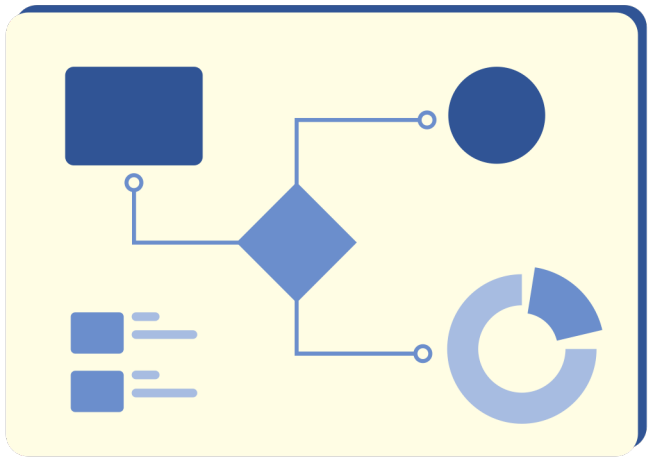
Algorithm Audit

CEN CENELEC

Statistical hypothesis testing

Risk management measures to mitigate the risk of indirect discrimination through high-risk AI profiling systems



July 2024

Overview

1 Introduction: risk profiling data pipeline and statistical hypothesis testing

2 Legal framework: in support of AI Act and EU non-discrimination law

3 Risk assessment and risk treatment

4 Residual risk acceptability

5 ISO risk standards: not enough

A Statistical background information

Activities NGO Algorithm Audit

Normative advice commissions
Advising on ethical issues emerging in concrete algorithmic practices through deliberation, resulting in *algoprudence* (jurisprudence for AI)

Technical tooling
Implementing and testing technical tools to detect and mitigate bias in data and algorithms, see bias detection tool, synthetic data generation

Knowledge platform
Bringing together knowledge and expertise to ignite the collective learning process for responsible algorithms, e.g., AI Policy Observatory and AI Act standards

Project work
Supporting public and private sector organisations with specific questions regarding responsible use of algorithms

Financially supported by

SIDN fonds
European Artificial Intelligence & Society Fund
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Algorithm Audit

CEN CENELEC

1. Introduction: risk profiling and statistical hypothesis testing

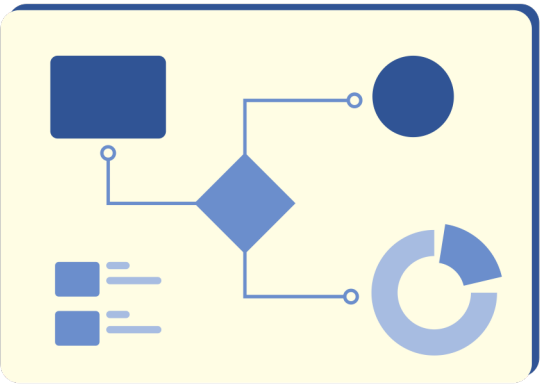
2. Legal framework: in support of AI Act and EU non-discrimination law

3. Risk assessment and risk treatment

4. Residual risk acceptability

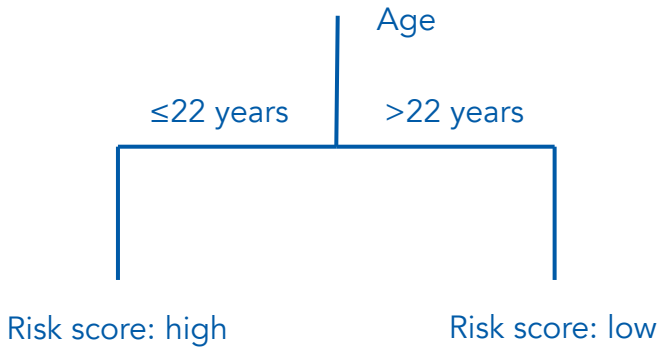
5. ISO risk standards: not enough

A. Statistical background information



What is risk profiling? And how does it relate to fundamental rights?

Profiling – example: student grant checks



Profiling defined in GDPR Art. 4 (4)

"Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements."

Risk profiling

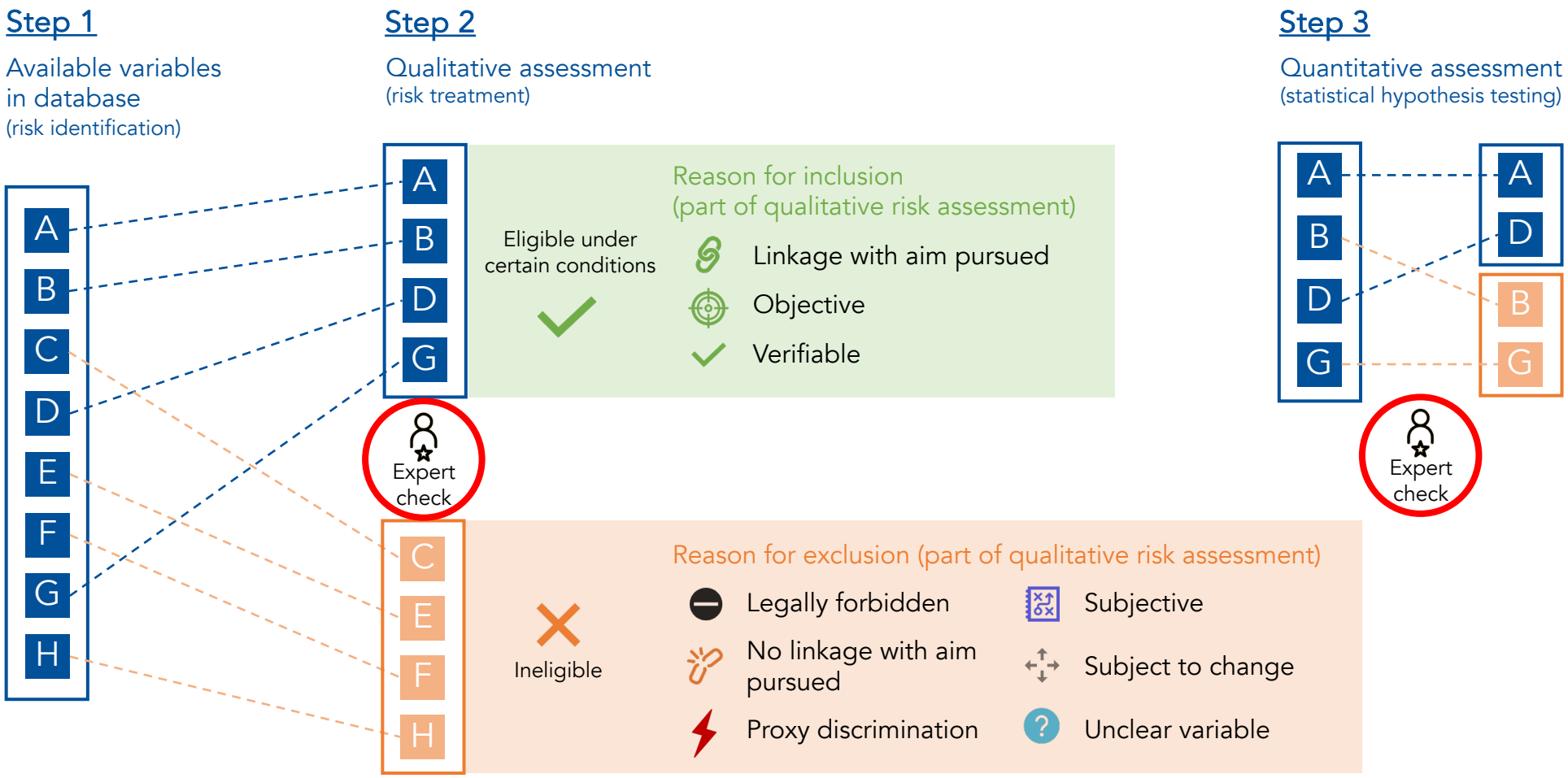
- Public and private organisations make use of risk profiling for enforcement and monitoring purposes
  - Rule-based risk profiling, e.g., unduly granted subsidies
  - ML-driven risk profiling, e.g., ad micro-targeting
- Differentiation is a feature, not a bug
- Differentiation brings the risk of (in)direct discrimination (risk as defined in AI Act terminology, see slide 9-10) through apparently neutral characteristics, such as ZIP code and type of SIM card

Risk profiling defined by The Netherlands Institute for Human Rights

"A set of one or more selection criteria based on which a certain risk of norm violation is assessed, and a selection decision is made."

Source: The Netherlands Institure for Human Rights, Discrimination through risk profiling (2021)
https://www.rijksoverheid.nl/documenten/rapporten/2021/11/30/discriminatie-door-risicoprofielen--mensenrechtelijk-toetsingskader

Quantitative and qualitative assessment of (in)eligible criteria to mitigate risk of indirect discrimination



Quantitative and qualitative assessment of (in)eligible criteria to mitigate risk of indirect discrimination

Step 3

Quantitative assessment
(statistical hypothesis testing)

Ineligible

B G

A D

Eligible under certain conditions

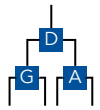
Step 4

Profile composition

Machine learning-based profiling → profile

e.g., *explainable boosting or extreme gradient (xg)boosting*

SMEs



Rule-based profiling → profile

Step 5

Expert check on final risk profile

Expert check

Note

- Focus on indirect discrimination because open system for justification (see slide 18)
- If no statistically significant relationship exists between profiling criterion and aim pursued ---> ineligible criterion
- There are solutions to deal with interaction variables, e.g., multiple hypothesis testing
- Normative judgement to assess risk acceptability issued by diverse group of stakeholders (see also N-doc N226)
- Concrete requirement for prescriptive standard, input for FRIA

Checking assumptions: statistical hypothesis testing as part of risk identification and risk treatment

Statistical hypothesis

H_0: fraud rate ≤22 years = fraud rate >22 years

H_A: fraud rate ≤22 years > fraud rate >22 years

Methodology

- Hypotheses should be tested on a sufficiently large random sample (see slide 19)
- Data points in random sample should be identical and independently distributed
- In this case population size 250k+, the statistical relationship can be assessed using Z-testing on random sample size 387
- Confidence level: 95% (p-value 0.05), power of test: 80%

Data

	Group size	#fraud	Percentage
≤22 years	288	9	3.2%
> 22 years	99	5	5.0%

Stats 101

- Pooled proportion: $p = \frac{9+5}{288+99} \approx 0.0362$
- Standard error of difference in proportions: $SE = \sqrt{p(1-p)\left(\frac{1}{288}+\frac{1}{99}\right)} \approx 0.0217$
- Calculate Z-score: $Z = \frac{0.032-0.050}{0.0217} \approx -0.888$
- p-value ≈ $0.1887 > 0.05$, no evidence for a statistically significant relationship --> ineligible profiling criterion

Statistical terminology also in:
- ISO 24153:2009
- ISO 3534-1:2006

Ineligible profiling criterion

Real world example of AI validation of public sector organisation in NLD (2024): https://algorithmaudit.eu/algoprudence/cases/aa202401_preventing-prejudice/

Risk measures should prevent algorithmic-driven indirect discrimination in the EU

10+ years of indirect discrimination by Dutch Executive Education Agency (DUO) could have been prevented by applying statistical hypothesis testing as a risk measure

Assumptions which were rejected by statistical hypothesis testing:

> Lower educated students commit more fraud
> Younger student commit more fraud

Two reports of Algorithm Audit were sent to Dutch Parliament causing apologies of the Dutch Minister of Education, Culture and Science for indirect discrimination in algorithmic-driven control process

Algorithm Audit

CEN CENELEC

1. Introduction: risk profiling and statistical hypothesis testing

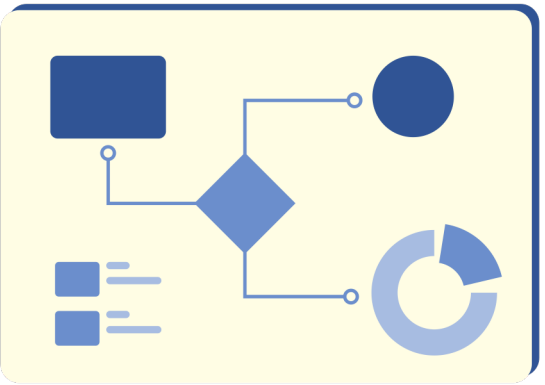
2. Legal framework: in support of AI Act and EU non-discrimination law

3. Risk assessment and risk treatment

4. Residual risk acceptability

5. ISO risk standards: not enough

A. Statistical background information



AI Act – In support of Art. 5, 9, 10 and Recital 42

In support of AI Act Article 9 – Risk management system

Art. 9 – Risk management system (selection)

2. (a) the identification and analysis of the known and ***the reasonably*** foreseeable risks

→ Indirect discrimination is a reasonably foreseeable risk of high-risk AI systems used for profiling

2. (b) the ***estimation and evaluation of the risks*** that may emerge when the high-risk AI system is used in accordance with its intended purpose

→ Intended purpose of profiling is differentiation

2. (d) the adoption of ***appropriate and targeted*** risk management measures ***designed to address the risks identified pursuant to point (a)***

→ Statistical hypothesis testing is an appropriate and targeted risk management measure

4. … ***with a view to minimising risks more effectively while achieving an appropriate balance in implementing the measures to fulfil those requirements***.

→ Random sampling and statistical hypothesis testing can be implemented in Step 2 of the profiling pipeline with reasonable efforts

In support of standardization request (SR) 1: risk management systems

AI Act terminology

Risk: Combination of the probability of occurrence of harm and the severity of that harm.

Harm: Injury or damage to the health, or damage to property or the environment, or infringement of fundamental rights.

Severity: For any risk to fundamental rights, the severity of risk includes consideration of the nature of the harm, the strength of the harm, the significance and scale of the harm in terms of the number of individuals whose rights are placed at risk, the irremediability of the harm and whether the rights at risk are those of rights-holder groups that enjoy additional or particular protections.

In support of AI Act Article 9 – Risk management system (c'tnd)

Art. 9 – Risk management system (selection)	
5. … ***relevant*** residual risk associated with each hazard, as well as the overall residual risk of the high-risk AI systems is judged ***to be acceptable***.	Relates to suitability test, key element of EU non-discrimination law, as described on slide 15
5. (a) … reduction of risks ***identified and evaluated pursuant to paragraph 2 in*** as far as ***technically feasible*** through adequate design and development of high-risk AI system	Technically feasible
5. (b) where appropriate, implementation of adequate mitigation and control measures ***addressing*** risks that cannot be eliminated	Layered approach, together with among others qualitative assessment and redress mechanisms
6. High-risk AI systems shall be tested for the purpose of identifying the most appropriate ***and targeted*** risk management measures.	Proposed risk management measures result of AI validation in practice in NLD
8. … Testing shall be carried out against ***prior*** defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system.	95% confidence interval, 80% power of test

AI Act terminology

Residual risk: risk remaining after risk control measures have been implemented

Hazard: potential source of harm

Acceptable risk: level of risk that is accepted in a given context based on the current values of society

Relates to suitability test, key element of EU non-discrimination law, as described on slide 15

In support of AI Act Article 5 – Prohibited AI practices, and Recital 42 Presumption of innocence

Art. 5 – Prohibited AI practices (selection)

1. (d) … *this prohibition shall not apply to AI systems used to support the human assessment of the involvement of a person in a criminal activity, which is already based on objective and verifiable facts directly linked to a criminal activity*

Recital 42 – Presumption of innocence

… natural persons should never be judged on AI-predicted behaviour based solely on their profiling … without a reasonable suspicion of that person being involved in a criminal activity based on objective verifiable facts and without human assessment thereof.

Legal obligation that risk profiling criteria are "objective and verifiable facts directly linked to a criminal activity"

Relates to proportionality test, key element of EU non-discrimination law, as described on slide 15

In support of AI Act Article 10 – Data and data governance

Art. 10 – Data and data governance	
2 (c) …relevant data-preparation processing operations, such as annotation, labelling, cleaning, **updating**, enrichment and aggregation;	Drawing random sample part of relevant data-preparation operation
2 (d) …the formulation of assumptions, in particular with respect to the information that the data are supposed to measure and represent;	Random sample needed to test assumptions
2 (f) …examination in view of possible biases **that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations;**	Related to representativeness of dataset and drawn random samples
2 (g) … appropriate measures to detect, prevent and mitigate possible biases identified according to point (f);	Colleting random sample appropriate measure to detect, prevent and mitigate indirect discrimination

European Convention of Human Rights (ECHR)

Contextualization of EU non-discrimination law in Dutch national context: Discrimination through risk profiling, The Netherlands Institute for Human Rights

EU non-discrimination law in the context of risk profiling

In the aftermath of the Dutch childcare benefit scandal, the Dutch Data Protection Authority (DPA) ruled that increased probability of higher scrutiny through risk profiling is considered as a 'particular disadvantage', i.e. a harm

1. Differentiation on the basis of race or nationality?

Adverse effect in comparison to others in a similar situation?
- No → No discrimination on the basis of race of nationality
- Yes ↓

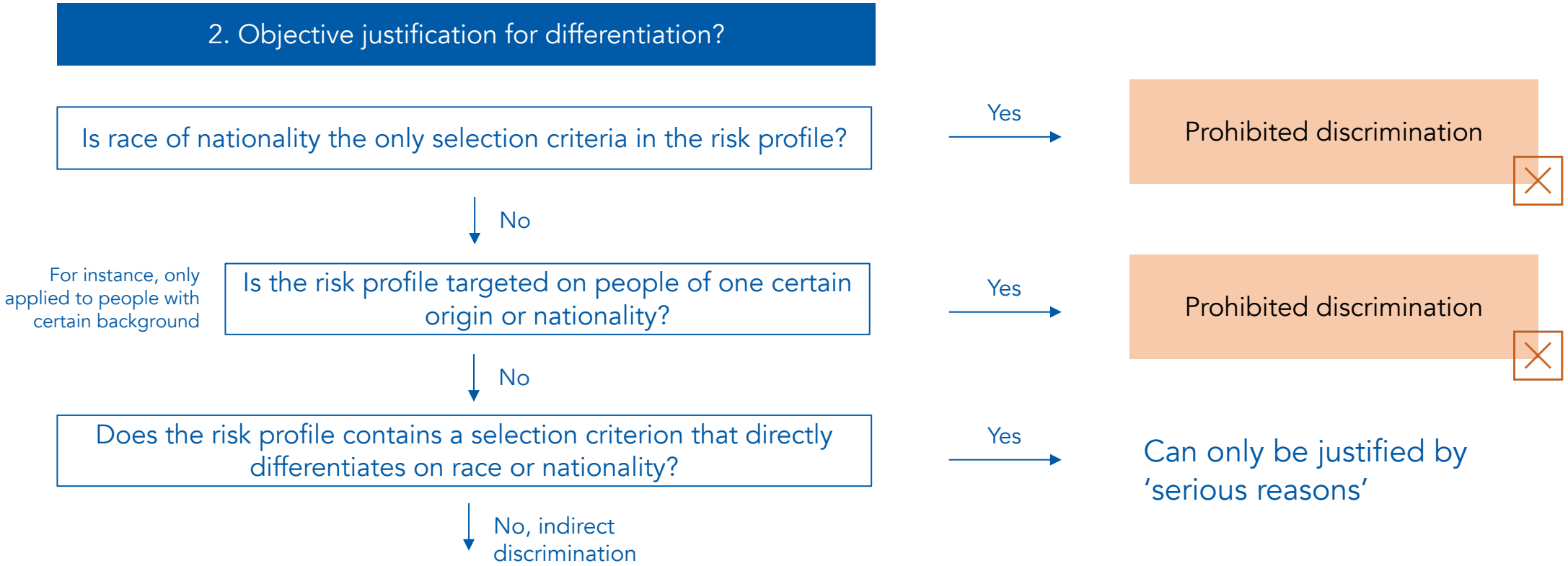
On the basis of race (skin color, origin, etnicity etc.) or nationality?
- No → No discrimination on the basis of race of nationality
- Yes ↓

Direct differentiation on the basis of race or nationality in the context of social protection?
- No → To be continued…
  Open list for objective justification
- Yes ↓

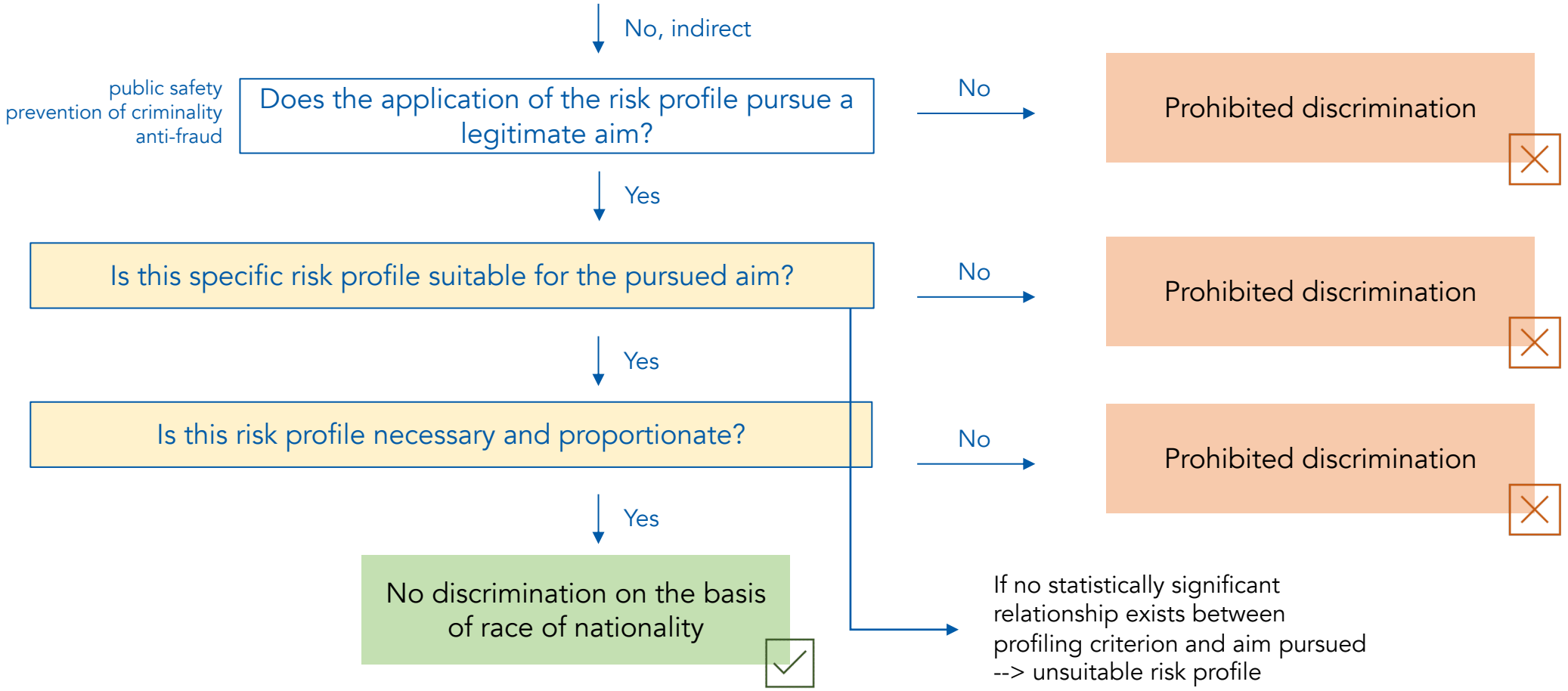
General Non-discrimination Directive
Closed list of justification

Source: The Netherlands Institure for Human Rights, Discrimination through risk profiling (2021)
https://www.rijksoverheid.nl/documenten/rapporten/2021/11/30/discriminatie-door-risicoprofielen--mensenrechtelijk-toetsingskader

EU non-discrimination law in the context of risk profiling



EU non-discrimination law in the context of risk profiling



Algorithm Audit

CEN CENELEC

1. Introduction: risk profiling and statistical hypothesis testing

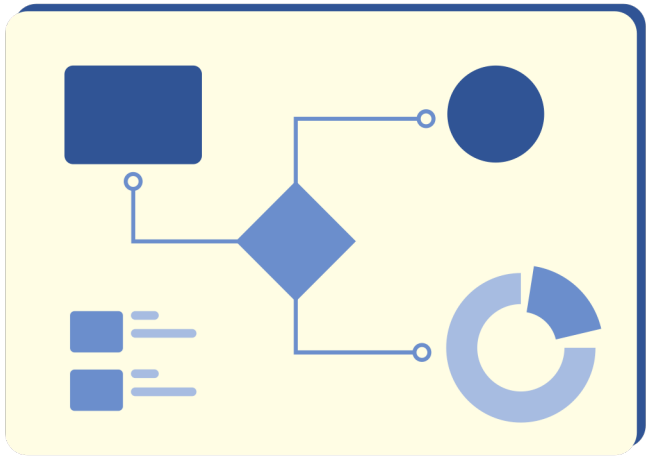
2. Legal framework: in support of AI Act and EU non-discrimination law

3. **Risk assessment and risk treatment**

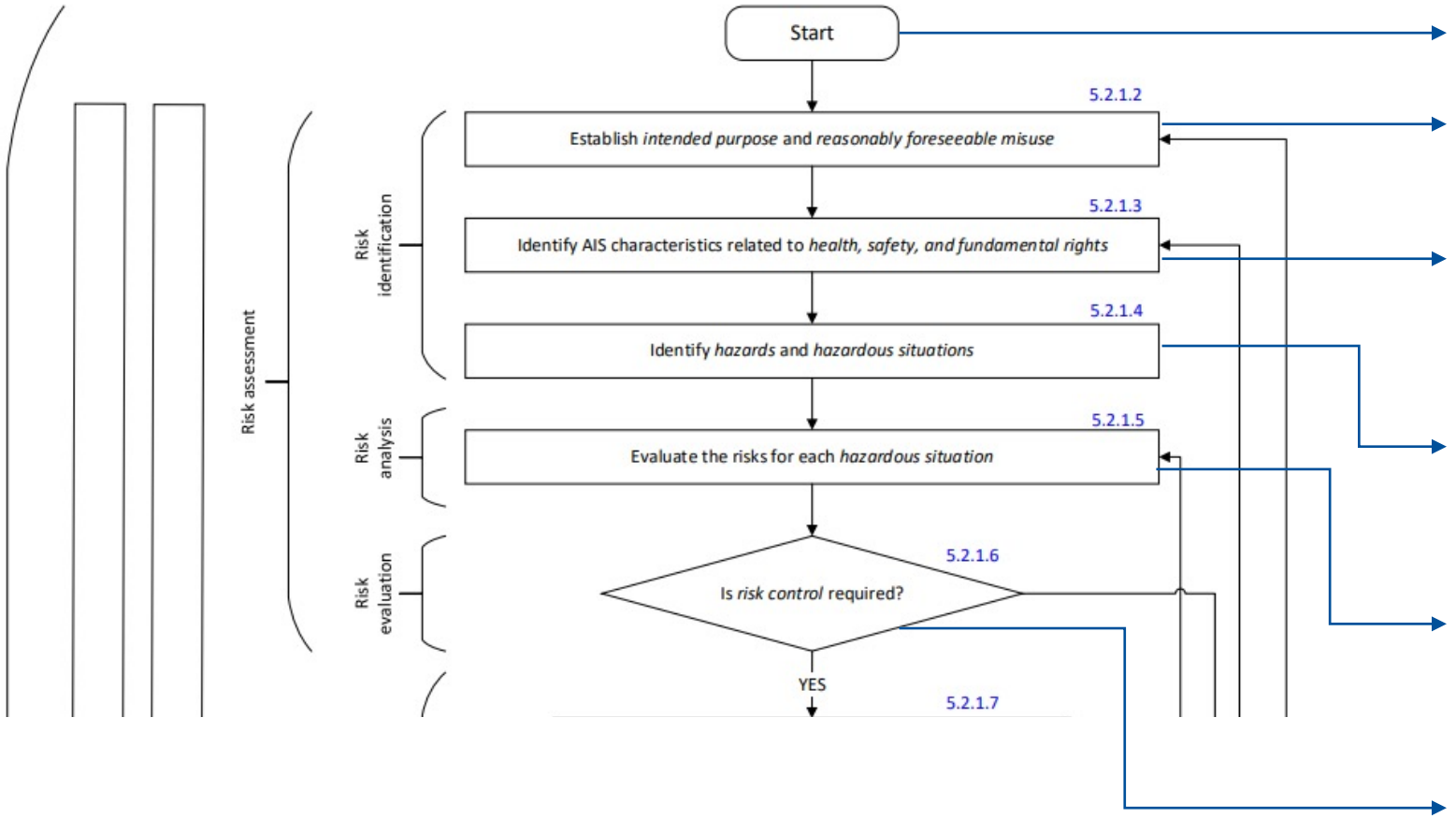
4. Residual risk acceptability

5. ISO risk standards: not enough

A. Statistical background information

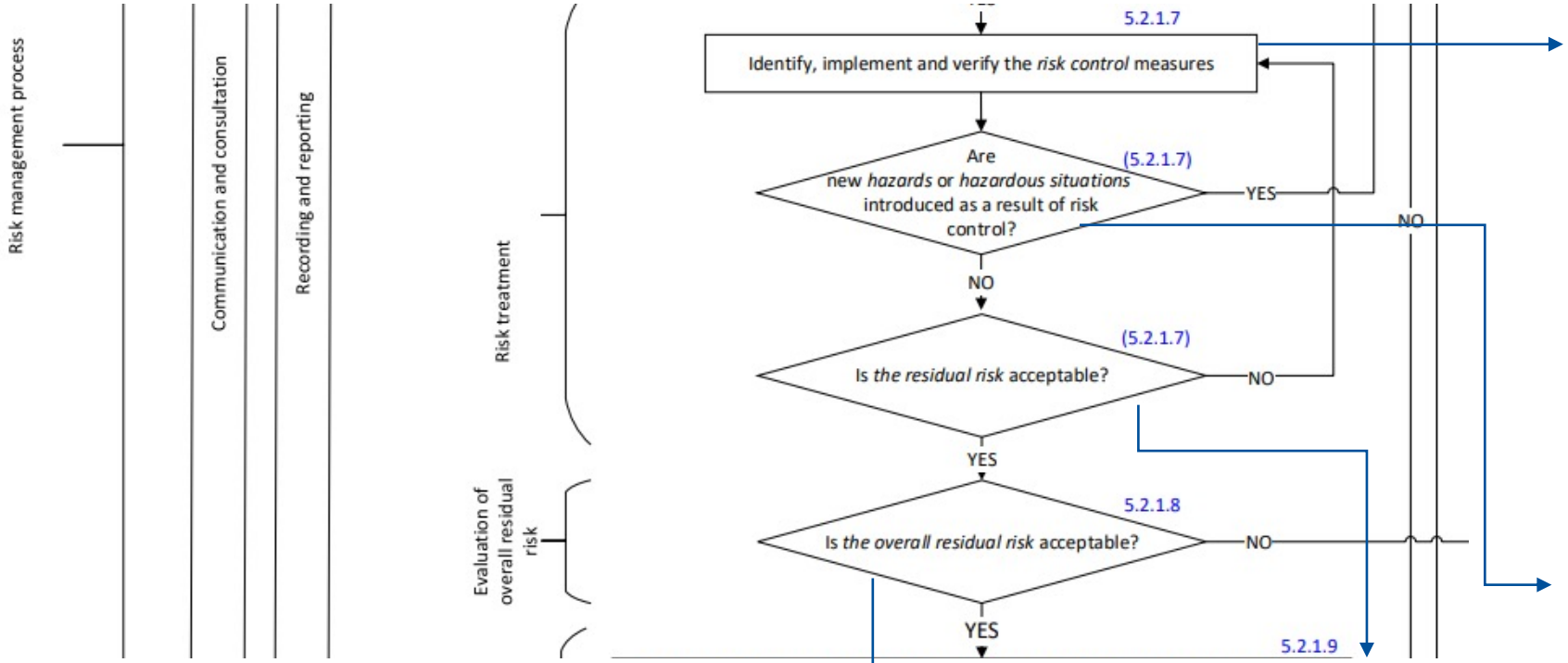


Risk process overview for profiling data pipeline



- High-risk AI profiling technique
- Random sample should be available according with Art. 10 Data governance
- Intended purpose: segmentation
- Reasonably foreseeable misuse: prohibited segmentation
- AIS characteristics related to fundamental rights: prohibited segmentation, i.e., indirect discrimination
- Hazard, hazardous situation: being subjected to discriminatory risk profile is a source of harm (hazard)
- Assess whether each considered profiling criteria (Step 1) poses an unacceptable risk of proxy discrimination
- If the risk of proxy discrimination cannot be excluded risk control is required

Risk process overview for profiling data pipeline



> Identified risk control measure: qualitative assessment (Step 2) and quantitative assessment (Step 3)
> Implement risk control measure: exclude ineligible profiling criteria
> Verification of risk control measure for qualitative assessment should be performed by diverse group of stakeholders (see N-doc N226)
> Verification of risk control measure for quantitative assessment is performed based on p-values

> Not likely as variables are only excluded, not included

> Excluding profiling criteria holding innocuous correlation with the aim pursued are excluded, decreasing the residual risk
> Assessing the acceptability of the (overall) residual risk level should be decided by diverse group of stakeholders (see N-doc N226)

Risk process overview for profiling data pipeline

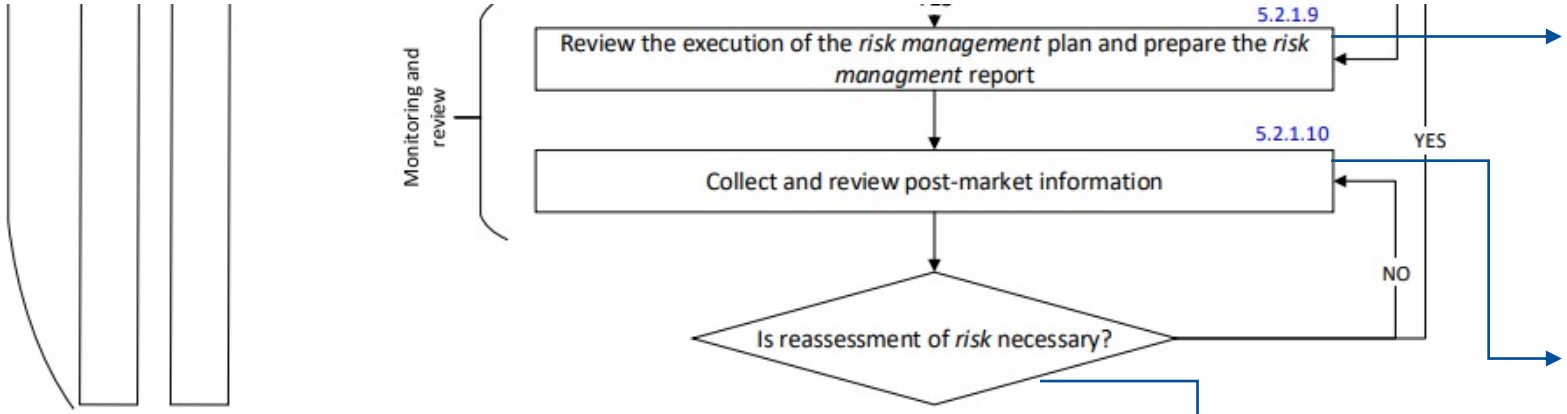


Figure 1: risk management process diagram for AI systems (informative). Note that this diagram is modified (simplified) compared with N281. References to sub-sections of risk assessment (5.2) are provided in blue.

> Qualitative and quantitative rationales for inclusion and exclusion of profiling criteria should be documented in risk management report, incl. execution of risk management plan

> For instance, post-market information could include supervised bias measurements as specified in AI Act Art. 10 (5)

> Risk should frequently be reassessed

Algorithm Audit

CEN CENELEC

1. Introduction: risk profiling and statistical hypothesis testing

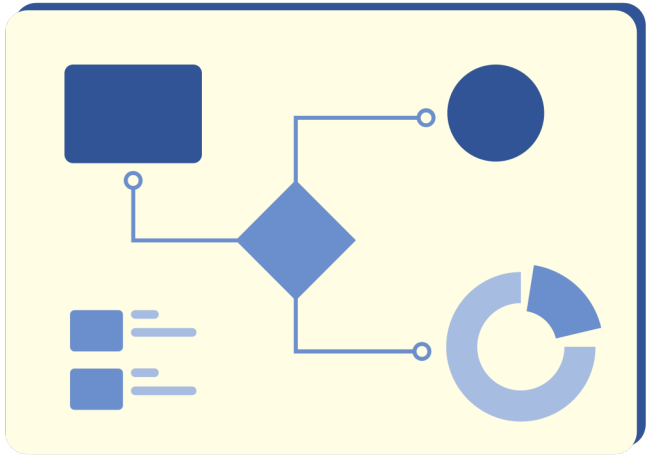
2. Legal framework: in support of AI Act and EU non-discrimination law

3. Risk assessment and risk treatment

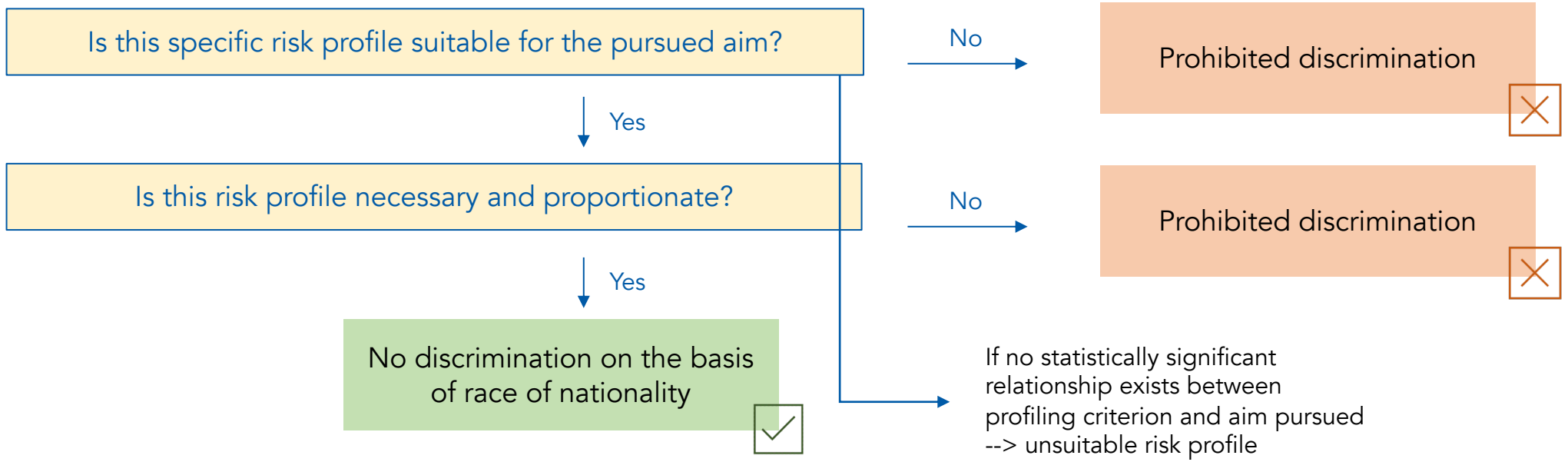
4. Residual risk acceptability

5. ISO risk standards: not enough

A. Statistical background information



EU non-discrimination law provides a framework to evaluate risk acceptability



> Applying qualitative and quantitative risk control measures (see slide 5-6) as a process to pass the suitability, necessity and proportionality test as required by EU non-discrimination law to conduct profiling

Who decides who decides? Normative decisions should be taken within democratic sight

Institutional actors…

- Courts
- Legislator
- Supervisory authorities

…and decentralized self-assessment

Ethical advice commission


Maarten van Asten, Alderman Finance, Digitalisation and Event Municipality of Tilburg


Munish Ramlal, Ombudsperson of Metropole region Amsterdam


Abderrahman Al Aazani, Representative of the Ombusperson of Rotterdam


Francien Dechesne, Associate Professor Law and Digital Technologies, Universiteit Leiden


Oskar Gstrein, Assistant Professor Governance and Innovation, Rijksuniversiteit Groningen

1. Initial written feedback on identified issue

2. Panel gathering

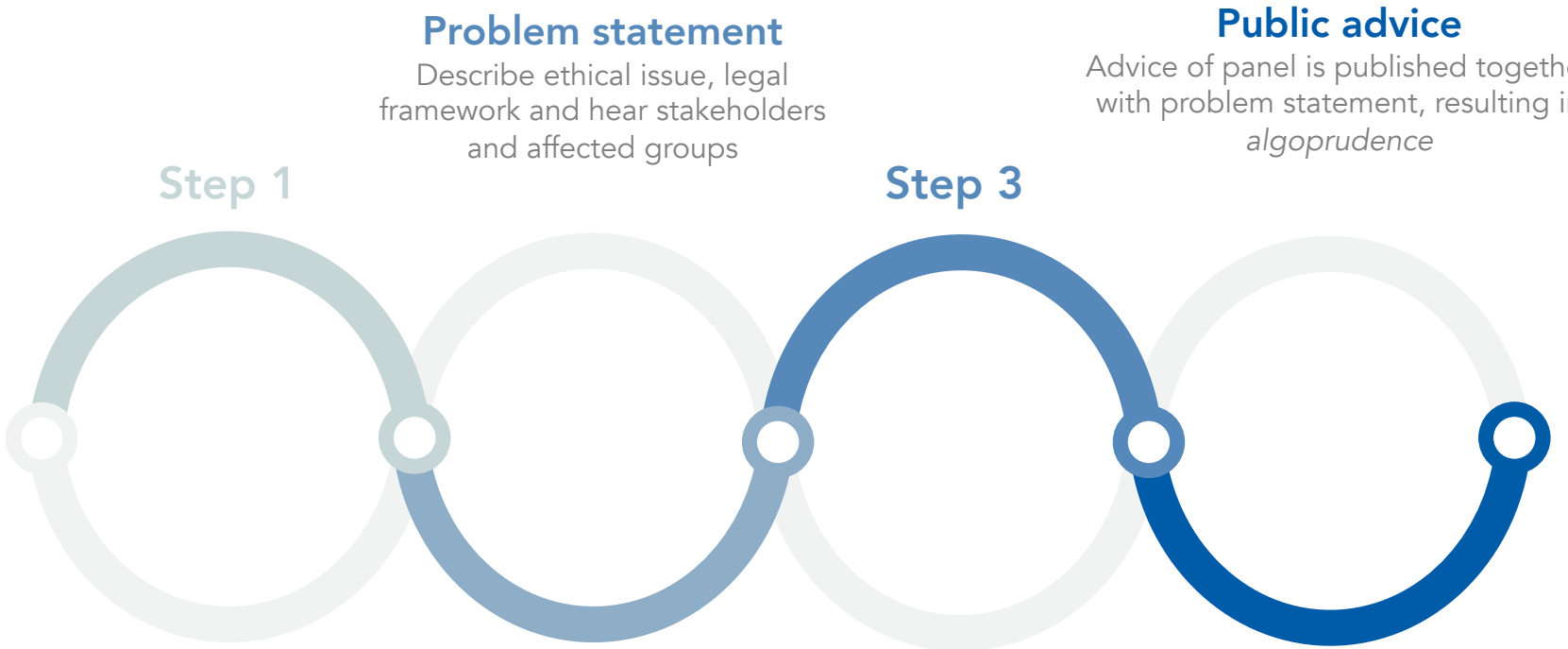
accepted state-of-the-art

diverse inclusive

deliberative transparent

See also N-doc N226

Evaluation process to resolve normative questions identified by a FRIA



Algorithm Audit

CEN CENELEC

1. Introduction: risk profiling and statistical hypothesis testing

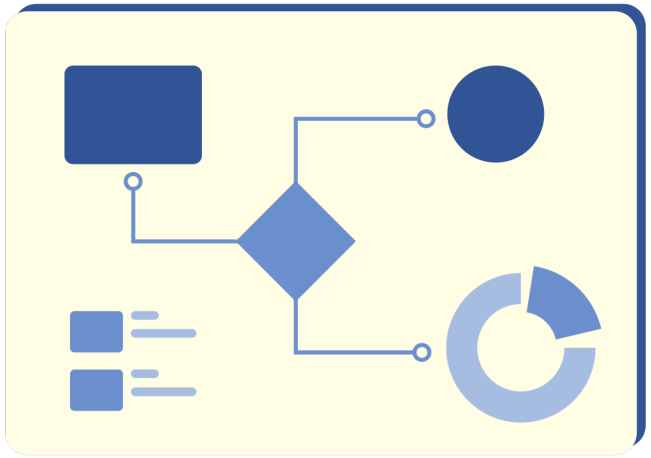
2. Legal framework: in support of AI Act and EU non-discrimination law

3. Risk assessment and risk treatment

4. Residual risk acceptability

5. ISO risk standards: not enough

A. Statistical background information



ISO risk standards are not enough to prevent the risk of indirect discrimination

ISO 31000 Risk management

- Risk management standard that provides general guidelines for risk management
- Offers principles, a framework, and a process for managing risk that can be used by any organization, regardless of its size, industry, or sector
- No focus on technology, neither profiling nor fundamental rights (such as non-discrimination) in the context of EU legislation

Note: statistical hypothesis testing is a cornerstone of empirical sciences, including social sciences and drug testing. It is included in the ICH E9: Statistical Principles for Clinical Trials, which is used by both the European Medicines Agency (EMA) and the U.S. Food and Drug Administration (FDA)

ISO 42001 AI management system

- Provides a baseline for risk identification, risk treatment and risk impact assessments for AI systems
- Not detailed enough to prevent fundamental rights violations through algorithmic profiling:
  - High-level description of AI risk assessment (6.1.2)
  - AI risk treatment (6.1.3) does not prescribe "appropriate and targeted risk management measures designed to address the risks identified" as mandated by Art. 9 AI Act
- Statistical hypothesis testing builds upon the AI risk management framework laid down in 420001, and tailors it to the EU context

ISO/IEC standards are not linked to fundamental rights as codified in EU law

ISO/IEC

> TS 4213 (performance metrics, ed.2 WIP)
> TR 24027 (metrics on bias)
> DTS 12791 (process requirements for bias)

Four main directives currently make up EU non-discrimination law:
> Race Equality Directive 2000/43/EC
> Framework Equality Directive
> Gender equality Directives 2004/113/EC and 2006/54/EC.

Additionally, primary law provisions include:
> Articles 2 and 3(3) of the Treaty on European Union
> Articles 8, 10, 19 and 157 of the Treaty on the Functioning of the European Union
> Articles 20, 21 and 23 of the Charter of Fundamental Rights of the EU (the Charter)

For requirements laid down by EU non-discrimination law for lawful risk profiling, see slide 16-18

Algorithm Audit

CEN CENELEC

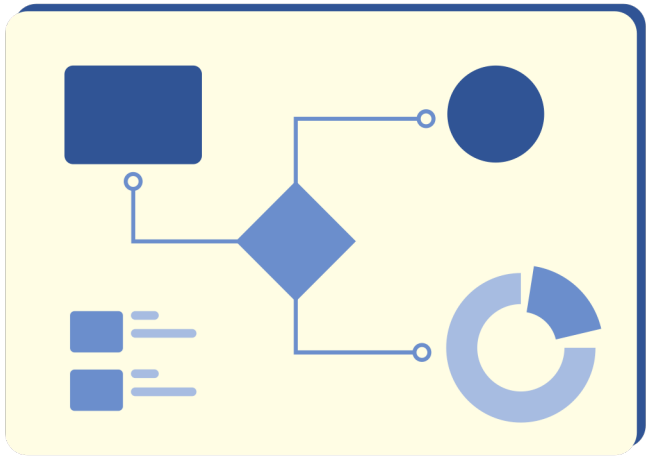
1. Introduction: risk profiling and statistical hypothesis testing

2. Legal framework: in support of AI Act and EU non-discrimination law

3. Risk assessment and risk treatment

4. Residual risk acceptability

5. ISO risk standards: not enough



A. Statistical background information

Theoretical answer to compute random sample size for m hypothesis tests

- **Goal:** Determine minimum random sample size to measure statistically significant difference in unduly use rate between students living ≤5km (group A) and >5km away (group B), and students ≤22 years old (group C) and >22 years old (group D)
- **Data needed:**
  - p_A, p_B: proportion unduly group ≤5km and proportion unduly group >5km
  - p_C, p_D: proportion unduly group ≤22 years old and proportion unduly group >22 years old
  - Two null and alternative hypothesis (one-sided):

    $H_0: p_A = p_B$ $\quad$ $H_0: p_C = p_D$

    $H_1: p_A > p_B$ $\quad$ $H_1: p_C > p_D$
  - Significance level (**α**): probability of false positive (accepting H_0 while H_1 is true), e.g., 0.05 or 0.01
  - Bonferroni correction for testing m hypothesis: $\alpha_{adjusted} = \frac{\alpha}{m}$
  - Power **(1 - β)**: probability of true positive (rejecting H_0 while H_0 is indeed false), e.g., 0.8 or 0.9
  - Ratio of sample sizes (k_1 and k_2): ratio of group sample sizes, i.e. $k_1 = \frac{N_A}{N_B}, k_2 = \frac{N_C}{N_D}$

Formula to compute sample size for m hypothesis tests

> Formula to determine random sample size (n_1) to test first hypothesis test and random sample size (n_2) to test second hypothesis test:

$$n_1 = \left(\frac{p_A(1-p_A)}{k_1} + p_B(1-p_B)\right) \cdot \left(\frac{\left(z_{1-\alpha_{adjusted}} + z_{1-\beta}\right)}{(p_A - p_B)}\right)^2,$$

$$n_2 = \left(\frac{p_C(1-p_C)}{k_2} + p_D(1-p_D)\right) \cdot \left(\frac{\left(z_{1-\alpha_{adjusted}} + z_{1-\beta}\right)}{(p_C - p_D)}\right)^2,$$

where:

$z_{1-\alpha_{adjusted}}$: critical value of the normal distribution at the $\alpha_{adjusted} = \frac{\alpha}{2}$ confidence level

$z_{1-\beta}$: critical value of the normal distribution at the $1-\beta$ confidence level

Sample size: $\mathrm{n} = k_1 n_1 + n_1 + k_2 n_2 + n_2$

Example: Random sample (n=387)

Sample 2014	Size of group	# unduly grants	Percentage
0km	8	0	0%
1m-1km	21	5	23,8%
1-2km	11	0	0%
2-5km	31	5	16,1%
5-10km	24	1	4,2%
10-20km	31	1	3,2%
20-50km	58	1	1,7%
50-500km	137	0	0%
onbekend	66	1	1,5%
Totaal	387	14	3,6%

Group A
$p_A = 14\%$
$N_A = 71$

Group B
$p_B = 1{,}3\%$
$N_B = 316$

Example: profiling criteria 'distance to parent(s)' group A: ≤5km, group B: >5km

- $p_A = 0.013$ (guestimation)
- $p_B = 0.14$ (guestimation)
- Hypothesis:

  H_0: $p_A = p_B$
  H_1: $p_A > p_B$

- Significance level (**α**): 0.05
- Power of test **(1 - β)**: 0.8
- Ratio group size (k): $\frac{71}{316} = 0.225$ (gebaseerd op willekeurige steekproef), $\frac{45.79k}{202.87k} = 0.226$ (gehele populatie)
- Size group B: 69
- Size group A: 0.225*69=16

--> Size random sample: 85

Note: only for 1 hypothesis

Example: profiling criteria 'distance to parent(s)' group A: ≤5km, group B: >5km

		p_A, p_B			
α	**(1 - β)**	1%, 2%	1%, 7%	1.3%, 14%	2%, 20%
0.05	0.8	$N_A = 3.933, N_B = 885$ n=4.818	$N_A = 188, N_B = 43$ n=231	N_A =16, N_B =69 n=85	$N_A = 48, N_B = 11$ n=59
	0.9	$N_A = 5.477, N_B = 1.226$ n=6.673	$N_A = 260, N_B = 59$ n=319	$N_A = 95, N_B = 22$ n=117	$N_A = 66, N_B = 15$ n=81
0.01	0.8	$N_A = 6.383, N_B = 1.437$ n=7.820	$N_A = 305, N_B = 69$ n=374	$N_A = 111, N_B = 25$ n=136	$N_A = 77, N_B = 18$ n=95
	0.9	$N_A = 8.279, N_B = 1.863$ n=10.142	$N_A = 395, N_B = 89$ n=484	$N_A = 144, N_B = 33$ n=177	$N_A = 100, N_B = 23$ n=123

Note:

- Higher confidence level --> larger random sample size
- Smaller difference --> larger random sample size

Algorithm Audit

Building public knowledge for ethical algorithms

www.algorithmaudit.eu

info@algorithmaudit.eu

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

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

Stichting Algorithm Audit is registered as a non-profit organisation at the Dutch Chambre of Commerce under license number 83979212