Ensuring Fairness with Transparent Auditing of Ethical 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 in the field of AI, there is a growing trend to integrate AI into the information flow of the decision making processes. However, the deployment of AI algorithms in this respect raises ethical considerations, particularly, regarding fairness. AI systems may exhibit biases, caused by various sources, that can potentially lead decision makers to derive unfair conclusions. For example, in 2016, journalists at ProPublica discovered that COMPAS, the algorithm system used by the American justice system to evaluate recidivism, displayed unfair treatment to different racial groups, favoring the majority groups while harming the minority groups; specifically, it violates a fairness measure called equalized odds. In recent years, researchers have devised a number of measures designed to evaluate the fairness of AI systems. We propose here a framework for auditing fairness of AI systems by including a third party auditor and an AI system provider and create a tool to facilitate the examining of such systems. It's pivotal for the integrity of the audits that the framework includes an independent and trusted arbiter for their objectivity and accountability; third party auditors are often further necessary for their specialized expertise in the relevant domains such as medicine and law. Auditors, equipped with our tool, can thoroughly review the AI systems for bias and fairness violations. The tool is open-sourced, easily accessible, 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, model makers, or the general public as a reference point when judging the fairness criterion of AI systems.**

***Index Terms*—keyword, keyword**

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