**Abstract**  
Fairness cannot be automated. As AI is omnipresent in digital society, there is an urgent need to review AI systems with respect to the qualitative requirements of law and ethics. To facilitate this endeavour, we propose a scalable, easy to use and open-source bias scan tool. This bias scan tool identifies potentially discriminated groups of similar users in AI systems (including proxy and intersectional discrimination) and requires no *a priori* information about existing disparities and sensitive attributes. As demonstrated on a BERT-based Twitter disinformation detection model, the bias scan tool identifies statistically significant disinformation classification bias on the basis of verified user profiles, the number of mentions and hashtags used in tweets. On a widely cited credit data set, statistically significant loan approval bias is observed on the basis of applicants’ job status, telephone registration and the amount of credit requested. These observations do not establish *prima facie* algorithmic discrimination. Rather, the identified disparities serve as a starting point to assess potential discrimination according to the context-sensitive legal doctrine, i.e., assessment of the legitimacy, proportionality, and indispensability of observed disparities. For this qualitative assessment, we propose an expert-led deliberative method to review identified quantitative disparities. In this manner, we bridge the gap between the statistical and legal reasoning paradigm to shed light into fair AI.

**Executive summary**The identified problem, the proposed solution (a quantitative bias scan tool and qualitative deliberation method to assess discriminatory AI) and forthcoming results on real-life data sets are outlined here.

**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 jurisdiction’s non-discrimination directives are applied to AI, one runs into difficulties: Under what circumstances can proxy-variables for protected characteristics 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, but could reinforce social inequality? And: How to arrive at well-founded quantitative thresholds to measure 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 assessing quantitative AI metrics 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 quantitative tool and a qualitative deliberation method to assess discriminatory AI. We believe a multi-disciplinary, well-informed and open debate is the best way forward to form normative judgements about algorithmic bias. 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. The tool can identify proxy discrimination, intersectional 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 method to review identified quantitative disparities in AI models, as detected for instance by a bias scan tool. As a first step, model metrics are collected in a standardized and automated manner through AI factsheets. Subsequently, quantitative fairness metrics are reviewed by an independent and diverse audit commission. Such an audit commission, composed of diverse experts from a wide range of backgrounds, shares after deliberation a normative advice on a specific issue. All cases and corresponding advice are made publicly available, thereby fostering public knowledge building and a the public debate on techno-ethical issues.

Our two-pronged quantitative-qualitative approach to assess algorithmic bias could be applied to all phases of the AI lifecycle, e.g., 1. pre-processing phase, 2. in-processing phase and 3. post-processing phase (see Figure 1). Our deliberative model is applicable as well to the conception phase of AI, since the need for an algorithmic approach for a task at hand is not always self-evident. To align with the scope of this challenge, we focus however on bias occurring in the in-processing and post-processing phase of the AI lifecycle.

Text

Description automatically generated with low confidence

Figure 1 – Conceptual breakdown of the AI lifecycle in four phases

**Results – Building public knowledge by qualitative interpretation of quantitative AI metrics**

In this report, three real-life case studies are discussed that illustrate the need for qualitative interpretation of quantitative metrics to safeguard equal treatment by AI. In the post-processing phase of the AI lifecycle, we apply our unsupervised bias scan tool to a BERT fake tweet detection tool and to a Random Forest/XGBoost loan approval tool to examine disparate group fairness metrics. 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 (type of SIM card) in a payment fraud prediction model as used by a large multinational e-commerce platform. We summarize the key insights of these case studies below.

**Unsupervised bias scan tool (quantitative method)**

**BERT disinformation detection** **model** – Our bias scan tool detects statistically significant disparities in a BERT disinformation detection model[[3]](#footnote-3) trained on verified Twitter data[[4]](#footnote-4). A post-hoc explanation method identifies algorithmic bias for disinformation classification on the basis of verified user profiles, the number of mentions and hashtags used in tweets.

**Loan approval model** – Our bias scan tool detects statistically significant disparities for a Random Forest/XGBoost loan approval model3 trained on the German Credit data set[[5]](#footnote-5). A post-hoc explanation method identifies potential algorithmic bias for the classification of negative outcomes on the basis of job status, whether a telephone is registered at the customer’s name and the amount of credit requested.

These automatically identified disparities in AI models pave the way for human experts to conduct a qualitative assessment whether the observed bias can be regarded as justifiable differentiation or as illegal discrimination.

**Qualitative assessment of proxy discrimination by an audit commission**

**Problem statement**  
For an implemented afterpay fraud prediction algorithm at a multinational e-commerce platform, NGO Algorithm Audit conducted a case study on proxy discrimination. The input variable 'Type of SIM card’ could act as a proxy variable for ethnicity. Since in Europe, Lebara and Lyca SIM-cards are relatively more often used by people with a Euro-African migration background due to low intercontinental call charges. So, afterpay fraud prediction algorithms including ‘Type of SIM card’ as an input variable might develop an unlawful bias. The company’s procedure on restricting afterpay services could then be perceived as discriminatory. On the other hand, companies do not want to disregard relevant knowledge retrieved from historical data to deal with afterpay fraud.

**Advice**

NGO Algorithm Audit’s independent audit commission advises[[6]](#footnote-6) against using type of SIM card as an input variable in algorithmic models that predict afterpay default and that block afterpay services for specific customers. As it is likely that type of SIM card acts as a proxy-variable for sensitive demographic categories, the model would run an intolerable risk of disproportionally excluding vulnerable demographic groups from the payment service. Absent reliable data that demonstrates otherwise, the ethical risk of including the SIM card variable outweighs potential benefits. The commission advises to consider a variety of alternatives in dealing with payment defaults.

The used methodology, outcomes and implications of the above case studies are discussed in more detail in this report.

**Conclusion**Quantitative methods, such as unsupervised bias scans tools, are helpful to discover potentially discriminated groups of similar users in AI systems in a scalable manner. The automated identification of disparity in AI models allows human experts to assess observed biases in a qualitative manner, subject to political, social and environmental traits. This two-pronged approach bridges the gap between the qualitative requirements of law and ethics, and the quantitative nature of AI. In making normative advice on identified ethical issues publicly available, over time a repository of “jurisprudence” emerges; from which data scientists and public authorities can distill best practices to build fairer AI.

What is NGO Algorithm Audit?

Algorithm Audit is an NGO that builds and shares public knowledge about ethical algorithms. Its main activity is to form independent audit commissions that give ethical advice on concrete algorithmic methods as used in the private and public sector. Additionally, in bringing together international experts from a range of disciplines and professional backgrounds, Algorithm Audit serves as a bottom-up European knowledge and advocacy platform for ethical automated decision-making.

NGO Algorithm Audit works together with partners under explicit conditions to avoid ethics washing. For instance, to maintain our independence, we depend only on public funding.

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)
3. https://github.com/NGO-Algorithm-Audit/AI\_Audit\_Challenge [↑](#footnote-ref-3)
4. Jing Ma, Wei Gao, and Kam-Fai Wong. 2017.Detect rumors in microblog posts using propagation structure via kernel learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). volume 1, pages 708-717. [↑](#footnote-ref-4)
5. German Credit Data from the UCI Repository of Machine Learning Databases [↑](#footnote-ref-5)
6. ### Type of SIM card as a predictor variable to detect payment fraud, NGO Algorithm Audit (2022).

   [↑](#footnote-ref-6)