**Table of contents**

**Summary** p.1

1. **Problem: The persistent gap between legal non-discrimination** p.2 **requirements and AI practice**EU, US and other jurisdictions’ non-discrimination law   
   Fair AI in practice
2. **Solution: A quantitative bias scan tool and qualitative deliberation** p.6Quantitative: Hierarchical Bias-Aware Clustering (HBAC)   
   Qualitative: NGO Algorithm Audit’s deliberative approach
3. **Results: Building public knowledge** p.7Case I – BERT disinformation detection model   
   Case II – Loan approval model   
   Case III – Proxy discrimination in in fraud detection algorithm

**Conclusion** p.9

**This submission is endorsed by:**

* A
* B
* C

Endorsement of this submission

This challenge is a collective effort of experts from a range of disciplines and professional backgrounds. Expertise from academia, industry and investigative journalism is brought together to deploy the bias scan tool and associated deliberative approach to define fair AI. This submission is an amalgam of statistics, philosophy, law, sociology, public policy and AI governance.

**Summary**  
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 time-consuming endeavour, we propose a scalable, easy to use, and open-source bias scan tool to identify 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, and is therefore able to detect potential proxy discrimination, intersectional discrimination and new types of differentiation that evade non-discrimination law.

As demonstrated on a BERT-based Twitter disinformation detection model, the bias scan tool identifies statistically significant disinformation classification bias against users with a verified profile and above average number of mentions and hashtags used in tweets. On the widely cited German 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 prohibited *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 of the aim pursued and whether the means of achieving that aim are appropriate and necessary. For this qualitative assessment, we propose an expert-led deliberative method to review identified quantitative disparities against the requirements of non-discrimination law and ethics. In our two-pronged quantitative-qualitative solution, scalable statistical methods work in tandem with the normative capabilities of human subject matter experts to define fair AI on a case-by-case basis.

What is NGO Algorithm Audit?

NGO Algorithm Audit 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, e.g., to maintain our independence, we depend only on public funding.

Problem: The persistent gap between legal non-discrimination requirements and AI practice

Problem statement  
At NGO Algorithm Audit, we observe a persistent gap between concrete AI practice and legal non-discrimination requirements. Whether international, EU or American 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.

## EU, US and other jurisdictions’ non-discrimination law and data protection directives

To ground our problem statement and proposed solution, we discuss legal challenges to review AI systems. We specifically focus on the conflicting requirements as formulated in non-discrimination and data protection law. Across international[[1]](#footnote-1), EU[[2]](#footnote-2) and American[[3]](#footnote-3) law, three challenges are identified that influence fair AI assessments:

1. **Protected grounds** – Equal protection laws prohibit agents from acting with “discriminatory purpose”[[4]](#footnote-4) based on a pre-defined list of protected characteristics. Such a list itemizes factors which are deemed socially unacceptable by society to differentiate upon, such as race, gender, nationality or religion. [Under current data protection directives, such as the EU’s General Data Protection Regulation (GDPR) and the mixture of US Data Privacy Laws];
2. **Direct and indirect discrimination** – Legal frameworks conceptually distinguish disparate treatment on protected groups (direct discrimination) and disparate impact (indirect discrimination) of protected groups. [Proxy and correlation challenge];
3. **Possible justification** – Disparate impact laws do not prohibit all group differences, the law only prohibits unjustified disparities. Put differently, either direct or indirect discrimination will be lawful if a legitimate aim objectively justifies disparities and the means of achieving that aim are considered appropriate and necessary [Deliberative method, normative, context-dependent, qualitative, case by case].

Hence, in our solution we aim to provide a solution to the following pitfalls to assess discriminatory AI:

1. indirectly harms people with protected characteristics, such as ethnicity or gender;
2. differentiation that does not harm people with protected characteristics, but that is still unfair.[[5]](#footnote-5)

## Fair AI in practice

[Something on current state of AI in public and private sector]   
  
Bias detection difficult when data on protected attributes is not available. Normative, human-led endavour to assess fair AI.

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.

Graphical user interface, text, timeline

Description automatically generated

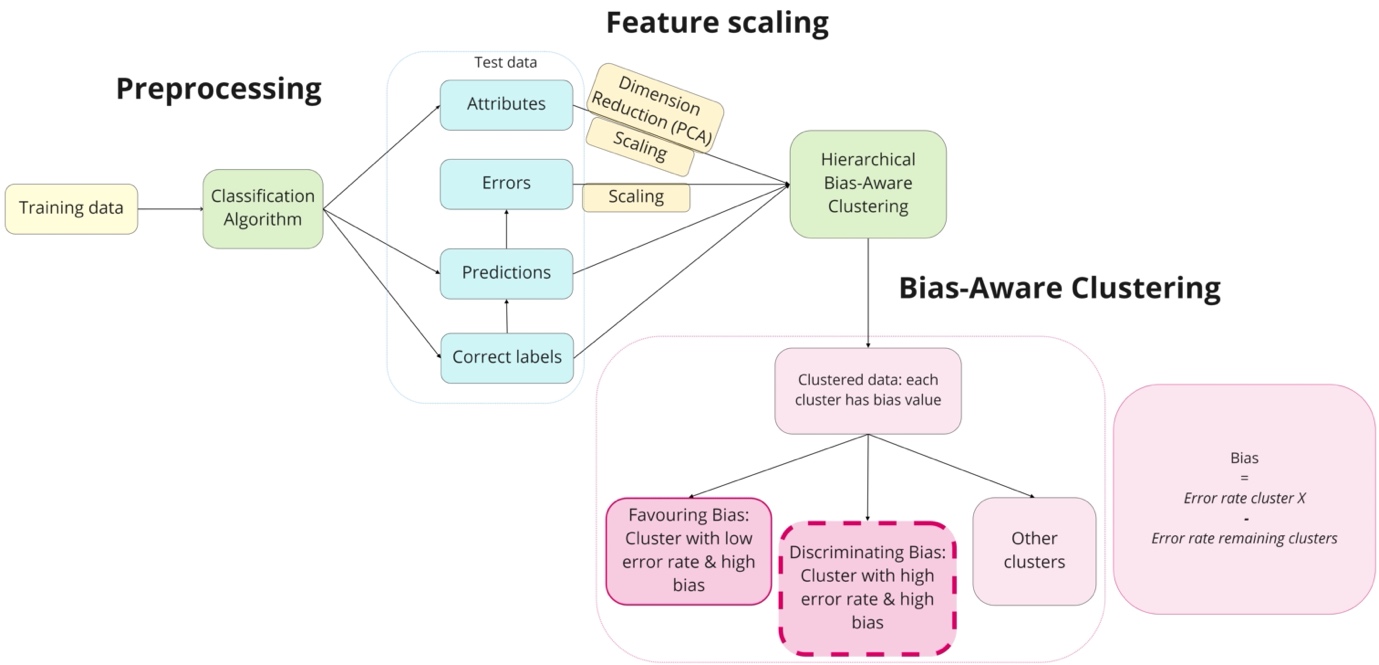
Figure 1 – Conceptual breakdown of the AI lifecycle in four phases

Solution – Fair AI through discussion: A deliberative way forward  
NGO Algorithm Audit proposes a quantitative bias scan tool and a qualitative deliberation method to assess Fair 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 – Bias scan tool

We present an open-source bias scan tool, based on the KMeans Hierarchical Bias Aware Clustering (HBAC) algorithm[[6]](#footnote-6), 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[[7]](#footnote-7).

#### Hierarchical Bias-Aware Clustering (HBAC)



## Qualitative – NGO Algorithm Audit’s deliberative approach

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.

# 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.

## Case I – BERT disinformation detection tool

Classifier model, data etc.

Our bias scan tool detects statistically significant disparities in a BERT disinformation detection model[[8]](#footnote-8) trained on verified Twitter data[[9]](#footnote-9). 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.

## Case II – Loan approval model

German Credit data set

Our bias scan tool detects statistically significant disparities for a Random Forest/XGBoost loan approval model3 trained on the German Credit data set[[10]](#footnote-10). 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.

## Case III – Proxy discrimination and a fraud detection model

**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[[11]](#footnote-11) 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.

1. The International Covenant on Civil and Political Rights, the International Covenant on Economic Social and Cultural Rights, and the International Covenant on the Elimination of All Forms of Racial Discrimination. [↑](#footnote-ref-1)
2. In the European Union (EU), the European Convention of Human Rights (ECHR) serves as the legal fundament against discrimination. Additional EU directives (2000/43/EC, 2000/78/EC, 2004/113/EC, and 2006/54/EC) provide context-specific protection, e.g., persons with disabilities, labor law, and good and services. [↑](#footnote-ref-2)
3. American Labor law, U.S. Constitution’s Fourteenth Amendment [↑](#footnote-ref-3)
4. See for instance, Washington v. Davis (1976). 426 U.S. 229 and U.S. Equal Employment Opportunity Commission https://tinyurl.com/29f7kj5b [↑](#footnote-ref-4)
5. 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-5)
6. Misztal-Radecka, Indurkya, Information Processing and Management. Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems (2021). [↑](#footnote-ref-6)
7. # 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-7)
8. https://github.com/NGO-Algorithm-Audit/AI\_Audit\_Challenge [↑](#footnote-ref-8)
9. 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-9)
10. German Credit Data from the UCI Repository of Machine Learning Databases [↑](#footnote-ref-10)
11. ### Type of SIM card as a predictor variable to detect payment fraud, NGO Algorithm Audit (2022).

    [↑](#footnote-ref-11)