Ensuring Fairness with Transparent Auditing of Quantitative Bias in AI Systems

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—With the rapid advancement of AI, there is a growing trend to integrate AI into decision-making processes. However, AI systems may exhibit biases that lead decision makers to draw unfair conclusions. Notably, the COMPAS system used in the American justice system to evaluate recidivism was found to favor racial majority groups; specifically, it violates a fairness standard called equalized odds. Various measures have been proposed to assess AI fairness. We present a framework for auditing AI fairness, involving third-party auditors and AI system provider, and we have created a tool to facilitate systematic examination of AI systems. Inclusion of independent and trusted auditors is pivotal for objectivity and accountability; third parties are often further necessary for their specialized expertise in relevant domains. Auditors, equipped with our tool, can thoroughly review AI systems for bias and fairness violations. The tool is open-sourced and publicly available. Unlike traditional AI systems, we advocate a transparent white-box and statistics-based approach. It can be utilized by third party auditors, AI developers, or the general public for reference when judging the fairness criterion of AI systems.**

***Index Terms*—AI, fairness, auditing**

I. Introduction

In recent years, the accelerating pace of artificial intelligence (AI) technologies have revolutionized numerous fields. In healthcare, AI-driven diagnostic systems streamline disease identification, while in finance, automated trading algorithms analyze market trends to execute optimal trades swiftly. This remarkable progress has reached diverse industries, sprouting out various applications for different purposes.

One of its most profound impact is how AI has transformed decision-making processes across sectors. By harnessing vast amounts of data and employing advanced algorithms, AI empowers organizations to make more informed and strategic decisions. From assessing job applicants to determining school admissions, AI-driven insights offer efficient and analytical advantages.

AI system, however, may be biased. Factors such as inherent biases in original data sets or flaws in algorithm designs could contribute to bias within AI systems. When left unchecked, the ramifications of such biases extend far beyond mere inaccuracies; it can lead to devastating consequences, such as amplifying systemic injustices, perpetuating group discrimination, and exacerbating societal inequalities.

Identify applicable funding agency here. If none, delete this.

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a tool for predicting recidivism—the tendency of criminals to reoffend. It is used in the criminal justice system in multiple states in the United States. In 2016, it was discovered in an investigation by the journalists at ProPublica that the COMPAS system is, in fact, unfair towards minority and disadvantaged groups.

Cases such as COMPAS underscore the critical importance of rigorously examining the fairness of AI systems by third parties. Fairness is fundementally a subjective social construct, heavily influenced by cultural context and deeply rooted in historical inequalities. However, there have been development in research leveraging statistical metrics to quantify fairness that provide transparent and objective insights.

By employing such mathematically rigorous methodologies, our research can offer a more deeper look into the fairness of AI systems, enabling third parties to make more informed and robust judgment. For example, it would reveal that the COMPAS system violates a fairness measure called equalized odds.

In addition to equalized odds, there are a number of fairness measures for AI systems. Our framework aims to help third parties apply these metrics and navigate complex datasets with ease, pinpointing instances of bias and discrimination.

With our tool, a third party such as ProPublica could clearly and easily demonstrates that COMPAS violates equalized odds according to their datasets. Our platform empowers organizations to conduct thorough assessments of AI fairness and uncover discrepancies in predictive models.

This functionality not only facilitates sound judgment but also fosters objectivity and transparency regarding AI systems, ultimately bolster the claim of fairness and justice in AI-enabled decision-making.

Our tool is written in Python and is offered as a Python package. It supports common datasets format such as csv. It is open-sourced and publicly available for downloads.

Column 1	Column 2
disparate impact	$\frac{P[\hat{Y}=1\|S\neq 1]}{P[\hat{Y}=1\|S=1]} \geq 1-\epsilon$
demographic parity	$\|P[\hat{Y}=1\|S=1]-P[\hat{Y}=1\|S\neq 1]\| \leq \epsilon$
equalized odds: false positive	$\|P[\hat{Y}=1\|S=1,Y=0]-P[\hat{Y}=1\|S\neq 1,Y=0]\| \leq \epsilon$
equalized odds: true positive	$\|P[\hat{Y}=1\|S=1,Y=1]-P[\hat{Y}=1\|S\neq 1,Y=1]\| \leq \epsilon$
equal opportunity	$\|P[\hat{Y}=1\|S=1,Y=1]-P[\hat{Y}=1\|S\neq 1,Y=1]\| \leq \epsilon$
accuracy eqaulity	$\|P[Y=\hat{Y}\|S=1]-P[Y=\hat{Y}\|S\neq 1]\| \leq \epsilon$
predictive parity	$\|P[Y=1\|S=1,\hat{Y}=1]-P[Y=1\|S\neq 1,\hat{Y}=1]\| \leq \epsilon$
equal calibration	$\|P[Y=1\|S=1,\hat{V}=v]-P[Y=1\|S\neq 1,\hat{V}=v]\| \leq \epsilon$
conditional statistical parity	$\|P[\hat{Y}=1\|S=1,L=l]-P[\hat{Y}=1\|S\neq 1,L=l]\| \leq \epsilon$
predictive equality	$\|P[\hat{Y}=1\|S=1,Y=0]-P[\hat{Y}=1\|S\neq 1,Y=0]\| \leq \epsilon$
conditional use accuracy equaliyt: true positive	$\|P[Y=1\|S=1,\hat{Y}=1]-P[Y=1\|S\neq 1,\hat{Y}=1]\| \leq \epsilon$
conditional use accuracy equality: true negative	$\|P[Y=0\|S=1,\hat{Y}=0]-P[Y=0\|S\neq 1,\hat{Y}=0]\| \leq \epsilon$
mean difference	$\|E[\hat{Y}\|S=1]-E[\hat{Y}\|S\neq 1]\| \leq \epsilon$

TABLE I: Example Table