Algorithm Audit

Bias scan report

Table of contents

Where to use the bias scan?

☁️ Available as AWS cloud application
https://www.algorithmaudit.eu/bias_scan/

💻 Github repository
https://github.com/NGO-Algorithm-Audit/AI_Audit_Challenge

Executive summary

Artificial intelligence (AI) is increasingly used to automate or support policy decisions that affects individuals and groups. It is imperative that AI adheres to the legal and ethical requirements that apply to such policy decisions. In particular, policy decisions should not be systematically discriminatory (direct or indirect) with respect to protected attributes such as gender, sex, ethnicity or race.

To achieve this, we propose a scalable, model-agnostic, and open-source bias scan tool to identify potentially discriminated groups of similar users in binary AI classifiers. This bias scan tool does not require *a priori* information about existing disparities and protected attributes, and is therefore able to detect possible proxy discrimination, intersectional discrimination and other types of differentiation that evade non-discrimination law. The tool is available as a web application, available on the website of NGO Algorithm Audit, such that it can be used by a wide public.

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, above average sentiment score and below average number of URLs used in their tweets. For a XGBoost loan approval model on the German Credit data set, statistically significant approval bias is observed on the basis of real estate ownership, negative account balance, unskilled job status and loans used to buy a new car or radio/television.

These observations do not establish prohibited *prima facie* 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-oriented deliberative method. Which allows policy makers, journalist, data subjects and other stakeholders to publicly 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.

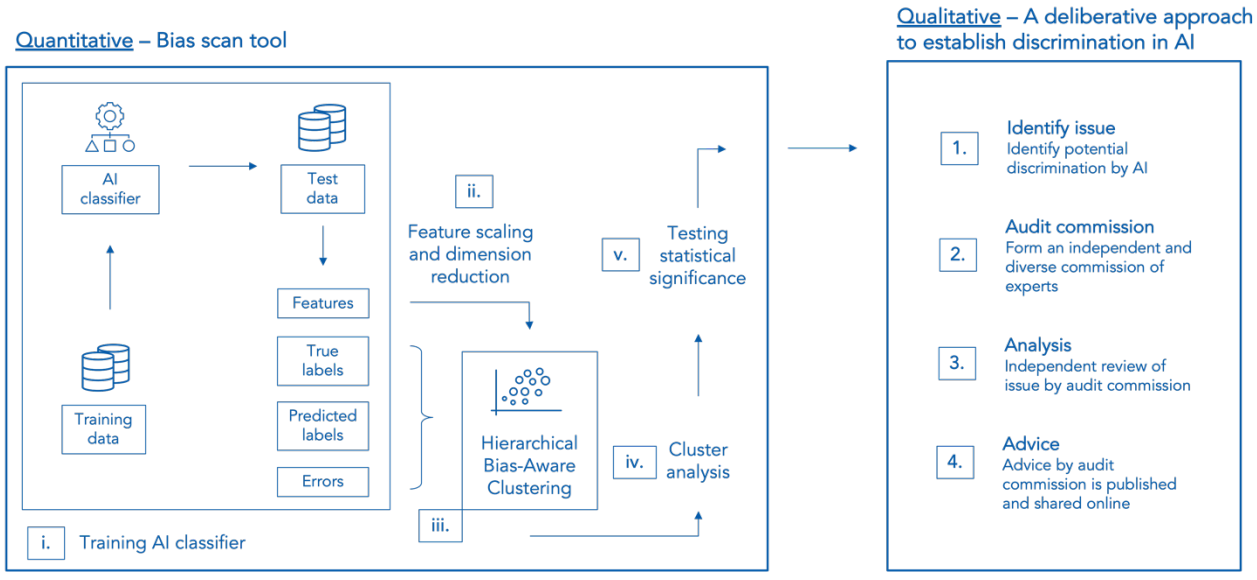

Figure 1 – Overview of quantitative-qualitative approach to assess fair AI with help of statistical bias scan tool and deliberative expert-led approach

Scope of the bias scan

The bias scan tool now works for binary AI classifiers, as used by public and private organizations on a day-to-day basis, such as profiling and ranking. We specifically focus on:

- Algorithms that indirectly harm people on the basis of protected characteristics, such as ethnicity or gender (indirect (proxy) discrimination also known as disparate impact);
- Algorithmic differentiation that does not harm people with protected characteristics, such as differentiation on the basis of web browser or house number, but such differentiation could still considered to be unfair, for instance as it reinforces social inequality.

Contributions to this bias scan

This bias scan tool is a collective effort of experts from a range of disciplines and professional backgrounds to assess the normative concept of 'fair AI'. Expertise from academia, industry, policy making and journalism resulted in this quantitative tool and associated deliberative approach. This bias scan serves as a starting point to demystify AI, i.e., to debate normative data modelling choices in an open and public manner. Stakeholders across society endorse this approach and support NGO Algorithm Audit to build and share public knowledge about ethical algorithms.

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. For more questions see our FAQ at https://www.algorithmaudit.eu/faq/.

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

1.1 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 laws 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 that involves policy makers, journalist, data subjects and other stakeholders.

1.2 Challenges arising from non-discrimination law and data protection legislation

To ground our problem statement and proposed solution, we discuss some legal challenges to assess bias in AI systems. We specifically focus on the requirements as formulated in non-discrimination law and data protection legislation. Across international[1], EU[2] and American[3] law, we discuss three challenges that influence the assessment of fair AI:

I. **Data availability on protected grounds** – Equal treatment laws prohibit agents from acting with "discriminatory purpose"[4] based on a pre-defined list of protected attributes. Protected attributes are deemed socially unacceptable by society to differentiate upon, such as race, gender, nationality, or religion. Current data protection directives, such as the European Union's (EU) General Data Protection Regulation (GDPR) and the mixture of US Data Privacy Laws[5], prohibit therefore often the use of protected attributes for general data processing purposes. In the EU, data on 'racial or ethnic origin' can only be collected for official statistical research. For instance, to assess potentially bias on the basis of race protected data might be available to test facial recognition software. In this submission, however, we focus on common AI applications deployed by public and private organizations, such as profiling and ranking, in which data on protected attributes is often absent on the basis of data protection laws. Hence, the first legal challenge we aim to address is that almost no

[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.
[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.
[3] American Labor law, U.S. Constitution's Fourteenth Amendment
[4] See for instance, Washington v. Davis (1976). 426 U.S. 229 and the U.S. Equal Employment Opportunity Commission https://tinyurl.com/29f7kj5b
[5] Hundreds of laws enacted at the federal and state levels serve to protect the personal data of U.S. residents.

organization is able to statistically measure algorithmic inequality with group fairness metrics absent data on protected attributes due to the requirements of equal treatment legislation to store and process such data;

II. **The proxy and correlation challenge** – Legal frameworks conceptually distinguish disparate treatment of protected groups (direct discrimination) and disparate impact on protected groups (indirect discrimination). As the use of protected attributes for AI applications is often prohibited on the basis of data protection laws (primarily the case in the EU), unequal algorithmic treatment involves predominantly disparate impact on protected groups through *proxy discrimination*. Proxies are apparently neutral characteristics, such as ZIP code, type of SIM card and literacy rate, that form groups that closely mirror protected groups[6]. Absent data on protected attributes (as discussed in **I.**) proxy discrimination in algorithmic systems can often not be measured with group fairness metrics. An urgent question is therefore: What personal characteristics can be considered as a proxy variable for protected attributes, and which of those variables should be excluded to prevent indirect discriminatory bias?

III. **Other types of discrimination and differentiation that evade non-discrimination law** – Scholarship has argued that granular analysis of personal and behavioral data entails heightened risk of intersectional discrimination[7] and new forms of differentiation that evade non-discrimination law[8]. Intersectional discrimination refers to a disadvantage based on two or more protected characteristics considered together, for example being a "black woman", a type of discrimination that the European Court of Justice has so far failed to adequately recognize[9]. New forms of differentiation refer to algorithms that differentiate upon new categories of people based on seemingly innocuous characteristics (*ad hoc bias)*, such as web browser preference or apartment number, or more complicated categories combining many data points. Such new types of differentiation could evade non-discrimination law, as these variables are no protected characteristics, but such differentiation could still be unfair, for instance if it reinforces social inequality.

IV. **Possible justification** – Non-discrimination and equal treatment laws do not prohibit all forms of disparate group differences; the law only prohibits unjustified disparities. Depending on the specific jurisdiction, direct and indirect discrimination are characterized by a (semi-)closed or (semi-)open list of protected grounds[10], possibilities for exemption and justification. Direct discrimination involves a narrow pool of justification available in direct discrimination cases. As opposed to an open-ended objective justification test applicable in indirect discrimination. Put differently, either

---

[6] Note that in some cases single proxy variables are closely related to a protected ground, from which the questions arises whether such cases should be classified as direct or indirect discrimination. Details on such cases are beyond the scope of this submission. Although, our proposed expert-oriented deliberative method to review disparate impact against the requirements of non-discrimination law provides a possible solution to deal with such questions.

[7] Algorithmic discrimination in Europe, Gerards and Xenidis (2021).

[8] 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).

[9] Judgment of 24 November 2016, David L. Parris v. Trinity College Dublin and Others, C-443/15, EU:C:2016:897.

[10] Algorithmic discrimination questions the boundaries of the exhaustive list of protected ground as defined, for instance, in Article 19 TFEU and sheds new light on the role and place of the non-exhaustive list of protected ground to be found in Article 21 of the EU Charter of Fundamental Rights. This debate is however beyond the scope of this work.

direct or indirect bias will be lawful if a legitimate aim objectively justifies disparities and the means of achieving that aim are considered appropriate and necessary. Assessment of these legal requirements is a qualitative task depending on the specific social, institutional and technical context of the case at hand.

In section **2.** *Solution: Fair AI through discussion – A deliberative way forward*, we present a quantitative and qualitative approach to mitigate the above four legal challenges to assess fair AI. We narrow down the scope of this submission to widely used AI applications, such as ML-based profiling and ranking, that are applied at a large scale to citizens and consumers on a day-to-day basis. For those AI applications, we focus on two categories of risks related to algorithmic decision-making: indirect (proxy) discrimination (**II.**) and unfair differentiation (**III.**) when data on protected attributes is not available (**I.**). Lastly, we focus on establishing prohibited algorithmic discrimination through the qualitative assessment of normative requirements as formulated in non-discrimination law, e.g., legitimacy test (**IV.**).

1.3 Bias along the AI lifecycle

Not only from a legal perspective, as well from a technical perspective fair AI can be assessed. We break down conceptually the AI lifecycle in four phases to describe how bias in AI might occur (see Figure 1).

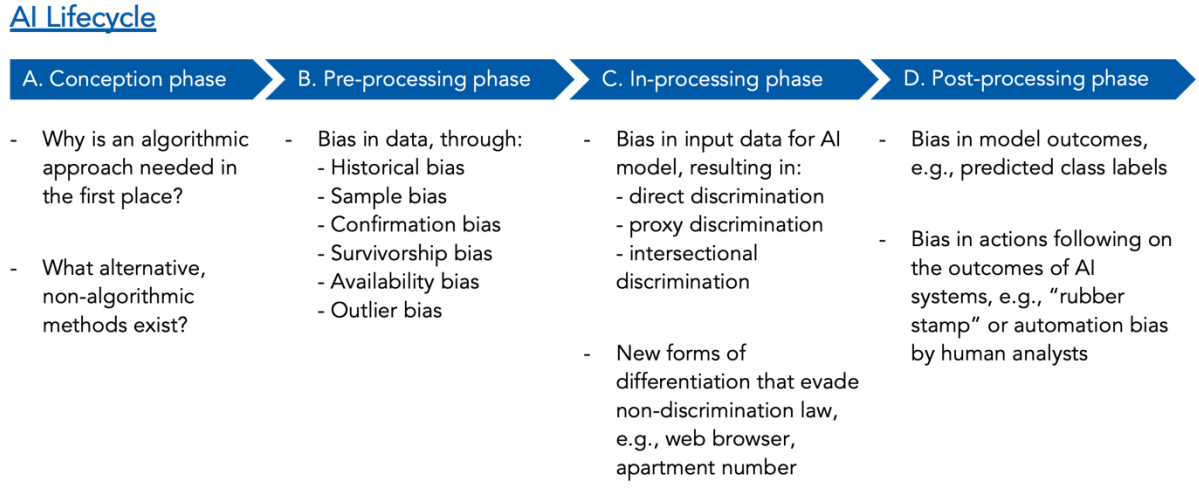

Figure 2 – Conceptual breakdown of the AI lifecycle in four phases

A. Conception phase – In assessing discriminatory and ethical risk pertaining to AI systems, a good practice is to start with the question in the conception phase of the AI lifecycle: Why is an algorithmic approach needed in the first place for the task at hand? For new deep learning methods, such as natural language understanding or computer vision algorithm based on foundation models, there might be a clear rationale for innovation purposes. For risk profiling methods, in the context of fraud protection in the public or private sector, such a rationale for the application of AI methods might not be self-evident.

B. Pre-processing phase – An immediate apparent ethical risk in the pre-processing phase of the AI lifecycle concerns biased data from which, for instance, selection criteria for risk profiles are distilled. *Historical bias* might stem from socio-cultural historical

inequalities which are mirrored in digital data collection processes. *Sample bias* refers to over- or underrepresentation of certain groups compared to the total population. *Confirmation bias* is the tendency to favor information, for instance in assigning class labels to build a supervised learning data set, that confirms prior beliefs or values. For a more complete lists of biases relevant for AI systems we refer to scientific literature[11], Google's Machine Learning Glossary on fairness[12] and Wikipedia's catalog of cognitive biases[13].

C. **In-processing phase** – For the in-processing phase, we focus on input data fed to an AI model. As discussed in more detail in legal challenges **II.** and **III.**, in this phase the issues of indirect (proxy) discrimination and unfair differentiation emerge.

D. **Post-processing phase** – In the last phase of the AI lifecycle, bias can occur in the AI model outcomes, e.g., predicted class labels, for instance due to impaired learning objectives of the AI system. A second risk is related to actions following up the outcome of the model. For instance, the "rubber stamp" or automation bias, i.e., human analysts that tend to believe the outcome of AI systems or follow their advice disproportionally often.

Although all biases occurring along the AI lifecycle are important to be detected and mitigated, we leave the pre-processing phase – specifically the assessment of the data quality – outside the scope of this submission. Rather, we focus on the qualitative assessment of AI's *raison d'être* (**A.** Conception phase) and on a *post-hoc* qualitative assessment of legal and ethical risks pertaining to the inclusion of certain data variables in AI models (**C.** In-processing phase and **D.** Post-processing phase). To identify what aspects of AI systems should be assessed qualitatively, we present a quantitative bias scan tool in the next section.

[11] Greenland, Sander. "Multiple-bias modelling for analysis of observational data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168.2 (2005): 267-306.
[12] https://developers.google.com/machine-learning/glossary/fairness#e
[13] https://en.wikipedia.org/wiki/List_of_cognitive_biases

2. Solution: Identifying potential discrimination in the sheer volume of AI data

NGO Algorithm Audit proposes a quantitative bias scan tool and a deliberative qualitative approach to address the challenges as described in section **1.** *Problem: The persistent gap between legal non-discrimination requirements and AI practice*. A quantitative approach is indispensable for monitoring AI-informed policy decisions because of the sheer data volume and technical complexity. A qualitative assessment by subject-matter experts is ultimately the only way to establish unlawful or unethical discrimination. Until jurisprudence on unwarranted algorithmic discrimination is available, we believe a multi-disciplinary, well-informed and open debate is the best way forward to form normative judgements about algorithmic bias. Our submission is therefore rooted in both the quantitative and qualitative reasoning paradigm to assess fair AI, which are both discussed below.

2.1 Quantitative – Bias scan tool

We present an open-source, model-agnostic bias scan tool, based on k-means Hierarchical Bias Aware Clustering (HBAC)[14], to discover potentially discriminated groups of similar users in AI systems. In contrast to many other fairness methods that detect bias, this bias scan tool uses unsupervised machine learning and thus does not require *a priori* information about existing disparities and protected attributes. This approach thus offers a solution to legal challenge **I.** (data on protected characteristics is often not available). In addition, by identifying similar groups of potentially discriminated users, the bias scan tool is (in theory) able to identify proxy discrimination, intersectional discrimination and new types of differentiation that evade non-discrimination law, thereby overcoming legal challenge **II.** (the proxy and correlation challenge) and legal challenge **III**. (other types of discrimination) In this report the HBAC bias scan tool is applied on homemade AI applications to assess its ability to detect discriminatory bias, i.e., the post-processing phase of the AI lifecycle. The outcome of the bias scans is discussed in section **3.** *Defining fair AI through the qualitative interpretation of quantitative metrics*. First, the steps involved in the bias scan tool are discussed at a conceptual level.

2.1.1 Bias scan: Unsupervised k-means Hierarchical Bias-Aware Clustering (HBAC)

The bias scan tool aims to identify groups for which a classification algorithm is systematically underperforming. Based on the k-means clustering algorithm, the HBAC method automatically splits the data points into clusters on the basis of their features. The objective of clustering algorithms is to maximise the *within-cluster similarity* and the *between-cluster dissimilarity*. Clusters are then compared with respect to classification errors in predicted outcomes. If there is a statistically significant difference in classification errors between clusters, then we have detected negative, potentially discriminatory, bias.

Using the HBAC bias scan tool proceeds in several steps: training an AI classifier, pre-processing the classifier predictions, identifying clusters, and analysing cluster disparities in

[14] Misztal-Radecka, Indurkya, Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems, *Information Processing and Management* (2021).

classification performance (see the quantitative part of Figure 2). More details on the steps of the HBAC pipeline are provided below.

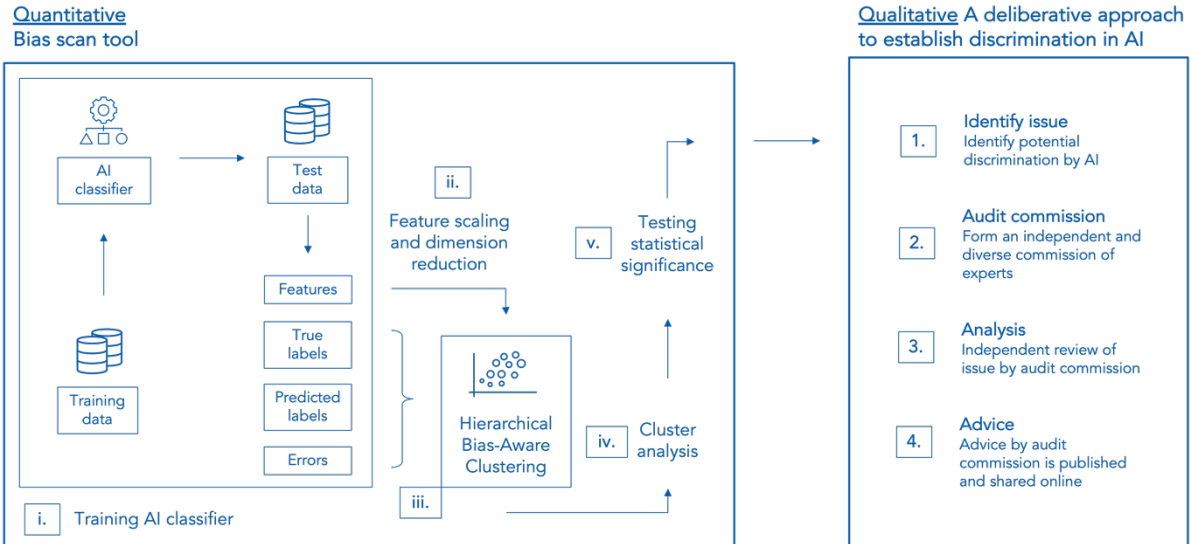

Figure 3 – Overview of proposed solution. A quantitative bias scan tool is combined with a qualitative deliberative approach to establish discriminatory AI.

i. **Training an AI classifier** – An AI classifier is trained on a data set to predict the labels of the observations using a user-specified learning objective. The classifier's predictions on the test data set serve as input data for the bias scan.

ii. **Feature scaling and dimension reduction** – Features in the test data set need to be scaled before being fed as inputs to the clustering algorithm. We scale all variables to have a standard deviation of one. Then all variables are assigned equal importance in the clustering algorithm when systematic under- or over performance of the classifier is calculated. A difficulty with clustering analysis using more than two variables is that the resulting clusters cannot be easily visualised to humans. We thus reduce the dimension of the input variables to a 2-dimensional representation and then show the clusters in a scatterplot on the two resulting dimension (see section **3.** *Defining fair AI through the qualitative interpretation of quantitative metrics*). We use Principal Components Analysis (PCA) for the dimension reduction. In short, PCA finds a lower-dimensional representation of several variables that explain a given fraction of the variance of the observations[15].

iii. **Hierarchical Bias-Aware Clustering** – The input data for HBAC consists of the scaled variables, predicted labels by the classifier, ground truth labels and classification errors. Classification errors are included in the clustering algorithm detect clusters for which the classifier makes worse predictions. The clustering is hierarchical in the sense that a nested relationship is constructed among the observations to form the groups. Recall that in clustering, observations in a data set are grouped into clusters such that observations in the same cluster are similar in terms of their features (high intra-cluster

[15] Technically, PCA finds a set of linear combinations (i.e., principal components) of the original variables. The principal components are constructed such that the first one maximises the variance explained in the observations using its combination of variables, the second one maximises the remaining variance explained using its combination, etc. Additional principal components are constructed analogously until there is no variance in the original variables left to be explained.

similarity), while observations in different clusters are dis-similar in terms of their features (low inter-cluster similarity). Forming clusters thus requires a metric for measuring similarity between observations. There exists no universally best metric, so the metric has to be chosen according to the problem context. Based on indirect evaluations of HBAC with various similarity metrics and clustering algorithms[16], we chose '1-Accuracy' as a similarity metric and k-means clustering for implementing the HBAC bias scan tool. Having specified a similarity metric and clustering method, the algorithm is run until a convergence criterion is met, e.g., run for a pre-specified number of iterations and then stops. Subsequently, the algorithm returns the identified clusters and their corresponding (averaged) classification errors. The performance of HBAC can be tailored via hyperparameter tuning. Details of our approach is provided in section **3.** *Defining fair AI through the qualitative interpretation of quantitative metrics.*

iv. **Analysis of identified clusters** – We are interested in the identified clusters with highest negative bias, i.e., cluster of which the AI classifier predicts more negative than positive labels. To analyze these clusters in a meaningful way, we first revert the scaling of the data features. To compare clusters, we then calculate for each feature the difference in its value in a discriminated cluster and all other clusters. One way to calculate this is by taking the average value of a feature in a discriminated cluster less its average value across all other clusters. Next, we test if these between-cluster differences are statistically significant.

v. **Testing statistical significance** – Since clustering is an unsupervised learning technique it is difficult to assess the reliability of identified disparities between clusters. How do we know whether the clusters represent true subgroups in the data, or whether they are simply a result of noise? To explore this, we investigate if the differences in the average values of features in a discriminated cluster and other clusters are statistically significant. Specifically, we use Welch's two-samples t-test for unequal variances as the variance of observations may differ between the compared clusters. The resulting p-values per cluster indicates if there is more evidence for the cluster than one would expect due to chance. However, there is no consensus on a single best approach to find features that can be sources of discrimination and thus require closer inspection[17]. Thus, the ability to detect bias in AI classifiers using quantitative methods stops here. To make progress, the identified disparities can serve as a starting point to assess potential discrimination according to the context-sensitive legal doctrine (see section **1.** *Problem: The persistent gap between legal non-discrimination requirements and AI practice*). For this qualitative assessment, we propose an expert-oriented deliberative method.

An implementation of the HBAC bias scan tool can be found in the Github repository created for this submission[18]. In the following, we discuss some limitations of the bias scan tool.

[16] The indirect evaluation properties include scalability, robustness, interpretability, parameter tuning complexity/sensitivity. See Muhammad, Auditing Algorithmic Fairness with Unsupervised Bias Discovery (2021) https://www.youtube.com/watch?v=g5I9MjxpWfk

[17] More details on unsupervised clustering methods can be found in Hastie et al. (2009).

[18] https://github.com/NGO-Algorithm-Audit/Bias_scan

2.1.2 Limitations of clustering and bias scan tools

Clustering can be a very useful and valid statistical tool if used properly. Some of the limitations associated with clustering are outlined below.

- HBAC finds statistically significant differences in the means of feature values between clusters. However, as the true clusters are unknown it is impossible to determine if the tested differences correspond to real patterns. We therefore recommend performing a sensitivity analysis by running the clustering with different parameter choices (e.g., similarity metrics, clustering algorithms and data samples) to check that the results are consistent. Finding similar clusters and getting similar results from the hypothesis tests with different parameter choices is evidence in favour of that the true clusters have been identified and that the tested differences correspond to real patterns.
- The assumption that the data has a hierarchical structure might be unrealistic. Hierarchical clustering has good performance if the true clusters are nested. As an example, suppose that the data consists of observations of people with an equal share of men and women, of which there is an equal share of American, Japanese, and French. We can imagine a scenario in which the best division of the people in terms of maximising within-cluster similarity is to split them into two groups defined by gender, or alternatively, to split them into three groups defined by nationality. In this scenario, the true clusters are not nested as the best division into three groups by nationality does not result from first taking the best division into two groups by gender and then splitting up those groups by nationality. Consequently, in this scenario the true clusters would not be well-represented by hierarchical clustering. In the context of our bias scan tool, nested structures do align well with notions of intersectional (or: multiple) discrimination, i.e., disadvantage based on two or more characteristics considered together, for example being a “black woman”. Similarly, hierarchical clustering conceptually fits one- or multi-dimensional proxy discrimination, i.e., disadvantage based on dog ownership and car type.
- K-means and hierarchical clustering will assign each observation to a cluster. Sometimes this might not be appropriate. For instance, suppose that most of the observations truly belong to a small number of (unknown) clusters, and that a small subset of the observations are outliers in the sense that they are different both from each other and from the remaining observations. As k-means and hierarchical clustering force every observation into a cluster, the found clusters may be heavily distorted due to the presence of the outliers that in reality do not belong to any shared cluster. In such a scenario, mixture models are an attractive approach as they are not sensitive to the presence of outliers. We refer readers to Hastie et al. (2009) for details.

Most importantly, care must be taken in how the results of a cluster analysis are reported. Being an exploratory tool, results should not be taken as the absolute truth about a data set. Rather, they should constitute a starting point for the development of a scientific hypothesis and further qualitative study, preferably with the help of subject matter experts.

2.2 Qualitative – A deliberative approach to define fair AI

The HBAC quantitative bias scan serves as a starting point to detect algorithmic bias. Ultimately, establishing discrimination is a qualitative, normative exercise that can only be performed by subject matter experts (SME).

We present a deliberative method to review identified quantitative disparities in AI models, as detected for instance by the HBAC bias scan tool. First, model metrics are collected in a standardized and automated manner through AI factsheets. Based on these context-specific information, a team of NGO Algorithm Audit drafts a problem statement and collects feedback on this document. Second, an independent and diverse audit commission is convened, consisting of diverse experts from a wide range of backgrounds. The commission members share initial written reactions on the problem statement, which will smoothen discussions during the gathering. Third, the identified issues are discussed during the commission gathering. Consensus among the commission members does not necessarily be reached. Disagreement on certain topics is valuable information on itself regarding the ethical issue at hand. Forth, based on the topics discussed during gathering best-practices are distilled and are shared online in NGO Algorithm Audit's case repository, available for all to learn from. More information on recent case studies can be found on https://www.algorithmaudit.eu/cases/.

Figure 4 – NGO Algorithm Audit's approach to perform case reviews

Related work (selection)

- Nasiriani et al.* propose a method to detect possible discrimination with hierarchical clustering. However, this approach requires pre-specified protected attributes, on which data is often not available (see legal challenge **I.**).
- Algorihthm Audit: Why, What and How?, Biagio Aragona
- Practical Fairness, Aileen Nielsen.

*Nasiriani, N., Squicciarini, A., Saldanha, Z., Goel, S., & Zannone, N. (2019). Hierarchical clustering for discrimination discovery: A top-down approach. In *Proceedings - ieee 2nd international conference on artificial intelligence and knowledge engineering, 2019* (pp. 187–194).

3. Results – Defining fair AI through the qualitative interpretation of quantitative metrics

In this report, two homemade case studies are discussed that illustrate the need for qualitative interpretation of quantitative metrics to safeguard fair treatment by AI. To demonstrate the applicability of our approach to the post-processing phase of the AI lifecycle, we apply our unsupervised bias scan tool to (1) a BERT classifier of fake tweets and, (2) to a XGBoost classifier of loan approvals to examine disparate group fairness metrics. To demonstrate the applicability of our approach to the in-processing phase of the AI lifecycle, we present a case study by NGO Algorithm Audit on the type of SIM card as a discriminatory proxy-variable for ethnicity in a payment fraud prediction model that was used by a large multinational e-commerce platform . We summarize the key insights of these case studies below.

3.1 Case I – BERT tweet disinformation classifier

A BERT-based disinformation classifier is trained on fake tweets from the Twitter15[19] data set, enriched with self-collected Twitter API data. Full detail on the training process can be found in this Github repository[20].

Results

The identified clusters by HBAC are displayed in Figure 5.

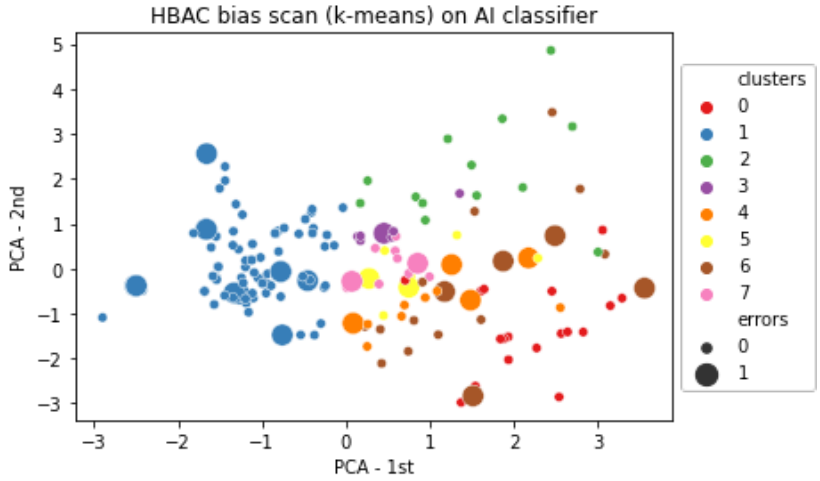

Figure 5 – Identified clusters by k-means HBAC scan. Cluster 4 has most negative bias: -0.27.

Statistical significant differences in features is measured between cluster with most negative bias (cluster 4, bias=-0.27) and rest of dataset.

[19] Liu, Xiaomo and Nourbakhsh, Armineh and Li, Quanzhi and Fang, Rui and Shah, Sameena, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (2015)
[20] https://github.com/NGO-Algorithm-Audit/Bias_scan/tree/master/case_studies

Feature	Difference	p-value
verified	0.53468	0.00000
sentiment_score	0.95686	0.00005
#URLs	-0.74095	0.00005

Table 1 – Statistical significant differences in features between the cluster with most negative bias and the rest of the dataset.

Tweets of users with a verified profile, above average sentiment score (based on the VADER[21] library) and below average number URLs used in their tweets are classified significantly more often as disinformation by the BERT-based classifier. Next, with the help of subject matter experts a qualitative assessment is needed to examine the measured quantitative disparities further. More details on this case study can be found here[22].

3.2 Case II – XGBoost loan approval classifier

A XGBoost loan application classifer trained on the widely cited German Credit data set[23]. The data was fed one-hot encoded to the k-means HBAC, as the clustering algorithm of the HBAC algorithm requires numerical data to calculate the distance between the data points. Full detail on the training process can be found in our Github repository[20].

Results

The identified clusters by HBAC are displayed in Figure 6.

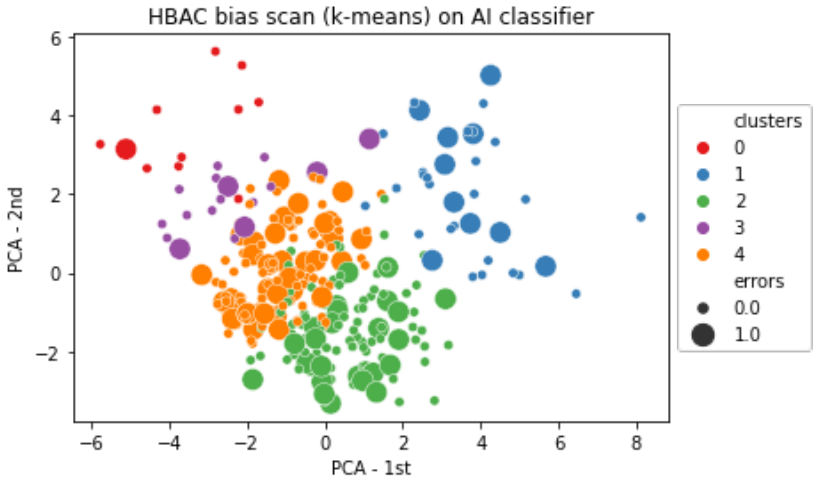

Figure 6 – Identified clusters by k-means HBAC scan. Cluster 4 has most negative bias: -0.05.

Statistical significant differences in features is measured between cluster with most negative bias (cluster 4, bias=-0.05) and rest of dataset. Note however that the measured negative cluster bias (-0.05) is not that larger. This should be taken into consideration when performing a qualitative assessment of suspected unfair treatment by the AI classifier. For sake of completeness differences in features are provided and statistically tested:

[21] https://github.com/cjhutto/vaderSentiment
[22] https://github.com/NGO-Algorithm-Audit/Bias_scan/blob/master/HBAC_scan/HBAC_BERT_disinformation_classifier.ipynb
[23] German Credit Data from the UCI Repository of Machine Learning Databases

Feature	Difference	p-value
hone registered	-1.33051	0.00000
unknown/no property	-0.68001	0.00000
self-employed/highly qualified employee	-0.65643	0.00000
free housing	-0.58238	0.00000
credit amount	-0.56406	0.00000

Table 2 – Statistical significant differences in features between the cluster with most negative bias and the rest of the dataset.

This means that loan applicants without real estate or unknown/no property, negative balance, unskilled job status, or those who want to use a loan to buy a new car or radio/television are significantly less often approved by the XGBoost classifier. Next, with the help of subject matter experts a qualitative assessment is needed to examine the measured quantitative disparities further. More details on this case study can be found.

3.3 Case III – Proxy discrimination in 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 ʾcould 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 ʻType of SIM card ʾas 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[24] 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.

[24] Type of SIM card as a predictor variable to detect payment fraud, NGO Algorithm Audit (2022).

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.