Exploring Indicators to Identify Bias in Artificial Intelligence Models

Omar Saadi

Thesis presented to the School of Technology and Management in the scope of the Master in Electrical and Computer Engineering.

Supervisors:

Prof. Ana Isabel Pereira

Prof. Raquel Ávila Muñoz

This document does not include the suggestions made by the board.

Bragança

2024-2025

Exploring Indicators to Identify Bias in Artificial Intelligence Models

Omar Saadi

Thesis presented to the School of Technology and Management in the scope of the Master in Electrical and Computer Engineering.

Supervisors:

Prof. Ana Isabel Pereira

Prof. Raquel Ávila Muñoz

This document does not include the suggestions made by the board.

Bragança

2024-2025

Dedication

To my family, for always being by my side, supporting every decision I've made with love, trust, and encouragement.

To the soul of my grandmother, Aicha, who was like a second mother to me, her kindness, strength, and unconditional love continue to guide me every day.

To my friends, who have been like family to me, especially during the times I've lived far from home. Your presence made those moments easier and full of meaning.

To all my professors, from the beginning of my studies until now, thank you for your patience, your guidance, and the knowledge you've shared with me.

Acknowledgment

This work was supported by the Instituto Politécnico de Bragança (IPB) and its dedicated professors, whose guidance and expertise made this thesis possible.

I am deeply grateful to my supervisor, Professor Ana Isabel Pereira, and co-supervisor, Professor Raquel Muñoz, for their patience, encouragement, and valuable insights during this journey.

I would also like to thank all the teaching and administrative staff at IPB, and everyone who contributed to my academic and personal growth during my time here.

Abstract

This thesis details the conception, development, and evaluation of the Bias Detector Tool, an open-source Python application specifically built in the context of this work to assess ethical aspects of artificial intelligence (AI) systems. The tool evaluates machine learning (ML) models and datasets across five key ethical dimensions: fairness, transparency, privacy, robustness, and accountability. It is intended to assist researchers, developers, and policy makers to identify ethical risks in AI systems.

The tool employs a modular pipeline that processes CSV-format datasets, applies established metrics using specialized libraries such as Fairlearn, Diffprivlib, and LIME and generates comprehensive output reports in TXT, JSON, and CSV formats. It offers a scoring mechanism that classifies each ethical indicator on a scale from 0 to 1 and provides both technical results and simplified interpretations for non-expert users.

To validate the tool's functionality and reliability, a test scenario was conducted in collaboration with Professor Raquel Ávila Muñoz, an expert in Equality, Diversity, and Inclusion at the Complutense University of Madrid. The evaluation compared datasets with limited ethical integrity such as those lacking diversity or metadata with well-structured datasets showing inclusive data practices. The tool successfully reflected these differences through the scoring system confirming its efficacy in identifying ethically problematic datasets.

In summary, this work contributes to the field of responsible AI by offering a practical, transparent, and user-friendly approach to ethical assessment. The tool is publicly available via GitHub, encouraging further adaptation and development by the research community.

Keywords: Data Mining, Machine learning, EDI (Equality, Diversity, Inclusion), Data Preprocessing, Ethical Indicators.

Resumo

Esta tese detalha a conceção, desenvolvimento e avaliação da Ferramenta de Detecção de Bias, uma aplicação Python de código aberto especificamente criada no contexto deste trabalho para avaliar os aspectos éticos dos sistemas de inteligência artificial. A ferramenta analisa modelos de machine learning e conjuntos de dados com base em cinco dimensões éticas principais: justiça, transparência, privacidade, robustez e responsabilidade. Ela foi criada para ajudar pesquisadores, desenvolvedores e formuladores de políticas a identificar riscos éticos em sistemas de IA.

A ferramenta utiliza um pipeline modular que processa dados no formato CSV, aplica métricas reconhecidas com o uso de bibliotecas especializadas como Fairlearn, Diffprivlib e LIME, e gera relatórios completos nos formatos TXT, JSON e CSV. O sistema de pontuação classifica cada indicador ético numa escala de 0 a 1, oferecendo tanto resultados técnicos quanto interpretações simplificadas para usuários não especializados.

Para validar o funcionamento e a confiabilidade da ferramenta, foi realizado um teste em colaboração com a Professora Raquel Ávila Muñoz, especialista em Igualdade, Diversidade e Inclusão na Universidade Complutense de Madrid. A avaliação comparou conjuntos de dados com baixa integridade ética como aqueles com pouca diversidade ou sem metadados com outros bem estruturados, que seguem boas práticas de inclusão. A ferramenta refletiu corretamente essas diferenças por meio de seu sistema de pontuação, confirmando sua eficácia na identificação de conjuntos de dados com problemas éticos.

Em resumo, este trabalho contribui para o campo da IA responsável, oferecendo uma abordagem prática, transparente e acessível para a avaliação ética. A ferramenta está disponível publicamente no GitHub, incentivando a sua adaptação e desenvolvimento

contínuo pela comunidade científica.

Palavras-chave: Mineração de Dados, Aprendizado de Máquina, EDI (Equidade, Diversidade e Inclusão), Pré-processamento de Dados, Indicadores Éticos.

Contents

List of Tables

List of Figures

Acronyms

AI	Artificial Intelligence
ML	Machine Learning
GDPR	General Data Protection Regulation
ART	Adversarial Robustness Toolbox
FGSM	Fast Gradient Sign Method
IBM	International Business Machines
EDI	Equality, Diversity, Inclusion
NA	Not Applicable
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
EU	European Union
NLP	Natural Language Processing
LLMs	Large Language Models
CI/CD	Continuous Integration / Continuous Delivery

Chapter 1

Introduction

In recent years, Artificial Intelligence **(AI)** has become a cornerstone technology, revolutionizing fields such as healthcare, finance, education, and criminal justice. While AI holds the promise of enhanced efficiency, accuracy, and decision-making capabilities, it also introduces significant challenges. Among these challenges, bias in AI models has emerged as a critical concern, threatening the fairness and trustworthiness of AI systems.

Bias in AI refers to systematic disparities in outcomes caused by imbalances in training data, algorithmic design, or deployment environments. Such biases can lead to discriminatory practices, perpetuate and amplify existing inequalities, and even result in legal or ethical violations. For instance, hiring algorithms have been criticized for favoring certain demographics over others, and facial recognition systems have shown reduced accuracy for minority groups, raising questions about the societal impact of AI.

The European Commission has recognized the need for responsible AI governance and has proposed ethical guidelines to address these challenges. Key principles such as fairness, transparency, and accountability are emphasized to ensure that AI systems respect fundamental human rights and prevent harm. However, identifying and mitigating bias in AI remains a complex task requiring a deeper understanding of its root causes and practical solutions [1].

The societal impact of biased AI systems is profound, as they can reinforce systemic

discrimination, erode public trust, and disproportionately harm marginalized communities. These risks highlight the urgent need for tools that can systematically detect and address ethical shortcomings in AI. Motivated by these concerns, this thesis aims to contribute a practical solution through the development of the Bias Detector Tool — a system designed to evaluate and promote fairness, transparency, and accountability in AI applications.

1.1 Goals

The primary goal of this thesis is to explore and validate indicators that can identify and classify bias in AI models. By applying methodologies and testing new approaches, this research aims to develop a robust framework for detecting and mitigating bias in AI systems. The specific objectives include:

1. Identifying key indicators of bias in AI models.
2. Testing the proposed indicators in real-world scenarios to evaluate their effectiveness.
3. Proposing solutions to mitigate identified biases and align AI practices with ethical guidelines, particularly those promoted by the European Commission.

The ultimate goal is to contribute to the development of fair, transparent, and trustworthy AI systems that uphold societal values and foster equitable outcomes for all users.

1.2 Document Structure

This thesis is organized into the following chapters:

- **Chapter 1: Introduction**

  The introductory chapter outlines the significance of AI in modern society, the challenges of bias in AI models, the goals and thesis structure.

- **Chapter 2: Concepts**

  This chapter introduces foundational concepts, such as what bias means in the context of AI models, the different types of bias and why detecting and addressing bias is essential.

- **Chapter 3: State of the Art**

  In this chapter, we review the existing literature on bias in AI, examining multiple aspects as per example situations where bias typically appears in AI systems, the effects of bias on individuals and society, solutions that have been proposed or implemented to mitigate bias, including tools, fairness metrics, case studies and proposed Solution which is an overview of a potential system that can detect bias in AI models by collecting and analyzing data from these models, highlighting its potential role in mitigating bias.

- **Chapter 4: Bias Detector Tool**

  This chapter presents the development of the Bias Detector Tool. It covers the motivation behind the tool, the design choices, the architecture, and the key functionalities. It also provides an explanation of how the tool assesses ethical indicators and detects bias in AI-related datasets.

- **Chapter 5: Test and Validation**

  This chapter focuses on testing and validating the proposed indicators and solutions through practical experiments on AI models. The effectiveness of the system will be evaluated based on its ability to identify and mitigate bias using Loan approval dataset.

- **Chapter 6: Critical analysis and future perspectives**

  This chapter provides a critical reflection on the Bias detector system, highlighting its key contributions, evaluating its strengths and limitations, and identifying paths for future enhancement.

- **Chapter 7: Conclusions**

  The final chapter summarizes the findings, discusses the implications for the future of AI, and provides recommendations for future research and development in ethical AI practices.

Chapter 2

Concepts

Bias in the context of AI and Machine Learning **(ML)** refers to systematic and unfair deviations in the outputs or decisions made by an algorithm. These biases can emerge at different stages of the data pipeline — from data collection and preprocessing to model training and deployment. When left unaddressed, such biases can compromise the fairness, reliability, and ethical acceptability of AI systems, potentially resulting in discriminatory outcomes, particularly for marginalized or underrepresented groups.

Biases in AI systems can be broadly categorized based on their origin, such as data bias, algorithmic bias, and societal bias. This taxonomy helps in understanding how and why these biases arise, and provides a foundation for developing targeted mitigation strategies [2].

2.1 Origin and Types of Bias

The ML lifecycle consists of three discrete stages for each there exist specific origin of bias [3]:

- bias that can originate from the data
- bias deriving from the ML models that are used
- bias by the ML engineers that develop and/or evaluate the produced ML models

An overview of the cause of bias that can occur in each of the three categories can be shown in Figure 2.1.

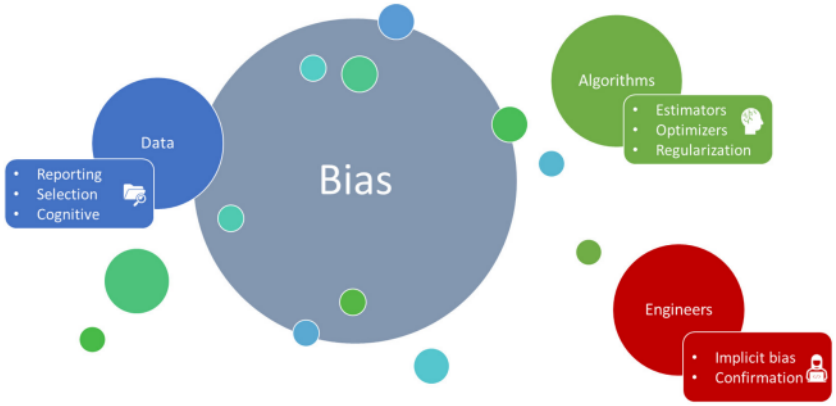


Figure 2.1: Origin of bias per ML life-cycle category [4]

The following are some of the most popular types of Bias [1], [5]:

- **Data**

  Definition: Bias introduced by imbalances in the training data [2].

  Examples: If an AI system is trained on data that mostly includes young people, it may perform poorly on older individuals because it has been fitted to patterns that primarily reflect a younger population. Similarly, if the data includes stereotypes or reflects historical inequalities, the model can replicate and perpetuate these biases [2].

- **Algorithmic**

  Definition: Bias stemming from the design or structure of the algorithm itself. Some algorithms may amplify existing data imbalances or introduce unintended consequences through the way they process information [6].

  Examples: A well-known case involved a commercial healthcare algorithm in the United States that prioritized patients for care management programs. The algorithm used healthcare cost as a proxy for medical need. Because Black patients historically had less access to healthcare and thus lower healthcare spending, the

algorithm systematically underestimated their risk scores, resulting in racial bias in patient prioritization [6].

- **Model Predictions**

  Definition: Bias that shows up in the model's predictions or outcomes, often due to both data and algorithmic biases [7].
  Examples: A biased hiring algorithm might favor applicants from certain backgrounds over others, even if qualifications are equal, due to historical data that favored certain groups [7].

- **User Interaction**
  Definition: Bias that occurs when the AI interacts with users in ways that reinforce stereotypes or unfairly categorize individuals [8].
  Examples: Facial recognition software has been shown to work less accurately for darker skin tones, which could lead to disproportionate misidentification in certain populations [8].

Bias can lead to ethical, legal, and social issues. If not identified and mitigated, it can cause AI models to unfairly disadvantage certain groups, perpetuate stereotypes, and even violate regulations such as those in the EU's General Data Protection Regulation (GDPR) [9] and AI Act [10], which emphasize fairness, accountability, and transparency. Identifying bias involves examining the training data, analysing the model's performance across diverse groups, and monitoring predictions for discrepancies. Mitigating bias often requires adjusting data collection methods, refining algorithms, and testing the model iteratively to ensure it performs fairly across different populations.

2.2 Indicators Promoted by the European Commission

To address bias in AI as per the European Commission's standards, we need indicators that align with principles of fairness, accountability, transparency, and robustness. The key ethical guidelines for trustworthy AI, as defined by the European Commission in documents such as the Ethics Guidelines for Trustworthy AI [10] and the AI Act [11].

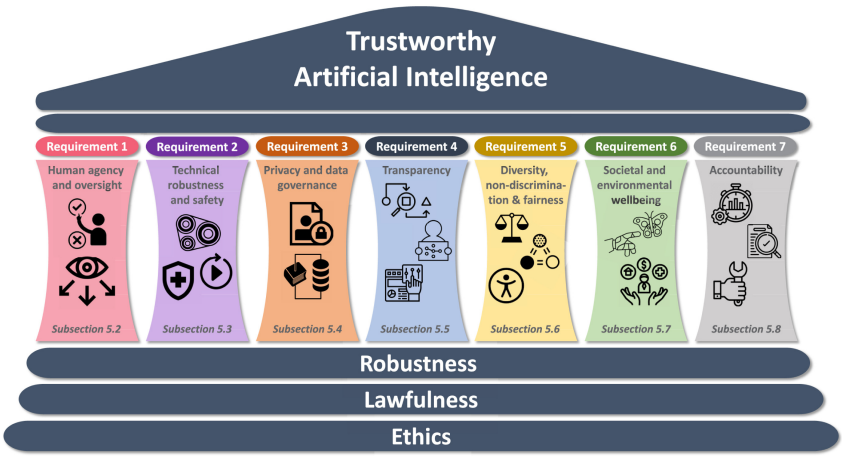


Figure 2.2: Pillars and requirements of Trustworthy AI [11]

Indicators promoted by the European Commission are divided like following.

2.2.1 Fairness

Demographic Parity: Ensuring that predictions are equally distributed across demographic groups (e.g., gender, ethnicity, age).

Equalized Odds: Checking if the model's accuracy is consistent across different demographic groups, especially in terms of false positive and false negative rates.

Calibration across groups: Ensuring that probability estimates (such as confidence levels in predictions) are reliable across different populations [12].

2.2.2 Transparency

Explainability: Explanation of how and why an AI model makes a specific decision is essential for transparency and trust. Explainability helps stakeholders understand model logic, especially in sensitive domains like healthcare or finance. Two common techniques used for this are SHape Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME).

SHAP uses principles from game theory to assign each feature an importance value based on its contribution to a prediction [13]. LIME explains individual predictions by creating a simple model that approximates the complex one locally [14]. Both approaches help expose how features influence outcomes, supporting accountability and bias detection.

Model Documentation: Documentation practices that detail the design, data, and intended use of the model, as recommended by the European Commission's AI Ethics Guidelines [11]. This includes having accessible, clear documentation about the model's purpose, limitations, and any known biases.

Auditability: The ability to audit AI decisions after they are made to verify fairness, correctness, and compliance with regulatory standards [15].

2.2.3 Accountability

Human Oversight: Ensuring that there is a mechanism for human intervention in case the AI system exhibits problematic behavior, such as incorrect predictions or systematic bias. This may involve implementing alert systems that flag anomalies, dashboards for real-time monitoring of decisions, or periodic manual review of output samples. For example, if the model produces decisions with disproportionate rejection rates for a specific demographic group, human auditors could investigate and take corrective action [16].

Responsibility Attribution: Clearly defining who is responsible for AI decisions, particularly in scenarios where biases could lead to negative impacts on individuals [17].

Regular Bias Audits: Implementing a routine check for bias to ensure the AI remains fair over time, especially if the model's data or algorithms are updated [18].

2.2.4 Robustness

Adversarial Testing: Testing the model against adversarial attacks to check for vulnerabilities that could lead to biased or unfair outputs.

Performance across conditions: Ensuring consistent performance under several conditions and for different subgroups, particularly in cases where data may be missing or imbalanced.

Stability of Results: Monitoring to confirm that predictions remain stable and consistent across different periods or environmental changes, which could affect certain demographic groups more than others [19].

2.2.5 Privacy

Data Privacy: Ensuring that personal data is anonymized or encrypted to protect user privacy, as required by the EU's General Data Protection Regulation (GDPR).

Data Minimization: Using only the data necessary for the AI's function to avoid collecting potentially biased or irrelevant information, which could influence fairness [20].

2.3 Risk-based Categorization of AI Systems

The EU AI Act, proposed by the European Commission, categorizes AI applications by risk into four levels: unacceptable, high-risk, limited risk, and minimal risk [10].

Systems deemed to present an unacceptable risk such as those enabling social scoring, real-time biometric surveillance in public spaces, or manipulative behavioral targeting are explicitly banned under the regulation, unless authorized by law for national security purposes.

These restrictions are enforced by EU regulatory authorities and designated national supervisory bodies responsible for AI compliance and oversight in each member state [21].

The second level, related to high-risk uses, will be subject to a conformity assessment before they can be deployed in the market. The conformity assessment looks at the quality

of the datasets to minimize risks and discriminatory outcomes, documentation and record keeping for traceability, transparency, and provision of information to users, human oversight, robustness, accuracy and cybersecurity provisions. The EU has defined a list of uses of AI that would be considered high-risk, such as access to employment, education and public services, management of critical infrastructures, safety components of vehicles, law enforcement and administration of justice.

The third level is devoted to limited risk uses and it will only have transparency obligations. For example, in the case of AI based chat-bots, users should be aware that they are interacting with a machine.

Finally, the fourth level includes minimal risk uses that will not be subject to any obligations tough the adoption of voluntary codes of conduct is recommended. This could enhance trust for adoption of AI and provide a lever for service differentiation, and thus a competitive advantage among service providers [21].

2.4 Limitations and Challenges of These Approaches

While these approaches have shown promising results in promoting fairness in AI, they are not without limitations and challenges. One major limitation is the potential for trade-offs between different types of fairness. For example, group fairness approaches may result in unequal treatment of individuals within a group, while individual fairness approaches may not address systemic biases that affect entire groups [22]. Additionally, it may be difficult to identify the most appropriate types of fairness in a given context, and how to balance them appropriately [23].

Another challenge is the difficulty of defining fairness itself. Different people and groups may have different definitions of fairness, and these definitions may change over time. This can make it challenging to develop AI systems that are considered fair by all stakeholders [24].

Furthermore, many of the current approaches to ensuring fairness in AI rely on statistical methods and assumptions that may not accurately capture the complexity of human

behavior and decision-making. For example, group fairness metrics may not consider intersectionality, or the ways in which different dimensions of identity such as race, gender, and socioeconomic status interact and affect outcomes [25].

Finally, there are concerns about the potential for unintended consequences and harmful outcomes resulting from attempts to ensure fairness in AI. For example, some researchers have found that attempts to mitigate bias in predictive policing algorithms may increase racial disparities in arrests [26].

Despite these challenges, the development of fair and equitable AI is an important and ongoing area of research. Future work will need to address these challenges and continue to develop new approaches that are sensitive to the nuances of fairness and equity in different contexts.

The EU has also raised concerns about the limits of current fairness tools. According to the European Commission and the European Data Protection Board, many technical solutions do not fully align with legal and ethical standards. For example, tools that claim to reduce bias may still violate fundamental rights if they lack transparency or fail to ensure informed consent. EU reports also highlight that fairness is context-dependent means what seems fair in one situation may not be fair in another. This means that relying only on automated fairness tools is risky and must be combined with human oversight, legal compliance, and ongoing evaluation to meet the EU's values of dignity, equality, and accountability [1].

Approach	Description	Examples	Limitations and Challenges
Group Fairness	Ensures that AI systems are fair to different groups of people, such as people of different genders, races, or ethnicities. Aims to prevent the AI system from systematically discriminating against any group. Can be achieved through techniques such as re-sampling, pre-processing, or post-processing the data.	1. Re-sampling techniques to create a balanced dataset. 2. Pre-processing or post-processing to adjust AI model output.	1. May result in unequal treatment of individuals within a group. 2. May not address systemic biases that affect individual characteristics. 3. Group fairness metrics may not consider intersectionality.
Individual Fairness	Ensures that AI systems are fair to individuals, regardless of their group membership. Aims to prevent the AI system from making decisions that are systematically biased against certain individuals. Can be achieved through techniques such as counterfactual fairness or causal fairness.	1. Counterfactual fairness ensuring the same decision regardless of race or gender.	1. May not address systemic biases that affect entire groups. 2. Difficulty determining which types of fairness are appropriate for a given context and how to balance them.
Transparency	Involves making the AI system's decision-making process visible to users.	Making AI system's decisions and processes understandable to users.	Different definitions of fairness among people and groups, and changing definitions over time.
Accountability	Involves holding the system's developers responsible for any harm caused by the system.	Developers held responsible for unfair decisions made by AI systems.	Determining responsibility and addressing potential harm.
Explainability	Involves making the AI system's decisions understandable to users.	Providing clear explanations of AI system's decisions.	Addressing the complexity of human behavior and decision-making.

Figure 2.3: Approaches to Fairness in Artificial intelligence [25]

Chapter 3

State of art

This chapter presents what has already been done to understand and handle bias in AI. It begins by showing where bias usually appears in AI systems. Then it explains how this bias can affect people, communities, and decisions in real life. After that, it gives an overview of the main tools and techniques currently used to reduce or manage bias. Finally, it highlights the limitations of these methods, what they cannot solve and why bias remains a serious challenge.

3.1 Situations Where Bias Appears

Bias in AI systems can manifest at several stages of their life-cycle, from data collection to real-world deployment. These biases reflect limitations in data representation, algorithm design, and application contexts, leading to unintended and often harmful consequences. Below, it will be discussed discuss key scenarios where bias arises, supported by academic insights and real-world examples.

3.1.1 Data Collection and Representation

Bias often originates in the datasets used to train AI models. If the training data is incomplete, unbalanced, or reflects historical inequities, the resulting models will perpetuate

these biases. One example of this situation comes from healthcare algorithms [6]. In 2019, a study revealed that an AI system used in United states. hospitals to identify patients needing extra medical care was significantly biased against Black patients. Although Black patients were just as sick as white patients, the algorithm predicted they needed less care. This happened because the system was trained on healthcare costs rather than actual health needs, and historically, less money has been spent on Black patients due to systemic disparities in access and treatment. As a result, the algorithm underestimated the health risks of many Black patients, reinforcing existing inequalities in medical care [6].

3.1.2 Algorithmic Design and Training

Algorithmic bias happens when ML models are built to achieve specific goals, like accuracy or speed, without considering fairness. As a result, these models can learn and repeat patterns that reflect discrimination or inequality.

A clear case of this issue can be seen in a hiring system once developed by Amazon [27]. The algorithm was trained using past hiring data where men were more likely to be hired. Because of this, the system began to favor male candidates and penalize resumes that included words related to women. This example shows how AI tools can reinforce old biases and make things worse for groups that are already underrepresented [27].

3.1.3 Real-World Deployment and Context

Even models that are well-trained and accurate in testing environments can still behave unfairly when used in real-world settings that are different from the original training context. This type of bias happens when the system continues patterns that already exist in the environment where it is deployed.

An example of this issue is predictive policing. These algorithms are often trained on historical crime data, which may already reflect unequal practices in law enforcement [28]. As a result, the system ends up focusing more on certain minority communities,

leading to repeated over-policing in the same areas. This creates a cycle where biased data leads to biased predictions, which then justify further biased actions [28].

3.1.4 Content Moderation

Bias in content moderation can occur when algorithms used to flag or remove content are influenced by the data or rules they were trained on. These systems are widely used by social media platforms to detect and manage harmful or inappropriate content. However, they often fail to consider differences in language, culture, or context. As a result, content from certain communities or regions is sometimes flagged more often than others, even when it doesn't violate any rules. This can lead to unfair representation and make some groups feel excluded or targeted. The use of keywords without understanding the full meaning or context can increase the risk of such errors.

One example of this situation is how content moderation systems have handled health-related posts and political discussions. On platforms like Instagram and Facebook, images of mastectomy scars or breastfeeding, which are shared for medical or educational reasons, have often been mistakenly removed or labeled as inappropriate due to automated nudity detection [29].

During election periods, moderation tools have also faced criticism for political bias. Posts using certain political hashtags or keywords were flagged more frequently, while similar content from different viewpoints was not [30]. These cases show how content moderation systems, if not carefully monitored and improved, can end up reinforcing bias instead of protecting users.

3.1.5 Biometric Identification

Biometric AI systems use highly sensitive personal data such as facial features, fingerprints, or iris scans to verify identity. While these technologies can improve digital security and efficiency, they also raise serious concerns about fairness, privacy, and ethics. The risk is especially high when the systems are deployed without clear user consent, strong data

protection, or transparent decision-making. If the algorithms behind these systems are not properly audited or designed with fairness in mind, they may unintentionally exclude certain groups or treat individuals unequally. These concerns are particularly important when the technology is used across borders, where laws and cultural expectations about privacy and ethics may vary [31].

A recent example of this issue is the Worldcoin project. This initiative was co-founded by Sam Altman, who is also known for his work at OpenAI, and combines digital identity verification with cryptocurrency access. The project introduced a device called the Orb, which scans a person's iris to create a unique identifier known as a World ID [32].

Although the system claims not to store personal data, many experts and regulators have raised concerns. The project has faced bans and investigations in several countries due to privacy and data protection risks. Critics have questioned whether users fully understood what biometric data was being collected and how it would be used. In some cases, even minors were reportedly scanned without proper measures. These events have sparked debate about fairness, consent, and equal access, showing how biometric systems, if not designed and deployed carefully, can unintentionally cause harm or deepen existing inequalities [33].

3.1.6 Accessibility and Language Processing

AI systems often fail to perform equally well for all users, especially when dealing with people who speak different languages, dialects, or accents. Many speech recognition technologies are trained using datasets that mostly include standard or widely spoken variations of a language [34]. As a result of this inequality, these systems often misinterpret or exclude people whose speech patterns differ from what the model expects. This creates a barrier to accessibility and may unintentionally discriminate against certain communities, limiting who can effectively interact with AI-powered tools.

One example of this situation is found in voice recognition assistants such as Siri, Alexa,

or Google Assistant. These systems have been shown to struggle with accurately understanding users who speak with strong regional accents or who use non-standard grammar. This problem is especially noticeable among speakers from under-represented linguistic backgrounds, whose voices may not have been properly included in the training data. The lack of diversity in voice samples results in lower recognition accuracy, causing frustration and reinforcing digital inequality for people around the world [35].

3.2 Bias impacts

Bias in AI systems has extensive and multifaceted effects, affecting social equity, economic opportunities, and institutional trust. These impacts are not only technical challenges but also societal and ethical concerns that demand immediate attention.

3.2.1 Social and Ethical Implications

AI systems often repeat the same biases found in the data they are trained on. When historical inequalities are present in the training data, models can learn to reflect and even amplify these patterns, leading to unfair or discriminatory outcomes [36]. One well-known example is facial recognition technology. Studies have shown that these systems are much less accurate when identifying women and individuals with darker skin tones. In some cases, the error rate for dark-skinned women reached up to 34.7%, while the error rate for light-skinned men was as low as 0.8% [8]. These gaps raise serious concerns, especially when such technology is used in sensitive areas like security, surveillance, or identification [8]. Bias can also appear in generative AI systems, which produce images, text, or audio. These models often reflect existing stereotypes learned from the data they are trained on. For example, when asked to generate images of professionals like doctors or CEOs, the results tend to show men far more often than women. This kind of output can reinforce narrow views of gender roles and under-represent people from minority groups. Even when unintentional, these biased results contribute to unequal representation in digital spaces and influence how people see each other and themselves in society [36].

3.2.2 Disparities

Bias in AI can make economic inequalities even worse. One area where this happens is in financial services. Many companies now use algorithms to help decide who gets access to credit, loans, or financial products [37]. But when these systems are trained on historical lending data that already contains unfair patterns, they often treat certain groups unfairly. For example, research has shown that some credit scoring models assign lower credit limits or reject loan applications more often for minority groups, even when those applicants have similar financial profiles to others. This makes it harder for people in those communities to access the financial support they need to build stable lives, start businesses, or invest in their future [38].

The same problem can be seen in hiring and recruitment tools. AI systems used to screen job applications or recommend candidates often favor people from majority groups. In some cases, these tools have been found to lower the rankings of resumes from women or candidates with names or experiences linked to minority backgrounds. This leads to fewer interviews, job offers, and promotions for those individuals. Over time, this creates a cycle where the same groups remain under-represented in leadership roles and higher-paying positions. By repeating past inequalities, these systems can widen the gap between groups in terms of income, opportunity, and long-term career growth [27].

3.2.3 Legal and Regulatory Risks

The use of biased AI systems creates serious legal risks and challenges for organizations. In many countries, there are strict anti-discrimination laws that protect individuals from unfair treatment based on race, gender. In the European Union, for example, the General Data Protection Regulation (GDPR) emphasizes fairness, transparency, and accountability in data processing [9]. If an AI system makes decisions that result in unequal treatment, the organization responsible could face fines, legal action or loss of public trust. These risks are especially high in industries like finance, healthcare, or law enforcement, where the impact of biased decisions can directly affect people's lives [39].

Biased systems can also damage public trust in technology. When people learn that AI tools are making unfair or harmful decisions, they may be less likely to use or support them. This can slow down the adoption of AI in important areas where it might otherwise bring real benefits.

A clear example of this problem has been seen in the use of facial recognition technology by police departments in the United States. Several cases have been reported where individuals, mostly African Americans, were wrongfully arrested because the technology misidentified them. These incidents have raised major concerns about fairness, transparency, and accountability, and they highlight the importance of responsible AI use that respects both legal standards and public expectations [40].

3.2.4 Psychological and Societal Effects

Bias in AI systems doesn't just affect decisions and opportunities, it also shapes how people think, feel, and interact with technology. When users notice or suspect that a system is biased or unfair, they are less likely to trust it. This erosion of trust can happen in many areas, especially when AI is used in personal and sensitive situations like job applications or loan approvals. People may begin to doubt whether the system is truly objective, or they may feel judged unfairly based on their background, gender, or race. These feelings of unfairness can affect public perception of AI as a whole, influencing whether people feel safe using automated services. Over time, this can reduce public confidence in AI, even in systems that are designed responsibly and with fairness in mind [6].

Biased AI can also reinforce social divisions that already exist. When algorithms repeatedly favor certain groups and overlook others, they contribute to unequal access to important resources like education, healthcare, or digital services. This can impact people's opportunities in life and widen the gap between privileged and underrepresented communities. These inequalities make it harder to achieve fairness across society, even when technical solutions are applied. If these patterns continue unchecked, they may create long-term consequences, not just in how technology works, but in how people relate

to each other in the digital world [41].

3.2.5 Biometric and Privacy Concerns

AI systems that use biometric data, such as facial recognition or iris scans, raise serious privacy and ethical concerns. These technologies collect sensitive personal information that, if not properly protected, can be misused or exposed. In many cases, systems have been launched without clear rules about how biometric data is stored, shared, or deleted. This has led to public backlash and legal investigations in several countries. Without strong protections, the use of biometric tools can lead to privacy violations and put individuals at risk, especially when the data is collected without informed consent or when users are not fully aware of how their information will be used [39].
There are also concerns about fairness and accessibility. Many biometric systems have been shown to perform less accurately for people from under-represented groups. For example, facial recognition technologies often struggle to correctly identify individuals with darker skin tones, leading to a higher risk of false matches. This creates a real danger of excluding people from important services or wrongly identifying them in high-stakes situations like security or law enforcement. These issues highlight the need for biometric AI systems to be carefully tested for bias and governed by strong privacy regulations [40].

The Table 3.1 shows how bias appears in different areas where AI is used. It gives examples from everyday sectors like hiring, banking, healthcare, and law enforcement. These examples help us see how AI tools can sometimes treat people unfairly. For example, by giving lower credit scores to certain groups or making more mistakes with medical diagnoses or facial recognition for specific populations. The table shows that bias in AI is not just a technical issue, it can seriously affect people's lives, especially those already at a disadvantage. That's why it's so important to build AI systems that are fair and safe for everyone.

Domain	Example	Impact
Employment	Recruitment tools	Gender and minority underrepresentation in hiring decisions
Financial Services	Credit scoring bias	Economic exclusion for disadvantaged groups
Healthcare	Biased diagnostic tools	Misdiagnosis or subpar care for marginalized populations
Law Enforcement	Facial recognition systems	Higher false arrest rates for African American individuals
Media	Generative AI outputs	Reinforcement of racial and gender stereotypes

Table 3.1: Examples of Bias Impacts Across Domains [42]

3.3 Overview of current approaches to mitigate Bias in AI

Researchers and practitioners have proposed several approaches to mitigate bias in AI. These approaches include pre-processing data, model selection, and post-processing decisions. However, each approach has its limitations and challenges, such as the lack of diverse and representative training data, the difficulty of identifying and measuring different types of bias, and the potential trade-offs between fairness and accuracy. Additionally, there are ethical considerations around how to prioritize different types of bias and which groups to prioritize in the mitigation of bias [25].

Despite these challenges, mitigating bias in AI is essential for creating fair and equitable systems that benefit all individuals and society. Ongoing research and development of mitigation approaches are necessary to overcome these challenges and ensure that AI systems are used for the benefit of all [25].

This section provides an in-depth look at some open-source tools and libraries that have been developed to help detect, measure, and mitigate bias in ML models. Each tool offers a unique set of features and methodologies, catering to different aspects of fairness in AI.

3.3.1 Fairlearn

Fairlearn is an open-source python library designed to bring fairness into the ML development process [43]. It helps developers understand how their models perform across different demographic groups and provides ways to reduce performance disparities. The library includes fairness metrics such as demographic parity and equalized odds, which are used to measure whether different groups receive equal treatment.
In addition to these metrics, Fairlearn offers bias mitigation strategies that work before, during, or after model training. These include optimization methods that adjust how the model learns to ensure fairer outcomes. Fairlearn integrates easily with popular ML libraries like scikit-learn, making it simple to integrate into existing workflows. It has been applied in multiple areas such as healthcare, finance and hiring systems [43].

3.3.2 AIF360 (AI Fairness 360)

Developed by IBM, AIF360 is aanother python library that helps detect and reduce bias in datasets and machine learning models. It supports a wide range of bias mitigation techniques, including pre-processing by modifying the dataset, in-processing by correcting the learning algorithm and post-processing by adjusting model outputs.
This tool has been applied in sectors such as finance, where fair credit scoring is essential, and in healthcare, where equitable treatment of patients is critical [44].

3.3.3 Themis-ML

Themis-ML is a ML library that focuses on fairness in predictive modeling [45]. It is built on top of scikit-learn and is particularly known for using adversarial training which is a method that encourages the model to make predictions that are not influenced by sensitive attributes like race or gender.
Themis-ML integrate easily with standard Python ML workflows which make it easy to use in real-world applications. It has been especially useful in recruitment systems and academic admissions processes, where decisions must be both accurate and fair [45].

3.3.4 TensorFlow Fairness Indicators

TensorFlow Fairness Indicators is a tool developed by Google to help visualize and evaluate how ML models perform for different groups of people [46]. It is particularly useful in production environments where models are used in large real-world systems.
The tool creates clear visualizations that show performance variation across demographic categories. It integrates with TensorFlow Extended (TFX) pipelines, allowing teams to monitor fairness during model development and after deployment [46].

3.3.5 Microsoft Responsible AI Toolbox

Microsoft's Responsible AI Toolbox brings together tools for improving fairness, transparency, and accountability in ML models [47]. It includes Fairlearn for fairness evaluation and mitigation, and InterpretML for explaining model predictions.
This toolbox is designed for enterprise-level applications and helps development teams identify and correct fairness issues while also providing insights into how their models make decisions. It supports the creation of responsible AI systems that are more understandable and reliable [47].

3.3.6 Adversarial Debiasing

Adversarial debiasing is a method where a ML model is trained alongside another model that tries to detect bias in the data [48]. The main model learns to make accurate predictions, while trying to prevent the second model called the adversary from figuring out sensitive attributes like race or gender. This encourages the system to ignore those attributes and focus only on the information that matters for the task.
Adversarial debiasing is commonly used in areas like finance, healthcare, and the social sciences, where protecting individuals from unfair treatment is especially important [48].

3.3.7 Sklearn-Contrib's FairLearn Extensions

This group of extensions to the popular scikit-learn library introduces additional techniques for addressing fairness [49]. It includes tools for adversarial bias removal and balanced data sampling, which can be used to create more representative datasets and fairer models.

These extensions are flexible and can be applied in multiple fields like healthcare, education and marketing where balanced outcomes and non-discriminatory predictions are required [49].

3.3.8 Custom Techniques

Sometimes, ready fairness tools aren't enough. In those cases, developers create their own methods to deal with specific problems in their data or models [25]. For example, they might add extra data to better represent groups that don't appear often which is called data augmentation, or they might write rules that make the model treat people more fairly while it learns. Another approach is to combine different models and average their results, which can help reduce bias.

These custom methods are especially useful in areas like healthcare, finance, or public policy where fairness is really important and mistakes can seriously affect people's lives. Creating solutions that fit the situation allows teams to handle fairness problems more effectively when standard tools can't do the job [25].

Chapter 4

Bias Detector Tool

This chapter introduces the Bias Detector Tool developed in this project. It explains what the tool is, how it works, and what it aims to achieve. The tool was designed to evaluate AI systems from an ethical point of view, focusing on key principles like fairness, privacy, robustness, transparency, and accountability.
In the following sections, It will be described the tool's structure, the Python libraries it uses, and how it is organized. It will be also explained what kind of input the tool needs, what kind of output it generates, and how a user can run it step by step.

4.1 Overview of the Tool

The Bias Detector Tool is a modular Python program designed to help evaluate ML models and datasets from an ethical perspective. It checks whether an AI system treats different groups fairly and respects key principles. This tool is especially helpful for researchers, data scientists, and developers who want to build responsible AI systems or check if their models might be unintentionally biased.
The main goal of the tool is not just to measure fairness but to provide a broader ethical picture. It evaluates models across five dimensions: fairness, transparency, privacy, robustness, and accountability. By analyzing these aspects, the tool helps users identify possible risks, understand their sources, and make informed decisions about improving

their systems.

The tool works automatically and is easy to use. Once a dataset is provided, the program processes the data, runs the evaluations, and generates clear, structured reports in formats that are readable both by humans and machines (TXT, JSON, and CSV). The reports include both technical scores and simple labels like "present," "not present," or "not applicable," so users without a technical background can still understand the results.

Each ethical indicator is calculated using a dedicated module:

- **Fairness** is measured by comparing prediction outcomes between different demographic groups like gender or race using metrics such as demographic parity difference. The tool checks whether the model treats groups equally regardless of sensitive attributes.

  Demographic Parity Difference measures whether different groups receive positive outcomes from a model at the same rate. It helps detect if a model is giving one group more favorable predictions than another [12].

- **Transparency** is measured using a method called **LIME**, which stands for Local Interpretable Model-Agnostic Explanations. The idea is to help users understand how the model makes decisions for individual cases.

  When the tool evaluates transparency, it picks a few individual predictions made by the model for example a loan approval for a specific person. Then, it slightly changes the input data like adjusting income or age and checks how the model's decision changes. By doing this many times, LIME finds out which features had the biggest impact on that particular decision [14].

- **Privacy** is checked by using a special kind of model called a differentially private model. This type of model is trained in a way that protects personal information [20].

  The tool does this by training two versions of the same model. One normal model without privacy protection and one private model with differential privacy using IBM's diffprivlib.

Then it compares how accurate each model is. If the private model is still almost as good as the normal one, that means the system can protect user privacy without losing too much performance. In that case, the tool gives a good privacy score.

- **Robustness** means how stable or strong a model is when the input data changes slightly. A good model should not be affected by small changes. The tool checks this by using a method called FGSM (Fast Gradient Sign Method). This method makes small changes to the input data, for example, slightly changing a number in a row and checks if the model still gives the same prediction [19].
  If a model changes its answer easily after just a small change in input, it's considered fragile and gets a lower robustness score. But if it keeps giving the correct results even after the input is modified a little, it shows that the model is strong and reliable.

- **Accountability** means whether the dataset and model allow people to understand and track what happened like who made a decision, when, and why. The tool checks for this by looking at the dataset's column names to see if there are features that support tracking and explanation [17].
  For example, if the dataset includes columns like timestamp, source, user id, or decision reason, it helps others see when and how a prediction was made. These kinds of features make the system easier to audit and explain later.

4.2 Libraries Used

The Bias Detector Tool is built entirely in Python and uses multiple well-known open-source libraries. Each library plays a specific role in helping the tool evaluate ethical aspects like fairness, privacy, transparency, and robustness.
To measure fairness, the tool uses a library called **Fairlearn**. It helps check whether different groups like men and women, or different age groups are treated equally by the model. Fairlearn provides special functions that calculate fairness scores and compare

how the model performs for each group.

For privacy, the tool relies on **Diffprivlib**, a library developed by IBM. This library helps to train models that follow the rules of differential privacy which means they can protect individual data by hiding it slightly during training. This helps make sure that no personal data can be traced back to a single person.

To improve transparency, the tool uses **LIME**. LIME is a technique that explains how a model makes a prediction. It works by showing which features like income or credit score had the biggest impact on the model's decision. This helps people understand why the model gave a certain result.

To test robustness, the tool uses Adversarial Robustness Toolbox (ART). This library helps generate tiny changes in the input data called adversarial examples to see if the model still work correctly. If the model changes its answer too easily, it means it's not very robust.

Other libraries where used in the tool include:

- **Pandas** and **NumPy** for handling data
- **Scikit-learn** for building and evaluating machine learning models
- **Matplotlib** for creating charts
- **CSV**, **JSON**, and **OS** for managing files and reports

The table 4.1 shows the main Python libraries used in the Bias Detector Tool. Each library plays a different role, from checking fairness and protecting privacy to building models and generating reports. These libraries work together to make the tool both powerful and easy to use for evaluating AI systems ethically.

Library	Purpose	What it Helps With
`fairlearn`	Fairness evaluation	Compares model performance across different demographic groups
`diffprivlib`	Differential privacy	Builds models that protect user data
`lime`	Model explainability	Shows which features influenced individual predictions
`art`	Adversarial robustness testing	Tests model stability against small input changes
`scikit-learn`	ML model training/testing	Builds and evaluates machine learning models
`pandas`, `numpy`	Data processing	Loads, organizes, and manipulates structured datasets
`matplotlib`	Visualization	Creates plots and graphs for result interpretation
`csv`, `json`, `os`	File handling	Manages input/output files and report generation

Table 4.1: Main libraries used in the Bias Detector Tool

4.3 Architecture of the tool

The Bias Detector Tool structured following a modular architecture that separates data loading, pre-processing, indicators evaluation, and report generation into clearly defined modules. This separation of concerns facilitates maintainability, extensibility and independent testing of each module.

The system follows a pipeline model. It processes each dataset in four main stages. The Figure 4.1 shows the overall structure of the Bias Detector Tool. It begins with loading the input dataset, followed by a data preprocessing step that prepares the information for analysis. Then, the system evaluates the dataset across five main indicators: fairness, transparency, privacy, robustness, and accountability. After these checks, the tool calculates a final score and creates a report to summarize the results.

Figure 4.1: Architecture of the Bias Detector Tool.

This section gives an explanation of each module of the Bias Detector Tool. Every part has a specific role starting from reading the input data all the way to checking for

ethical issues like fairness or privacy to contribute to the overall compliance assessment process.

4.3.1 Data Loading : data_loader.py

The Bias Detector Tool starts by reading the dataset, which is in CSV format. This is done by the **data_loader.py** module. It uses the pandas library to load the data and begins with some basic checks, for example, it verifies whether the file exists and if it can be read without errors. If the file passes these checks, it is loaded into a **DataFrame**, a format that is easy to work with in the next steps. During this stage, the tool also uses logging to keep a record of the process. This means that if something goes wrong or needs to be reviewed later, users can trace back what happened. The main function used here is **load_csv(file_path)**.

4.3.2 Data Preprocessing : data_preprocessor.py

Once the dataset is loaded, the tool moves on to the preprocessing step, which is handled by the **data_preprocessor.py** module. This part of the tool is responsible for cleaning and preparing the data so that it can be properly analyzed. It starts by keeping only the relevant columns such as gender, race, or income and filters out anything that isn't useful, like ID fields. It then uses **LabelEncoder** from the **scikit-learn** library to convert any text-based values like male or female into numbers, so the ML models can work with them.

The module also handles missing values by either filling them in or removing them, depending on what's most appropriate for the dataset. Extra attention is given to make sure important columns like gender, race, and timestamps are preserved, since these are critical for evaluating ethical indicators like fairness and accountability. After everything is cleaned, the processed data is saved, and a **clean DataFrame** is returned. The main functions in this module are **preprocess_data(df)** and **save_processed_data(df, path)**.

Figure 4.2 shows a sample of the dataset after preprocessing. At this stage, the data has been cleaned and prepared for evaluation.

ApplicantII	Name	Age	Occupatio	EducationI	MaritalSta	MonthlyIn	LoanAmou	LoanType	CreditScor	ApprovalSt	Applicatio	race	sex
0	5	0.272528	7	3	2	0.654377	0.802053	3	0.33647	0	3	4	0
1	0	-1.39781	4	0	1	-0.91391	-0.19021	3	0.186132	0	1	3	1
2	6	1.239563	9	2	0	-0.33258	-1.13963	4	1.073839	0	13	2	0
3	12	0.184616	2	1	1	1.181956	-0.02958	4	-0.66578	1	6	1	1
4	7	0.008791	4	2	1	-0.6221	-1.11715	0	-1.46042	0	16	0	0
5	5	1.151651	0	1	1	-1.00492	-0.05172	0	-0.83759	1	5	2	1
6	3	-0.07912	7	2	2	-0.08388	1.075155	2	-0.61567	1	14	0	1
7	6	-0.78242	6	3	0	0.692913	-1.67814	4	-0.27204	1	8	1	1
8	9	0.887914	4	3	0	0.601898	1.170799	0	-1.55349	1	2	2	1
9	11	-1.74945	6	1	2	0.813335	-0.97886	3	-0.1217	0	17	3	0
10	7	1.327476	1	2	1	-1.21103	0.860025	1	-0.5584	1	0	0	0
11	6	0.71209	3	1	1	-0.99681	-0.8792	0	1.252812	0	11	2	1
12	11	-1.39781	0	2	1	-0.07348	1.260788	2	-0.05727	1	10	0	0
13	4	1.063739	8	2	1	1.680887	-1.63624	0	1.424627	0	15	0	1
14	10	-0.16703	4	2	2	-0.92963	0.989662	2	1.288607	0	7	0	0
15	1	-1.48572	0	3	2	0.199813	1.149204	4	1.460421	1	4	1	1
16	6	-1.39781	2	1	0	1.665929	0.411781	2	-1.81121	1	12	3	1
17	0	0.184616	9	2	0	-1.25008	-1.18248	1	1.095316	0	18	3	0
18	2	0.624177	3	0	1	1.339393	0.313072	0	0.329311	0	19	2	0
19	8	0.800002	5	2	1	-1.41208	0.850693	1	-0.49397	0	9	1	1

Figure 4.2: Processed data after encoding

4.3.3 Fairness Evaluation : fairness_calculator.py

The fairness evaluation is handled by the **fairness_calculator.py** module. Its goal is to check whether the model treats different demographic groups equally. It uses the **fairlearn** library to calculate a metric called **Demographic Parity Difference**.

This metric looks at how often the model gives positive predictions like loan approval to each group, for example, comparing the approval rate for men versus women. If the difference between the groups is small, the model is considered more fair.

The module first tries to automatically detect which column in the dataset represents the protected attribute such as gender or race and which one is the binary target such as approved or not approved. The output is a dictionary that includes the calculated fairness score. This helps determine if the model is favoring one group over another.

4.3.4 Transparency Evaluation : transparency_calculator.py

Transparency is all about helping users understand how the model makes its decisions. This module, handled by **transparency_calculator.py,** begins by training a simple **logistic regression model** on the data. It assumes that the last column is the target variable. Once the model is trained, it uses **LIME**, a tool that explains individual predictions by showing which features like income or age, etc had the most influence on the model's output.

The module then calculates the model's accuracy to measure how well it performs, and uses the explanations from LIME to assess how interpretable the model is. Based on these factors, it produces a transparency score, which reflects how clearly the system's decision-making process can be understood. The final output includes this **transparency score**, along with the accuracy value and a list of the most important features. These results help show whether the model is not only correct, but also understandable.

4.3.5 Robustness Evaluation : robustness_calculator.py

To see how stable the model is, this module runs a test using **the Adversarial Robustness Toolbox (ART)**. It creates small changes in the test data using a method called **Fast Gradient Sign Method (FGSM)**. The goal is to see if the model still performs well even when the input is slightly modified. The output includes two accuracy scores, one from the original data and one from the adversarial data. If the difference is big, the model may not be very robust.

4.3.6 Privacy Evaluation : privacy_calculator.py

This module checks if the model can keep user data private. It uses **diffprivlib**, a library made by **IBM**, to train a model using differential privacy. This means the model is trained in a way that hides personal details. The module then checks how accurate the private model is. The output is the accuracy score of the privacy-preserving model. If it performs well while keeping data private, it gets a better **privacy score**.

4.3.7 Accountability Evaluation : accountability_calculator.py

The accountability check is done by the **accountability_calculator.py** module. This part doesn't use a complex ML model. Instead, it looks through the column names in the dataset to search for specific keywords, such as "**timestamp**", "**reason**" and "**audit**." These terms help the system determine whether the dataset contains elements that support auditability, traceability, and explainability which are all important for understanding how decisions are made and being able to track them later.

The Figure 4.3 shows how the tool searches for accountability related keywords in the dataset. It defines lists of keywords for three aspects checks whether any column names contain these keywords.

```
# Define keyword lists
audit_keywords = ["audit", "log", "audit_flag"]
explain_keywords = ["explain", "reason", "justification", "explanation"]
trace_keywords = ["time", "timestamp", "history", "trace"]

# Check if any column name contains any of the keywords for each aspect.
auditability = any(any(keyword in col for keyword in audit_keywords) for col in cols_lower)
explainability = any(any(keyword in col for keyword in explain_keywords) for col in cols_lower)
traceability = any(any(keyword in col for keyword in trace_keywords) for col in cols_lower)

logging.info(f"Auditability detected: {auditability}")
logging.info(f"Explainability detected: {explainability}")
logging.info(f"Traceability detected: {traceability}")
```

Figure 4.3: Accountability features

For each of these three dimensions, the module checks whether the corresponding feature is present or not. It then converts these results into a numeric **accountability score**, based on how many of the indicators are found. This score is saved and later used in the final report, where it contributes to the ethical evaluation of the AI system.

4.3.8 Report Generation : report_generator.py

In the final step, the tool brings everything together in the module **report_generator.py** and prepares the results for the user. It creates three types of reports: a structured **CSV file**, a machine-readable **JSON file** and a human-readable **TXT file**. The CSV file lists the scores for each of the five ethical indicators: **fairness**, **transparency**, **privacy**,

robustness, and **accountability**. These scores are given on a scale **from 0 to 1**, where a score closer to 1 means better performance for that ethical dimension.

Along with the individual scores, the tool also calculates a final average score. This score provides a conclusion of how ethically the system performs. If any indicator couldn't be evaluated due to missing or unclear data, it is marked as **not applicable (NA)** and excluded from the average. By offering results in multiple formats, the tool ensures that both technical and non-technical users can easily understand the system's ethical strengths and identify areas that may need improvement.

Figure 4.4 shows a sample JSON report generated by the tool. It includes detailed scores for each ethical indicator.

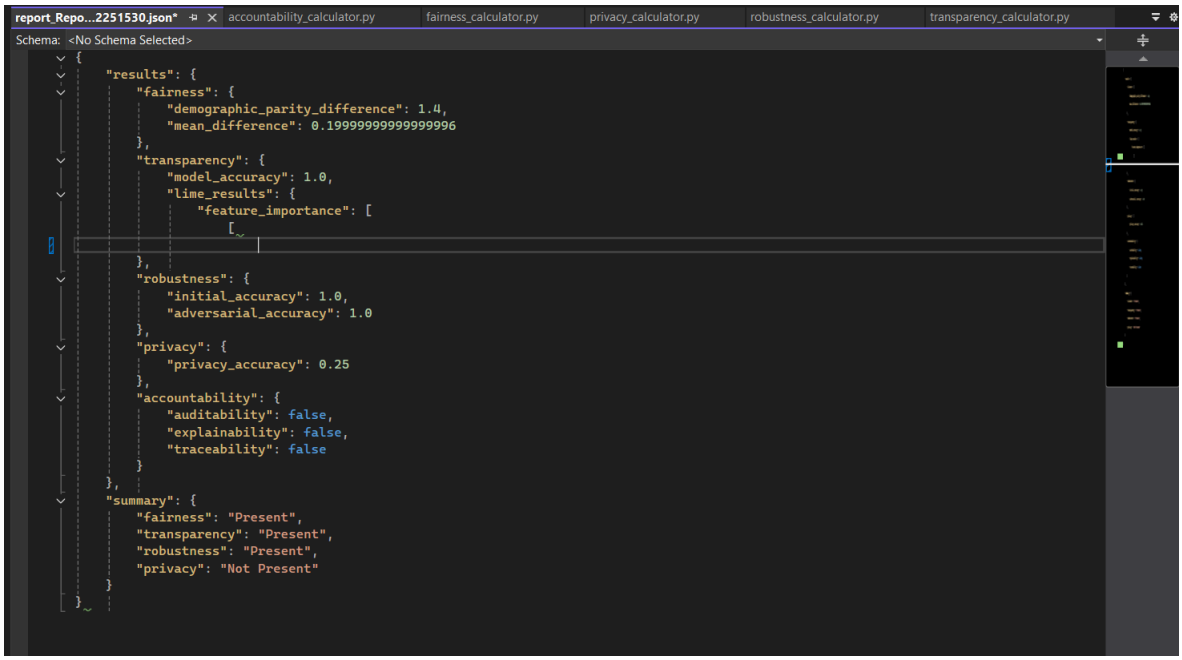


Figure 4.4: JSON result

4.3.9 Main File : main.py

The **main.py** script acts like the brain of the Bias Detector Tool. It controls how the entire system runs by coordinating the different parts step by step. First, it takes a folder that contains the **input CSV files**. Then, it calls each module one after the other — starting from loading the data, preprocessing it, running all the evaluations for fairness,

privacy, transparency, robustness, and accountability, and finally generating the reports. This script also takes care of handling errors. If something goes wrong during the process, it logs the issue so the user can understand what happened and fix it later. At the end of the process, it creates and saves all the output files, a **CSV file** with the final scores, a **JSON report** for technical review and a **TXT summary** that is easy for non-technical users to read.

Figure 4.5 shows the central logic of the main.py file, which manages the entire bias detection pipeline and calls each module in sequence.

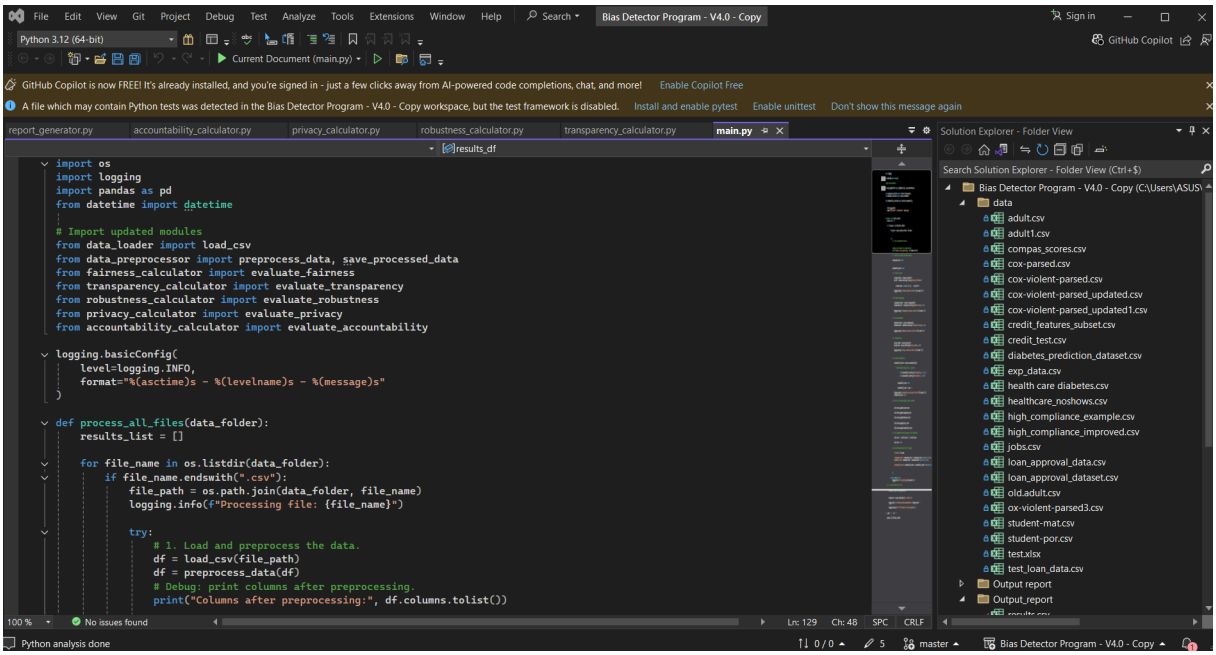


Figure 4.5: main.py

4.4 Tool Execution Process

To run the Bias Detector Tool, the first step is to prepare the data. This tool works with structured **CSV files**, so the datasets that the user want to analyze needs to be placed in the **data folder** of the project. These files often include columns such as gender, race, timestamps, or income, which are commonly used to detect ethical concerns like bias or unfair treatment. However, the tool is designed to work with a wide range of features

depending on the context and available data.
Once the data is in place, the user need to run the **main.py** script. This script is like the central brain of the tool, it automatically finds each CSV file in the folder, loads the data using the **data_loader** module, cleans and prepares it with the **data_preprocessor**, and then sends it to the evaluation modules.
Each module focuses on one ethical aspect:

- Fairness checks if all groups are treated fairly.
- Transparency shows how understandable the model's decisions are.
- Privacy tests how well the model protects user data.
- Robustness measures how stable the model is when input changes.
- Accountability looks for features like audit and trace logs in the data.

After these evaluations, the tool creates three types of reports. These files are saved in the Output_report folder.
The tool uses a scoring system from 0 to 1 for each ethical indicator. A score of 0.7 or higher means the indicator is "present", while anything below 0.7 is marked "not present". If some information is missing or not applicable, for example, no gender column in the dataset, that indicator is marked as "not applicable". These parameters are set in the current version of the tool but can be changed depending on the user specific needs and it can be done with the file **config.json**.
Finally, the tool also calculates a final average score, which gives a an overvieview of how ethically the system behaves. Higher average means more ethical system .
Before running the tool, necessary Python libraries like fairlearn, diffprivlib, and lime need to be installed. These can be added using **pip install**. Once everything is set, the program runs automatically and is simple enough to use, even for users who don't have a strong technical background.
The complete source code for the Bias Detector Tool is publicly available on GitHub. This

repository includes all modules, example datasets, and instructions for installation and usage. Users and researchers can review the implementation, run the tool locally, or adapt it to their own use cases. The repository is accessible at: https://github.com/omarsaadii/Bias-Detector-Tool.

Chapter 5

Test and validation

This chapter explains how the Bias Detector Tool was tested to make sure it works as expected. The goal was to check whether the tool can correctly identify ethical problems in datasets, such as missing information or limited diversity. To do this, It will be used two different type of datasets, one with poor structure and reduced representation, and another with proper formatting and a broader range of categories. It will be also collaborate with an expert in Equality, Diversity, and Inclusion, **Professor Raquel Ávila Muñoz** from the **Complutense University of Madrid**. Her role was to review the tool's results and confirm whether the labels assigned to the datasets,indicating the presence or absence of bias, were appropriate based on ethical and inclusion standards. Her positive feedback helped validate the tool's reliability in real-world evaluation scenarios.

5.1 Use Case Demonstration

To test the efficiency of the Bias Detector Tool, It will be designed a simple and controlled experiment using two types of datasets [50] one labeled with bias and another without bias identified. The goal was to see if the tool could clearly detect differences in ethical quality between poorly structured data and well-prepared, diverse data.

5.1.1 Datasets Description [50]

The Datasets with bias were intentionally modified to include common problems that often appear in real-world data. For example, It will be removed or replaced important fields like timestamp, source, and audit with generic column names such as K, L, and M. These changes simulated poor data practices that lead to low scores. It will be also reduced the number of race categories to only two, creating a lack of diversity. This setup helped test whether the tool would assign lower ethical scores to datasets with missing context and limited representation.

Figure 5.1 shows an example of datasets with bias used to test the Bias Detector Tool. This dataset is intentionally designed to simulate poor data practices and expose fairness and accountability issues. Multiple problems make this dataset ethically problematic.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ApplicantII	Name	Age	Occupatio	Education	MaritalSta	MonthlyIn	LoanAmou	LoanType	CreditScor	ApprovalSt	Applicatio	race	sex
2	1	Person 1	25	Engineer	Bachelor	Married	5000	20000	Personal	750	Approved	########	1	1
3	2	Person 2	30	Teacher	Master	Single	3000	10000	Education	680	Rejected	########	2	0
4	3	Person 3	40	Doctor	Doctorate	Divorced	8000	30000	Personal	800	Approved	########	1	0
5	4	Person 4	22	Student	High Schoo	Married	1000	5000	Education	620	Rejected	########	2	1
6	5	Person 5	35	Manager	Bachelor	Single	4000	15000	Business	700	Approved	########	1	1
7	6	Person 6	45	Engineer	Bachelor	Married	6000	25000	Personal	720	Approved	########	1	0
8	7	Person 7	50	Teacher	Master	Single	7000	10000	Education	770	Rejected	########	2	1
9	8	Person 8	60	Doctor	Doctorate	Divorced	3000	20000	Personal	690	Approved	########	2	0
10	9	Person 9	28	Student	High Schoo	Married	2000	8000	Education	710	Rejected	########	1	0
11	10	Person 10	33	Manager	Bachelor	Single	4500	12000	Business	740	Approved	########	1	1
12	11	Person 11	44	Engineer	Bachelor	Married	3500	18000	Personal	780	Approved	########	2	1
13	12	Person 12	21	Teacher	Master	Single	2500	6000	Education	660	Rejected	########	1	0
14	13	Person 13	27	Doctor	Doctorate	Divorced	4000	14000	Personal	650	Approved	########	2	0
15	14	Person 14	31	Student	High Schoo	Married	5500	22000	Education	670	Rejected	########	1	1
16	15	Person 15	55	Manager	Bachelor	Single	7500	26000	Business	760	Approved	########	2	1
17	16	Person 16	48	Engineer	Bachelor	Married	3800	9000	Personal	720	Approved	########	1	0
18	17	Person 17	39	Teacher	Master	Single	2900	7000	Education	680	Rejected	########	2	1
19	18	Person 18	32	Doctor	Doctorate	Divorced	4700	21000	Personal	730	Approved	########	2	0
20	19	Person 19	41	Student	High Schoo	Married	5100	17000	Education	810	Rejected	########	1	1
21	20	Person 20	29	Manager	Bachelor	Single	3200	11000	Business	640	Approved	########	1	0

Figure 5.1: Loan dataset example (loan_approval_data.csv)

First, the column names are too generic or poorly labeled. This reduces the dataset's transparency and traceability, as it is unclear what values like "1" or "2" mean, or who verified or submitted the application. Second, the values in the race column are limited to only two categories, which reflects low diversity and fails to align with **Equality, Diversity, and Inclusion (EDI)** principles.

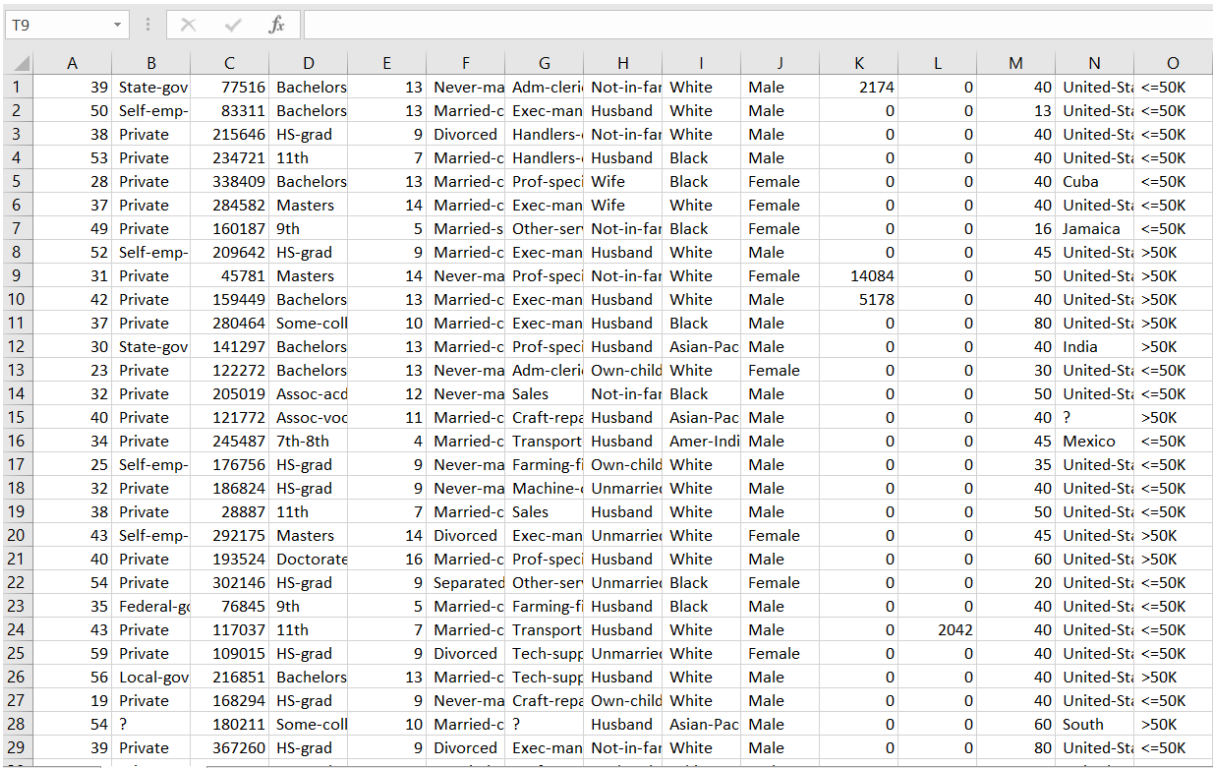
Finally, the imbalance in categories within the race and sex columns may lead ML models to treat certain groups unfairly. This is a direct fairness concern, as models trained on this data might perform better for the majority group and worse for underrepresented groups.

EDI stands for Equality, Diversity, and Inclusion. These three concepts are important when building fair and respectful AI systems. Equality means treating people fairly and making sure no one is left out or disadvantaged because of their gender, race, age, or background. Diversity is about having a mix of different people and experiences represented in the data, so the AI system doesn't just learn from one group. Inclusion means making sure that everyone who is represented in the data is also treated with care and their differences are respected.

In data practices, applying EDI means using clear and meaningful labels, including a variety of identity categories, and not ignoring small or minority groups. For example, a dataset that includes clear column names like timestamp, submitted_by or reviewer_id makes it easier to understand who handled the data and when, which improve transparency and accountability.

Similarly, a more inclusive dataset will have race and gender categories that reflect real-world diversity, such as Black, White, Asian, Latin, Indian or Mixed and options beyond binary gender like non-binary or prefer not to say. These practices are considered good because they reduce bias, allow for better representation of minority groups and help ensure that models trained on this data perform more fairly across all demographics.

Figure 5.2 shows another example of a poorly labeled dataset that was used in testing the Bias Detector Tool.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	39	State-gov	77516	Bachelors	13	Never-ma	Adm-cleri	Not-in-fa	White	Male	2174	0	40	United-St	<=50K
2	50	Self-emp-	83311	Bachelors	13	Married-c	Exec-man	Husband	White	Male	0	0	13	United-St	<=50K
3	38	Private	215646	HS-grad	9	Divorced	Handlers-	Not-in-fa	White	Male	0	0	40	United-St	<=50K
4	53	Private	234721	11th	7	Married-c	Handlers-	Husband	Black	Male	0	0	40	United-St	<=50K
5	28	Private	338409	Bachelors	13	Married-c	Prof-speci	Wife	Black	Female	0	0	40	Cuba	<=50K
6	37	Private	284582	Masters	14	Married-c	Exec-man	Wife	White	Female	0	0	40	United-St	<=50K
7	49	Private	160187	9th	5	Married-s	Other-ser	Not-in-fa	Black	Female	0	0	16	Jamaica	<=50K
8	52	Self-emp-	209642	HS-grad	9	Married-c	Exec-man	Husband	White	Male	0	0	45	United-St	>50K
9	31	Private	45781	Masters	14	Never-ma	Prof-speci	Not-in-fa	White	Female	14084	0	50	United-St	>50K
10	42	Private	159449	Bachelors	13	Married-c	Exec-man	Husband	White	Male	5178	0	40	United-St	>50K
11	37	Private	280464	Some-coll	10	Married-c	Exec-man	Husband	Black	Male	0	0	80	United-St	>50K
12	30	State-gov	141297	Bachelors	13	Married-c	Prof-speci	Husband	Asian-Pac	Male	0	0	40	India	>50K
13	23	Private	122272	Bachelors	13	Never-ma	Adm-cleri	Own-child	White	Female	0	0	30	United-St	<=50K
14	32	Private	205019	Assoc-acd	12	Never-ma	Sales	Not-in-fa	Black	Male	0	0	50	United-St	<=50K
15	40	Private	121772	Assoc-voc	11	Married-c	Craft-repa	Husband	Asian-Pac	Male	0	0	40	?	>50K
16	34	Private	245487	7th-8th	4	Married-c	Transport	Husband	Amer-Indi	Male	0	0	45	Mexico	<=50K
17	25	Self-emp-	176756	HS-grad	9	Never-ma	Farming-fi	Own-child	White	Male	0	0	35	United-St	<=50K
18	32	Private	186824	HS-grad	9	Never-ma	Machine-(	Unmarriec	White	Male	0	0	40	United-St	<=50K
19	38	Private	28887	11th	7	Married-c	Sales	Husband	White	Male	0	0	50	United-St	<=50K
20	43	Self-emp-	292175	Masters	14	Divorced	Exec-man	Unmarriec	White	Female	0	0	45	United-St	>50K
21	40	Private	193524	Doctorate	16	Married-c	Prof-speci	Husband	White	Male	0	0	60	United-St	>50K
22	54	Private	302146	HS-grad	9	Separated	Other-ser	Unmarriec	Black	Female	0	0	20	United-St	<=50K
23	35	Federal-g	76845	9th	5	Married-c	Farming-fi	Husband	Black	Male	0	0	40	United-St	<=50K
24	43	Private	117037	11th	7	Married-c	Transport	Husband	White	Male	0	2042	40	United-St	<=50K
25	59	Private	109015	HS-grad	9	Divorced	Tech-supp	Unmarriec	White	Female	0	0	40	United-St	<=50K
26	56	Local-gov	216851	Bachelors	13	Married-c	Tech-supp	Husband	White	Male	0	0	40	United-St	>50K
27	19	Private	168294	HS-grad	9	Never-ma	Craft-repa	Own-child	White	Male	0	0	40	United-St	<=50K
28	54	?	180211	Some-coll	10	Married-c	?	Husband	Asian-Pac	Male	0	0	60	South	>50K
29	39	Private	367260	HS-grad	9	Divorced	Exec-man	Not-in-fa	White	Male	0	0	80	United-St	<=50K

Figure 5.2: Dataset with bias example (adult.csv)

This dataset that was intentionally modified to test how well the Bias Detector Tool responds to poor data practices. In this case, important columns that are normally used to evaluate accountability and robustness such as source, timestamp, and audit were replaced with random columns like “K”, “L”, and “M”. This change simulates a situation where the dataset lacks traceable or verifiable information.

On the other hand, the dataset without bias identified was carefully structured and followed best practices. It included clear column names and meaningful data, along with a richer and more balanced set of race categories. This dataset was expected to perform well across most of the ethical indicators.

Figure 5.3 shows an example of a well-prepared dataset used to test the Bias Detector Tool. Unlike the biased examples, this dataset includes clear, properly labeled columns with meaningful information such as race, sex, timestamps, and decision sources. These

fields are essential for evaluating ethical indicators.

J59 | Hispanic

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	id	name	first	last	compas_s	sex	dob	age	age_cat	race	juv_fel_co	decile_sco	juv_misd_	juv_other_	priors_cou	days_b_sc	c_jail_in	c_jail_out	c_case_nu	c_offense_	c_arrest_d	c_days_fr	c_charge_	c_
2	1	miguel her	miguel	hernandez	########	Male	########	69	Greater th	Other	0	1	0	0	0	-1	########	########	13011352	########		1	(F3)	Ag
3	1	miguel her	miguel	hernandez	########	Male	########	69	Greater th	Other	0	1	0	0	0	-1	########	########	13011352	########		1	(F3)	Ag
4	2	michael ry	michael	ryan	########	Male	########	31	25 - 45	Caucasian	0	5	0	0	0									
5	3	kevon dixc	kevon	dixon	########	Male	########	34	25 - 45	African-An	0	3	0	0	0	-1	########	########	13001275	########		1	(F3)	Fe
6	4	ed philo	ed	philo	########	Male	########	24	Less than	African-An	0	4	0	1	4	-1	########	########	13005330	########		1	(F3)	Pc
7	4	ed philo	ed	philo	########	Male	########	24	Less than	African-An	0	4	0	1	4	-1	########	########	13005330	########		1	(F3)	Pc
8	4	ed philo	ed	philo	########	Male	########	24	Less than	African-An	0	4	0	1	4	-1	########	########	13005330	########		1	(F3)	Pc
9	4	ed philo	ed	philo	########	Male	########	24	Less than	African-An	0	4	0	1	4	-1	########	########	13005330	########		1	(F3)	Pc
10	4	ed philo	ed	philo	########	Male	########	24	Less than	African-An	0	4	0	1	4	-1	########	########	13005330	########		1	(F3)	Pc
11	5	marcu bro	marcu	brown	########	Male	########	23	Less than	African-An	0	8	1	0	1				13000570	########		1	(F3)	Pc
12	6	bouthy pie	bouthy	pierrelouis	########	Male	########	43	25 - 45	Other	0	1	0	0	2				12014130CF10A		########	76	(F7)	ar
13	7	marsha mi	marsha	miles	########	Male	########	44	25 - 45	Other	0	1	0	0	0	0	########	########	13022355	########		0	(M1)	Ba
14	8	edward rid	edward	riddle	########	Male	########	41	25 - 45	Caucasian	0	6	0	0	14	-1	########	########	14002304	########		1	(F3)	Pc
15	8	edward rid	edward	riddle	########	Male	########	41	25 - 45	Caucasian	0	6	0	0	14	-1	########	########	14002304	########		1	(F3)	Pc
16	9	steven ste	steven	stewart	########	Male	########	43	25 - 45	Other	0	4	0	0	3	-1	########	########	13012216CF10A		########	1	(F3)	ar
17	9	steven ste	steven	stewart	########	Male	########	43	25 - 45	Other	0	4	0	0	3	-1	########	########	13012216CF10A		########	1	(F3)	ar
18	9	steven ste	steven	stewart	########	Male	########	43	25 - 45	Other	0	4	0	0	3	-1	########	########	13012216CF10A		########	1	(F3)	ar
19	10	elizabeth t	elizabeth	thieme	########	Female	########	39	25 - 45	Caucasian	0	1	0	0	0	-1	########	########	14004524	########		1	(M1)	Ba
20	11	darrell bla	darrell	blackburn	########	Male	########	20	Less than	Caucasian	0	10	0	1	0	-1	########	########	13016981	########		1	(F3)	Pc
21	11	darrell bla	darrell	blackburn	########	Male	########	20	Less than	Caucasian	0	10	0	1	0	-1	########	########	13016981	########		1	(F3)	Pc
22	11	darrell bla	darrell	blackburn	########	Male	########	20	Less than	Caucasian	0	10	0	1	0	-1	########	########	13016981	########		1	(F3)	Pc
23	12	jamie good	jamie	good	########	Female	########	26	25 - 45	Caucasian	0	5	0	0	0	-1	########	########	14012075	########		1	(F3)	Pc
24	12	jamie good	jamie	good	########	Female	########	26	25 - 45	Caucasian	0	5	0	0	0	-1	########	########	14012075	########		1	(F3)	Pc
25	12	jamie good	jamie	good	########	Female	########	26	25 - 45	Caucasian	0	5	0	0	0	-1	########	########	14012075	########		1	(F3)	Pc
26	13	bo bradac	bo	bradac	########	Male	########	21	Less than	Caucasian	0	3	0	0	1	428	########	########	13000017	########		308	(F2)	Ins
27	14	benjamin f	benjamin	franc	########	Male	########	27	25 - 45	Caucasian	0	4	0	0	0	-1	########	########	13016402	########		1	(F3)	Pc
28	15	ellyaher la	ellyaher	lanza	########	Male	########	23	Less than	African-An	0	6	0	0	3	0	########	########	13018837	########		0	(M1)	Ba
29	16	kortney co	kortney	coleman	########	Female	########	37	25 - 45	Caucasian	0	1	0	0	0	0	########	########	13000053	########		0	(M1)	Ba

Figure 5.3: Dataset without bias identified example

What makes this dataset more appropriate for ethical evaluation is that it reflects a diverse population with multiple race categories not just binary, includes traceable data, and maintains consistency in structure. Because of these qualities, the Bias Detector Tool was able to correctly assess it as ethically good, assigning high scores in most indicators. Both datasets were evaluated using the exact same tool pipeline and at the same time. This made it possible to compare results objectively and see how each dataset performed across the five ethical dimensions: fairness, privacy, robustness, transparency, and accountability.

5.1.2 Results

After testing both types of datasets,those considered with bias and those considered without bias identified, the results clearly showed how the Bias Detector Tool reacts to ethical strengths and weaknesses in data. Datasets that lacked diversity, had missing or generic data, or showed poor transparency like **loan_approval_data.csv** or **adult.csv** that will be used in our test received low scores across most of the indicators. These scores were expected because it will either has a simplified the race categories or remove of important fields that help with auditability and traceability.

On the other hand, datasets like **cox-violent-parsed**, which include a richer variety of demographic information and well labeled columns, have much higher score. This clear difference in scoring shows that the tool successfully detects when the dataset is well-structured and respectful of ethical values or problematic, which helps users identify areas that need improvement before building AI models.

Figure 5.4 shows the results generated by the Bias Detector Tool. The rows highlighted in red represent datasets that performed poorly on ethical indicators, such as **adult.csv** and **loan_approval_data.csv**. These were considered biased examples. The row in green represents a dataset without bias identified that we used in our test **cox-violent-parsed.csv**, which scored well on most indicators, reflecting strong ethical structure and diversity.

	A	B	C	D	E	F	G
1	File Name	Fairness Score	Transparency Score	Robustness Score	Privacy Score	Accountability Score	Final Compliance Score
2	adult.csv	NA	0.466	0.466	0.004	NA	0.312
3	compas_scores.csv	0.966	0.384	0.392	0.114	NA	0.464
4	cox-parsed.csv	0.902	0.929	0.848	0.717	NA	0.849
5	cox-violent-parsed.csv	0.966	0.963	0.966	0.906	NA	0.950
6	cox-violent-parsed_updated.csv	0.333	1.000	1.000	1.000	0.667	0.800
7	cox-violent-parsed_updated1.csv	1.000	1.000	1.000	0.000	0.667	0.733
8	credit_features_subset.csv	NA	0.654	0.654	0.276	0.333	0.479
9	credit_test.csv	NA	0.981	0.947	0.399	0.333	0.665
10	diabetes_prediction_dataset.csv	0.921	0.959	0.950	0.086	NA	0.729
11	exp_data.csv	1.000	NA	NA	NA	0.667	0.833
12	health care diabetes.csv	NA	0.747	0.766	0.519	NA	0.677
13	healthcare_noshows.csv	0.928	0.345	0.345	0.000	NA	0.405
14	high_compliance_example.csv	0.959	0.000	0.000	0.000	0.667	0.325
15	high_compliance_improved.csv	0.985	0.000	0.000	0.000	0.667	0.330
16	jobs.csv	0.600	0.000	0.000	0.500	NA	0.275
17	loan_approval_data.csv	0.250	0.250	0.250	0.500	NA	0.313
18	loan_approval_dataset.csv	NA	0.009	0.009	0.015	NA	0.011
19	ox-violent-parsed3.csv	0.966	NA	NA	NA	0.667	0.816
20	student-mat.csv	NA	NA	NA	NA	0.667	0.667
21	student-por.csv	NA	NA	NA	NA	0.667	0.667
22	test_loan_data.csv	NA	0.250	0.250	0.250	NA	0.250

Figure 5.4: Results of test datasets

The results shown in the Figure above reflect how the Bias Detector Tool evaluated different datasets based on five ethical indicators: fairness, transparency, robustness, privacy, and accountability. These indicators help us understand whether the data supports responsible AI practices. Based on these results, a clear distinction emerges between the datasets labeled as “good” and those labeled as “bad.” The “good” datasets demonstrate

higher structural quality, diversity, and ethical transparency, while the "bad" datasets often exhibit poorly labeled columns, reduced representational diversity, and limited metadata — all of which increase the risk of biased or unfair outcomes in AI systems.

To support the reliability of the testing approach, the datasets used in this experiment were reviewed by **Professor Raquel Ávila Muñoz**, an expert in Equality, Diversity, and Inclusion from the Complutense University of Madrid.

After examining the Datasets with bias and dataset without bias identified, she confirmed that the labels assigned to them were appropriate based on their structure, content, and level of diversity. Although she did not see the tool's output directly, her feedback validated the design of the tool and supported the idea that the evaluation was by using meaningful and realistic data examples.

5.2 Final Considerations

The testing process demonstrated that the Bias Detector Tool is capable of identifying key ethical strengths and weaknesses in different datasets. By comparing biased datasets that are missing data, limited diversity and poor structure with well-prepared datasets, the tool consistently produced results that aligned with expectations. The low scores for datasets lacking ethical indicators confirmed the tool's ability to flag common ethical issues.

Additionally, the feedback of **Professor Raquel Ávila Muñoz** added an extra layer of confidence to the evaluation. Her confirmation that the test datasets were appropriately labeled helped validate the design and relevance of the test case.

This chapter confirms that the Bias Detector Tool can be a valuable support for researchers and developers aiming to improve the ethical quality of their data before deploying ML models.

Chapter 6

Critical Analysis and Future Perspectives

6.1 Summary of Contributions

The introduced modular tool performs evaluation of datasets and ML models over ethical dimensions which include fairness together with transparency and privacy and robustness plus accountability.

The system presents assessment results in multiple file formats which include TXT, CSV as well as JSON aside from the individual module outcome generation. The rating scale of the system extends between 0 and 1 while showing descriptive marks which include "present" or "not applicable" for better system usage.

The study included an entire use case following test data process to show how the pipeline works from preprocessing through evaluation to report production.

The developed framework provides flexible ethical AI assessment capabilities which help research activities and practical implementations.

6.2 Strengths of the System

Multiple beneficial characteristics in the bias detection system lead to its powerful performance abilities along with flexible adaptation.

The modular design ensures that individual parts can be maintained while preserving the entire framework through independent component changes.

Reputable open source libraries, including fairlearn, diffprivlib, LIME, and ART, provide system integrity and assure reliability, reproducibility, and alignment with community standards.

The system provides multiple output formats which enable both technical users and non-technical users to understand the results through accessible TXT summaries and structured CSV files.

Interpret-ability through LIME along with accountability scoring and handling of missing data scenarios enables ethical AI evaluation through a practical and comprehensive approach.

6.3 Limitations and Challenges

During development and testing multiple restrictions were identified about how the system operates despite reaching its core purposes.

One major limitation of the system is how it tries to detect important columns, like protected attributes (such as race or gender) and the target variable (what the model is predicting). Right now, the tool uses simple rules based on column names to guess which ones to use. But this approach isn't very reliable, because different datasets often use different names for the same things. As a result, the tool might miss key columns or make wrong assumptions, which can lead to errors in the analysis.

The current release lacks the capacity to provide bias mitigation techniques since it only detects biases rather than implementing or recommending solutions. The tool functions primarily as a testing instrument that measures but fails to provide a direct solution.

Integration into production environments was not implemented, and the tool currently operates as a standalone pipeline. Real-time monitoring or deployment into CI/CD (Continuous Integration/Continuous Delivery) systems would require further development.
The system deals exclusively with binary classification while handling structured tabular data types. The system lacks capability to evaluate fairness in non-tabular data including text and image formats.

6.4 Ethical Reflections

The measurement system brings clarity to AI ethical aspects yet people should distinguish between automated processes and human evaluation. The ability to generate fairness scores through tools does not grant such tools capability to determine actual fairness standards.
Ethical problems in AI are not easy to fix. Many datasets become biased even unintentionally, this often happens simply because of how the data was collected. Also, even if a system achieves perfect fairness in numbers, that doesn't always mean its decisions are truly ethical. The real world is complex, and when AI systems ignore the context behind the data, their automated evaluations can become misleading. Instead of helping, they might create confusion or make the wrong decision.
For example, the labels "present" and "not present" in the tool's report are useful for giving a quick summary, but they only capture part of the picture. Ethical issues like fairness, privacy, and accountability are complex, and simple yes-or-no answers don't reflect that complexity. In real-life areas like banking, healthcare, or the justice system, decision-makers often have to define different needs and priorities. A basic label isn't always enough to understand the impact of an AI system in these sensitive fields.
The compliance score is sometimes misunderstood as a stamp of approval, when in reality it's just a signal. People who are not familiar with how the system works might think that a high score means the model is completely safe, fair, and unbiased in every situation—but that's not true. A high score simply means the model met certain requirements based on

the data and conditions it was tested under. It doesn't guarantee ethical performance in all real-world cases.

This tool is meant to help, not replace, human decision-making. It supports data scientists, engineers, and policy teams by pointing out possible problems in the data or model before they make any final decisions. While the tool can highlight risks and give useful insights, ethical choices still need to be made by people. Responsible AI depends on human judgment, not just automation.

6.5 Future Work

The system delivers an excellent ethical assessment framework yet additional development opportunities exist for its expansion.

The most critical future step requires a shift from bias detection to implementing effective mitigation strategies. Current version of the system show instances of unfairness but next versions will actually implement bias mitigation solutions by using adjusting algorithms and post-processing methods and adversarial debiasing techniques. Users would then have the means to reduce bias in addition to targeting their locations.

The system still needs smarter ways to detect sensitive features like gender, race, or age. Right now, it tries to guess these based on built-in rules and the column names in the dataset—but this approach doesn't always work well, especially when column names are unclear or inconsistent. In the future, using more advanced techniques like Natual Language Processing (NLP), Deep Learning, or Large Language Models (LLMs) could help the tool better understand the structure and meaning of the data. This would make it more reliable and effective across different datasets.

There's also the potential to make the tool more accessible. Adding a graphical user interface or a web dashboard could make it easier for non-technical users to run analyses and read reports without needing to touch code.

A next step for the tool is to make it work in real time. Right now, the tool has to be run

manually, but future versions could be connected directly to the model training and deployment process. This would let the tool automatically check for fairness and robustness as models are being built and updated—without needing to run it separately. Adding this kind of automation into a CI/CD pipeline would make the tool more powerful and easier to use in real-world projects.

Adding more types of indicators would make the tool even more useful. For example, future versions could support multi-class classification, allow fairness checks for regression models, and include extra features to test how robust a system really is. It could also be adapted for specific industries like healthcare or finance, where ethical rules are strict and clearly defined.

These upgrades would make the tool more powerful and practical, especially in environments where ethical AI isn't just an option but a real requirement.

Chapter 7

Conclusion

The research developed a modular evaluation software which evaluates machine learning systems through several ethical aspects. The main purpose was developing a practical tool which supports responsible AI development through ethical measures of fairness transparency robustness privacy and accountability.

The developed system incorporated trusted open-source libraries as part of its modular structure to produce output that can be understood by both humans and machines. The overall system functionality was demonstrated through a loan approval process that illustrated the tool's operation starting from data input until ethical scoring and report creation.

This work not only demonstrated technical achievements but pointed out essential barriers to ethical AI implementation concerning machine limitations and score interpretation unsureness and requiring human interpretation combined with field-specific knowledge.

Though it meets its diagnostic objectives the system generates opportunities to build further advancements. The tool requires further development through implementation of bias suppression protocols alongside support for different data bases and in-time deployment to provide increased effectiveness.

The evaluation of AI ethics transcends technical considerations into a social obligation for society. The project moves toward practical systems which deliver intelligence alongside accountability and fairness in order to assist ethical evaluations of AI.

Bibliography

[1] E. Commission, *White paper on artificial intelligence*, https://ec.europa.eu/commission/presscorner/detail/en/fs_20_282, Accessed: 2024-12-15, 2020.

[2] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. 2019, Preliminary Draft. DOI: 10.2139/ssrn.2477899.

[3] T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, *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, p. 15, 2023. DOI: 10.3390/bdcc7010015. [Online]. Available: https://doi.org/10.3390/bdcc7010015.

[4] A. Kumar, V. Aelgani, R. Vohra, and J. Suri, "Artificial intelligence bias in medical system designs: A systematic review," *Multimedia Tools and Applications*, 2023. DOI: 10.1007/s11042-023-15345-6. [Online]. Available: https://doi.org/10.1007/s11042-023-15345-6.

[5] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group, 2016, ISBN: 978-0-553-41881-1.

[6] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[7] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, 2020, pp. 469–481. DOI: 10.1145/3351095.3372828.

[8] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*, 2018, pp. 77–91.

[9] European Union, *General data protection regulation (gdpr)*, https://gdpr-info.eu/, Accessed: 15 December 2024, 2016.

[10] E. Commission, *The artificial intelligence act (ai act)*, Proposed regulation laying down harmonized rules on artificial intelligence, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[11] H.-L. E. G. on Artificial Intelligence, *Ethics guidelines for trustworthy ai*, European Commission, High-Level Expert Group on AI, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[12] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 3315–3323.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, pp. 4765–4774, 2017.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[15] D. Leslie, "Understanding artificial intelligence ethics and safety," The Alan Turing Institute, Tech. Rep., 2019. DOI: 10.5281/zenodo.3240529. [Online]. Available: https://doi.org/10.5281/zenodo.3240529.

[16] L. Floridi, J. Cowls, M. Beltrametti, *et al.*, "Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018. DOI: 10.1007/s11023-018-9482-5.

[17] V. Dignum, "Responsible artificial intelligence: Designing ai for human values," *ITU Journal: ICT Discoveries*, vol. 2, no. 1, 2019. [Online]. Available: https://www.itu.int/en/journal/002/Pages/default.aspx.

[18] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 429–435. DOI: 10.1145/3306618.3314244.

[19] M. Brundage, S. Avin, J. Clark, *et al.*, "Toward trustworthy ai development: Mechanisms for supporting verifiable claims," in *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 272–282. DOI: 10.1145/3375627.3375830.

[20] P. Voigt and A. Von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer International Publishing, 2017, ISBN: 978-3-319-57959-7. DOI: 10.1007/978-3-319-57959-7.

[21] European Commission, *Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act)*, COM/2021/206 final. Official definition of risk levels in the EU AI Act, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[22] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016, Accessed: February 19, 2025. DOI: 10.2139/ssrn.2477899.

[23] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Algorithmic fairness," *AAAI Conference on Artificial Intelligence*, vol. 1, no. 1, pp. 1–11, 2018, Accessed: February 19, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v32i1.11375.

[24] C. Dwork, N. Immorlica, A. D. Kalai, and M. Raghavan, "Decoupled classifiers for fair and efficient machine learning," *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 119–133, 2018, Accessed: February 19, 2025. [Online]. Available: https://arxiv.org/abs/1806.00692.

[25] A. Name, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *arXiv preprint*, vol. arXiv:2304.07683, 2023, Accessed: February 19, 2025. [Online]. Available: https://arxiv.org/pdf/2304.07683.

[26] A. G. Ferguson, "Predictive policing and reasonable suspicion," *Emory Law Journal*, vol. 62, no. 2, pp. 259–325, 2012, Accessed: February 19, 2025. [Online]. Available: https://scholarlycommons.law.emory.edu/elj/vol62/iss2/1.

[27] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," *Reuters*, 2018, Accessed: 2024-12-15. [Online]. Available: https://www.reuters.com/article/amazon-ai-recruiting-idUSKCN1MK08G.

[28] O. Academic, *Oxford academic website*, https://academic.oup.com, Accessed: 15 December 2024, n.d.

[29] J. C. Wong. "How a cancer group thwarted facebook's censorship: Square breasts." Accessed: 18 May 2025, The Guardian. (Oct. 2016), [Online]. Available: https://www.theguardian.com/technology/2016/oct/20/facebook-bans-breast-cancer-video-square-breasts.

[30] A. Bridgman, E. Merkley, P. J. Loewen, *et al.*, "The causes and consequences of covid-19 misperceptions," *MISINFORMATION REVIEW*, 2020, Accessed: 18 May 2025. [Online]. Available: https://misinforeview.hks.harvard.edu/wp-

content/uploads/2020/06/formatted_causes_consequences_bridgman-et-al4.pdf.

[31] E. A. Whitley and R. van Brakel, "A typology of biometric data usage in the european union," *Computer Law Security Review*, vol. 41, p. 105 530, 2021. DOI: 10.1016/j.clsr.2021.105530.

[32] Worldcoin, *World app and proof of personhood: Digital identity and cryptocurrency management*, https://worldcoin.org, Accessed: 15 December 2024, n.d.

[33] European Data Protection Board, *Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement*, Accessed: 15 December 2024, 2023. [Online]. Available: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052022-use-facial-recognition-technology-area_en.

[34] A. Koenecke, A. Nam, E. Lake, *et al.*, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020. DOI: 10.1073/pnas.1915768117.

[35] A. Koenecke, A. Nam, E. Lake, *et al.*, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020. DOI: 10.1073/pnas.1915768117.

[36] L. Nicoletti and D. Bass, "Humans are biased: Generative ai is even worse," *Bloomberg Technology+Equality*, 2023, Accessed: 15 December 2023. DOI: 10.5555/bloomberg.2023.3432.

[37] M. Hurley and J. Adebayo, "Credit scoring in the era of big data," *Yale Journal of Law and Technology*, vol. 18, no. 1, pp. 148–216, 2016. [Online]. Available: https://digitalcommons.law.yale.edu/yjolt/vol18/iss1/5/.

[38] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[39] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a standard for identifying and managing bias in artificial intelligence," *NIST Special Publication*, vol. 1270, 2022.

[40] E. Ferrara, "The butterfly effect in artificial intelligence systems: Implications for ai bias and fairness," *SSRN Electronic Journal*, 2023. DOI: 10.2139/ssrn.4614234.

[41] L. Sweeney, "Discrimination in online ad delivery," *Communications of the ACM*, vol. 56, no. 5, pp. 44–54, 2013. DOI: 10.1145/2447976.2447990.

[42] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021. DOI: 10.1145/3457607.

[43] S. Bird, M. Dudík, R. Edgar, B. Fresquez, and H. Wallach, *Fairlearn: A toolkit for assessing and improving fairness in ai*, Accessed: 19 February 2025, 2020. [Online]. Available: https://fairlearn.org/.

[44] R. K. E. Bellamy, K. Dey, M. Hind, *et al.*, *Ai fairness360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*, Accessed: 19 February 2025, 2019. DOI: 10.1147/JRD.2019.2942287.

[45] T.-M. Developers, *Themis-ml: A fairness-focused extension of scikit-learn*, https://github.com/cosmicBboy/themis-ml, Accessed: February 19, 2025, 2023.

[46] M. Abadi, P. Barham, J. Chen, *et al.*, *Tensorflow: Large-scale machine learning on heterogeneous systems*, https://www.tensorflow.org, Accessed: February 19, 2025, Software available from https://tensorflow.org, 2015.

[47] M. Research, *Microsoft responsible ai toolbox*, https://www.microsoft.com/en-us/ai/responsible-ai, Accessed: February 19, 2025, 2023.

[48] H. AI, *Adversarial debiasing - bias mitigation in machine learning*, https://holisticai.readthedocs.io/en/latest/getting_started/bias/mitigation/inprocessing/bc_adversarial_debiasing_adversarial_debiasing.html, Accessed: February 19, 2025, 2023.

[49] scikit-learn-contrib, *Fairlearn extensions: Additional fairness techniques in the scikit-learn ecosystem*, `https://github.com/scikit-learn-contrib`, Accessed: February 19, 2025, 2023.

[50] Kaggle, *Kaggle*, `https://www.kaggle.com/`, Accessed 25 June 2025.