Auditing a Dutch Public Sector Risk Profiling Algorithm Using an Unsupervised Bias Detection Tool

Floris Holstege
University of Amsterdam
Amsterdam
Netherlands

Mackenzie Jorgensen
Northumbria University
Newcastle
United Kingdom

Kirtan Padh
TU Munich
Munich
Germany

Jurriaan Parie*
Algorithm Audit
The Hague
Netherlands

Krsto Proroković
Algorithm Audit
The Hague
Netherlands

Joel Persson
Algorithm Audit
The Hague
Netherlands

Lukas Snoek
Algorithm Audit
The Hague
Netherlands

Abstract

Algorithms are increasingly used to automate or aid human decisions, yet recent research shows that these algorithms may exhibit bias across legally protected demographic groups. However, data on these groups may be unavailable to organizations or external auditors due to privacy legislation. This paper studies bias detection using an unsupervised bias detection tool when data on demographic groups are unavailable. We collaborated with the Dutch Executive Agency for Education to audit an algorithm that was used to assign risk scores to college students at the national level in the Netherlands between 2012-2023. Our audit covers more than 250,000 students across the country. The unsupervised bias detection tool highlights known disparities between students with a non-European migration background and students with a Dutch or European-migration background. Our contributions are two-fold: (1) we assess bias in a real-world, large-scale, and high-stakes decision-making process by a governmental organization; (2) we provide the unsupervised bias detection tool in an open-source library for others to use to complete bias audits. Our work serves as a starting point for a deliberative assessment by human experts to evaluate potential discrimination in algorithmic decision-making.

1 Introduction

Between 2012 and 2023, a simple rule-based risk profiling algorithm from the Dutch Executive Agency for Education (DUO) contributed to indirect discrimination in a control process to check whether students were unduly allocated a college grant [20]. A rule-based algorithm, consisting of three apparently neutral characteristics of the students (type of education, age, and distance to parents), was used to assign higher risk scores to students enrolled in vocational education, younger students, and students who lived near their parental address. The seemingly impartial risk assessment led to unequal treatment of students with a non-European migration background, remaining unnoticed for over 10 years and affecting more than 350,000 students [19]. [2] "Risk" in this context, and throughout the text, refers to the risk of "undue" or improper granting of college grants to students.

*Corresponding author: j.parie@algorithmaudit.eu

[2]Between 2012-2023, more than 350,000 students were assigned a risk score by the algorithm. Our analysis focuses on the student population in two years: 2014 and 2019, which consists of a total of 298,882 students.

Given this, in 2024, the Dutch Minister for Education, Culture and Science apologized on behalf of the Dutch government for indirect discrimination experienced by students with a non-European migration background and announced a EUR 61 million compensation plan [19]. This recent discriminatory algorithm scandal in the Netherlands [32] highlights that not only complex statistical methods, such as machine learning, but also simple rule-based algorithms can pose significant risks of embedding discrimination in algorithmic decision-making processes. In a recent interview, the chair of the Dutch Data Protection Authority warned that discriminatory rule-based algorithms are "around every corner" in executive branches of the Dutch government [49]. This type of risk profiling is prohibited by the European Convention of Human Rights (ECHR), the Charter of Fundamental Rights of the European Union (EU) [27], the EU General Data Protection Regulation (GDPR) [24] and the EU AI Act [26], as well as the national Public Administration Law [18]. Limited institutional enforcement capabilities, along with a lack of awareness, have led to inadequate oversight of algorithmic decision-making processes [42].

To combat bias in algorithmic systems, the academic community has presented a multitude of ways to define, detect, and mitigate different notions of bias at different points along the algorithmic pipeline [54, 3, 9, 41]. Most bias mitigation methods require data on demographic groups to prevent discriminatory bias. These demographic groups are typically protected attributes under discrimination law, such as race, gender and age [2].[3] However, data on demographic groups are often not available to organizations or external algorithm auditors [1, 7]. Unsupervised learning can aid in this challenge by detecting groups in the data for which a given notion of equal treatment is violated, avoiding the need for ex-ante specification of demographic or legally protected groups [43].

In this paper, we use an unsupervised bias detection tool to re-audit the risk profiling algorithm from DUO for bias. The original audit, conducted by Algorithm Audit in 2023, relied on aggregated demographic data on country of birth and country of origin provided by the Dutch national statistics office, Statistics Netherlands [51], and is referred to as "supervised bias detection." For both audits, DUO provided data processed by their algorithm that is not publicly accessible. By combining results from the supervised and unsupervised audit, we assess whether the clusters identified by the unsupervised bias detection tool, which represent potentially unfairly treated groups, correspond to disadvantaged demographic groups resulting from the supervised analysis. This allows for a comparison between an unsupervised learning approach and a supervised approach. We emphasize that the clustering procedure itself does not rely on the available protected demographic data: we only use the aggregated demographic data to illustrate, post hoc, that the flagged clusters align with negatively affected groups. This additional analysis is therefore a validation of the unsupervised clustering methodology, not a replacement for supervised bias detection.

In our context, the bias metric[4] requires equal probabilities of "high risk" classification for students with a Dutch or European migration background and those with a non-European migration background. We justify the use of this metric since it assesses the adverse effect of the risk profiling algorithm on different demographic groups: a crucial aspect for investigating indirect discrimination under European non-discrimination law. We find that the cluster with the highest bias returned by the unsupervised bias detection tool is characterized by students who are more often enrolled in vocational education, live close to their parental address, and are relatively more often 15-18 or 25-50 years old. This aligns with demographic groups found to be negatively affected by the risk profiling algorithm, based on aggregated statistics regarding students' non-European migration background.

Our paper makes two contributions: (1) A **large scale audit of indirect discrimination**: The DUO audit presents a high-quality and socially relevant dataset, providing insights into the assessment of indirect discrimination in algorithmic decision-making. This study thus contributes to the advancement of public knowledge on auditing algorithmic systems for indirect discrimination. (2) An **open-source**

[3]The exact protected grounds are enshrined in European non-discrimination law, such as the Treaty on the Functioning of the European Union (TFEU), the EU Charter of Fundamental Rights, and specific directives like the Racial Equality Directive (2000/43/EC) and the Employment Equality Directive (2000/78/EC). Note that some provisions do not mention any particular grounds of discrimination, such as Article 20 of the EU Charter of Fundamental Rights which includes openly formulated clauses such as "everyone is equal before the law." The main protected attributes include: race and ethnicity, gender and sex, religion or belief, disability, age, sexual orientation, nationality, and marital or family status. We will refer to these attributes as *demographic groups.*

[4]Throughout our paper, we use the term bias metric to refer to what is commonly called a fairness metric in the algorithmic fairness literature. The bias metric we define reflects the fairness metric of demographic parity [21].

tool: The unsupervised bias detection method is available as an open-source Python package[5] and a web application.[6] The latter allows non-technical users to apply the tool to their own datasets in order to detect potential bias. The underlying algorithm we provide includes methodological improvements that substantially reduce the false positive rate.

Overall, our paper showcases the potential of unsupervised learning to aid in auditing algorithmic decision-making processes through a real-world use case. We argue that unsupervised bias detection serves as a starting point for a deliberative assessment by human experts to evaluate potential discrimination and unfairness in the process when demographic groups are unavailable.

2 Literature review

We review a selection of the literature on bias detection and non-discrimination law, focusing on three topics: (1) challenges in detecting and establishing indirect discrimination, (2) bias detection methods with and without access to demographic data, and (3) auditing algorithmic systems for bias.

2.1 Challenges in detecting and establishing discrimination

Under European non-discrimination law, DUO's control process is classified as indirect discrimination because a neutral practice disproportionately disadvantaged non-European migrant students [14]. In algorithms, discrimination often arises from the *proxy and correlation challenge* [28]. Proxies refer to characteristics that are correlated with a demographic group, leading to disadvantaged outcomes for this group. For example, the Court of Justice of the European Union (CJEU) found that using part-time employment as a basis for differentiation in the context of occupational pension schemes constituted indirect discrimination, since more than 80% of part-time workers in Spain in 2012 were women [15]. In some cases, courts may accept that differential treatment of a demographic group is acceptable. Under EU law, indirect discrimination can be justified if it is demonstrated that a legitimate aim is being pursued, and the means of achieving that aim are appropriate and necessary [11, 12, 13, 23]. Our paper highlights how unsupervised methods can aid in the challenge of assessing indirect discrimination in algorithmic decision-making processes when demographic group data is not available.

2.2 Bias detection with and without access to demographic groups

In the algorithmic fairness literature, identifying systematic differences in the treatment of individuals or groups compared to others is commonly referred to as *bias detection*. Most bias detection methods assume access to demographic groups [22, 30, 39, 37, 4, 16]. However, in practice, these labels are often unavailable due to privacy legislation [52]. For instance, Article 9 in the EU's GDPR restricts the collection of "special categories of data" [24], which include data on ethnicity, religion, health, and sexual orientation.[7] The EU AI Act introduces a potential exception to the prohibition on processing special category data: "to the extent that it is strictly necessary for the purpose of ensuring bias detection and correction" [26]. Nonetheless, ongoing legal uncertainties regarding the interplay between the AI Act and the GDPR underscore the need for methods to assess bias in algorithmic systems without relying on demographic data. Algorithms can be biased even in the absence of demographic data, particularly through proxy discrimination, where a set of attributes acts as a proxy for the protected attribute [2].

Few alternatives exist to approaches that rely on demographic group data. One class of methods circumvents the limited availability of such data by adopting a Rawlsian definition of fairness [48] that maximizes utility for the most disadvantaged group [31, 40, 8]. However, the results of these methods do not always adhere to more traditional parity-based group fairness metrics [35]. A set of alternatives are unsupervised learning methods for fairness [44, 38]. These methods tend to modify traditional clustering algorithms to handle fairness criteria explicitly. Our work contributes to this latter category. Unsupervised learning methods are particularly helpful to examine the effects of

[5] `https://github.com/NGO-Algorithm-Audit/unsupervised-bias-detection`

[6] `https://algorithmaudit.eu/technical-tools/bdt/#web-app`

[7] Age and gender are not considered special categories of data, see Article 9(1) GDPR. Although, a caveat could exist [53]. The other way around, "political opinions," "trade union membership," "genetic," and "biometric" data are special categories of data but are not protected by EU non-discrimination directives.

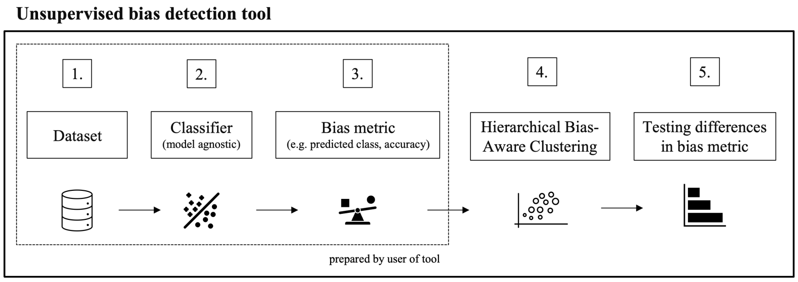


Figure 1: Schematic overview of the steps involved in applying the unsupervised bias detection tool. The required information is a dataset, a classifier's outputs, and a bias metric. Part of the dataset is used to train the Hierarchical Bias-Aware Clustering algorithm [43]. Another part of the dataset is used to test whether differences in the bias metric across clusters are statistically significant.

apparently neutral provisions that might result in indirect discrimination when demographic groups are not known beforehand.

2.3 Auditing algorithmic systems

Inspired by established practices in non-algorithmic disciplines, auditing has the potential to mitigate ethical and legal risks in algorithmic systems through both internal and external oversight [47]. Legislators across various jurisdictions are implementing regulations to impose audit frameworks for these systems. For instance, since 2023, New York City has required automated employment decision-making tools to undergo audits by independent third-party auditors [45]. Similarly, since 2024, the EU Digital Services Act mandates annual independent third-party audits for online platforms and search engines with more than 45M annual users [25]. In the coming years, the EU AI Act places predominantly self-regulatory obligations on producers and deployers of high-risk AI systems [26].

Implementing effective audits for algorithmic systems poses practical and cultural challenges. Auditors frequently encounter obstacles such as restricted data access, differing views on what constitutes a legitimate auditor [29], and uncertainty about which fairness metrics best fit the context [10]. Furthermore, auditing reports are not always required to be disclosed, reducing transparency and hindering the development of public knowledge regarding auditing standards. By discussing the audit of DUO's risk profiling algorithm, we contribute to public knowledge on best practices for auditing algorithmic systems.

3 Insights into the unsupervised bias detection tool

This section introduces the theoretical background of the unsupervised bias detection tool. We use examples from our DUO case study to clarify the methodology. We highlight that our research provides a realistic case-study of applying the tool to investigate its usefulness when one lacks protected attribute data. Our study is of both academic and practical interest since the initial algorithm is easy to apply but lacked justification, and because our data is both comprehensive and from a high-stakes policy use-case. Figure 1 provides an overview of all the steps in applying the tool. We also provide a brief overview of our second contribution, the open source bias detection tool.

3.1 Notation and problem description

Suppose that we have a total of N students in the dataset, labeled $u_1, u_2, \ldots u_N$. Each row in the dataset reflects characteristics $\boldsymbol{x}_i \in \mathbb{R}^d$ (e.g., education level, age, distance to parents) of a student u_i receiving a grant. Suppose also that we have a risk profiling algorithm $f : \mathbb{R}^d \to \{0, 1\}$ (e.g., the risk profiling algorithm used by DUO) that takes the features of a user as input and outputs whether a user should be deemed high risk of unduly being given a grant. There exists a true value of this variable of interest, denoted y_i, denoting whether the grant was, in fact, duly given (or not).

We measure the degree to which an algorithm is biased by using a bias metric M. For a user u_i, the value of the metric $m_i = M(f(\boldsymbol{x}_i), y_i)$ is some function of the prediction $f(\boldsymbol{x}_i)$ and (potentially) y_i.

To give some concrete examples, this could be a measure of performance, such as accuracy (requiring $f(\boldsymbol{x}_i), y_i$). Alternatively, it could measure demographic parity (requiring only $f(\boldsymbol{x}_i)$), equality of opportunity, etc. Without loss of generality, we assume that a higher average value for m_i for one group compared to another group indicates bias towards the former group.

For $k \subset \{1, 2, \ldots, N\}$, let $U_k = \{u_i\}_{i \in k}$ be a subset of all applicants U. We denote by $\bar{M}(U_k)$ the mean of the metric m_i among all in U_k. Our goal is to test whether the difference between $\bar{M}(U \setminus U_k)$ and $\bar{M}(U_k)$ is statistically significant. If we had access to a demographic group for which we would like to investigate bias, we would simply define the subset U_k to be, for example, different genders, races, or intersections thereof. However, since the demographic group is unavailable in our setting, we leverage clustering algorithms to separate the data into subsets. Let $\mathcal{K} = \{1, ..., K\}$ denote the set of clusters, and let $\boldsymbol{x}_{i,k}, y_{i,k}, m_{i,k}$ denote the data associated with user u_i belonging to cluster $k \in \mathcal{K}$.

3.2 The clustering algorithm

In this paper, we use the Hierarchical Bias-Aware Clustering methodology from Misztal-Radecka and Indurkhya [43], hereafter referred to as "HBAC." HBAC is an iterative clustering algorithm that produces a partition of the original dataset to isolate potentially unfairly treated groups into distinct clusters. We use HBAC as a starting point for the DUO audit, as it represents a standard method for unsupervised bias detection in recommendation systems [17]. Additionally, we propose methodological enhancements to refine the algorithm for real-world algorithmic auditing.

The HBAC algorithm first considers all observations in the dataset to belong to a single cluster and then iteratively splits the observations into distinct clusters. At each iteration, the algorithm identifies the cluster with the highest standard deviation of a specified bias metric, which serves as an indicator of imbalance between clusters. This cluster is then split into two subclusters. The process repeats for a predefined number of iterations, $max\_iterations$. The pseudocode for the algorithm is provided in Algorithm 1 in Appendix A.[8] Compared to the original algorithm from Misztal-Radecka and Indurkhya [43], the proposed method supports k-modes (enabling clustering over categorical variables) alongside k-means.

The original algorithm from Misztal-Radecka and Indurkhya [43] does not specify how to assign clusters to data points that are not part of the original training dataset. We propose using the centroids of each cluster. The datapoint is assigned to the cluster whose distance from the centroid is the smallest. In the case of k-means and k-modes, the centroids are defined as, respectively, the mean and mode for each of the characteristics $\boldsymbol{x}_{i,k}$ [34, 36]. The Euclidean distance (k-means) or Hamming distance (k-modes) is used to determine which centroid is closer.

3.3 How to test differences in bias between clusters

In the original paper, no detailed methodology is outlined for testing whether there is a statistically significant difference in the bias metric between the clusters identified by the HBAC algorithm. It is not immediately evident that the most deviating cluster, with respect to the bias metric, is statistically significantly different from data points outside this cluster. To address this, we formulate a null hypothesis that assumes no difference in bias between the most deviating cluster and the rest of the dataset. Using a bootstrapped null-distribution, we test for differences between clusters, allowing for the construction of a valid statistical hypothesis test. We further suggest, for appropriate statistical testing, to apply sample splitting and adjustments for multiple hypothesis testing. We define the aforementioned tests in Appendix D. Following that, in Appendix E, we illustrate the test results showing that the combination of the Bonferroni correction [33] and sample splitting prevents us from wrongly concluding that there is a difference in the bias metric while there is none. Thus, we propose methodological improvements to the original HBAC algorithm that minimize false bias detection when using cluster analysis.

3.4 The open-source bias detection tool

The unsupervised bias detection tool is available on Github. [9] It implements the HBAC algorithm for detecting potentially discriminated user groups, along with the methodological improvements

[8] The Appendix is available at `https://arxiv.org/abs/2502.01713`.

[9] `https://github.com/NGO-Algorithm-Audit/unsupervised-bias-detection`

described in Section 3.3. Currently, the library supports *k*-means and *k*-modes algorithms for splitting the data into clusters. The modular design of the tool ensures that additional splitting methods can be integrated with minimal effort. To enhance accessibility, the tool is embedded within a web application, [10] enabling people with no programming background to use it.

4 Experiment setup: applying the bias detection tool to DUO's risk profiling

We were granted access to ground truth labels for demographic groups in the form of aggregated statistics on students' migration backgrounds from Statistics Netherlands [51]. This case study is therefore well-suited to evaluate the effectiveness of the unsupervised bias detection tool in detecting bias from DUO's risk profiling algorithm. If these ground-truth group-level data had not been available, the unsupervised bias detection tool discussed in our paper could have served as a method to examine *potential* bias in the DUO verification of whether a student lived at their stated address.

4.1 Description of the DUO case study

From 2012 to 2023, DUO aimed to verify that college grants were allocated to students who genuinely lived away from their parents. Some students claimed to have an independent living status to qualify for a grant, even though they did not actually reside at the address they provided [20]. The process used to verify whether a student lived at the stated address is known as the *CUB process* (Controle Uitwonendenbeurs) which translates to "Check Non-Resident Grant." As part of the CUB process, students were assigned a risk score ranging between 0 and 180 by a rule-based algorithm (further details are in Appendix B).[11] This score was subsequently binned into six risk categories: very high (1), high (2), average (3), low (4), very low (5) to unknown (6).[12] Students in categories (1) and (2) were considered "high risk" of being unduly allocated the grant. Based on the assigned risk category, human analysts of DUO followed operational instructions to subject students to an investigation. Although students from all risk categories could potentially be investigated, the majority of investigations focused on higher-risk categories [20].

After concerns were raised regarding potential bias in the CUB process affecting students with a non-European migration background, DUO requested Statistics Netherlands to provide aggregated statistics [13] on students' countries of origin and countries of birth [14] who received a college grant for the years 2014, 2017, 2019, 2021, and 2022 [51]. From these data, the non-European migration background of students can be retrieved in aggregate. We do not have individual-level data on students' migration backgrounds.

This paper focuses on the years 2014 and 2019, covering a total of 298,882 students who received college grants. We specifically focus on the CUB population from 2014 ($n = 248,649$) as this was the last year all types of students—MBO 1-2 and MBO 3-4 (vocational education), HBO (higher professional education), and WO (university education)—were awarded a college grant. We do not combine data from the different years since in 2019 only MBO students were eligible for a college grant. The results are replicated for the CUB 2019 dataset ($n = 50,233$ students) in Appendix C.

DUO's risk profiling algorithm differentiated on the basis of three variables. First, the type of education the student is following, categorized as MBO 1-2, MBO 3-4 (vocational education), HBO (higher professional education), and WO (university education). Second, the age of the student, categorized as: 15-18, 19-20, 21-22, 23-24, and 25-50 years. Third, the distance between the address of the parents and the student, measured through the 6-digit postal code of the registered addresses and categorized as: 0km, 1m-1km, 1-2km, 2-5km, 5-10km, 10-20km, 20-50km, 50-500km, and unknown. We note that 0km indicates that the student and their parent(s) share the same postal code but not the same house number.

[10] `https://algorithmaudit.eu/technical-tools/bdt/#web-app`

[11] The Appendix is available at `https://arxiv.org/abs/2502.01713`.

[12] Risk category 6 includes students with a risk score of 0. Risk categories 5, 4, 3, 2, and 1 correspond to students with risk scores in the ranges of 1-19, 20-39, 40-59, 60-79, and 80-180 respectively (see Appendix B).

[13] Through protocols established by Statistics Netherlands, the aggregated population statistics can be computed securely. Data were analyzed in a privacy-preserving manner within a secure environment. Groups are rounded to the nearest ten. Results pertaining to groups smaller than 10 people are not published.

[14] Students with a non-European migration background are defined as those who were either born outside the Netherlands and Europe [50], or who have at least one parent born outside these regions.

In 2014, 34,050 students were assigned the risk category 6 (unknown), e.g., because the distance to their parents was unknown. We exclude these students from the clustering analysis, since the algorithm did not make a prediction on their risk of undue allocation of the grant, and these cases were deprioritized in the control process [20]. Clustering is applied to the remaining dataset of $n = 214,599$ students.

4.2 Applying the unsupervised bias detection tool to the CUB 2014 dataset

Data preprocessing: We transform the 3 categorical profiling characteristics to one-hot encoded variables. We define no ordinal relationship between categories. We use the k-modes clustering within Algorithm 1 as defined in Appendix A.

Bias metric: As stated before, our metric requires that the probability of being classified as "high risk" by the risk profiling algorithm (based on $f(x_i)$) is equal for students with a European migration background and students with a non-European migration background. Thus, if these probabilities are equal, this condition satisfies our bias metric and corresponds to demographic parity [21]. We consider this to be an appropriate bias metric, as classifications by the risk profiling algorithm contributed to discrimination in the CUB process [19]. In addition, the assessment of the adverse effect of the risk profiling algorithm on different demographic groups is a crucial aspect to investigate indirect discrimination under European non-discrimination law. An alternative definition of bias could involve examining whether college grants were duly awarded (y_i). However, ground truth labels indicating actual outcomes are only available for a small subset of students who underwent manual review by DUO (n=3,179). Consequently, we use predicted risk scores ($f(x_i)$) instead, enabling the application to a substantially larger population (n=298,882). We emphasize that the HBAC algorithm can be used for any bias metric of interest.

Hyperparameters: The minimum number of samples per cluster is selected by iterating over 5,000, 10,000, 20,000, and 30,000 samples and selecting the one which maximizes the average Calinksi-Harabasz index [5],[15] measured via 5-fold cross-validation. We set the maximum number of iterations sufficiently high ($max\_iterations = 1,000$), such that Algorithm 1 in Appendix A continues until the minimum number of samples per cluster is reached. The initial centroids are determined via the method proposed by Cao et al. [6].

Testing procedure: We split our dataset and reserve 20% of it for testing the differences in the bias metric per cluster. We fit the unsupervised bias detection tool (and determine hyperparameters) on the remaining 80% of the data. Since our bias metric is binary, we use a Pearson χ^2-test [46].

5 Bias audit results and discussion of DUO's risk profiling algorithm

In this section, we outline the findings of our bias audit for the CUB 2014 dataset. We include results from the CUB 2019 dataset in Appendix C.[16] Through analyzing both CUB datasets, we show that the unsupervised bias detection tool detected bias in the risk profiling algorithm that aligned with the indirect discrimination of non-European migrants found in the initial audit.

5.1 Analyzing identified clusters

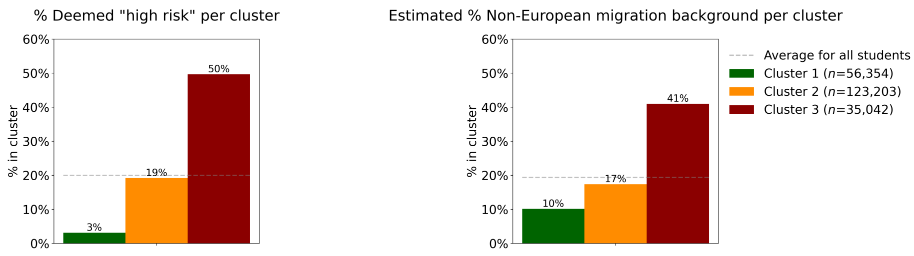
The unsupervised bias detection tool returns three clusters for the CUB 2014 dataset. The characteristics of these clusters are analyzed below.

Cluster size: Cluster 1 consists of $56,354$ (26.3%) students, where cluster 2 consists of $123,203$ (57.4%) students and cluster 3 of $35,042$ (16.3%) students.

Testing the difference in bias metric per cluster: We find stark differences in the bias metric across the three clusters. In the cluster with the highest bias (cluster 3) as shown in Figure 2a, 50% of the students are classified as "high risk" by the risk profiling algorithm, versus 3% and 19% respectively in clusters 1 and 2. These differences are statistically significant at the 0.1% level, as

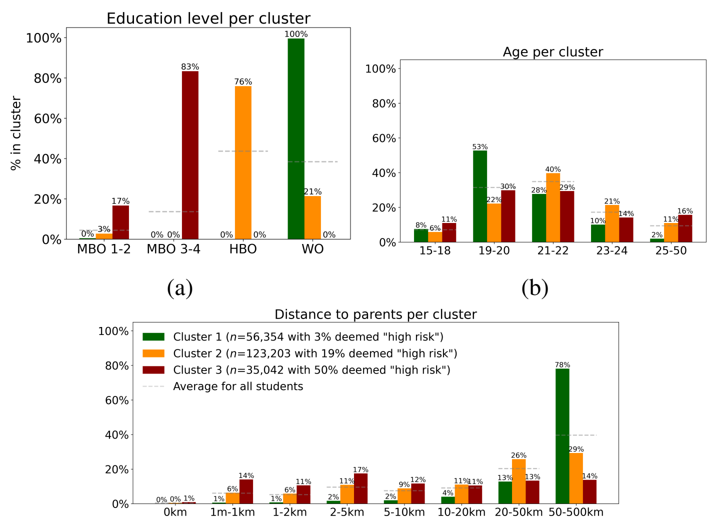
[15] This index measures the ratio of the *between* cluster variance over the *within* cluster variance. Intuitively, we want the clusters to have little variation in the bias metric *within* each cluster, but large variation *between* the clusters. This means that the data points are well separated into distinct clusters.

[16] The Appendix is available at `https://arxiv.org/abs/2502.01713`.



(a) (b)

Figure 2: The percentage of students (a) deemed as "high risk" and (b) estimated as having a non-European migration background for identified clusters based on the student population of 2014, excluding students for which the risk profile was unknown ($n = 214,599$).



(a) (b)

(c)

Figure 3: The percentage of students in each cluster for the three characteristics used in the risk profiling algorithm: (a) type of education, (b) age, and (c) distance to parents within the student population of 2014, excluding students for which the risk profile was unknown ($n = 214,599$). For each subgroup, the average is indicated by a dashed line. Cluster 1 is green, cluster 2 is yellow, and cluster 3 is red.

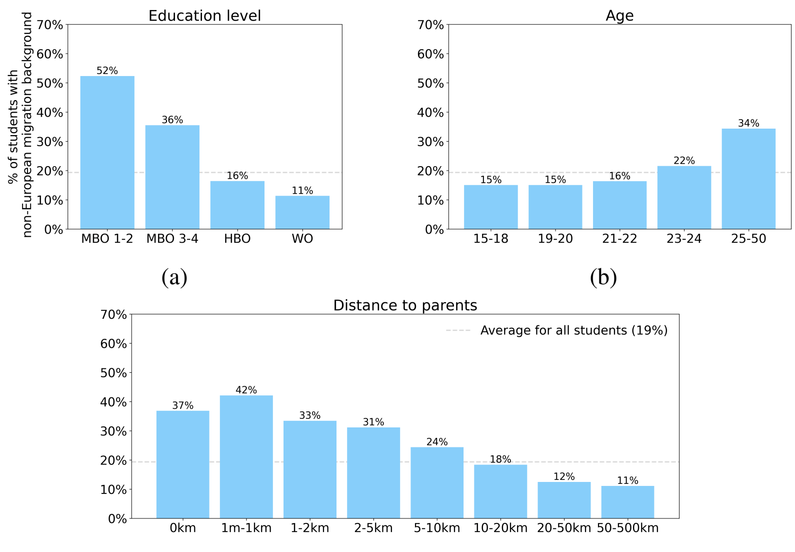
indicated in Table 3 in Appendix D, where we report differences between $\bar{M}(U \backslash U_k)$ and $\bar{M}(U_k)$ for the validation set separated for testing, including the p-values.

Student characteristics per cluster: Figure 3 illustrates the characteristics of students within each cluster. Students in cluster 3, associated with the highest average bias, are more likely to be enrolled in vocational education (MBO 1-2, MBO 3-4), be overrepresented in age groups 15-18 and 25-50, and live on average closer to their parents. In contrast, students in cluster 1, which exhibits the lowest average bias, exclusively attend university, are more likely to be 19-20 years old, and typically live far from their parents. Students in cluster 2 tend to follow higher professional education (HBO), do not display distinct patterns regarding age and live slightly more often 20-50km away from their parents.

5.2 Comparing results with aggregate statistics on non-European migration background

We now examine the extent to which the identified clusters align with the non-European migration background data, using the aggregated data provided by Statistics Netherlands [51].[17] We exclude cases where the distance is unknown, since the algorithm did not apply then.

[17]The aggregated statistics are rounded to the nearest ten, hence why there are $n = 248,650$ reported students instead of $248,649$.



(a) (b) (c)

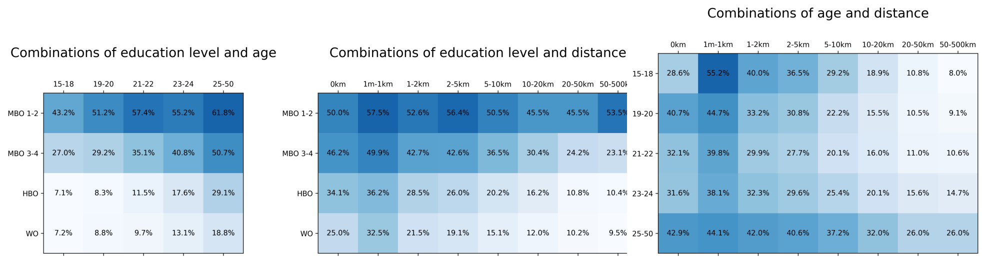
Figure 4: The percentage of students with a non-European migration background for the three used profiling characteristics: (a) type of education, (b) age and (c) distance to parents for students in the CUB 2014 dataset ($n = 248,650$); this number is rounded to the nearest 10 and excludes students where the distance is unknown.

Relationship between clusters and non-European migration background: As individual-level demographic data on students' origin are unavailable, we estimate the percentage of students with a non-European migration background per cluster using aggregate statistics. This percentage is determined using the weighted average of the proportion of students per cluster, based on the percentage of students with a non-European migration background for each combination[18] of the three characteristics [51]. There are stark differences in the proportion of students with a non-European migration background across clusters. Figure 2b shows that it is estimated that 41% of students in cluster 3 have a non-European migration background, compared to 17% in cluster 2 and 10% in cluster 1. These results indicate that clusters found by the unsupervised bias detection tool are related to the demographic group of interest.

Proxies for non-European migration background: To shed light on the relationship between clusters and non-European migration background, we investigate how characteristics of the clusters (and the risk profiling algorithm) serve as proxies for non-European migration background. This is illustrated in Figure 4. Students who are enrolled in vocational education (MBO 1-2, MBO 3-4), in the age groups 23-24 and 25-50, and living less than 10km away from their parents on average are more likely to have a non-European migration background, corresponding to the characteristics of students in cluster 3. In contrast, university students (WO), young students, and students living far from their parents are less likely to have a non-European migration background. These characteristics correspond to students in cluster 1; this also aligns with the risk profile logic: younger students and those living closer to their parents are assigned higher risk scores [20].

Combinations of profiling characteristics as proxies for non-European migration background: In addition to examining single characteristic relationships, we also investigate how specific combinations of characteristics act as a proxy for having a non-European migration background. Figure 5a highlights that young students in university education (WO) are less likely to have a non-European migration background. This result aligns with the characteristics of cluster 1, which consists exclusively of university (WO) students and contains relatively more students aged 19-20. Relatively older students (older than 23) who follow vocational education (MBO 1-2, MBO 3-4) are also more likely to have a non-European migration background.

[18]With 4 types of education, 5 age groups and 8 distance categories, there are in total 160 possible combinations. Each possible combination is assigned to a unique cluster, e.g. if two students belong to the same combination of education level, age, and distance to parents, they belong to the same cluster.



(a) (b) (c)

Figure 5: The percentage of students with a non-European migration background for three (bivariate) combinations of the used profiling characteristics: (a) type of education and age, (b) type of education and distance to parents, and (c) age and distance to parents for the CUB 2014 dataset ($n = 248,650$), excluding students where the distance is unknown.

Figure 5b indicates that students who follow vocational education (MBO) and live relatively close to their parents are more likely to have a non-European migration background. This pattern is reflected in cluster 1, which contains MBO 1-2 and MBO 3-4 students, who tend to live closer to their parents. Figure 5c shows a strong overrepresentation of students with a non-European migration background in the 15-18 age group who live 1m-1km from their parents. This pattern is somewhat reflected in cluster 3, which contains a higher proportion of 15-18 year old students and students living 1m-1km away from their parents.

5.3 Limitations

Limitations of clustering to audit algorithmic systems: The unsupervised bias detection tool alone cannot establish prohibited discrimination, as human expertise is essential to evaluate bias within the context of relevant legislation. However, the tool can serve as a starting point for such deliberative assessments. For instance, the clusters that the tool identifies could prompt further investigation into how the characteristics of individuals relate to demographic groups.

Limitations of applying the unsupervised bias detection tool to the DUO use case: There are several challenges associated with using aggregated data to evaluate the proportion of students with a non-European migration background in the clusters identified by the tool. For example, this approach requires estimating the percentage of students with a non-European migration background and prevents precise analysis of the joint relationship between the bias metric, clusters and demographic groups. However, this is an unavoidable limitation due to the sensitive nature of the data.

6 Conclusion

As demonstrated through the DUO case study, the unsupervised bias detection tool identifies clusters that correspond to groups of students found to be unequally treated by a risk profiling algorithm. The clustering results highlight proxy characteristics associated with a non-European migration background: vocational education (MBO) and living close to parents, as examined through aggregate statistics provided by Statistics Netherlands. By improving unsupervised learning methods which are behind the tool we provide, we enhance public knowledge on auditing algorithmic decision-making processes for bias, in the absence of demographic data. In addition, the unsupervised bias detection method has been implemented as an open-source tool in the form of a Python package, accompanied by a graphical interface in a web application to encourage adoption by a non-technical audience. The tool provides a scalable method that helps private entities, as well as internal and external auditors, to obtain inferences about seemingly neutral rules or practices that might embed unwanted biases. We hope that the contributions from this paper will serve the community by assisting in the detection of biases in algorithmic decision-making systems.

Acknowledgments and Disclosure of Funding

For the initial research for this project, Floris Holstege, Mackenzie Jorgensen, Kirtan Padh, and Krsto Proroković worked under a six month-long funded contract at Algorithm Audit, supported by a SIDN Fund grant for the development of digital commons (Digitale gemeenschapsgoederen). The work of Mackenzie Jorgensen was also supported by the U.K. Research and Innovation under Grant EP/S023356/1 in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (for the majority of the time she worked on this project) and by the Engineering and Physical Sciences Research Council [Grant Number EP/Y009800/1 AI KP0003] through funding from Responsible Ai UK.

References

[1] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 249–260. doi:10.1145/3442188.3445888

[2] Carolyn Ashurst and Adrian Weller. 2023. Fairness Without Demographic Data: A Survey of Approaches. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–12.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. 4–1 pages.

[5] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.

[6] Fuyuan Cao, Jiye Liang, and Liang Bai. 2009. A new initialization method for categorical data clustering. *Expert Syst. Appl.* 36, 7 (Sept. 2009), 10223–10228. doi:10.1016/j.eswa.2009.01.060

[7] Centre for Data Ethics and Innovation. 2023. *Enabling responsible access to demographic data to make AI systems fairer*. Technical Report. Department for Science, Innovation and Technology.

[8] Junyi Chai and Xiaoqian Wang. 2022. Self-supervised fair representation learning without demographics. *Advances in Neural Information Processing Systems* 35 (2022), 27100–27113.

[9] Alexandra Chouldechova and Aaron Roth. 2020. A Snapshot of the Frontiers of Fairness in Machine Learning. *Commun. ACM* 63, 5 (2020), 82–89.

[10] Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The measure and mismeasure of fairness. *The Journal of Machine Learning Research* 24, 1 (2023), 14730–14846.

[11] Council of the European Union. 2000. *Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32000L0043

[12] Council of the European Union. 2000. *Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0078

[13] Council of the European Union. 2004. *Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/dir/2004/113/oj/eng

[14] Court of Justice of the European Union. 1986. Case 170/84. Bilka - Kaufhaus GmbH v Karin Weber von Hartz. *EU:C:1986:204* (1986).

[15] Court of Justice of the European Union. 2012. Case 385/11. Isabel Elbal Moreno v. Instituto Nacional de la Seguridad Social (INSS) and Tesorería General de la Seguridad Social (TGSS). *ECLI:EU:C:2012:746* (2012).

[16] Eoin Delaney, Zihao Fu, Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2024. Oxonfair: A flexible toolkit for algorithmic fairness. *Advances in Neural Information Processing Systems* 37 (2024), 94209–94245.

[17] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2024. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction* 34, 1 (2024), 59–108.

[18] Dutch Parliament. 1994. Dutch Public Administration Law. https://wetten.overheid.nl/BWBR0005537/2025-01-01

[19] Dutch Parliament. 2024. Follow-up study CBS and Algorithm Audit into the audit of the live-away grant. (2024). https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2024Z08699&did=2024D20430

[20] Dutch Parliament. 2024. Internal investigation into the audit of the live-away grant. (2024). https://www.tweedekamer.nl/kamerstukken/detail?id=2024D07565&did=2024D07565

[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.

[22] Marc N. Elliot, Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. 2009. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv Outcomes Res Method* 9, 2 (2009), 69–83.

[23] European Parliament and the Council. 2006. *Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/dir/2006/54/oj/eng

[24] European Parliament and the Council. 2016. *Regulation (EU) 2016/679 of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of such Data (General Data Protection Regulation)*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016R0679

[25] European Parliament and the Council. 2022. *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services (Digital Services Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065

[26] European Parliament and the Council. 2024. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[27] European Parliament, the Council, and the Commission. 2000. *Charter of Fundamental Rights of the European Union (EU) (2000/C 364/01)*. Official Journal of the European Communities. https://www.europarl.europa.eu/charter/pdf/text_en.pdf

[28] Janneke Gerards and Raphaële Xenidis. 2021. Algorithmic discrimination in Europe – Challenges and opportunities for gender equality and non-discrimination law. *European Commission: Directorate-General for Justice and Consumers, Publications Office* (2021). doi:doi/10.2838/544956

[29] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. 2024. Auditing Work: Exploring the New York City algorithmic bias audit regime. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1107–1120. doi:10.1145/3630106.3658959

[30] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems* 29 (2016).

[31] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[32] Melissa HEikkilä. 2022. Dutch scandal serves as a warning for Europe over risks of using algorithms. *Politico* (2022).

[33] Yosef Hochberg and Ajit Tamhane. 1987. *Multiple comparison procedures*. Wiley, New York.

[34] Zhexue Huang. 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery* 2, 3 (1998), 283–304. doi:10.1023/A:1009769707641

[35] Rashidul Islam, Huiyuan Chen, and Yiwei Cai. 2024. Fairness without Demographics through Shared Latent Space-Based Debiasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 12717–12725.

[36] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. 2023. *Unsupervised Learning*. Springer International Publishing, Cham, 503–556. doi:10.1007/978-3-031-38747-0_12

[37] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (2018), 113–174.

[38] Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. 2021. UDIS: Unsupervised Discovery of Bias in Deep Visual Recognition Models. In *British Machine Vision Conference (BMVC)*.

[39] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. *Advances in Neural Information Processing Systems* 30 (2017).

[40] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness Without Demographics through Adversarially Reweighted Learning. *Advances in neural information processing systems* 33 (2020), 728–740.

[41] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[42] Anne Meuwese, Jurriaan Parie, and Ariën Voogt. 2024. How 'algoprudence' can contribute to responsible use of ML-algorithms. *Dutch Journal for Legal Professionals* 10 (2024).

[43] Joanna Misztal-Radecka and Bipin Indurkhya. 2021. Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems. *Information Processing & Management* 58, 3 (2021), 102519. doi:10.1016/j.ipm.2021.102519

[44] Neda Nasiriani, Anna Squicciarini, Zara Saldanha, Sanchit Goel, and Nicola Zannone. 2019. Hierarchical Clustering for Discrimination Discovery: A Top-Down Approach. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. 187–194. doi:10.1109/AIKE.2019.00041

[45] New York City Council. 2023. Automated Employment Decision Tools (AEDT). https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page

[46] Karl Pearson. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50, 302 (1900), 157–175.

[47] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 557–571. doi:10.1145/3514094.3534181

[48] John Rawls. 2001. *Justice as fairness: A restatement*. Harvard University Press.

[49] Frank Rensen. 2024. 'Bij zo'n beetje elke tegel die we lichten, ontdekken we discriminerende algoritmen bij de overheid', zegt de Autoriteit Persoonsgegevens. *de Volkskrant* (2024).

[50] Statistics Netherlands. 2024. Country of origin. *Website* (2024). https://www.cbs.nl/nl-nl/maatwerk/2024/08/landenindeling-van-de-variabele-herkomstland-2022

[51] Statistics Netherlands. 2024. Ontvangers uitwonendenbeurs; herkomst, 2014, 2017, 2019, 2021 en 2022. *Maatwerktabel* (2024). https://www.cbs.nl/nl-nl/maatwerk/2024/21/ontvangers-uitwonendenbeurs-herkomst-2014-2017-2019-2021-en-2022

[52] Marvin Van Bekkum and Frederik Zuiderveen Borgesius. 2023. Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law & Security Review* 48 (2023), 105770.

[53] Marjolein van den Brink, Peter Dunne, and Jenny Julén Votinius. 2018. *Trans and intersex equality rights in Europe–a comparative analysis*. Publications Office of the European Union.

[54] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *CHI Conference on Human Factors in Computing Systems*. 1–14.