Description as given on AI Audit Challenge website https://hai.stanford.edu/policy/ai-audit-challenge

**What’s this about?**

* we lack the necessary tools to independently analyze and audit
* applied tools that can assess whether deployed AI systems exhibit bias or carry potential for discrimination.
* developing better tools for AI governance and in bridging the worlds of engineers and regulators, of technology and policy.

**What’s our objective?**

* to assess AI systems to determine whether they engage in prohibited discrimination.
* instead have chosen to prioritize impact through applied investigations, tools, and demonstrations.
* a particularly valuable area on which to concentrate is harmful bias in reference to protected categories
* How open source and deployed AI systems deal with protected characteristics and classes?
* Is it possible to identify indirect discrimination, through proxies and inferences?

Open-source modelssuch as GPT-NeoX-20B, BERT, GPT-J, YOLO, and PanGu-α.

Deployed systems in use by the public and private sector, such as COMPAS, GPT-3 and POL-INTEL.

* It is critical that policymakers and technology developers work in tandem.

**How is this project different from conferences like ACM FAccT?**

* In this case, our first target risk is bias and illegal discrimination.
* The outputs we are interested in are software, code, and/or tools that allow people to test publicly available algorithms and deployed models for illegal bias and discrimination, in ways that are useful and actionable for the people most likely to use such tools—namely, regulators, civil society, and journalists.

Evaluation criteria

* **Insights:** What did we learn using the tool?
* **Alignment:**How well anchored is the audit with legal and policy needs?
* **Impact:** How many people would benefit from the tool?
* **Ease of use:** Is the tool usable for our target audience?
* **Scalability:** Can the tool be used at scale and/or used in different contexts?
* **Replicability:** Can the results be replicated by other users using the same systems?
* **Documentation:** How well-explained are the findings?
* **Sustainability:** Is the tool financially and environmentally sustainable?

**Executive summary**

Problem – How to assess unfair differentiation and illegal discrimination in AI systems?  
At NGO Algorithm Audit, we observe a persistent gap between concrete AI practice and legal non-discrimination requirements. Whether EU, US, or other region’s non-discrimination directives are applied to AI, one runs into difficulties: Under what circumstances can proxy-variables for protected characteristics, such as ethnicity, can justifiably be used? How to deal with AI systems that differentiate on the basis of characteristics that do not significantly correlate with protected grounds? And: How can we arrive at well-founded quantitative thresholds for measuring the fairness of AI? Answers require normative choices to be made on a case-by-case basis that are subjected to local social, political, and environmental factors. We therefore see an urgent need for reviewing AI’s quantitative features against the qualitative requirements of law and ethics, in a public and case-based manner.

Solution – Fairness through discussion: A deliberative way forward  
NGO Algorithm Audit proposes a deliberative method to assess discriminatory AI. We believe a multi-disciplinary, well-informed and open debate is the best way forward to review AI statistics according to the normative requirements of non-discrimination law. Hence, our submission is rooted in both the quantitative and qualitative reasoning paradigm to assess fair AI:

* **Quantitative:** We present an open-source bias scan tool, based on the KMeans Hierarchical Bias Aware Clustering (HBAC) algorithm[[1]](#footnote-1), to discover potentially discriminated groups of similar users in AI systems. This bias scan tool does not require *a priori* information about existing disparities and sensitive attributes. So, the tool can identify proxy discrimination and new types of differentiation that evade non-discrimination law. For instance, differentiation based on browser type and house number (no protected characteristics), but which could still be perceived as unfair differentiation since it could reinforce social inequality[[2]](#footnote-2);
* **Qualitative:** We present a deliberative procedure to assess identified discrimination in AI models, for instance as detected by the KMeans HBAC bias scan tool. This includes AI factsheet that summarizes what model information is needed for a diverse and independent commission of multi-disciplinary experts to assess the case at hand. Our work is based on best practices from real-life case studies as conducted by NGO Algorithm Audit.

Our two-pronged quantitative-qualitative approach could be applied to all phases of the AI lifecycle, e.g., 1. conception phase, 2. pre-processing phase, 3. in-processing phase and 4. post-processing phase. To align with the scope of this challenge, we focus on the in-processing and post-processing phase of AI models.

In this report, four case studies are discussed that illustrate the need for qualitative interpretation of quantitative AI metrics. In the post-processing phase of the AI lifecycle, we apply our unsupervised bias scan tool to a BERT fake tweet detection tool and a RandomForest loan approval tool. In the in-processing phase of the AI lifecycle, we elaborate on a case study conducted by NGO Algorithm Audit concerning a discriminatory proxy-variable for ethnicity in a payment fraud prediction model as used by a large multinational e-commerce platform. We briefly describe the insight generated from those three case studies.

Unsupervised bias scan tool (quantitative method)

BERT fake news detection tool:

* Loan approval tool:

Audit commission (qualitative method)

* Type of SIM card proxy discrimination:



Figure 1 – NGO Algorithm Audit’s working method to audit AI

1. Misztal-Radecka, Indurkya, Information Processing and Management. Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems (2021) [↑](#footnote-ref-1)
2. # Zuiderveen Borgesius and Gerards, Colorado Technology Journal. Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence (2022).

   [↑](#footnote-ref-2)