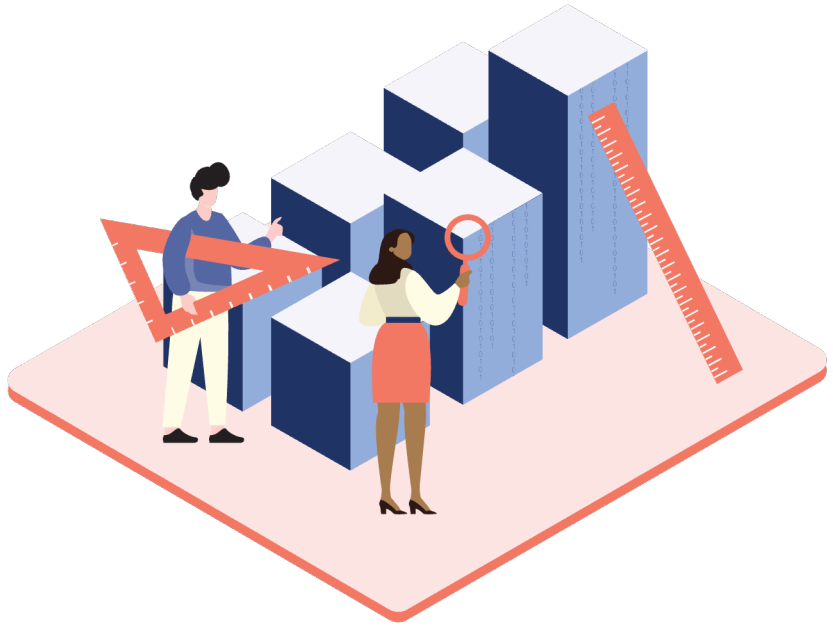
Algorithm Audit

ETH zürich

Auditing algorithmic systems in practice



real-world example

May 23rd 2025

Overview

1. What is Algorithm Audit?
2. Real-world audit: public sector risk profiling algorithm
   i. Background information
   ii. Frameworks for auditing
   iii. Supervised bias testing
   iv. Unsupervised bias testing
3. Q&A

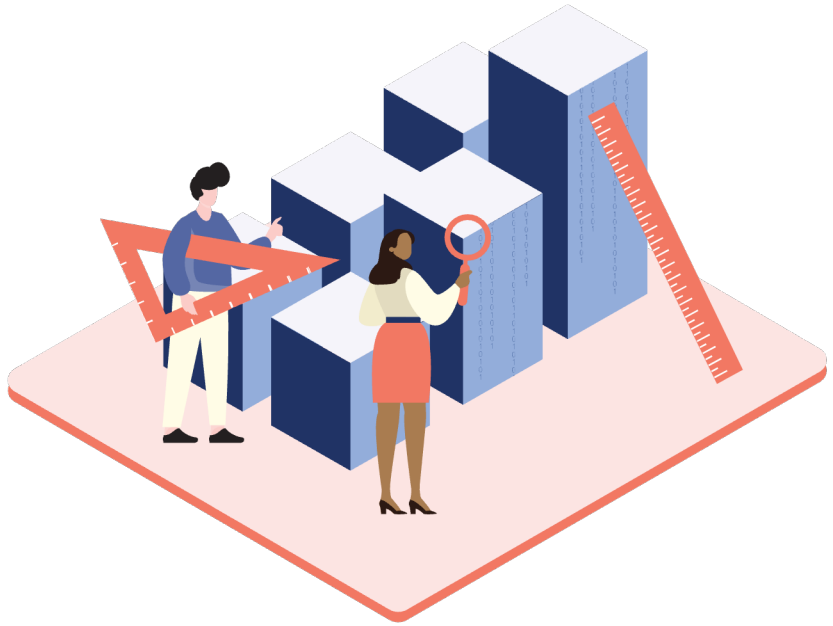
Algorithm Audit

ETH zürich

1. What is Algorithm Audit?

2. Real-world audit: public sector risk profiling algorithm

3. Q&A



Overview of core activities NGO Algorithm Audit

Core activities

Knowledge platform

Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, e.g., AI Policy Observatory and white papers

Normative advice commissions

Advising on ethical issues that arise in concrete algorithmic practice through deliberative and diverse normative advice commissions, resulting in algoprudence

Technical tools

Implementing and testing technical tools to detect and mitigate bias, e.g., bias detection tool and synthetic data generation

Project work

Support for specific questions from public and private sector organisations regarding responsible use of AI

Collaborating with

Council of Europe / Conseil de l'Europe
CEN
StandICT.eu
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties
Ministerie van Justitie en Veiligheid
Dienst Uitvoering Onderwijs, Ministerie van Onderwijs, Cultuur en Wetenschap
Gemeente Amsterdam
City of Utrecht
gemeente Deventer
European Artificial Intelligence & Society Fund
SIDN fonds

Algorithm Audit

ETH zürich

1. What is Algorithm Audit?

2. **Real-world audit: public sector risk profiling algorithm**

3. Q&A

i. Background information

Background information

- **Jun'23:** News article claiming approx. 97% of students in appeal procedure (sample n≈300) have a migration background
- **Political action:** Minister suspends risk profiling algorithm
  - External investigation, PwC
  - Internal investigation, Algorithm Audit
- **Focus of audit**
  - Qualitative: interviews with employees and document due diligence
  - Quantitative: data analysis (population statistics)

Studenten met migratieachtergrond opvallend vaak beschuldigd van fraude, minister wil systeem grondig nagaan

Sumeyye Ersoy

Salwa van der Gaag

Onderwijsminister Robbert Dijkgraaf wil "grondig nagaan" of de fraudecontroles door Dienst Uitvoering Onderwijs (DUO) met studiefinanciering "wel echt eerlijk zijn". De minister reageert daarmee op een onderzoek van NOS op 3 en Investico.

Studenten met een migratieachtergrond worden opvallend vaak beschuldigd van fraude met studiefinanciering, is het beeld van een rondgang langs ruim dertig advocaten die studenten bijstaan in bezwaar- en beroepsprocedures tegen DUO. De dienst is verantwoordelijk voor alle studiefinanciering en leningen aan studenten.

De afgelopen tien jaar ondersteunden de advocaten in totaal 376 studenten die beschuldigd werden van frauderen met de uitwonende beurs. Bij 367 van hen, in vrijwel alle gevallen dus, ging het om studenten met een migratieachtergrond. Dijkgraaf spreekt van een "verontrustend signaal" in een

Algorithm Audit

Step 0 – Start population

Student population living away from parent(s)

Step 1 – Risk profile

- Type of education
- Age
- Distance to parent(s)

high risk

180

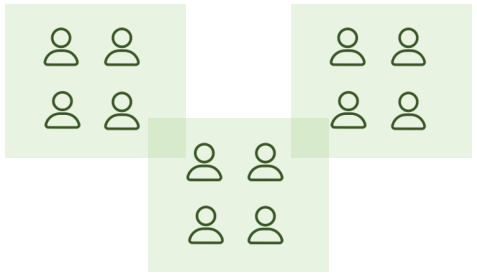
60

...

0

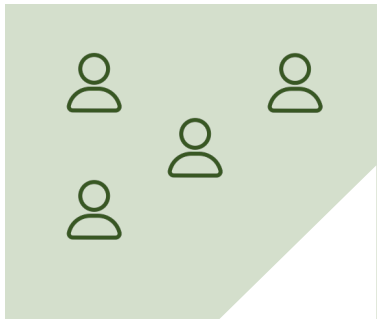
risk scores

Step 2 – Division per region



Step 3 – Desk research: Human analyst

Step 4 – External: House visit



Step 5 – Decision and next steps

Unduly population

Step 6 – Appeal procedure

Appeal population 97,6%

ii. Frameworks for auditing

Overview of candidate frameworks

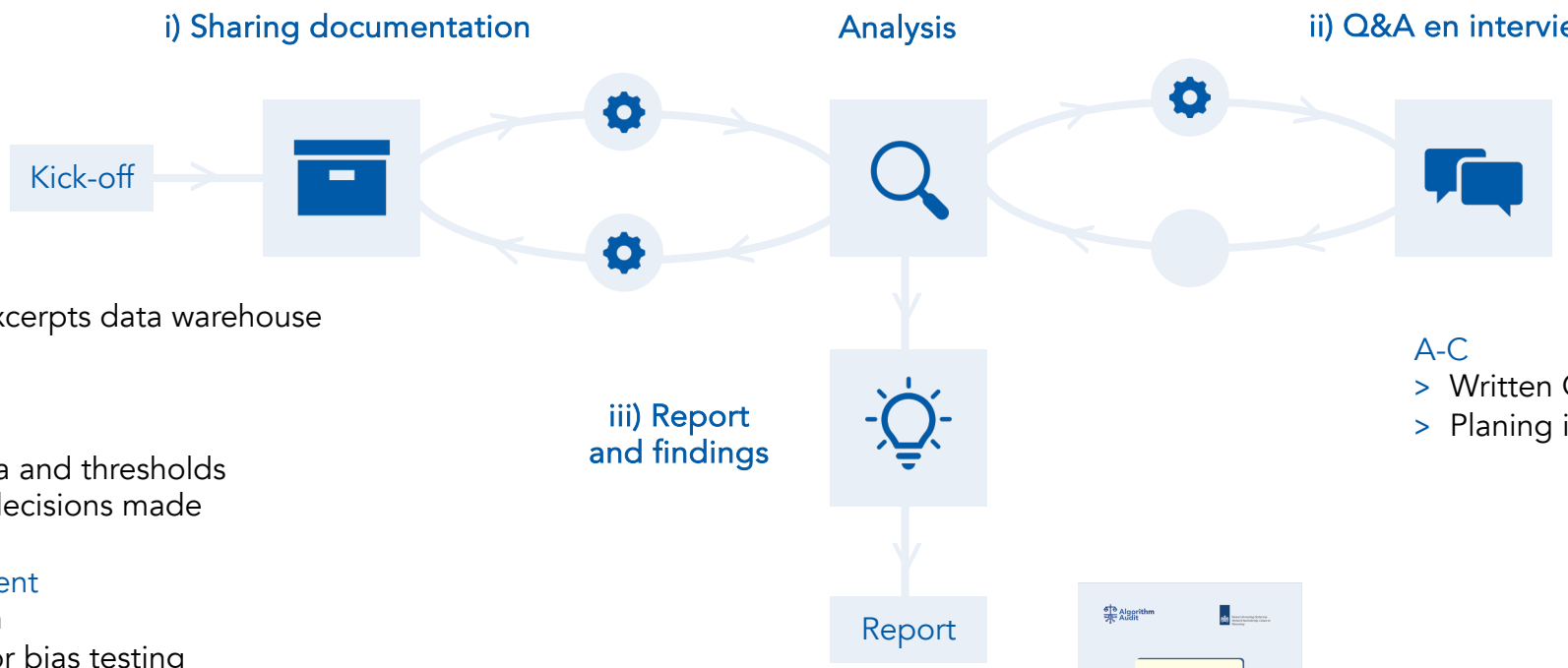
Framework (selection)	Used for research
CEN-CENELEC AI Act standaarden	
ISO 42001 standaard Management AI	
Toetsingskader algoritmes (2021) AR	
Handreiking non-discriminatie by design (2021) BZK *Discriminatie door risicoprofielen* CvRM (2021)	
Onderzoekskader algoritmes (2023) ADR	
Impact Assessment Mensenrechten en Algoritmes (IAMA) (2021) BZK + Utrecht Data School	
Algoritmekader BZK	not yet available

Example of structured review

1. *Sturing & Verantwoording*
2. *Privacy*
3. *Data & Model*
4. *Informatiebeveiliging*

Deelgebied	Risico	#	Beheersmaatregelen	Bron
Doelstelling	Algoritme functioneert niet in lijn met geformuleerde doelstellingen.	DM.1	De doelstelling van het algoritme is concreet uitgewerkt tot functionele eisen voor het algoritme. De mate waarin aan deze eisen is voldaan is bepaald.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.1 t/m 1.6
Prestatie	Het model presteert suboptimaal voor de taak die uitgevoerd moet worden.	DM.2	De keuze voor het model en de hyperparameters zijn beargumenteerd en vastgelegd.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2
	Het model wordt toegepast terwijl niet aan de voorwaarden voor het model voldaan wordt. Prestatie zoals op de testset is niet gegarandeerd.	DM.3	De grenzen van de toepasbaarheid van het model zijn gedocumenteerd. De voorwaarden waaronder het model gebruikt kan worden en waaronder niet, zijn duidelijk.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.1, 1.2
	Bij het in productie nemen van het model of bij latere evaluatie is het niet duidelijk of het model voldoende presteert.	DM.4	De functionele eisen zijn uitgewerkt tot adequate en meetbare prestatiecriteria. De gestelde criteria zijn behaald.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.1 t/m 1.6
	De prestatie van model lijkt hoger dan het in werkelijkheid is.	DM.5	De train-, test- en validatieset zijn gescheiden.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2, 2.1§100
	Door onjuiste training van het model presteert het model in de praktijk minder goed dan bij de tests.	DM.6	Bij de keuze voor training- en testdata in de ontwikkelfase is gelet op under- en overfitting.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2, 2.1§100
	Het model presteert in productie niet goed door niet representatieve trainings-/testdata.	DM.7	De doelpopulatie is vastgesteld. Er is gecontroleerd dat de testdata representatief is voor de data van de verschillende subgroepen die in de productiedata voorkomen.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2, 2.1§100
	Door veranderingen in de data presteert het model niet meer zoals verwacht.	DM.8	De output en performance van het model worden geëvalueerd bij veranderingen in de data.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2, 1.3

Overview of due dilligence cycle



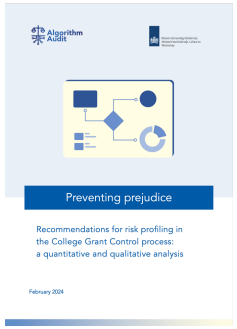
A. Source data
> Access to csv excerpts data warehouse
> Data dictionary

B. Risk profile
> Selection criteria and thresholds
> Motivation for decisions made

C. Bias measurement
> Null distribution
> Methodology for bias testing

Q&A v1
30 questions

A-C
> Written Q&A
> Planing interviews



Example of structured review

1. *Sturing & Verantwoording*
2. *Privacy*
3. *Data & Model*
4. *Informatiebeveiliging*

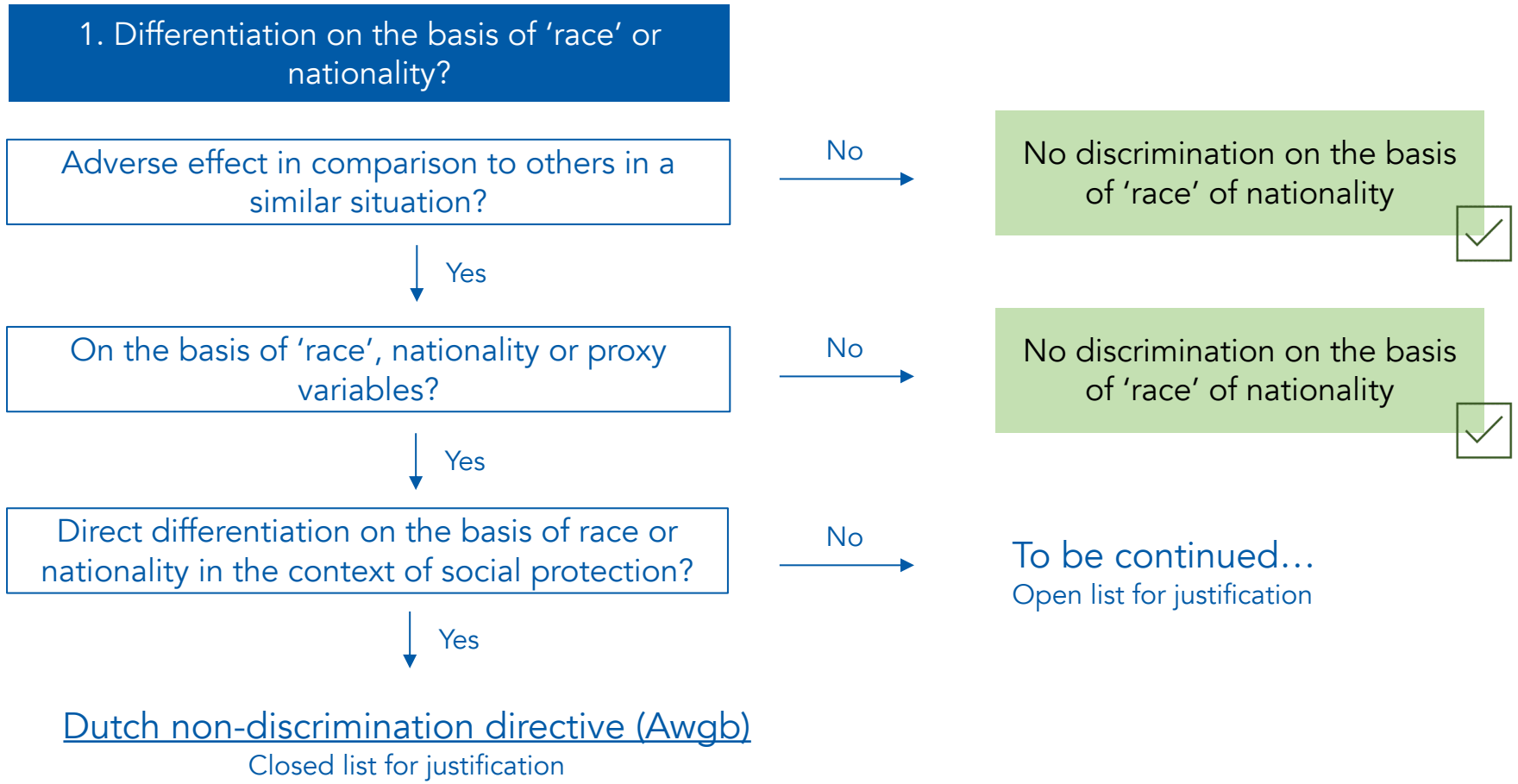
Deelgebied	Risico	#	Beheersmaatregelen	Bron	
Doelstelling	Algoritme functioneert niet in lijn met geformuleerde doelstellingen.	DM.1	De doelstelling van het algoritme is concreet uitgewerkt tot functionele eisen voor het algoritme. De mate waarin aan deze eisen is voldaan is bepaald.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.1 t/m 1.6	~
Prestatie	Het model presteert suboptimaal voor de taak die uitgevoerd moet worden.	DM.2	De keuze voor het model en de hyperparameters zijn beargumenteerd en vastgelegd.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2	✗
	Het model wordt toegepast terwijl niet aan de voorwaarden voor het model voldaan wordt. Prestatie zoals op de testset is niet gegarandeerd.	DM.3	De grenzen van de toepasbaarheid van het model zijn gedocumenteerd. De voorwaarden waaronder het model gebruikt kan worden en waaronder niet, zijn duidelijk.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.1, 1.2	✓
	Bij het in productie nemen van het model of bij latere evaluatie is het niet duidelijk of het model voldoende presteert.	DM.4	De functionele eisen zijn uitgewerkt tot adequate en meetbare prestatiecriteria. De gestelde criteria zijn behaald.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.1 t/m 1.6	✗
	De prestatie van model lijkt hoger dan het in werkelijkheid is.	DM.5	De train-, test- en validatieset zijn gescheiden.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2, 2.1§100	nvt
	Door onjuiste training van het model presteert het model in de praktijk minder goed dan bij de tests.	DM.6	Bij de keuze voor training- en testdata in de ontwikkelfase is gelet op under- en overfitting.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2, 2.1§100	✗
	Het model presteert in productie niet goed door niet representatieve trainings-/testdata.	DM.7	De doelpopulatie is vastgesteld. Er is gecontroleerd dat de testdata representatief is voor de data van de verschillende subgroepen die in de productiedata voorkomen.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2, 2.1§100	✓
	Door veranderingen in de data presteert het model niet meer zoals verwacht.	DM.8	De output en performance van het model worden geëvalueerd bij veranderingen in de data.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.2, 1.3	✗

Example of structured review (c'tnd)

- 1. *Sturing & Verantwoording*
- 2. *Privacy*
- 3. *Data & Model*
- 4. *Informatiebeveiliging*

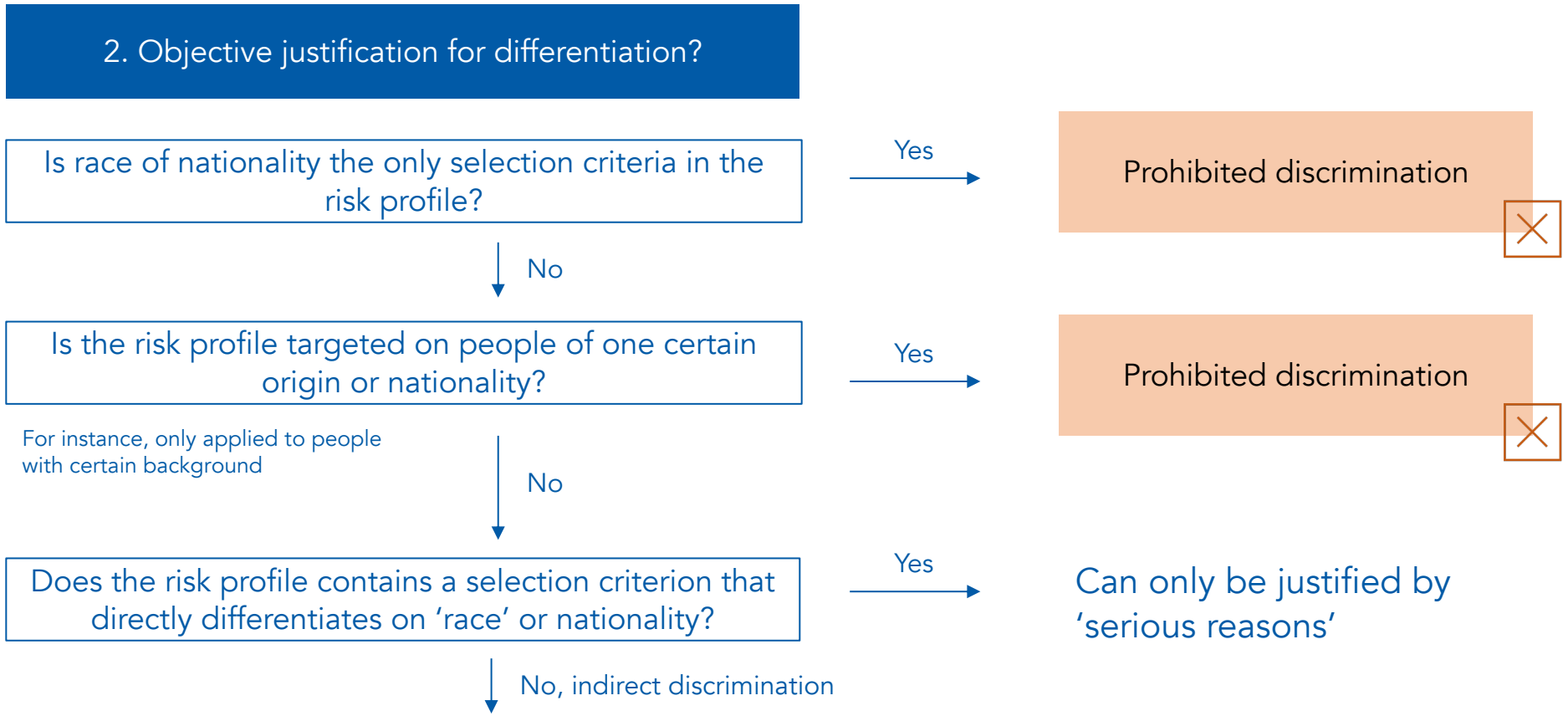
Deelgebied	Risico	#	Beheersmaatregelen	Bron	
Bias en discriminatie	Het model creëert onwenselijke systematische afwijking voor specifieke personen, groepen of andere eenheden (bias/discriminatie)	DM.16	De definitie van de verschillende groepen en de gewenste prestatie van het model voor deze groepen zijn opgenomen in de functionele eisen.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.5	✗
		DM.17	De mate van geaccepteerde bias in de uitkomst is opgenomen in de functionele eisen en uitgewerkt in meetbare prestatiecriteria.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.5	✗
		DM.18	De methoden om bias te voorkomen, detecteren en corrigeren zijn vastgelegd.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.5	✗
		DM.19	De mate van bias in de data, dataverzameling en het model zijn in kaart gebracht.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.5	✗
		DM.20	Tijdens de ontwikkeling van het model is beoordeeld of er een verschil bestaat tussen de prestatie van het model tussen verschillende subgroepen. De prestatiemetrieken afleidbaar uit de confusionmatrix zijn vergeleken voor deze subgroepen.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.5	✗
		DM.21	De uitkomstbias van productiedata is beoordeeld voor de verschillende subgroepen en voldoet aan de prestatiecriteria.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.5	✗
	Bias in het algoritme leidt tot discriminatie.	DM.22	Bij de geconstateerde bias is beoordeeld of deze op discriminatie duidt.	EC/AI HLEG April 2019 - Hoofdstuk II. 1.1, 1.5	✗

Legal framework: EU non-discrimination law in a nutshell



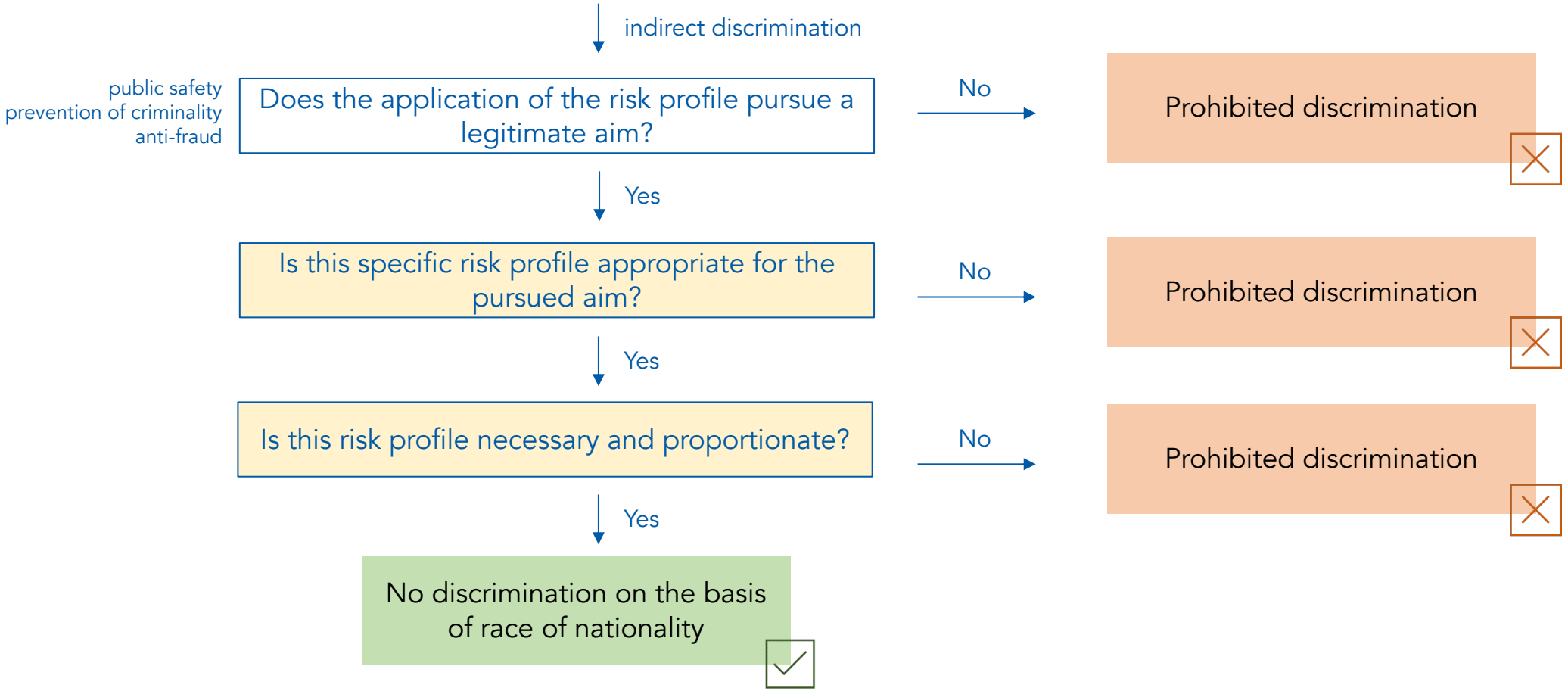
Source: Netherlands Institute for Human Rights, Discrimination through risk profiling (2025)
https://publicaties.mensenrechten.nl/publicatie/4093c026-ae41-4c1d-aa78-4ce0e205b5de

Legal framework: EU non-discrimination law in a nutshell (c'tnd)



Source: Netherlands Institute for Human Rights, Discrimination through risk profiling (2025)
https://publicaties.mensenrechten.nl/publicatie/4093c026-ae41-4c1d-aa78-4ce0e205b5de

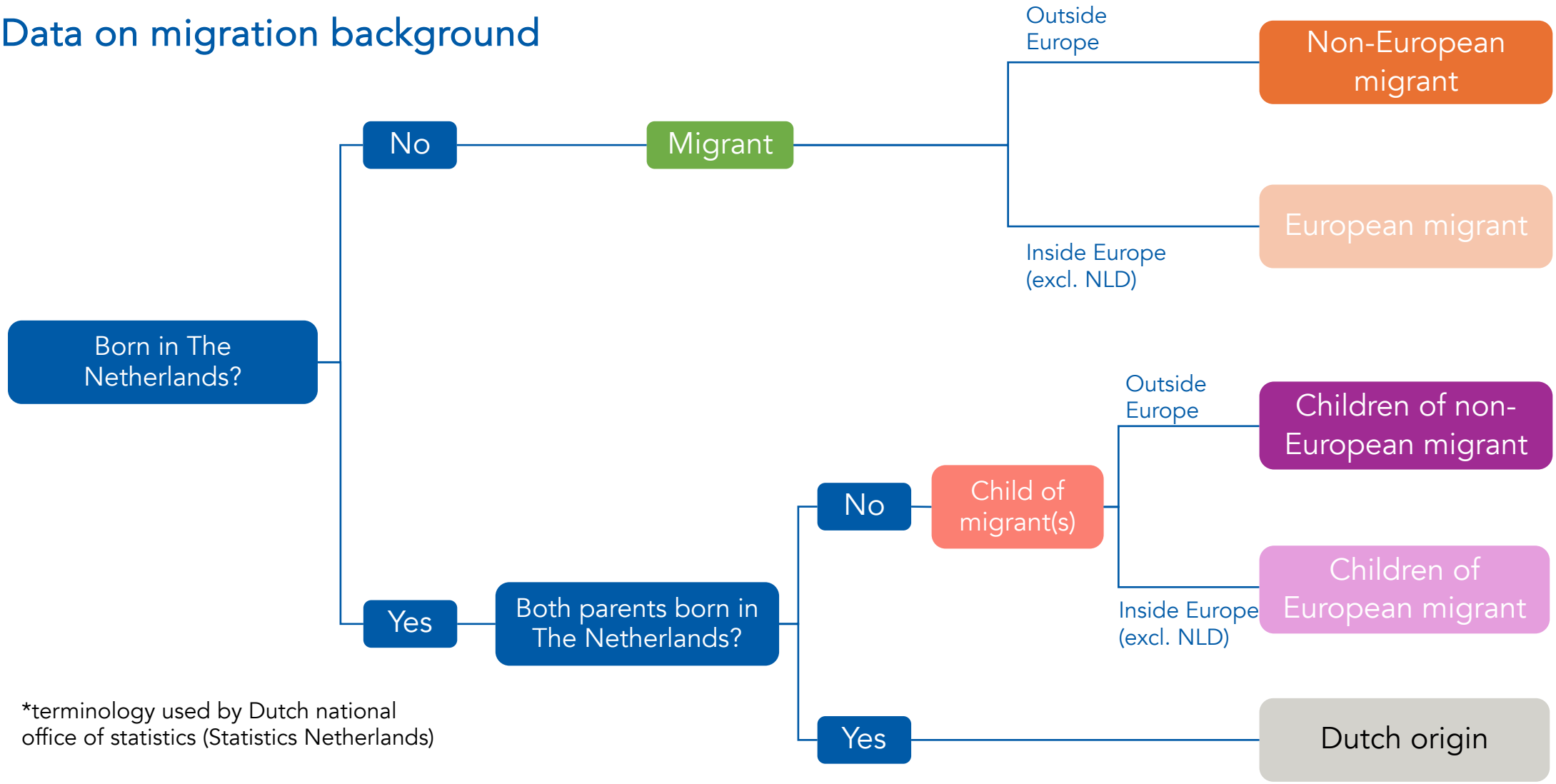
Legal framework: EU non-discrimination law in a nutshell (c'tnd)



Source: Netherlands Institute for Human Rights, Discrimination through risk profiling (2025)
https://publicaties.mensenrechten.nl/publicatie/4093c026-ae41-4c1d-aa78-4ce0e205b5de

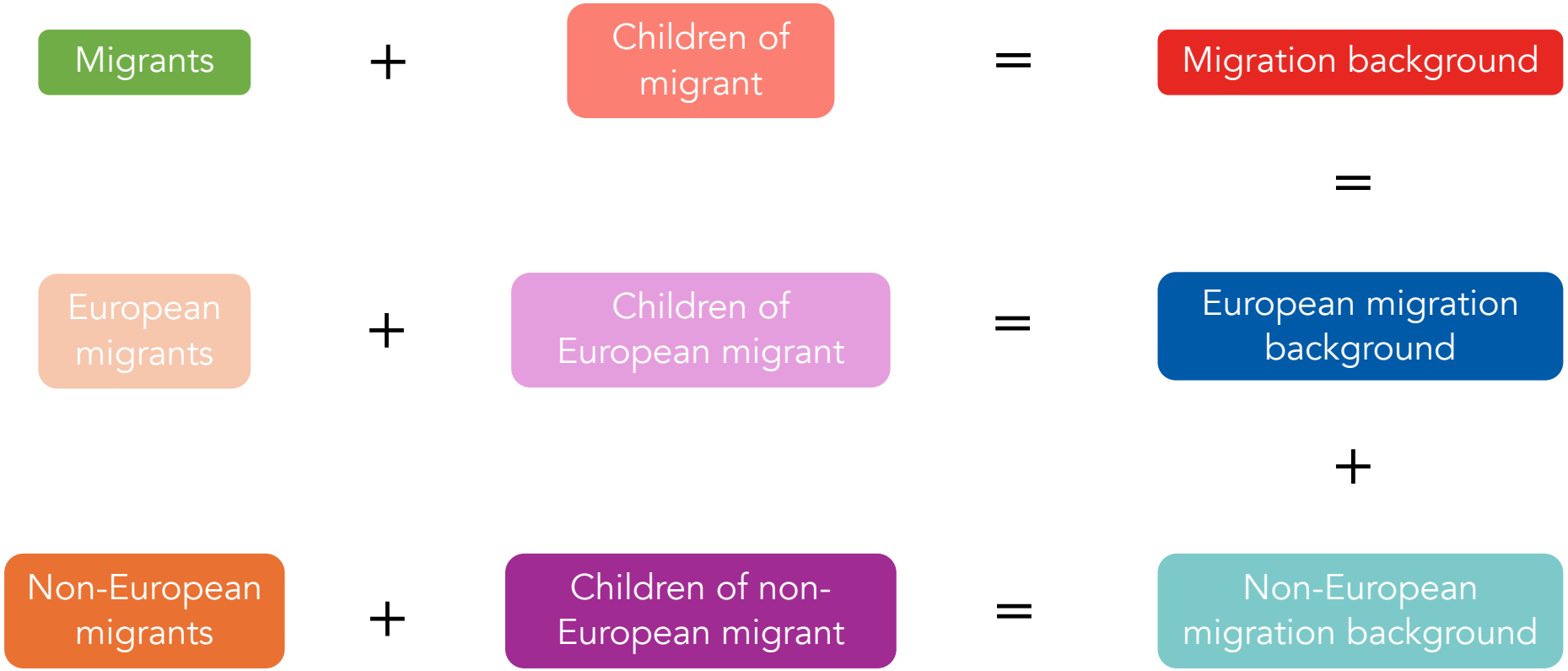
iii. Supervised bias testing

Data on migration background



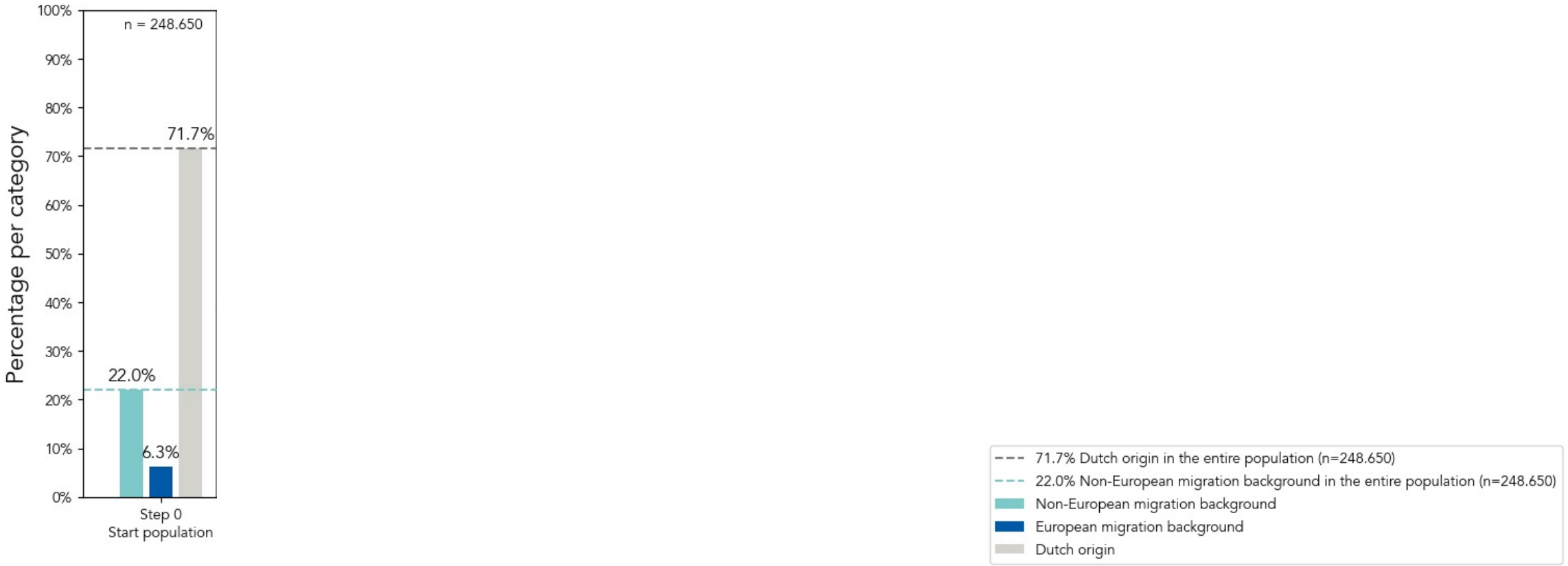
*terminology used by Dutch national office of statistics (Statistics Netherlands)

Data on migration background



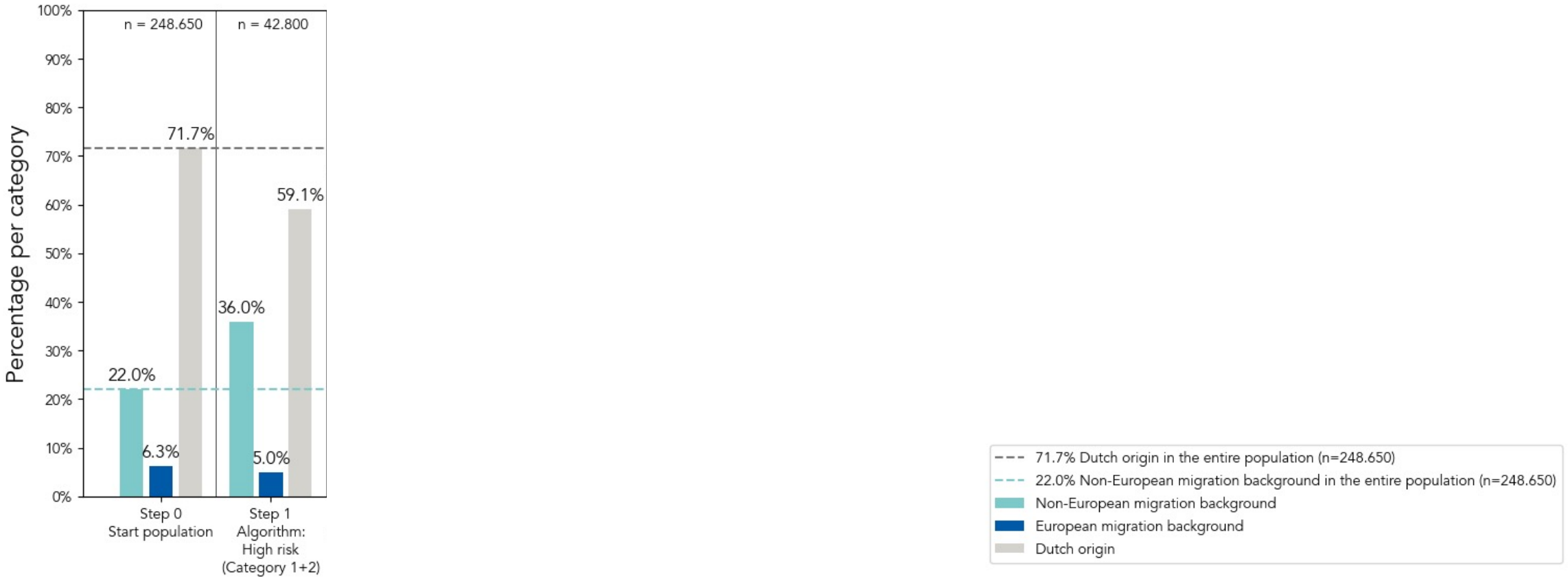
Algorithm Audit

Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2014 (n=248.650)

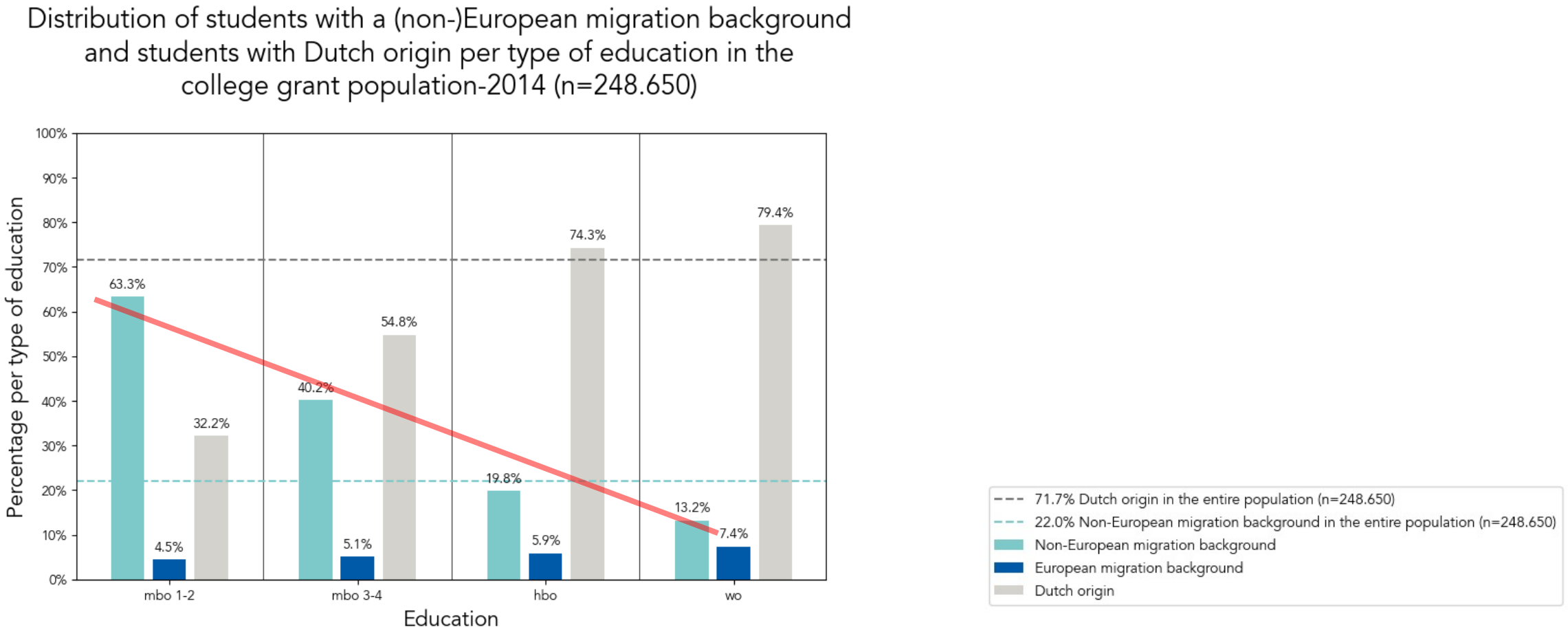


Algorithm Audit

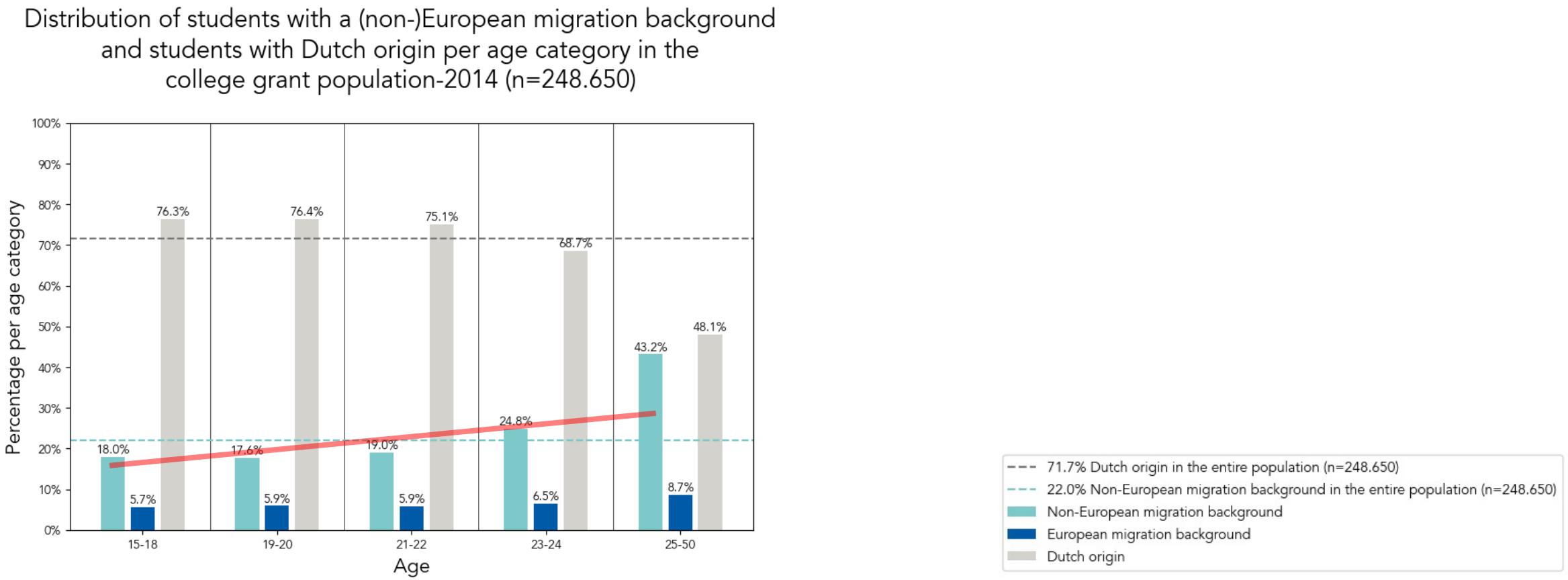
Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2014 (n=248.650)



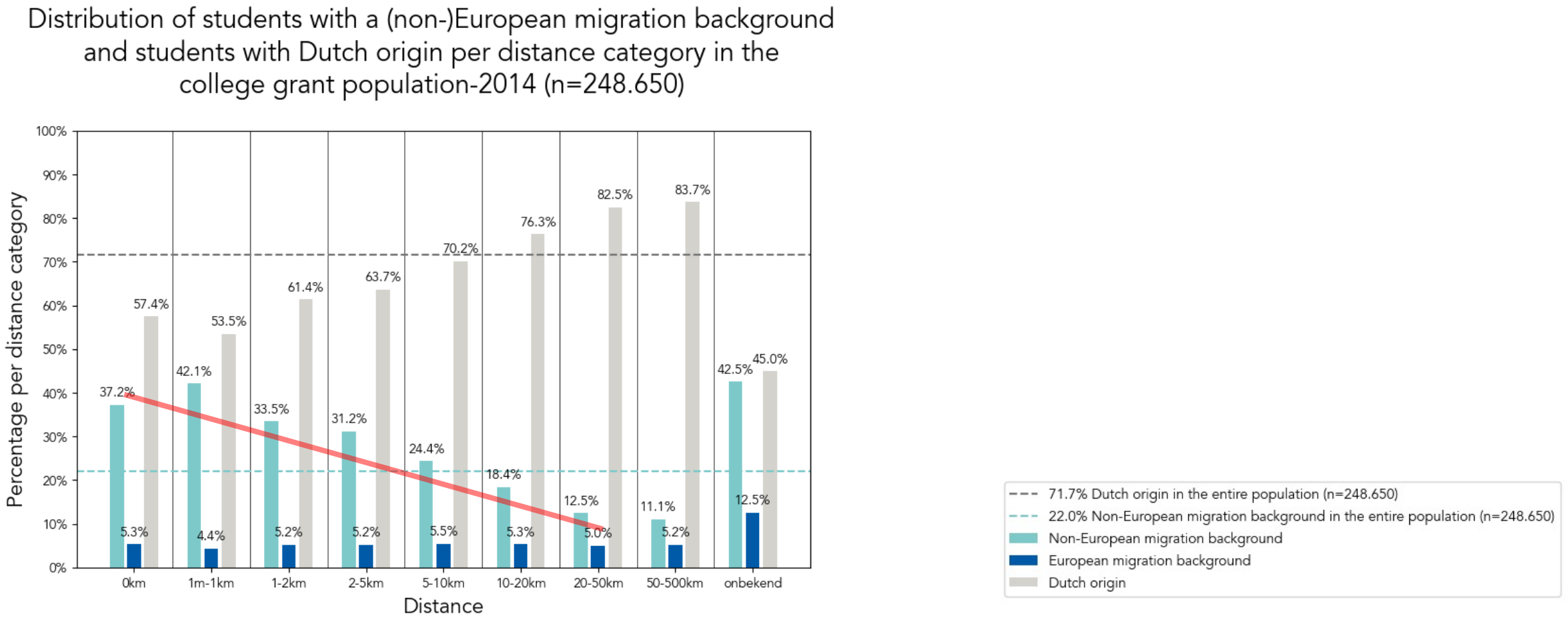
Proxy nature of 'type of education' with respect to non-European migration background



Proxy nature of 'age' with respect to non-European migration background



Proxy nature of 'distance to parent(s)' with respect to non-European migration background

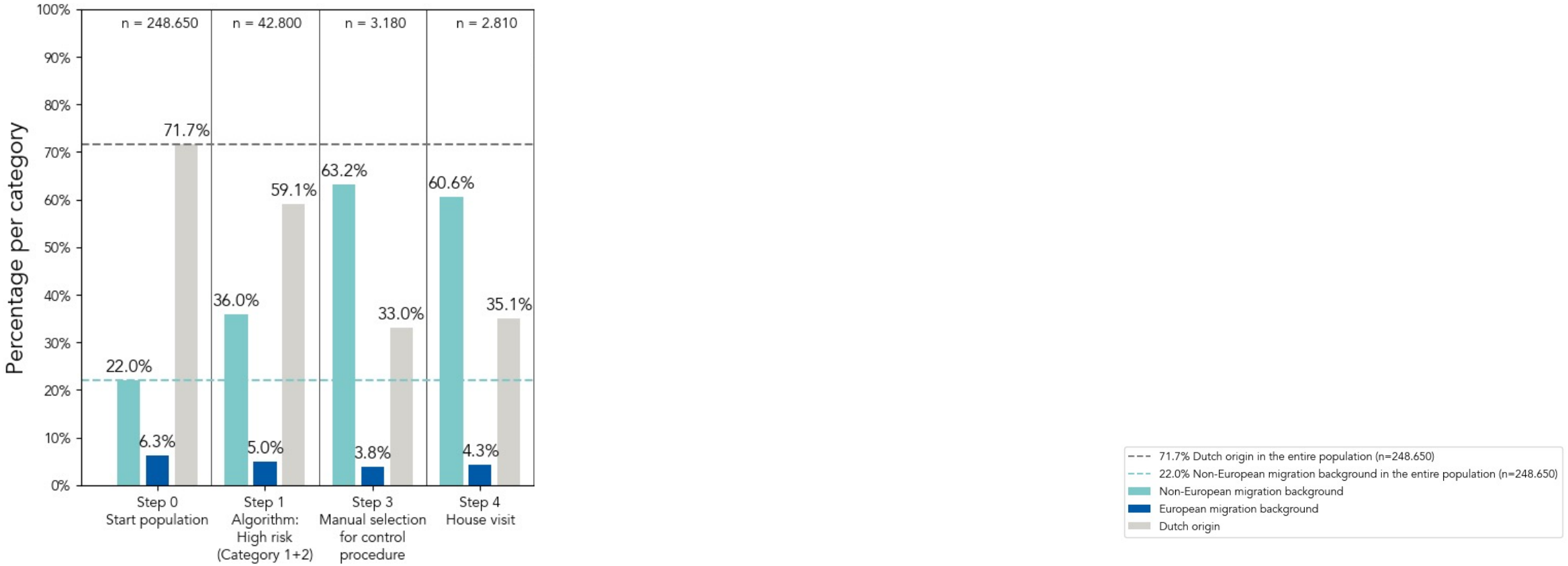


Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2014 (n=248.650)

Algorithm Audit

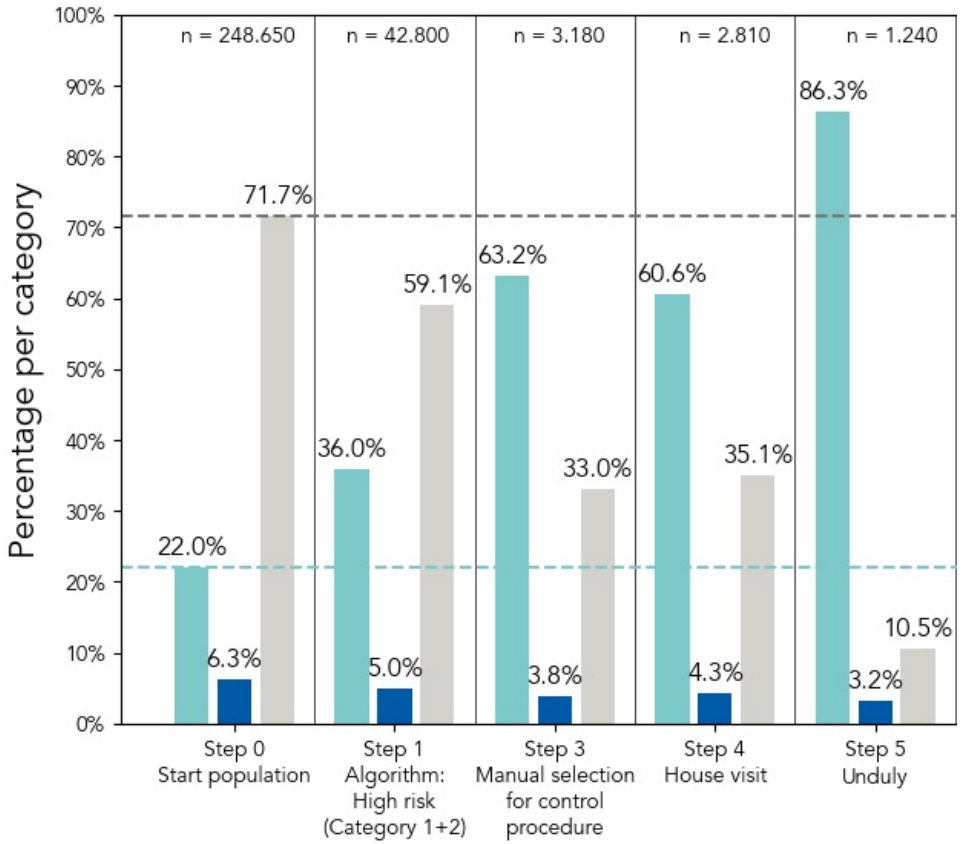
Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2014 (n=248.650)

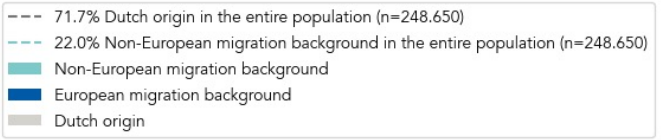
Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2014 (n=248.650)



Algorithm Audit

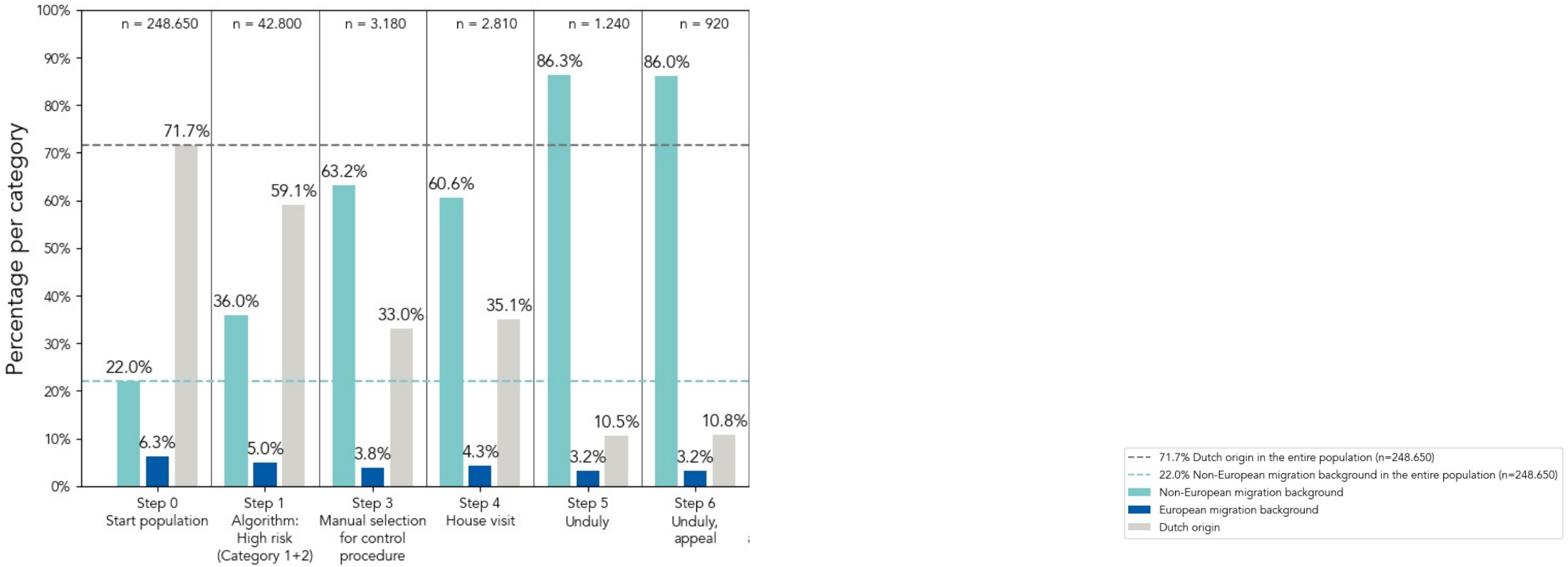
Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2014 (n=248.650)





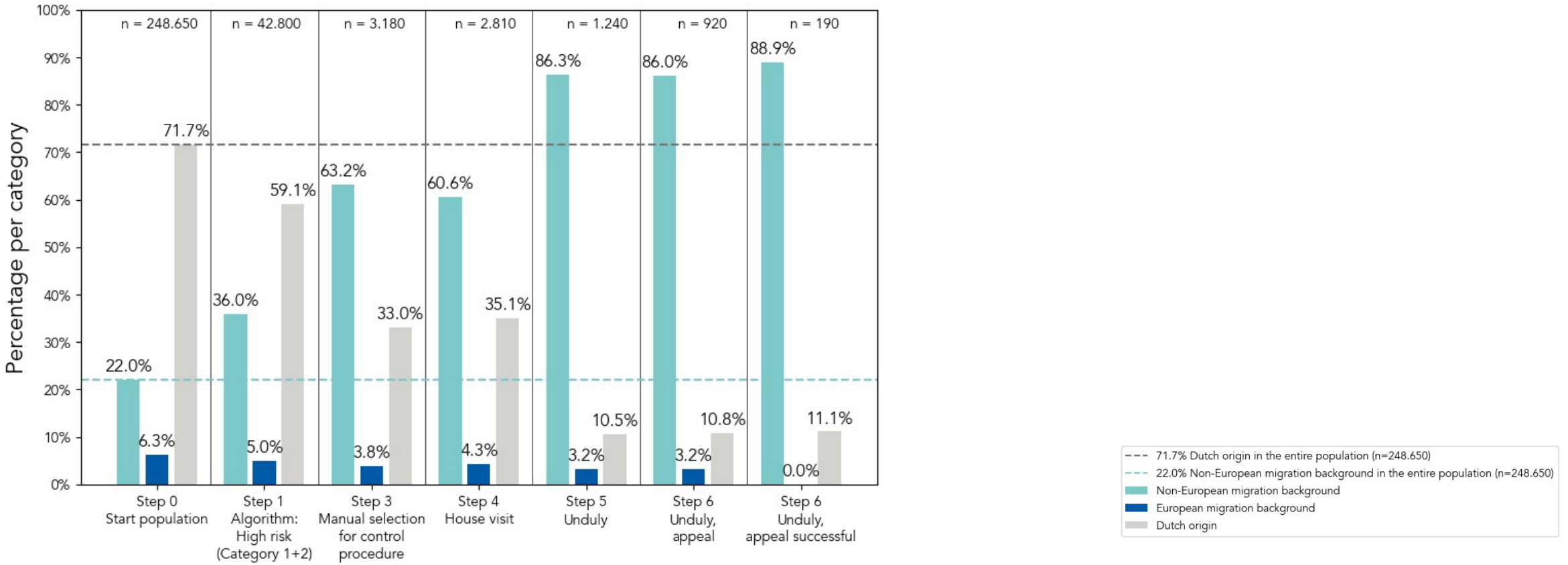
Algorithm Audit

Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2014 (n=248.650)



Algorithm Audit

Distribution of students with a (non-)European migration background and students with Dutch origin per step of the CUB process for the college grant population-2014 (n=248.650)



Based on quantitative insights controlling state actors were able to take a normative position

Apologies to disadvantaged students by Dutch state

- Announced by Minister of Education, Culture and Science in Nov'24
- Compensation of 10.000+ students
- Costs: >€61M

Action within Dutch Parliament

- Parlementary papers 2023/24 D21614
  Type of education as protected attribute
- Parlementary papers 2023/24 24724 nr. 229
  General periodic inspection
- Parlementary papers 2023/24 24724 nr. 231
  Scientific methodology for algorithms

Tweede Kamer der Staten-Generaal 2

Vergaderjaar 2023-2024

24 724 Studiefinanciering

Nr. 231 MOTIE VAN HET LID BECKERMAN C.S.
Voorgesteld 21 maart 2024

De Kamer,

gehoord de beraadslaging,

constaterende dat DUO bij het opsporen van fraude met de uitwonendenbeurs gebruikmaakte van een algoritme waar mbo-niveau en familiaire-laties als risico-indicatoren zijn gebruikt, zonder dat er enige onderbouwing is tussen deze indicatoren en fraude;

constaterende dat handhavers aan de hand van deze informatie zelf keuzes hebben moeten maken over selectie, waardoor er ruimte is ontstaan voor selectie op basis van vooringenomenheid;

overwegende dat hierdoor een situatie is ontstaan waarbij sprake kon zijn van een discriminerende feedbacklloop, waarin vrijwel alleen studenten met een migratieachtergrond zijn beschuldigd van fraude;

verzoekt de regering ervoor te zorgen dat DUO in het vervolg gaat werken met wetenschappelijk onderbouwde algoritmes, zodat wordt voorkomen dat er één groep wordt uitgepikt,

en gaat over tot de orde van de dag.

Beckerman
Ergin
Stultiens

Tweede Kamer, vergaderjaar 2023–2024, 24 724, nr. 231

Case law

- Case law: Rechtbank Overijssel 29 oktober 2024, ECLI:NL:RBOVE:2024:5627
- Evidence obtained through the application of a discriminatory risk profiling algorithm and control process is deemed unlawful

de Rechtspraak

All data and source code available on Github

iv. Unsupervised bias testing

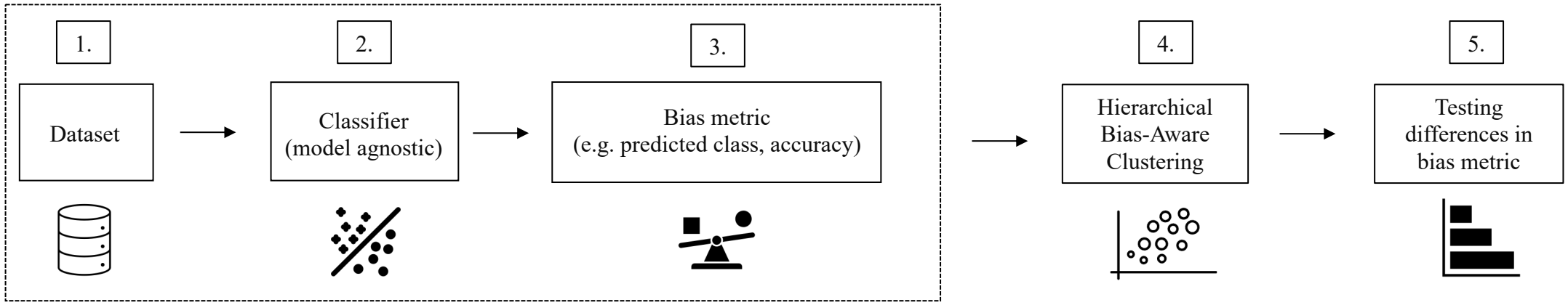
What to do if no demographic data was available?

Anomaly detection

> Identifying groups where an algorithm or AI system (classifier) shows variations in performance

> No access needed to demographic data to form groups, for instance, using clustering

Unsupervised bias detection pipeline



prepared by user of tool

Pseudo code of HBAC clustering algorithm

Algorithm 1: Hierarchical Bias-Aware Clustering

Input: A dataset $X = \{x_1, \ldots, x_N\}$ and bias metric $\{m_1, \ldots m_n\}$. Set the *max_iterations* and a minimum of samples per cluster $n_{\min}$.

Output: A partition $\{C_1, \ldots C_k\}$

Define the partition = $\{X\}$
for $i \leftarrow 1$ **to** *max_iterations* **do**
    Set C to be the cluster in partition with the highest standard deviation of metric M among those that have not been selected in any previous iteration.
    Split C into two clusters $C^{'}$ and $C^{''}$ using k-means or k-modes
    if $\max(\bar{M}(C^{'}), \bar{M}(C^{''})) \geq \bar{M}(C) \quad \wedge \quad |C^{'}| \geq n_{\min} \quad \wedge \quad |C^{''}| \geq n_{\min}$ **then**
        Remove C from partition
        Add $C^{'}$ and $C^{''}$ to partition
    end
end

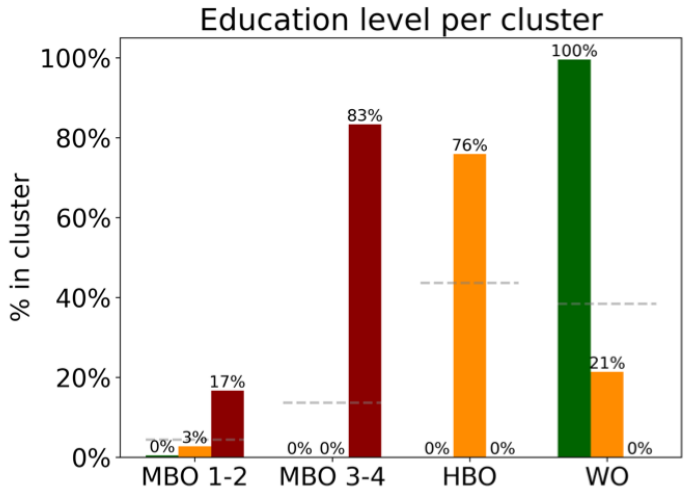
Available as pip package
unsupervised-bias-detection

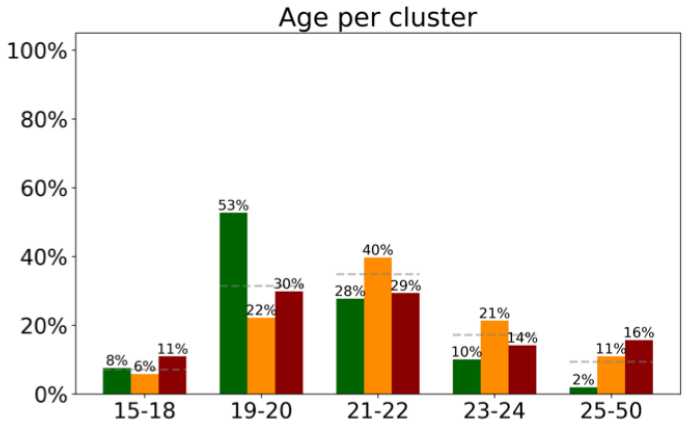
Available as local-first web on
Algorithm Audit's website

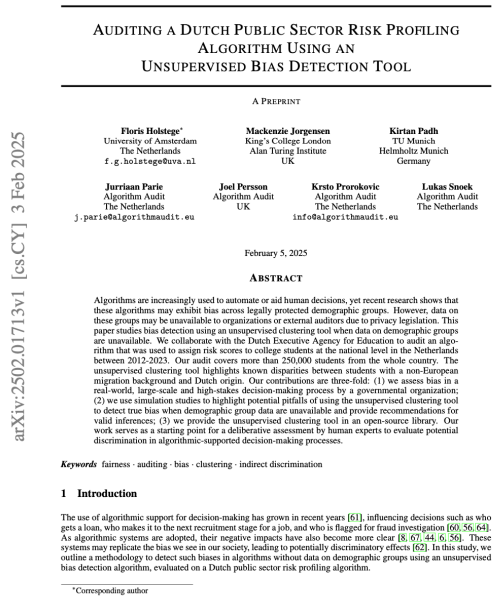
Results applying HBAC on DUO dataset

Model parameters

- Bias metric: being classified as 'high risk' by profiling algorithm
- 2014 data: 248,649 students
- Categorical data: HBAC using k-modes
- 80-20 out-of-sample fitting
- Described in paper







AUDITING A DUTCH PUBLIC SECTOR RISK PROFILING ALGORITHM USING AN UNSUPERVISED BIAS DETECTION TOOL

A PREPRINT

Floris Holstege*
University of Amsterdam
The Netherlands
f.g.holstege@uva.nl

Mackenzie Jorgensen
King's College London
Alan Turing Institute
UK

Kirtan Padh
TU Munich
Helmholtz Munich
Germany

Jurriaan Parie
Algorithm Audit
The Netherlands
j.parie@algorithmaudit.eu

Joel Persson
Algorithm Audit
UK

Krsto Prorokovic
Algorithm Audit
The Netherlands
info@algorithmaudit.eu

Lukas Snoek
Algorithm Audit
The Netherlands

February 5, 2025

ABSTRACT

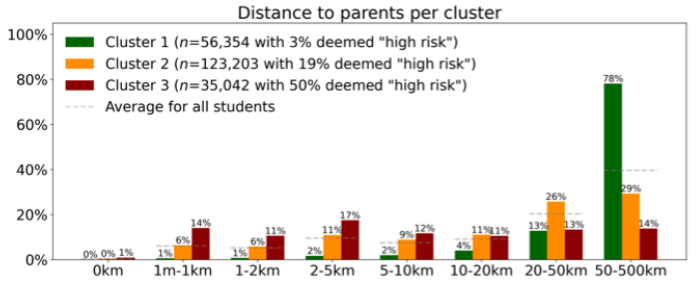
Algorithms are increasingly used to automate or aid human decisions, yet recent research shows that these algorithms may exhibit bias across legally protected demographic groups. However, data on these groups may be unavailable to organizations or external auditors due to privacy legislation. This paper studies bias detection using an unsupervised clustering tool when data on demographic groups are unavailable. We collaborate 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 from the whole country. The unsupervised clustering tool highlights known disparities between students with a non-European migration background and Dutch origin. Our contributions are three-fold: (1) we assess bias in a real-world, large-scale and high-stakes decision-making process by a governmental organization; (2) we use simulation studies to highlight potential pitfalls of using the unsupervised clustering tool to detect true bias when demographic group data are unavailable and provide recommendations for valid inferences; (3) we provide the unsupervised clustering tool in an open-source library. Our work serves as a starting point for a deliberative assessment by human experts to evaluate potential discrimination in algorithmic-supported decision-making processes.

Keywords fairness · auditing · bias · clustering · indirect discrimination

1 Introduction

The use of algorithmic support for decision-making has grown in recent years [61], influencing decisions such as who gets a loan, who makes it to the next recruitment stage for a job, and who is flagged for fraud investigation [60, 56, 64]. As algorithmic systems are adopted, their negative impacts have also become more clear [8, 67, 44, 6, 56]. These systems may replicate the bias we see in our society, leading to potentially discriminatory effects [62]. In this study, we outline a methodology to detect such biases in algorithms without data on demographic groups using an unsupervised bias detection algorithm, evaluated on a Dutch public sector risk profiling algorithm.

*Corresponding author



Algorithm Audit

ETH zürich

1. What is algoprudence?

2. Real-world audit: public sector risk profiling algorithm

3. **Q&A**

Algorithm Audit

www.algorithmaudit.eu

info@algorithmaudit.eu

https://www.linkedin.com/company/algorithm-audit/

https://github.com/NGO-Algorithm-Audit

NGO Algorithm Audit is registered in
The Netherlands Chamber of Commerce under number
83979212