**Table of contents**

**Executive Summary** p.1

1. **Problem: The persistent gap between legal non-discrimination** p.2 **requirements and AI practice**Problem statement

Challenges arising from non-discrimination law and data protection legislation  
Bias along the AI lifecycle

1. **Solution: Fair AI through discussion – A deliberative way forward** p.6Quantitative: Hierarchical Bias-Aware Clustering (HBAC)   
   Qualitative: NGO Algorithm Audit’s deliberative approach
2. **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 shed light in the normative concept of ‘fair AI’. This submission is an amalgam of statistics, philosophy, law, sociology, public policy and AI governance. And serves as a starting point to demystify AI to make normative data modelling choices together with society.

**Executive summary**  
As AI is omnipresent in digital society, there is an urgent need to review algorithms 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 possible proxy discrimination, intersectional discrimination and other 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 an unverified profile and an above average number of mentions and hashtags used in tweets. On the 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* 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 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.

In this submission we focus on common AI applications applied 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 (disparate impact or indirect (proxy) discrimination);
* 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.

Scope of submission

1. **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 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.

## **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]](#footnote-1), EU[[2]](#footnote-2) and American[[3]](#footnote-3) law, we discuss three challenges that influence the assessment of fair AI:

1. **Data availability on protected grounds** – Equal treatment laws prohibit agents from acting with “discriminatory purpose”[[4]](#footnote-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]](#footnote-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 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;
2. **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]](#footnote-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?
3. **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]](#footnote-7) and new forms of differentiation that evade non-discrimination law[[8]](#footnote-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]](#footnote-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.
4. **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]](#footnote-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: 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.**).

## **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).

Text

Description automatically generated

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

1. **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.
2. **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]](#footnote-11)**, Google’s Machine Learning Glossary on fairness**[[12]](#footnote-12)** and Wikipedia’s catalog of cognitive biases**[[13]](#footnote-13)**.
3. **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.
4. **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, 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 the qualitative assessment of legal and ethical risks pertaining to the inclusion of certain data variables to AI models (**C.** In-processing phase). To identify what aspects of AI systems should be assessed qualitatively, we present a quantitative bias scan tool in the following section.

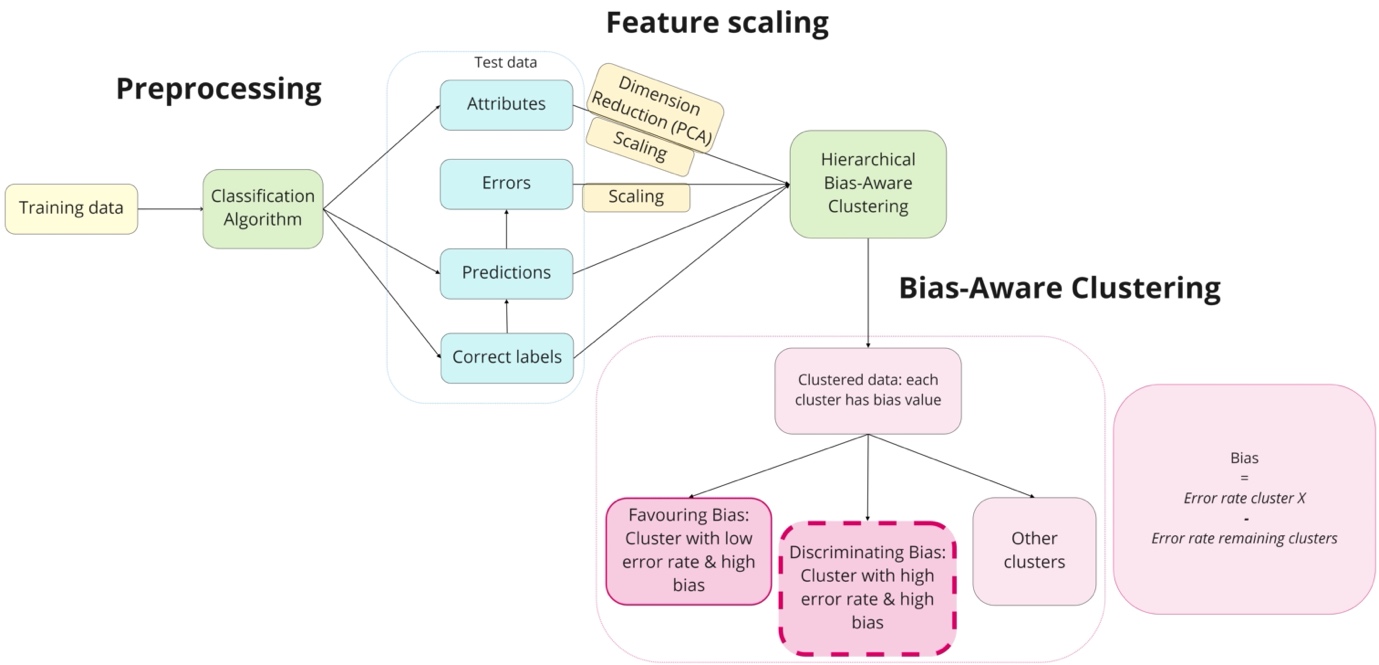
**2. Solution – Fair AI through discussion: A deliberative way forward**NGO Algorithm Audit proposes a quantitative bias scan tool and a qualitative deliberation method to assess Fair AI. [quantitative]  
We believe a multi-disciplinary, well-informed and open debate is the best way forward to form normative judgements about algorithmic bias. So, 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 k-means Hierarchical Bias Aware Clustering (HBAC) algorithm[[14]](#footnote-14), 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 (unsupervised learning). The tool can therefore 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[[15]](#footnote-15).

#### Bias scan: Unsupervised K-Means Hierarchical Bias-Aware Clustering (HBAC)

When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other. To make this concrete, we must define what it means for two or more observations to be similar or different. Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.



#### Pre-processing

#### Feature scaling

If the variables are scaled to have standard de- viation one before the inter-observation dissimilarities are computed, then each variable will in effect be given equal importance in the hierarchical clustering performed.

We might also want to scale the variables to have standard deviation one if they are measured on different scales; otherwise, the choice of units (e.g. centimeters versus kilometers) for a particular vari- able will greatly affect the dissimilarity measure obtained.

PCA looks to find a low-dimensional representation of the observa- tions that explain a good fraction of the variance;

#### Clustering

Clustering looks to find homogeneous subgroups among the observations.

To perform K-means clustering, we must first specify the desired number of clusters K; then the K-means algorithm will assign each observation to exactly one of the K clusters. The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible. Because the K-means algorithm finds a local rather than a global opti- mum, the results obtained will depend on the initial (random) cluster as- signment of each observation. To perform K-means clustering, we must decide how many clusters we expect in the data. The problem of selecting K is far from simple.

One potential disadvantage of K-means clustering is that it requires us to pre-specify the number of clusters K. Hierarchical clustering is an alter- native approach which does not require that we commit to a particular choice of K. Hierarchical clustering has an added advantage over K-means clustering in that it results in an attractive tree-based representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n.

The assumption of hierarchical structure of data might be unrealistic. For instance, suppose that our observations correspond to a group of people with a 50–50 split of males and females, evenly split among Americans, Japanese, and French. We can imagine a scenario in which the best division into two groups might split these people by gender, and the best division into three groups might split them by nationality. In this case, the true clusters are not nested, in the sense that the best division into three groups does not result from taking the best division into two groups and splitting up one of those groups. Consequently, this situation could not be well-represented by hierarchical clustering.

Linkage defines the dissimilarity between two groups of observations [single, average, complete, centroid] What type of linkage is used?

Why k-means is the right clustering methods [Selma scriptie]

#### Validating clusters obtained

Any time clustering is performed on a data set we will find clusters. But we really want to know whether the clusters that have been found represent true subgroups in the data, or whether they are simply a result of clustering the noise. For instance, if we were to obtain an independent set of observa- tions, then would those observations also display the same set of clusters? This is a hard question to answer. There exist a number of techniques for assigning a p-value to a cluster in order to assess whether there is more evidence for the cluster than one would expect due to chance. However, there has been no consensus on a single best approach. More details can be found in Hastie et al. (2009).

#### Limitations of clustering

Both K-means and hierarchical clustering will assign each observation to a cluster. However, sometimes this might not be appropriate. For instance, suppose that most of the observations truly belong to a small number of (unknown) subgroups, and a small subset of the observations are quite different from each other and from all other observations. Then since K- means and hierarchical clustering force every observation into a cluster, the clusters found may be heavily distorted due to the presence of outliers that do not belong to any cluster. Mixture models are an attractive approach for accommodating the presence of such outliers. These amount to a soft version of K-means clustering, and are described in Hastie et al. (2009).

We have described some of the issues associated with clustering. However, clustering can be a very useful and valid statistical tool if used properly. We mentioned that small decisions in how clustering is performed, such as how the data are standardized and what type of linkage is used, can have a large effect on the results. Therefore, we recommend performing clustering with different choices of these parameters, and looking at the full set of results in order to see what patterns consistently emerge. Since clustering can be non-robust, we recommend clustering subsets of the data in order to get a sense of the robustness of the clusters obtained. Most importantly, we must be careful about how the results of a clustering analysis are reported. These 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 study, preferably on an independent data set.

## 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.

## So what?

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.

# 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[[16]](#footnote-16) trained on verified Twitter data[[17]](#footnote-17). 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[[18]](#footnote-18). 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[[19]](#footnote-19) 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 the U.S. Equal Employment Opportunity Commission https://tinyurl.com/29f7kj5b [↑](#footnote-ref-4)
5. Hundreds of laws enacted at the federal and state levels serve to protect the personal data of U.S. residents. [↑](#footnote-ref-5)
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. [↑](#footnote-ref-6)
7. Algorithmic discrimination in Europe, Gerards and Xenidis (2021). [↑](#footnote-ref-7)
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). [↑](#footnote-ref-8)
9. Judgment of 24 November 2016, David L. Parris v. Trinity College Dublin and Others, C-443/15, EU:C:2016:897. [↑](#footnote-ref-9)
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. [↑](#footnote-ref-10)
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. [↑](#footnote-ref-11)
12. https://developers.google.com/machine-learning/glossary/fairness#e [↑](#footnote-ref-12)
13. https://en.wikipedia.org/wiki/List\_of\_cognitive\_biases [↑](#footnote-ref-13)
14. Misztal-Radecka, Indurkya, Information Processing and Management. Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems (2021). [↑](#footnote-ref-14)
15. # 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-15)
16. https://github.com/NGO-Algorithm-Audit/AI\_Audit\_Challenge [↑](#footnote-ref-16)
17. 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-17)
18. German Credit Data from the UCI Repository of Machine Learning Databases [↑](#footnote-ref-18)
19. ### Type of SIM card as a predictor variable to detect payment fraud, NGO Algorithm Audit (2022).

    [↑](#footnote-ref-19)