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 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 quantitative metrics about AI 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 against 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. 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. This includes an AI factsheet to collect model metrics in a standardized and automated manner. Assessing discrimination is however a normative and context dependent exercise. We therefore discuss a deliberative method used by NGO Algorithm Audit to review differential treatment of AI by an independent and diverse commissions of multi-disciplinary experts on a case-to-case basis.

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. 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 the in-processing and post-processing phase of the AI lifecycle.

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 guarantee equal treatment by AI models. 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 XGBoost loan approval tool to examine quantitative disparities between groups. For 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 take-aways of the 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. 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 the tweet.

Loan approval model – Our bias scan tool detects statistically significant disparities for a XGBoost loan approval model3 trained on the German Credit data set. A post-hoc explanation method identifies potential algorithmic bias for fake news classification on the basis of the amount of job status, registered telephone and credit amount 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 unfair differentiation or as illegal discrimination.

Qualitative assessment of discrimination through audit commission

Based on 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 demographic groups. If this is the case, afterpay fraud prediction algorithms might develop an ethnic, religious or demographic bias. The company’s procedure on restricting afterpay services could then be perceived as discrimination. On the other hand, companies do not want to disregard relevant knowledge retrieved from historical data to deal with afterpay fraud.

NGO Algorithm Audit’s independent audit commission advises 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.

We elaborate in more detail on the used methodology, outcomes and implications of the above case studies in this report.

What is NGO Algorithm Audit?

Algorithm Audit is a young NGO that builds and shares public knowledge about ethical algorithms. Our 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.

Note: We work together with partners under explicit conditions to avoid ethics washing. For instance, to maintain our independence, we do not depend on funding from organizations that are subject to our evaluations.

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)