individual perspectives. What one group considers fair may not align with other people's understanding, leading to debates about determining what is fair and what is not. This subjectivity and variation in viewpoints complicates efforts to develop standardized fairness metrics, as no single approach can universally capture the diverse and often conflicting notions of fairness present across different social, legal, and institutional contexts.

Beyond its subjectivity, fairness is also highly context-dependent. The same algorithm might be considered fair in one application but biased in another, depending on the societal, legal, and institutional constraints surrounding it. For instance, fairness considerations in hiring algorithms differ from those in criminal justice risk assessments, necessitating tailored approaches rather than generic solutions. Moving forward, privileged and unprivileged groups change depending on the application domain, as well as the fairness criteria that are taken into account.

2.2.2 Bridging Perspectives

Bridging the socio-legal and technical perspectives on fairness is a significant challenge. Guidelines and descriptive methodologies exist to address fairness compliance from a social-legal perspective, but their approach offer broad guidelines without defining practical fairness measurements, leaving interpretation to technical experts ([CMS+25]). The lack of alignment between these viewpoints makes it difficult to translate abstract fairness principles into concrete computational methods. This also leads to a proliferation of metrics, each measuring slightly different aspects of fairness, reflecting the diverse priorities and domain perspectives.

A fundamental obstacle to this integration is the differing language used by socio-legal experts and technical people. It is difficult to reach an agreement if even a concept or term can assume different meanings depending on the perspective. This linguistic division creates a difficult barrier to interdisciplinary collaboration, leading to misunderstandings even when working towards shared goals.

These perspectives are shaped also by distinct academic and methodological backgrounds. Legal and ethical frameworks tend to be verbose and highly context-specific, relying on various interpretations and case-by-case analyses. In contrast, technical disciplines prioritize concrete steps and pragmatic aspects.

Chapter 3

The Meta-Methodology

3.1 The Roles(?)

3.1.1 Stakeholders Awareness

3.2 The Q/A Mechanism

3.3 Decision Support System

3.4 Assisting AI System Building

Chapter 4

Design

4.1 Architecture

Chapter 5

Implementation

Chapter 6

Conclusions

Bibliography

[CCMO23] Roberta Calegari, Gabriel G. Castañé, Michela Milano, and Barry O'Sullivan. Assessing and enforcing fairness in the AI lifecycle. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6554–6562. ijcai.org, 2023.

[CMS+25] Giovanni Ciatto, Mattia Matteini, Laura Sartori, Maria Rebrean, Catelijne Muller, Andrea Borghesi, and Roberta Calegari. AI-fairness: the FairBridge approach to practically bridge the gap between socio-legal and technical perspectives. In *Proceedings of the 58th Hawaii International Conference on System Sciences*, page 6499, 2025. (In press).

[IML23] Zahid Irfan, Fergal McCaffery, and Róisín Loughran. Evaluating fairness metrics. In Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo, editors, *Advances in Bias and Fairness in Information Retrieval - 4th International Workshop, BIAS 2023, Dublin, Ireland, April 2, 2023, Revised Selected Papers*, volume 1840 of *Communications in Computer and Information Science*, pages 31–41. Springer, 2023.

Acknowledgements

Optional. Max 1 page.