Algorithm Audit

Technical documentation

Table of contents

Where to use the bias scan tool?

Available as web application
https://www.algorithmaudit.eu/bias_scan/

Source code available on GitHub
https://github.com/NGO-Algorithm-Audit/Bias_scan

Executive summary

Artificial intelligence (AI) is increasingly used to automate or support policy decisions that affect 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 unfair. AI can be unfair, for instance, through differentiation linked to protected attributes, such as gender, sex, ethnicity or race, i.e., direct or indirect discrimination. Or, through differentiation upon new categories of people (defined by a high-dimensional mixture of features) that evades non-discrimination law. Such new types of differentiation could be perceived as unfair, for instance if it reinforces socio-economic inequality.

To address these risks, we propose a scalable, model-agnostic, and open-source bias scan tool to identify potentially unfair treated 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 (higher-dimensional) differentiation. The tool is available as a web application on the website of Algorithm Audit, such that it can be used by the entire AI auditing community.

As demonstrated on a self-trained BERT Twitter disinformation detection model, the bias scan tool identifies statistically significant disinformation classification bias against users with a verified profile, higher number of followers and higher user engagement score. These results are supported by sensitivity testing for 162 reasonable hyperparameter configurations of the unsupervised bias scan tool. Resulting in 1,000+ clusters, which aggregation statistics confirm our findings.

These observations do not establish prohibited *prima facie* discrimination. Rather, the identified disparities serve as a starting point to assess potential unfair treatment 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 *proportional* and *necessary*. For this qualitative assessment, we propose an expert-led, deliberative method. Informed by the quantitative results of the bias scan, we raise questions about the BERT disinformation classifier to an independent audit commission of AI experts, to form a normative judgement about unfair treatment by this specific classifier.

All documentation relating to this case study is publicly available. In this way, we enable policy makers, journalists, data subjects and other stakeholders to review the normative judgements Algorithm Audit's commissions arrive at. 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 (see Figure 1).

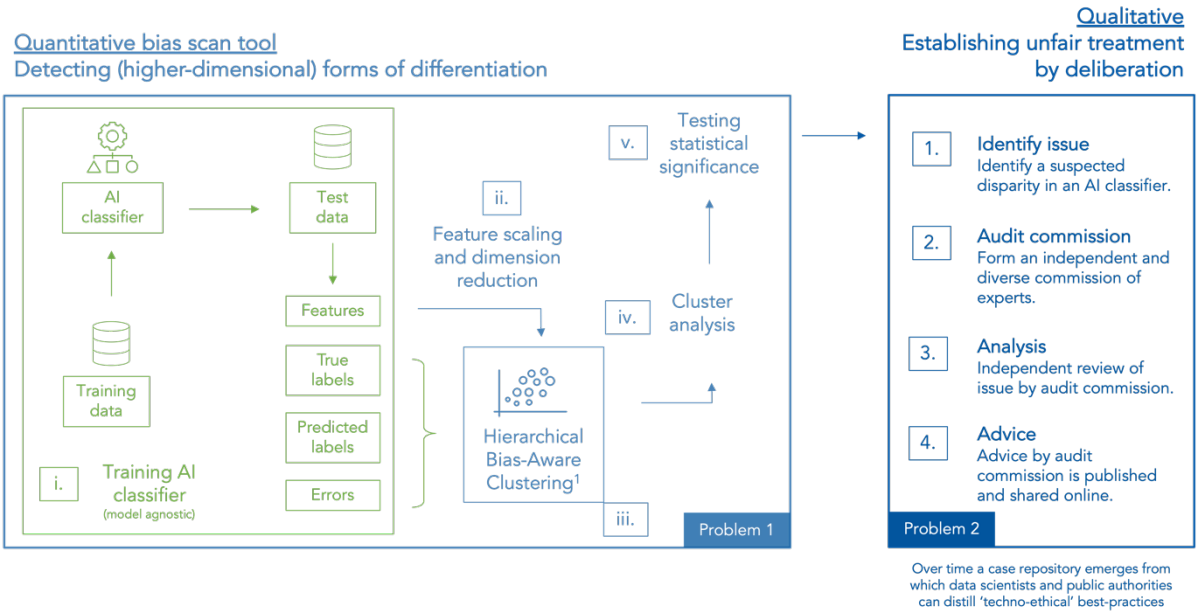

Figure 1 – *Overview of the quantitative-qualitative approach to assess fair AI with help of the unsupervised bias scan tool and deliberative expert-led approach.*

Scope of this bias scan

The quantitative bias scan tool currently only 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, e.g., indirect higher-dimensional (proxy) discrimination also known as *ad hoc bias*;
- Algorithmic differentiation that is not related to protected characteristics, such as differentiation on the basis of web browser or house number, but could still be perceived as unfair, for instance as it reinforces socio-economic inequality.

Contributions to this bias scan tool

This bias scan tool is a collective effort of experts from a range of disciplines and professional backgrounds, to shed light into the abstract concept of 'fair AI'. We bring together expertise from academia, industry, policy making and journalism to strengthen this quantitative-qualitative 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 Algorithm Audit to build and share public knowledge about ethical algorithms. A full list of endorsement can be found in the Appendix.

Work of NGO Algorithm Audit

Audit commissions	Advising on ethical issues emerging in concrete algorithmic practices
Technical tooling	Implementing and testing technical tools to detect and mitigate bias
Advocacy	Contributing to public debate on responsible use of algorithms
Knowledge sharing	Sharing techno-ethical insights with society, policy makers and others

Supported by

EUROPEAN ARTIFICIAL INTELLIGENCE FUND

NL AI Coalition

SIDN fonds

1. Problem description

At 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.

In section 1.1, quantitative challenges are described to detect (higher-dimensional) forms of algorithmic differentiation. Section 1.2 elaborates upon qualitative challenges to establish unfair treatment by binary AI classifiers. Section 1.3 concludes with an overview where our quantitative-qualitative approach intervenes in the AI lifecycle.

1.1 Problem 1 (quantitative) – Detecting higher-dimensional forms of algorithmic differentiation

The quantitative reasoning paradigm of AI poses challenges to assess fair treatment. Three pressing issues are described.

I. Sheer data volume – The human mind is not equipped to deal with large volumes of numbers. Statistical tools are therefore indispensable to detect disparities in the sheer volume AI uses to make predictions.
II. Higher-dimensional forms of differentiation (detecting *ad hoc bias*) – Scholarship has argued that granular analysis of personal and behavioral data entails heightened risk of intersectional discrimination[1] and new forms of differentiation that evade non-discrimination law[2]. 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[3]. New forms of differentiation refer to algorithms that differentiate upon new categories of

[1] Algorithmic discrimination in Europe, Gerards, J. and Xenidis, R. (2021).
[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).
[3] Judgment of 24 November 2016, David L. Parris v. Trinity College Dublin and Others, C-443/15, EU:C:2016:897.

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.

III. Data availability on protected grounds and group fairness metrics – 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. This work focusses however 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 third issue is that no 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.

1.2 Problem 2 (qualitative) – A persistent gap between general legal requirements and concrete AI practice

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[6], EU[7] and American[8] law, we discuss twee challenges that influence the assessment of fair treatment by AI agents:

IV. The proxy and correlation challenge – At a conceptual level, legal frameworks make a distinction between disparate treatment of protected groups (direct discrimination) and disparate impact on protected groups (indirect

[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.

[6] 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.

[7] 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.

[8] American Labor law, U.S. Constitution's Fourteenth Amendment

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[9]. Absent data on protected attributes (as discussed in III.) 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?

V. 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 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: A quantitative method to inform qualitative bias testing*, we present a quantitative-qualitative approach that mitigates the above five challenges to assess fair treatment by AI classifiers. We focus on two categories of risks related to fair treatment:

- indirect (higher-dimensional) forms of discrimination (II. and IV.)
- unfair differentiation and *ad hoc bias* (II.).

In doing so, we incorporate the quantitative challenge of sheer data volume (I.) and the legal challenge of legitimacy testing (V.).

[9] 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.

[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.

1.3 Bias along the AI lifecycle

Not only from a legal perspective, as well from a technical perspective fair algorithmic treatment can be studies. In the fair machine learning literature, bias in the AI lifecycle occurs in four phases (see Figure 2).

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? There might be a clear *raison d'être* 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].

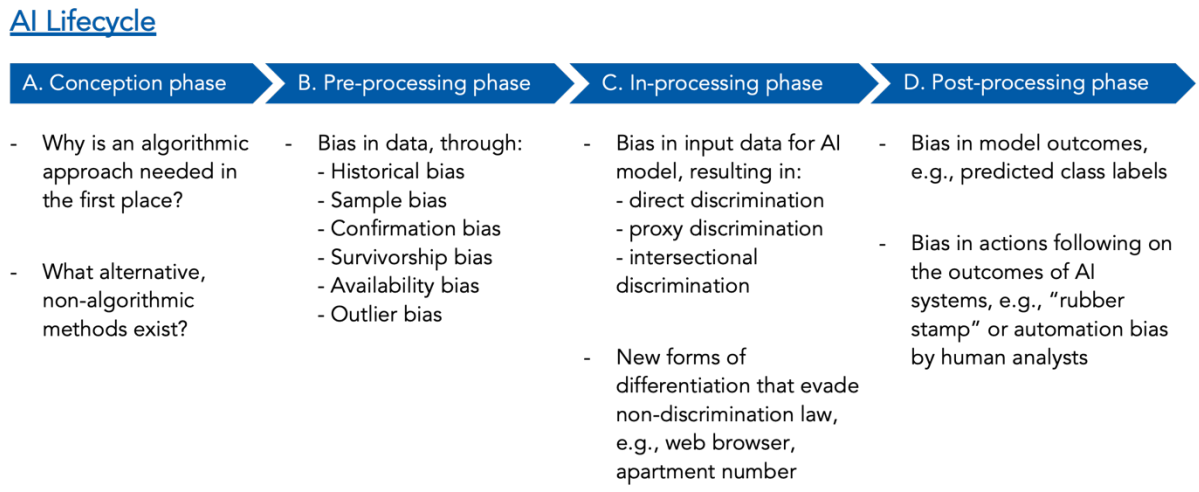

Figure 2 – *Conceptual breakdown of the AI lifecycle in four phases.*

[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

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 (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.

2. Solution: A quantitative method to inform qualitative bias testing

Algorithm Audit proposes a quantitative bias scan tool and a qualitative deliberative approach to address the challenges as described discussed in Section 1. *Problem description*. A quantitative approach is indispensable for monitoring AI-informed policy decisions due to the sheer data volume and it's technical complexity. At the same time, a qualitative assessment by subject-matter experts is 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.

2.1 Unsupervised bias scan tool to *detect* differentiation (quantitative)

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 III. (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, *ad hoc bias* and new types of differentiation that evade non-discrimination law, thereby overcoming quantitative challenge II.(higher-dimensional form of differentiation) and legal challenge IV. (the proxy and correlation challenge). In this report the HBAC bias scan tool is applied on a self-trained binary AI classifier 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. *Case study: Testing our quantitative-qualitative approach*. 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

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

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

classification performance (see the quantitative part of Figure 3). More details on the steps of the HBAC pipeline are provided below.

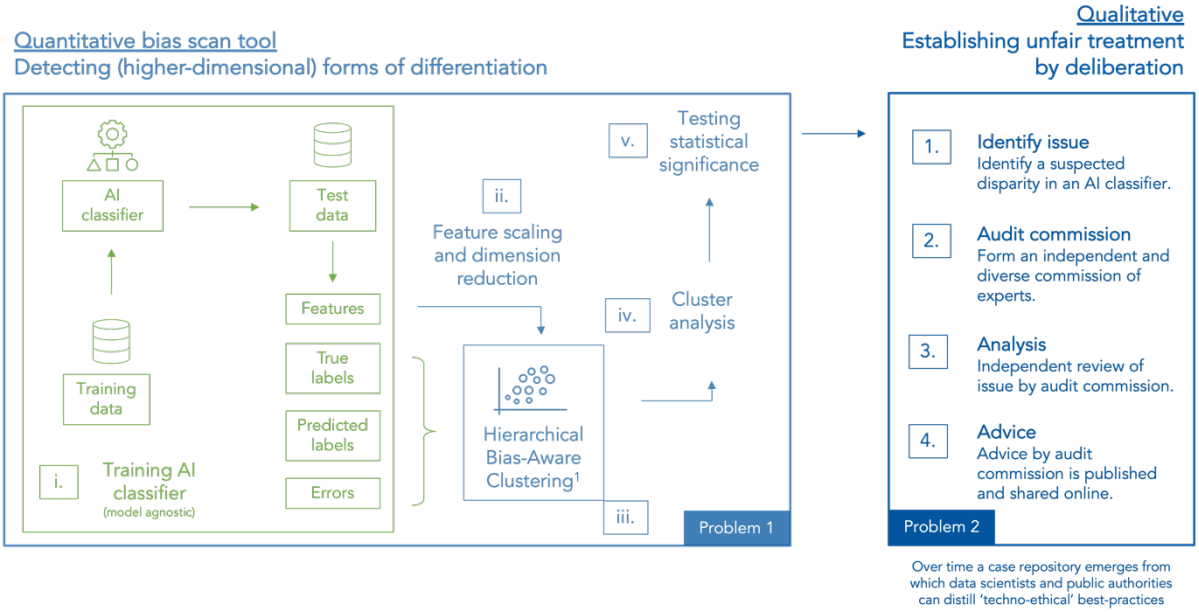

Figure 3 – *Overview of proposed solution. A quantitative bias scan tool to detect differentiation is combined with a qualitative deliberative approach to establish discrimination.*

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.
iii. Hierarchical Bias-Aware Clustering (HBAC) – 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 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[15], we experiment with '1-Accuracy', '1-False Positive Rate (FPR)' and '1-False Negative Rate (FNR)' 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. *Case study: Testing our quantitative-qualitative approach.*

iv. Analysis of identified clusters – We are interested in the identified clusters with highest bias, i.e., cluster of which the AI classifier predicts more negative than positive labels, False Positives (FPs) or False Negatives (FNs). 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. The results are displayed in confidence interval plots. Note, however, that there is no consensus on a single best approach to find features that can be sources of

[15] 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

discrimination and thus require closer inspection[16]. 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 description*). 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[17].

The k-means HBAC algorithm uses various hyperparameters. Parameters prevent HBAC to find only clusters with a small amount of datapoints, for which it is hard to find meaningful features. An overview and description of all hyperparameters is given in Table 1. An example of hyperparameter tuning is given in Section *3.1 Identifying disparities in a BERT-based Twitter disinformation classifier (quantitative).*

Number of initial clusters	The desired number of initial clusters oft he k-means clustering algorithm.
Maximum number of iterations	The HBAC algorithm is terminated after the maximum number of iteration threshold is reached, or after no clusters are found that have a higher discrimination bias when compared to the clusters of the previous iteration.
Minimal splitable cluster size	Number of elements that need to be in the cluster to be eligible for a next cluster split.
Minimal acceptable cluster size	Number of elements in a new candidate cluster during splitting to be accepted as a new cluster.

Table 1 – Hyperparameters of the HBAC algorithm.

In the next section, we discuss some limitations of the bias scan tool.

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

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

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

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 A deliberative approach to *establish* unfair treatment (qualitative)

The 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 (SMEs).

We present a deliberative method to review identified quantitative disparities in AI models, as detected for instance by the quantitative bias scan. First, model metrics are collected. Based on this context-specific information, a team of 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. This process is displayed in Figure 4. 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 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 – *Algorithm Audit's approach to conduct case studies.*

Related work

Various studies are performed to discover hidden bias using (hierarchical) clustering algorithms:

- 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 III.).

HBAC complements current bias scan tools by its' ability to automatically detect potentially discriminated groups by AI classifiers.

*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, aike 2019* (pp. 187–194).

3. Case study: Testing our quantitative-qualitative approach

To test our quantitative-qualitative approach, a case study is conducted for a real AI classifier. In Section 3.1, the HBAC bias scan is applied on a self-trained BERT disinformation classifier. In Section 3.2, questions are formulated that will be raised to an audit commission consisting of AI experts. All relevant documentation relating to this case study can be found on GitHub[18].

3.1 Identifying disparities in a BERT-based Twitter disinformation classifier (quantitative)

For this case study, we review a BERT disinformation classification algorithm[19] which is trained on the Twitter1516 dataset[20], enriched with self-collected Twitter API data[21]. The dataset consists of 1,057 verified true and false tweets, 3 user features (verified profile, #followers, user engagement) and 5 content features (length, #URLs, #mentions, #hashtags, sentiment score). We run two bias scans. In Scan 1, the bias metric is defined by the False Positive Rate (FPR). In Scan 2, the bias metric is defined by the False Negative Rate (FNR). FPR relates to true content predicted to be false, proportional to all true content. FNR relates to false content predicted to be true, proportional to all false content. In sum:

Scan 1. $\text{Bias} = \text{FPR}_{\text{cluster}} - \text{FPR}_{\text{rest of dataset}}$

Scan 2. $\text{Bias} = \text{FNR}_{\text{cluster}} - \text{FNR}_{\text{rest of dataset}}$.

The full bias scan pipeline is displayed in Figure 5.

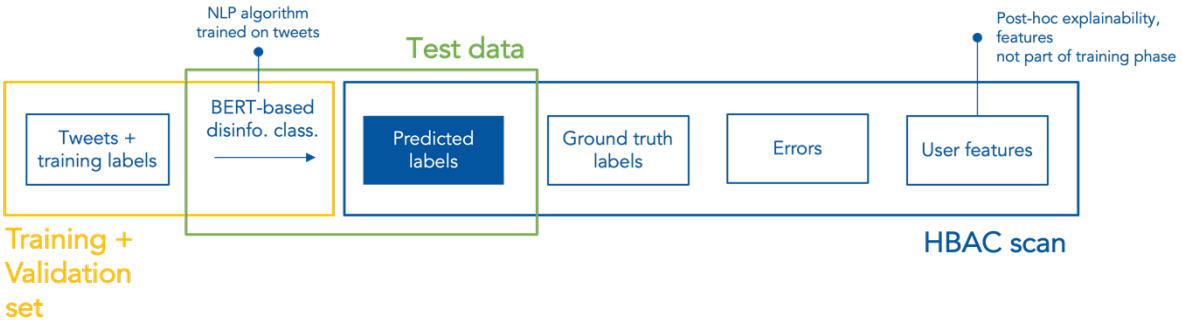

Figure 5 – *Bias scan pipeline*

For Scan 1, the cluster for which the disinformation classifier is underperforming the most (bias=0.08, n=249) is characterized by the features displayed in Figure 6. The

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

[19] More information about the self-trained BERT-based classification algorithm can be found here: https://github.com/NGO-Algorithm-Audit/Bias_scan/blob/master/case_studies/BERT_disinformation_classifier

[20] 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)

[21] More information on the data collection process can be found here: https://github.com/NGO-Algorithm-Audit/Bias_scan/blob/master/data/Twitter_dataset/Twitter_API_data_collection.ipynb

feature difference is the average in means between the disparately treated cluster and the rest of the dataset. Hypothesis testing[22] indicates that on average, user that:

- are verified, have higher #followers, user engagement and #URLs;
- use less #hashtags and have lower tweet length

have more true content classified as false (false positives).

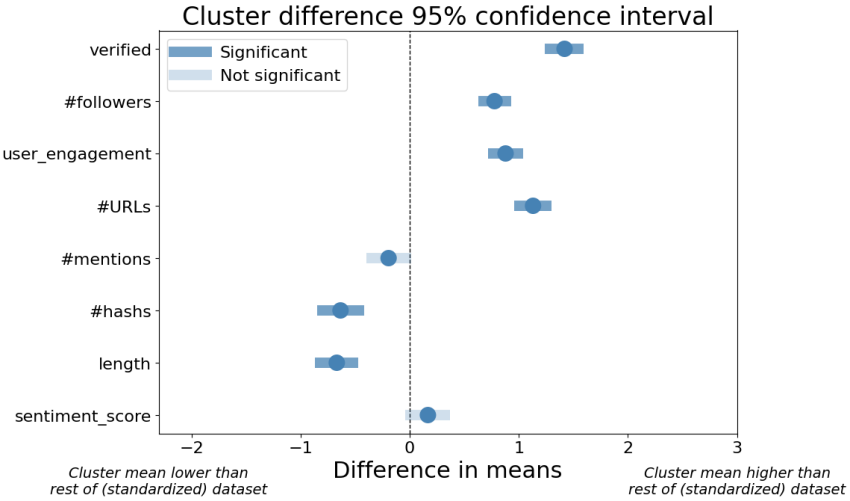

Figure 6 – Identified quantitative feature disparities in cluster with highest bias. For this bias scan, bias is defined by the False Positive Rate.

For Scan 2, the cluster for which the disinformation classifier is underperforming the most (bias=0.13, n=46) is characterized by the features displayed in Figure 7. Hypothesis testing indicates that on average, user that:

- use more #hashtags and have higher sentiment score;
- are non-verified, have less #followers, user engagement and tweet length

have more false content classified as true (false negatives).

[22] Here, the hypothesis tested is that there is no difference in feature means of the cluster and the pooled feature means of other clusters. These differences are statistically significant even after performing a Bonferroni correction to adjust for false discoveries due to multiple hypothesis testing.

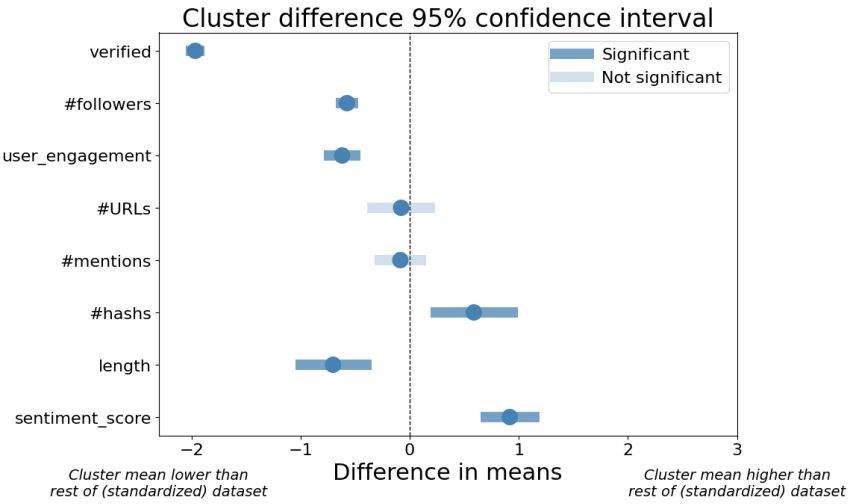

Figure 7 – Identified quantitative feature disparities in cluster with highest bias. For this bias scan, bias is defined by the False Negative Rate.

These results might indicate (higher-dimensional) unfair treatment by the disinformation classifier. More information on the identified clusters and robustness tests of the results can be found on GitHub[23].

3.2 Audit commission: Assessing potentially unfair treatment by an AI classifier (qualitative)

The identified disparities in Section 3 do not establish prohibited *prima facie* discrimination. Rather, the identified disparities serve as a starting point to assess potential unfair treatment according to the context-sensitive qualitative doctrine. To assess unfair treatment, we question:

i) Is there an indication that one of the statistically significant features, or a combination of the features, stated in **Error! Reference source not found.**-3 a re critically linked to one or multiple protected grounds?

ii) In the context of disinformation detection, is it as harmful to classify true content as false (false positive) as false content as true (false negative)?

iii) For a specific cluster of people, is it justifiable to have true content classified as false 8 percentage points more often? For a specific cluster of people, is it justifiable to have false content classified as true 13 percentage points more often?

[23] https://github.com/NGO-Algorithm-Audit/Bias_scan/blob/master/HBAC_scan/HBAC_BERT_disinformation_sensitivity_testing.ipynb

iv) Is it justifiable that the disinformation classification algorithm is too harsh towards users with verified profile, more #followers and higher user engagement and too lenient towards users with non-verified profile, less #followers and lower user engagement?

The audit commissions convenes in Jan-Feb 2023, to elaborate on the questions i)-iv).

Auditing disinformation detection algorithms

As of December 2022, Article 28 of the European Digital Services Act (DSA) subjects very large online platforms (VLOPs) to annual independent auditing of their services and risk mitigation measures. Open-source AI auditing tools, such as this bias scan tool, help to detect and mitigate (higher-dimensional) forms of unfair treatment in disinformation detection and other AI (ranking or recommender) systems.

With this case study, Algorithm Audit aims to provide qualitative guidelines how statistical methods can be used to monitor unfair treatment by AI systems. Without clear guidance from data protection boards, supervisors, researchers and algorithmic regulatory body, misinterpretation of quantitative metrics stands in the way of independent quantitative and qualitative oversight of the risk of biased AI systems. Algorithm Audit provides normative advice by convening audit commissions to shed light on identified ethical issues.

Future improvements of the bias scan tool

Technical implementation of bias scan tool:

- Apply MLOps best-practices to improve robustness
- Improve user experience

Methodological approach

- DBScan alternative to k-means HBAC

Case study:

- Apply bias scan tool to new use case, for instance Natural Language Processing algorithm, like Fundamental Rights Agency case study https://fra.europa.eu/en/publication/2022/bias-algorithm

4. Conclusion

Quantitative methods, such as unsupervised bias scans tools, are helpful to discover potentially unfair treated groups of similar users in AI systems in a scalable manner. Automated bias scans of AI classifiers allow human experts to assess found disparities 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 on the website of Algorithm Audit, over time a repository of "techno-ethical jurisprudence" emerges; from which data scientists and public authorities can distill best practices to build fairer AI.

Appendix – Contributors and endorsements

Algorithm Audit's bias scan tool team

- Jurriaan Parie, Trustworthy AI consultant at Deloitte
- Ariën Voogt, PhD-candidate in Philosophy at Protestant Theological University of Amsterdam
- Joel Persson, PhD-candidate in Applied Data Science at ETH Zürich.

Endorsements

Journalism

- Gabriel Geiger, Investigative Reporter Algorithms and Automated Decision-Making at Lighthouse Reports
- AA
- BB

Industry

- Laurens van der Maas, Data Scientist at AWS;
- CC
- DD

Academia

- Anne Meuwese, Professor in Public Law & AI at Leiden University
- Hinda Haned [to be confirmed], Professor in Data Science at University of Amsterdam
- Marlies van Eck, Assistant Professor in Administrative Law & AI at Radboud University
- Emma Beauxis-Ausselet [to be confirmed], Associate Professor Ethical Computing at University of Amsterdam
- Vahid Niamadpour, PhD-candidate in Linguistics at Leiden University
- Floris Holstege, PhD-candidate in Explainable Machine Learning at University of Amsterdam

Civil society organisations

- XX
- YY
- Simone Maria Parazzoli, Intern at the OECD Observatory of Public Sector Innovation (OPSI).