A Survey of tools for Trustworthy AI

Varshith Vattikuti Neil Handa

December 6, 2024

Abstract

This document explores the development of trustworthy AI, addressing ethical, legal, and transparency challenges in AI deployment. It categorizes tools for AI trustworthiness into technical, procedural, and educational, highlighting frameworks like GDPR and the EU AI Act. Recommendations include granular transparency standards, real-time auditing, explainability frameworks, and creator-consent systems to ensure AI systems are ethical, accountable, and socially beneficial.

1 Preface

In recent years, numerous frameworks, declarations, and principles have emerged from various organizations worldwide to guide the development of trustworthy artificial intelligence (AI). These frameworks outline the essential goals and desired outcomes for trustworthy AI systems, including safety, fairness, transparency, accountability, and privacy. However, they often lack specific instructions on how to achieve these objectives in practical scenarios. This gap highlights the importance of tools for trustworthy AI. These tools encompass a range of methods, techniques, mechanisms, and practices designed to measure, evaluate, communicate, improve, and enhance the trustworthiness of AI systems and applications.

2 Preliminary Considerations

The advancement of AI has been swift in recent years, with its applications spreading across various domains. AI holds the potential to bring substantial benefits to society, including enhancements in productivity, innovation, health, education, and overall well-being. However, the rapid progress of AI also presents significant risks and challenges—social, ethical, legal, economic, and technical—that must be addressed to ensure AI is trustworthy. As a result, AI has become a focal point for stakeholders worldwide, leading to numerous discussions and initiatives aimed at developing and deploying AI responsibly and ethically.

Generally, AI systems and applications are considered trustworthy when they can be developed and deployed reliably, without causing harm to individuals, groups, or society. Although there is no universally accepted definition of trustworthy AI, various stakeholders, including governments and international organizations, have proposed their own definitions. These definitions often overlap and are characterized by principles such as fairness, transparency, accountability, privacy, safety, and explainability.

Tools for trustworthy AI are specific approaches or methods designed to enhance the trustworthiness of AI. These tools bridge the gap between high-level AI principles and their practical implementation. They encompass methods, techniques, mechanisms, and practices that help

measure, evaluate, communicate, improve, and enhance the trustworthiness of AI systems. The ultimate goal of these tools is to provide developers, policymakers, and other stakeholders with the resources needed to ensure that AI is developed and deployed responsibly and ethically.

Key Questions to Address

As we delve into the development and deployment of trustworthy AI, several critical questions arise that must be addressed to ensure the responsible and ethical use of AI systems. These questions highlight the essential considerations for building trust and confidence in AI:

- How can we ensure that AI systems are aligned with human values and respect human rights?
- How transparently are the decisions and actions of AI systems explained?
- How can we prevent and mitigate potential harms of AI, such as bias and discrimination?
- How safe, secure, and reliable is an AI system?

Addressing these questions is crucial for fostering trust and confidence in AI among consumers and the public. By focusing on these aspects, we can work towards developing AI systems that are not only technologically advanced but also ethically sound and socially beneficial.

3 Trustworthy AI and Tools for Trustworthy AI in This Context

Trustworthy AI involves developing systems that are reliable and do not harm individuals or society. While there is no universal definition, key characteristics include fairness, transparency, accountability, privacy, safety, and explainability. Various frameworks, such as the European Commission's Ethics Guidelines and the OECD AI Principles, outline these principles but often lack practical guidance.

Tools for trustworthy AI bridge this gap by providing methods to measure, evaluate, and enhance AI trustworthiness. These tools help developers and policymakers ensure that AI is developed and deployed responsibly and ethically.

UK government[9]	National Institute of Standards and Technology (US)[10]	European Commission[11]	Organisation for Economic Co-operation and Development[12]
Five principles: • Safety, security and robustness • Appropriate transparency and explainability • Fairness • Accountability and governance • Contestability and redress	Seven characteristics: • Valid and reliable • Safe • Secure and resilient • Accountable and transparent • Explainable and interpretable • Privacy-enhanced • Fair – with harmful bias managed	Three components: • Lawful • Ethical • Robust Four ethical principles: • Respect for human autonomy • Prevention of harm • Fairness • Explicability Seven requirements: • Human agency and oversight • Technical robustness and safety • Privacy and data governance • Transparency • Diversity, non-discrimination and fairness • Societal and environmental well-being • Accountability	Five principles: • Inclusive growth, sustainable development and well-being • Human-centred values and fairness • Transparency and explainability • Robustness, security, and safety • Accountability

Figure 1: Characteristics of Trustworthy AI from different stakeholders [21]

4 Classification of Existing Tools

Some tools are technical, providing solutions through code or algorithms that can be applied to AI models or datasets to ensure their trustworthiness. Many of these technical tools are developed by large private sector companies like IBM, Google, and Microsoft. A significant number of these tools are open-source, which promotes their adoption and allows for collaborative problem-solving and bug fixing.

Category	**Tool**
Fairness	AT&T SIFT
	Microsoft Fairlearn
	IBM AI Fairness 360
Transparency	IEEE Standard for Transparency of Autonomous Systems
	Google Model Card Toolkit
Explainability	Google Cloud Explainable AI service
	IBM AI Explainability 360 Toolkit
	Microsoft InterpretML
Robustness	IBM Adversarial Robustness 360 Toolkit

Table 1: Technical Tools for Trustworthy AI

Other tools are procedural, offering compliance-based solutions where AI models are evaluated and tested to determine their trustworthiness. Unlike technical tools, which see high participation from the private sector, procedural tools are developed by a diverse range of stakeholders, including governments and trade unions. These tools aim to ensure that AI systems are implemented ethically and inclusively.

Category	**Tool**
Inclusive Implementation	German Trade Union Confederation's Good Work by Design
	Negotia AI Governance System
	Google People + AI Guidebook
Ethical Implementation	IBM Everyday Ethics for AI
	IEEE Ethics Certification Program for Autonomous and Intelligent Systems
	IEEE Trusted Data & AI Systems Playbook for Finance Initiative
	Denmark Algorithm Test
Transparent and Explainable Implementation	Microsoft Datasheets for Datasets
	IBM AI Factsheets 360
	UK Information Commissioner's Office "Explaining decisions made with AI"

Table 2: Procedural Tools for Trustworthy AI

Additionally, there are educational tools designed to raise awareness about trustworthy AI among specific stakeholders or the general public. These tools can be tailored to different audi-

ences, ranging from broad public education to targeted groups affected by AI implementation, such as small and medium-sized enterprises (SMEs) or workers.

Category	Tool
Businesses	Denmark Data Ethical Dilemma Game
Workplace Actors	Negotia AI Governance System
2*General Public	Finland AI Course "Elements of AI"
	VIRT-EU Service Package

Table 3: Educational Tools for Trustworthy AI

5 Observational Analysis of Ethical and Legal Challenges

5.1 Data Sourcing and Transparency

The training of LLMs relies on massive datasets, which are often aggregated from sources like Common Crawl, GitHub, and Project Gutenberg. While these datasets provide a diverse corpus, they frequently include copyrighted, biased, or sensitive material. For instance, Common Crawl, used by GPT-4, is a publicly available web scraping database comprising billions of web pages. However, OpenAI has not disclosed which specific sites were included, nor the filtering criteria used to remove harmful content [1]. This lack of transparency raises concerns about compliance with intellectual property laws and the ethical implications of using unverified data.

The case of Getty Images v. Stability AI exemplifies the risks associated with non-consensual data usage. Stability AI allegedly used copyrighted images without obtaining licenses, leading to significant legal challenges [2]. Similarly, Doe v. GitHub highlighted the risks of insufficient dataset traceability, as developers accused GitHub's Copilot of reusing open-source code without adhering to licensing terms [3].

5.2 Explainability and Decision Transparency

A critical aspect of AI transparency is the explainability of decision-making processes. For example, a hypothetical credit scoring model with 22 core decision-making factors might need to disclose the relative weights of each factor (e.g., income, credit history, debt-to-income ratio) to meet transparency requirements. However, the EU AI Act does not specify exactly how many factors must be explained for compliance, creating a significant gap in enforcement [4]. A practical example of this issue is found in the mortgage approval algorithms deployed by large financial institutions, where decision pathways often remain opaque, even to regulators. Explainability becomes particularly challenging in complex systems like LLMs, where outputs are influenced by intricate, non-linear interactions between numerous factors. In healthcare, for example, diagnostic AI tools often fail to disclose the weight assigned to specific patient variables, such as age or comorbidities, leaving clinicians with insufficient information to validate or challenge the system's recommendations. Similarly, in credit scoring, an over-reliance on opaque models has led to accusations of algorithmic bias, particularly against minority groups [11].

This ambiguity allows developers to provide superficial explanations that do not address user concerns or regulatory expectations. Academic studies by Mittelstadt et al. emphasize that meaningful transparency requires not only technical disclosures but also interpretability for users and that technical disclosures alone are insufficient for meaningful transparency [5]. However, a 2023

survey conducted by the Institute for Ethical AI and Machine Learning revealed that fewer than 20% of AI developers provide multi-level explanations, which accommodate the varying technical expertise of stakeholders [14]. Several frameworks and tools have been developed to address the gap in explainability. For example, the SHapley Additive exPlanations (SHAP) method has been widely adopted in machine learning for attributing model outputs to individual input features. A study by Lundberg et al. (2020) demonstrated the effectiveness of SHAP in explaining complex medical diagnostic systems, enabling clinicians to understand the contributions of individual patient variables to diagnostic outcomes [15]. However, the adoption of such tools remains inconsistent, particularly in commercial applications.

5.3 Traceability and Record-Keeping

Article 14 of the EU AI Act emphasizes the need for AI developers to maintain comprehensive records of datasets, algorithms, and modifications to ensure traceability and accountability. These provisions are particularly important for high-risk systems, where errors or biases can have significant societal impacts. However, the lack of standardized logging practices undermines the effectiveness of this requirement.

Meta's LLaMA model serves as a prime example. In its release documentation, Meta disclosed that LLaMA was trained on publicly available repositories, including GitHub, but did not specify how sensitive or proprietary data was filtered out. This omission raises concerns about compliance with intellectual property laws and the potential misuse of copyrighted material. GitHub itself hosts over 370 million repositories, with more than 72% of them using open-source licenses, each with unique usage restrictions [9]. The lack of a clear audit trail in models like LLaMA makes it difficult to determine whether such licenses were respected, a critical gap in traceability. Moreover, a study by the Center for AI and Digital Policy (2023) revealed that less than 15% of AI developers provide detailed documentation of dataset provenance and algorithmic changes, even for high-risk systems [10].

In another case, GitHub's Copilot, which also trained on publicly available repositories, faced backlash for generating code snippets that allegedly reused copyrighted content. A study by Stanford University (2023) found that 40% of Copilot's outputs contained code directly traceable to publicly available repositories, raising questions about compliance with licensing terms and the effectiveness of existing traceability mechanisms [11].

The challenges are compounded by the absence of standardized tools for record-keeping. While some organizations have adopted frameworks like Data Version Control (DVC) or Model Cards for documenting changes, their use remains limited. A survey by McKinsey & Company (2022) found that fewer than 10% of organizations deploying AI models employ version control systems for their datasets and algorithms, leaving significant gaps in traceability [12]. This lack of standardized logging practices also creates issues in identifying the origin of biases in AI systems. For instance, in Amazon's AI hiring tool, which was discontinued in 2018 due to gender bias, the lack of detailed records on the dataset and algorithmic changes made it difficult to pinpoint the source of the issue. Such incidents underscore the need for rigorous traceability mechanisms to prevent and address ethical violations in AI deployment [13].

6 Regulatory Efforts

6.1 General Data Protection Regulation: Global Impact and Ambiguity

The GDPR requires organizations to disclose how personal data is collected, processed, and stored, granting users the right to access, correct, and delete their data. However, enforcement has been inconsistent, with many companies exploiting ambiguities to minimize compliance costs. While

GDPR violations can result in fines of up to 4% of a company's annual global revenue, these penalties are often insufficient to deter large firms. For example, Meta faced a €1.2 billion fine in 2023 for transferring EU user data to the United States in violation of GDPR rules. Despite the record-breaking penalty, Meta's total revenue exceeded €116 billion that year, making the fine less than 1% of its annual income and failing to incentivize meaningful compliance [16]. However, the more major issue is that GDPR allows companies to process user data without explicit consent if they can demonstrate a "legitimate interest." This clause has been exploited by firms like Google, which argued that its ad personalization practices fell under this category. In a 2021 ruling by the French data regulator CNIL, Google was fined €100 million for violating cookie consent rules, but the lack of clear guidelines allowed the company to delay significant changes for months while appealing the decision [17].

6.2 High-Level Expert Group on AI: Ethical Guidelines

The High-Level Expert Group on AI, established by the European Commission in 2018, developed the "Ethics Guidelines for Trustworthy AI" in 2019. These guidelines outlined seven key principles, including transparency, accountability, and fairness, aimed at fostering ethical AI development. However, the guidelines were voluntary and lacked enforceable mechanisms, limiting their practical impact. For example, the principle of transparency required developers to document their AI systems comprehensively, but the absence of auditing requirements allowed companies to provide superficial documentation. The limited scope of the guidelines became evident in cases like Amazon's AI hiring tool, which was discontinued in 2018 after it was found to exhibit gender bias. The lack of enforceable transparency measures prevented regulators from addressing the root cause of these biases effectively [13].

6.3 Other Regulatory Efforts and Their Limitations

The EU AI Act is also informed by broader international and industry-specific frameworks, such as the United Nations' "Personal Data Protection and Privacy Guidelines" and the United States' sectoral AI regulations, including the FDA's guidelines for AI in medical devices. These efforts have faced similar enforcement challenges. The UN guidelines emphasize the need for cross-border data protection but lack the enforcement mechanisms required to hold multinational corporations accountable. For instance, in 2021, a study revealed that 70% of AI systems deployed in developing countries used datasets that violated the guidelines by including sensitive demographic information without proper anonymization [18]. The FDA requires manufacturers of AI-powered medical devices to demonstrate the safety and effectiveness of their systems through explainability and bias testing. However, companies like IBM Watson faced criticism for overstating the capabilities of their AI oncology systems, which provided inaccurate recommendations in some cases. The lack of pre-approval auditing for datasets allowed such systems to be marketed without addressing these flaws [19].

7 Recommendations to Enhance Compliance

7.1 Granular Transparency Standards

Developers should disclose source URLs, repository IDs, and metadata for each dataset used during training. For example, OpenAI's GPT-4, which relies on Common Crawl, should provide a detailed breakdown of the specific websites included and the criteria used to filter harmful or irrelevant content. Such granular disclosures would prevent the inclusion of copyrighted or biased material, aligning with both the ethical and legal standards outlined in the Act. To ensure compliance, policymakers must establish mechanisms for auditing and accountability. Developers

should be required to maintain auditable logs of all datasets used during the training process, updated in real time and accessible to regulators upon request. Open-source tools such as Data Version Control (DVC) can automate this process, creating a comprehensive history of dataset modifications. Policymakers should also establish independent audit bodies tasked with verifying the accuracy and completeness of these disclosures. For example, these bodies could utilize tools like Google's Dataset Search to cross-reference declared data sources with the datasets actually used.

7.2 Real-Time Auditing Systems

The EU AI Act currently emphasizes pre-deployment assessments, but the dynamic nature of AI systems necessitates continuous monitoring to address evolving risks. To ensure ongoing compliance, real-time auditing systems must be integrated into the regulatory framework. These systems would enable developers and regulators to identify and address issues such as bias, demographic disparities, and data leakage during a system's operational lifecycle. Developers should be required to submit monthly compliance reports detailing system outputs, detected biases, and corrective measures. These reports can be automated using tools such as Google's Fairness Indicators, which analyze disparities in system outputs across demographic groups and track anomalies over time. Policymakers should facilitate this process by establishing secure APIs for developers to submit these reports to regulatory bodies. This would streamline compliance reporting while maintaining robust oversight. In addition to compliance reports, continuous monitoring systems should employ tools like IBM's AI Fairness 360 toolkit to detect and flag biases in real time. These open-source frameworks provide metrics for evaluating fairness, ensuring that systems adhere to Article 14's traceability requirements.

7.3 Tiered Explainability Framework

High-risk systems, such as those used in healthcare or credit scoring, should disclose all decision-making factors, including the relative weights assigned to each input and the interactions between them. For instance, a credit scoring model might reveal how factors such as debt-to-income ratio, credit history, and employment stability influence its outputs. This level of detail ensures that stakeholders can identify potential biases or inaccuracies in the system's logic. In contrast, lower-risk systems, such as recommendation engines, can meet compliance by explaining only the most impactful factors. For example, these systems could focus on the top 75% of factors influencing their outputs, reducing the compliance burden without compromising accountability. Policymakers can enforce this framework by standardizing explainability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). These tools provide consistent metrics for attributing model outputs to individual input features, enabling meaningful scrutiny.

7.4 Creator-Consent Frameworks

A recurring issue in AI development is the use of datasets without explicit consent from creators. While Article 13 requires developers to ensure transparency in data usage, it does not mandate mechanisms for verifying creator consent. This gap can be addressed by establishing a centralized consent database, modeled on Flickr's Creative Commons licensing system. Such a database would allow creators to specify usage terms for their works, ensuring that only ethically sourced data is used in AI training. Developers must verify their datasets against this database before deployment, with non-compliance resulting in penalties such as fines or suspension of deployment licenses. Blockchain technology can further enhance this process by creating tamper-proof records of creator consent, ensuring that datasets are both transparent and traceable. The implementation

of such frameworks would require collaboration between policymakers, industry stakeholders, and content creators. Policymakers should provide incentives for developers to adopt these systems, such as tax breaks or reduced compliance fees. Additionally, regular audits should verify that datasets align with the permissions granted by creators, further reinforcing accountability.

8 Framework for Trustworthy AI Tools

The goal of the present framework is to structure information and facilitate comparison between different tools used for various purposes and in diverse contexts. It does not aim to provide any qualitative assessment. The seven key dimensions of the framework are:

- **Tool Overview**: Includes the tool's name and other important information
- **Tool Background**: Includes the details of the organization,stakeholder group and country/ region where the tool was created. It also includes the date/year in which the tool was introduced.
- **Tool Classification**: Classifies the tool as one of the 3 major categories - technical, procedural or educational as discussed above. Also specifies the type of tool that is toolkit/ certification/ guidelines/ standards etc.
- **Applicability**: Includes points like target users, stakeholders groups and geographical scope of the tool under consideration.
- **Principle Alignment**: Considers the tool's relevance to the Trustworthy AI Principles and the various characteristics for trustworthy AI.
- **Adoption Feasibility**: Evaluates the maturity of the tool and how up-to-date it is in terms of the resources required and legal conditions for use.
- **Implementation Motivators**: Highlights the expected benefits from using the tool and the various methods that can help in its adoption.

Type	Field	Definition
Tool description	Name	The name of the tool
	Link	A link to an up-to-date document
	Description	A brief summary of the tool and its purpose
Tool origin	Organisation	The organisation that developed the tool
	Stakeholder group	The stakeholder group from which the initiative originates
	Country	The country or region where the initiative originated
	Date of publication	Date the tool was published in its first version
	Contact email	Email of the contact person for the tool (not for public use)
Tool categorisation	Type of Approach	High-level category of the tool
	Type of Tool	Category of the tool
Scope	Technology platform	The technology platform(s) that the tool can be used for
	Target stakeholder group	The stakeholder group where the tool is expected to be implemented
	Primary and secondary policy area	The policy area(s) where the tool is expected to be implemented
	Geographical scope	The country or region that the initiative targets
	Target users of the tool	Users who are expected to use the tool to implement a project
	Impacted stakeholders	Groups of people that will be impacted by the implementation
Alignment with international AI Principles	Relevance to international AI Principles	Grade relevance to international AI Principles
Potential for adoption	Maturity of the tool	Project phase the tool is currently in
	Degree tool is kept up to date	How the tool is kept up to date with evolving standards, requirements, etc.
	Degree of free use of the tool	Legal conditions for using the tool
	Required resources to implement	The extent to which certain resources are needed to implement/use the tool
	Stakeholders involved	Stakeholders who will be involved in the implementation and operation of the tool
Implementation incentives	Expected benefits	Expected benefits from using the tool
	Enforcement mechanisms	Enforcement mechanisms attached with the usage of this tool

Table 4: Detailed Table for Tool Description and Categorization

Example Tool : AI Fairness 360 Toolkit

The AI Fairness 360 (AIF360) toolkit, developed by IBM, is an open-source library designed to help detect and mitigate bias in machine learning models. It provides a comprehensive suite of fairness metrics and bias mitigation algorithms applicable across various stages of the AI lifecycle. The toolkit is widely adopted for promoting fairness and equity in AI systems, aligning with global ethical AI principles.

Type	Field	Definition
Tool description	Name	AI Fairness 360
	Link	https://aif360.res.ibm.com/
	Description	A library for assessing and mitigating bias in machine learning models.
	Organisation	IBM
Tool origin	Stakeholder group	AI/ML researchers and developers
	Country	United States
	Date of publication	October 2018 https://ieeexplore.ieee.org/document/8843908
	Contact email	-
Tool categorisation	Type of Approach	Open-source software library
	Type of Tool	Technical tool for bias mitigation
Scope	Technology platform	Python-based data science environments
	Target stakeholder group	Developers and policymakers
	Primary and secondary policy area	Fairness and transparency in AI
	Geographical scope	Global
	Target users of the tool	Developers building AI systems
	Impacted stakeholders	End users and marginalized groups
	AI system lifecycle stage(s) covered	Data preparation, model training, evaluation
Alignment with international AI Principles	Relevance to international AI Principles	Supports fairness, accountability, transparency
	Maturity of the tool	Production-ready
Potential for adoption	Degree tool is kept up to date	Actively maintained
	Degree of free use of the tool	Fully open-source
	Required resources to implement	Python programming knowledge
	Stakeholders involved	Developers, organizations
Implementation incentives	Expected benefits	Identifies and mitigates bias in AI models
	Enforcement mechanisms	Diagnostic and mitigation tool, voluntary adoption

Table 5: Overview of the AI Fairness 360 Toolkit by IBM

The dataset for which the above toolkit was used is the **Adult Census Income** dataset, which is commonly used to predict whether an individual's annual income exceeds $50,000 based on census data.

The **Adult Census Income** dataset contains the following protected attributes:

- **Race**:
  - Privileged group: *White*
  - Unprivileged group: *Non-white*
- **Sex**:
  - Privileged group: *Male*
  - Unprivileged group: *Female*

Key Points: Protected Attributes

- A **protected attribute** is a characteristic that divides a population into groups, often based on sensitive factors such as *race*, *gender*, *caste*, or *religion*.
- These attributes are used to ensure that outcomes (e.g., loan approvals, job offers) achieve **parity** across groups, promoting fairness and reducing bias.
- Protected attributes are **context-specific**. For example:
  - In a U.S.-based dataset, **race** and **gender** might be key protected attributes.
  - In a different context, like educational access in India, **caste** might be more relevant.
- **Parity of Outcomes:**
  - When evaluating fairness, outcomes for these groups should be as equal as possible unless there is a justified reason for differences.
  - For instance: If an AI system is used to approve loans, the approval rate should not unfairly favor one gender or race over another.
- **Why Are Protected Attributes Important?**
  - AI systems can unintentionally perpetuate or amplify societal biases present in training data.
  - Identifying and focusing on protected attributes allows fairness interventions to ensure equitable outcomes for all groups.

8.1 Overview of Bias Metrics

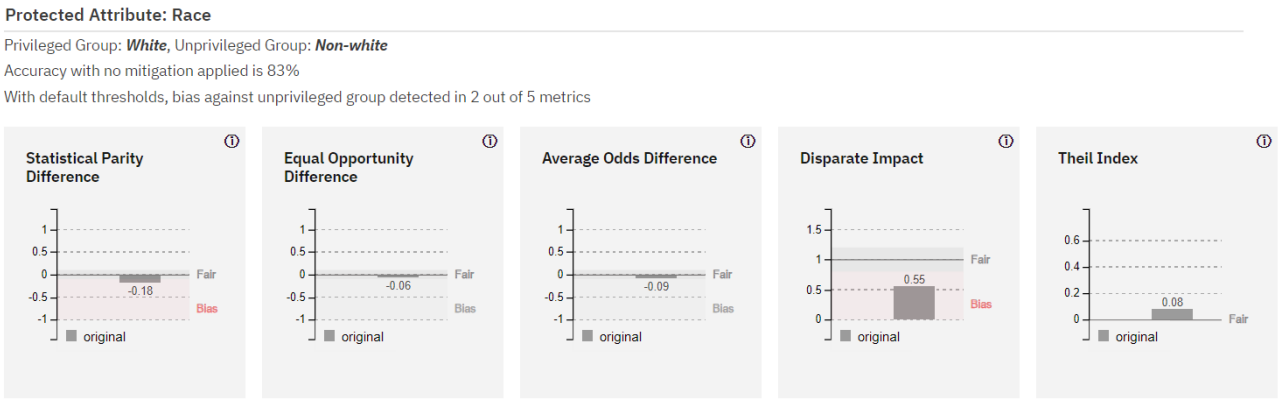

Figure 2: Bias metrics from the AI Fairness 360 Toolkit (Demo available on website)

8.1.1 Disparate Impact

- **Definition:** Measures the difference in outcomes between privileged and unprivileged groups. A high value indicates favoritism towards the privileged group.
- **Example:** In a job recruitment AI, if the system selects predominantly white applicants while rejecting qualified non-white candidates, it demonstrates a disparate impact. The model shows a score of **0.55** for this metric.

8.1.2 Average Odds Difference

- **Definition:** Assesses the disparity in predicted positive outcomes between groups. A negative score suggests the privileged group has a higher likelihood of favorable outcomes.
- **Example:** In a credit scoring system, if white applicants receive a 70% approval rate and non-white applicants only 61%, the model shows bias. The average odds difference is **-0.09**.

These metrics indicate the need for interventions to ensure equitable treatment and promote fairness in AI systems.

8.2 Bias Mitigation Algorithms

A variety of algorithms can be used to mitigate bias. The choice of which to use depends on whether we want to fix the data (**pre-process**), the classifier (**in-process**), or the predictions (**post-process**). More information can be found here.

The key bias mitigation algorithms are:

- **Reweighting**
  Weights the examples in each (group, label) combination differently to ensure fairness before classification.

- **Optimized Pre-Processing**
  Learns a probabilistic transformation that can modify the features and the labels in the training data.

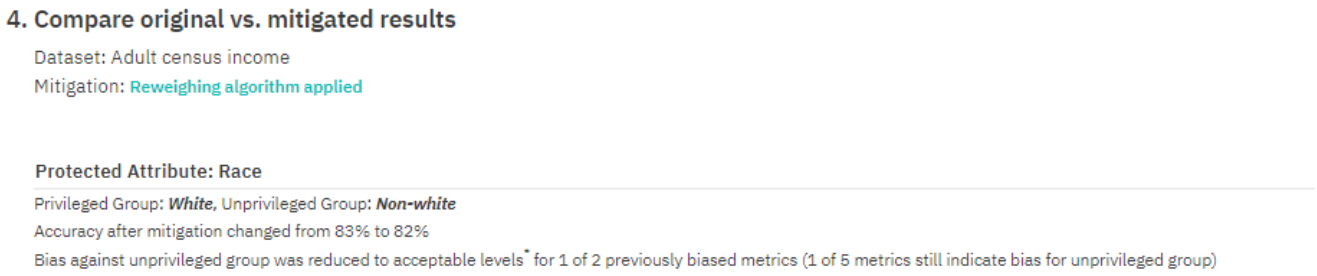

Figure 3: Results after using tool

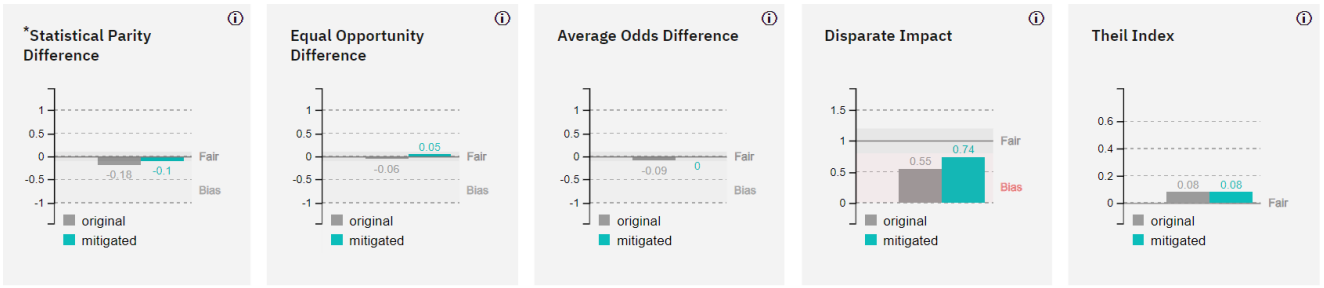

Figure 4: Comparison of bias before and after using tool (Demo available on website)

- **Adversarial Debiasing**
  Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions.

- **Reject Option Based Classification**
  Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

The source code for the AI Fairness 360 (AIF360) toolkit is publicly available on GitHub. The repository includes the implementation of various bias metrics and mitigation algorithms, along with detailed documentation and examples.

GitHub Repository: https://github.com/Trusted-AI/AIF360

8.3 Application of IBM AI Fairness 360 Toolkit

The **IBM AI Fairness 360 (AIF360)** toolkit was employed to assess and mitigate bias in machine learning models for two distinct datasets: the *Adult Census Income Dataset* and the *COMPAS Dataset*. These datasets are commonly used to demonstrate fairness challenges and solutions in AI systems. Below are the details:

- **Adult Census Income Dataset:**
  - **Dataset Link:** UCI Machine Learning Repository
  - **Use Case:** Predict whether an individual's income exceeds $50,000 based on demographic and work-related attributes.
  - **Protected Attributes:** *Race* and *Gender*.
  - **Key Observation:** Bias metrics revealed disparities in outcomes between privileged and unprivileged groups, which were mitigated using the *Reweighing* algorithm.

- **COMPAS Dataset:**
  - **Dataset Link:** ProPublica COMPAS Dataset
  - **Use Case:** Predict the likelihood of recidivism within two years for individuals.
  - **Protected Attribute:** *Race.*
  - **Key Observation:** The analysis uncovered racial disparities in predicted outcomes, which were addressed using the *Reweighing* algorithm to ensure equitable predictions.

GitHub Repository: The implementation details and results for these datasets can be found in the repository: https://github.com/neilhanda83/IBM-Fairness–Census-Income-COMPAS.

References

1. Floridi, L., et al. "AI Transparency: Ethical and Legal Perspectives." Journal of AI Ethics, 2018.
2. Getty Images v. Stability AI. Case No. 22-CV-8374, 2023.
3. Doe v. GitHub, Inc. Case No. 21-CV-3245, 2023.
4. Mittelstadt, B. "Principles for AI Transparency." Ethics in Information Technology, 2019.
5. Hugging Face. "Model Cards Documentation," 2024.
6. Google AI. "Fairness Indicators." 2023.
7. Trail of Bits. "PrivacyRaven Documentation." 2024.
8. High-Level Expert Group on AI. "Ethical Guidelines for Trustworthy AI," 2019.
9.GitHub. "State of the Octoverse 2023." GitHub, 2023.
10. Center for AI and Digital Policy. "Transparency in AI Development: 2023 Report." CAIDP, 2023.
11. Narayanan, A., et al. "Copyright and AI: A Study of Copilot's Code Generation." Stanford AI Ethics Lab, 2023
12. McKinsey & Company. "AI Development Practices Survey 2022." McKinsey, 2022.
13. Reuters. "Amazon Abandons AI Hiring Tool That Showed Bias Against Women." Reuters, 2018.
14. Institute for Ethical AI and Machine Learning. "State of AI Explainability," 2023.
15. Lundberg, S., et al. "SHAP: A Unified Approach to Interpreting Machine Learning Models." Journal of Machine Learning Research, 2020.
16. Financial Times. "Meta Fined €1.2 Billion for GDPR Violations." 2023.
17. Reuters. "Google Fined €100 Million by CNIL for Cookie Violations." 2021.
18. United Nations. "Personal Data Protection and Privacy Guidelines." 2021.
19. STAT News. "IBM Watson's AI Oncology Flaws Highlight Need for Rigorous Pre-Approval Auditing." 2020.
20. Tools For Trustworthy AI. OECD DIGITAL ECONOMY PAPERS June 2021
21. Examining the landscape of tools for trustworthy AI in the UK and the US. RAND Corporation May 2024