Algorithm Audit

Public standard – Meaningful human intervention for risk profiling algorithms

Preventing decision-making based solely on profiling

May 2025

Summary

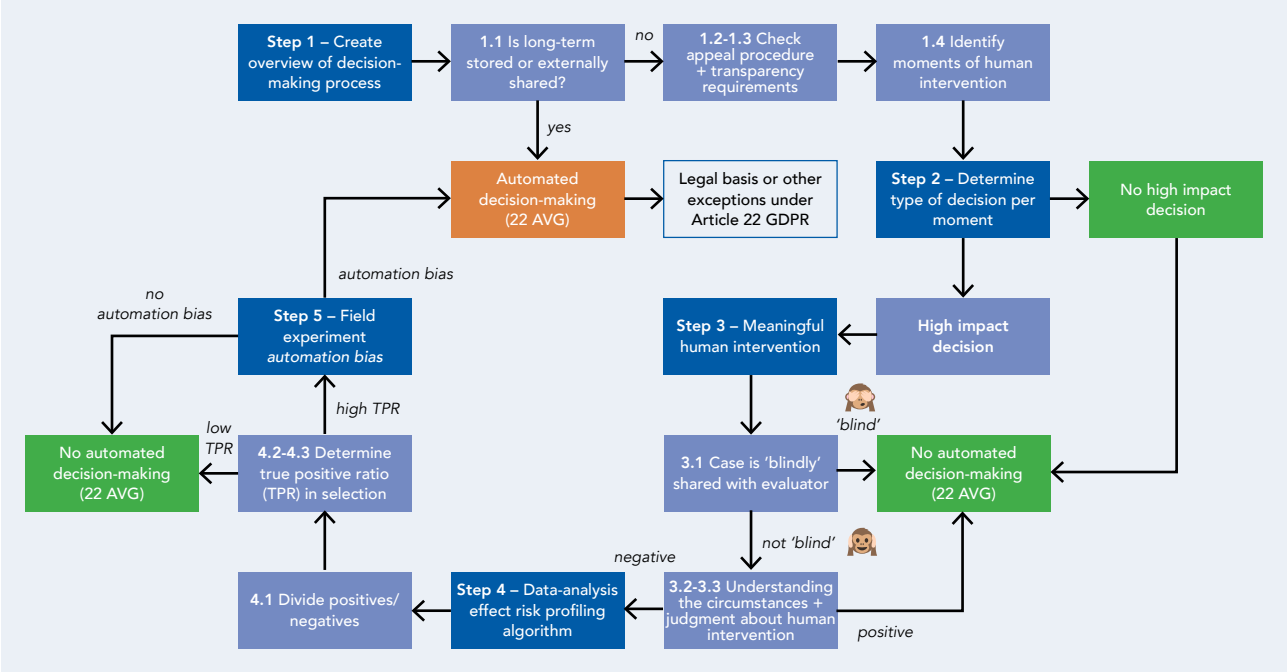
A fully automated decision is prohibited under Article 22(1) of the General Data Protection Regulation (GDPR). This document provides a pragmatic step-by-step guide on how to navigate this prohibition in the context of risk profiling algorithms. The steps consolidate advice from previously published documents and incorporate practical experience from Algorithm Audit's work on human-algorithm interaction in both public and private sectors. Central to this standard is the concept of 'blind' assessment – where evaluators do not know whether a case was selected randomly or was selected by a risk profiling algorithm. In addition to such qualitative safeguards, the document explains how data analysis can support the prevention of prohibited automated decision-making. This public standard bridges recent case law, particularly the Schufa ruling by the Court of Justice of the European Union, with practical implementation of algorithms. This step-by-step guide is also relevant for complying with Article 14 of the AI Act, which mandates human oversight when deploying AI systems. The standard focuses exclusively on the application of risk profiling algorithms, both in public and private domains. Other forms of automated decision-making and profiling alone fall outside the scope of this public standard.

This document serves as an extension to Algorithm Audit's Public Standard on Risk Profiling.[1] The step-by-step guide below will be integrated into Q7 of the open-source AI Act Implementation Tool.[2]

Overview step-by-step guide

Follow the steps below to prevent automated decision-making in risk profiling. Steps 1-5 are explained in detail later in this document.

[1] Public standard Profiling algorithms, Algorithm Audit (2024).
[2] AI Act Implementation Tool, Algorithm Audit (2025); Implementing the AI Act– Definition of an AI-system, Algorithm Audit (2025).



Description step-by-step guide

With the Schufa ruling, the Court of Justice of the European Union (CJEU) clarifies how the prohibition on automated decision-making, as established in Article 22 of the General Data Protection Regulation (GDPR), should be interpreted in the context of risk profiling. The EU's highest court determines that decision-making based solely on profiling occurs when: 1) a decision is made, 2) it is based solely on profiling, and 3) It has legal effects or otherwise significantly affects the individual concerned.[3] Steps 1-5 assess whether this cumulative requirement is met.

Assessing whether decision-making is based solely on profiling is not merely a qualitative exercise. The impact of profiling algorithms on the decision-making process also has an empirical dimension. In the Schufa ruling, the CJEU notes that there is an *"automated establishment of a probability value based on personal data"* and that *"an insufficient probability value leads, in almost all cases, to the refusal of that bank to grant the loan applied for."*.[4] Based on research into practical applications of algorithm-driven decision-making, particularly in the grant control process of the Dutch Executive Agency for Education (DUO),[5] and a machine learning-driven risk profiling algorithm applied by a commercial car sharing platform[6], in Step 4 an empirical method is discussed how can be determined to what extent the outcome of a risk profiling algorithm is followed by evaluators. This empirical insight can inform the assessment of whether, in *"almost all cases,"* the advice of the algorithm is followed by a decision-maker. Step 5 outlines an empirical method to determine whether automation bias is present in the decision-making process.

For this public standard, the Schufa ruling[7], the Consultation Document on Meaningful Human Intervention by the Dutch Data Protection Authority (AP)[8], the advice on Article 22 GDPR and automated selection tools by the AP [9], the advice on Automated selection techniques by Pels Rijcken[10], guidelines from the European Data Protection Board (EDPB)[11] and legal-scientific literature[12] [13] have been consulted.

3 ECLI:EU:C:2023:957, case C-634/21, Court of Justice of the European Union (2023).
4 Supra note 3, considerations 47 and 48.
5 Addendum Preventing prejudice, Algorithm Audit (2024).
6 To be published algoprudential case.
7 Supra note 3.
8 Consultation Meaningful human intervention, Dutch Data Protection Authority (AP) (2025).
9 Advice Article 22 GDPR and automated selection tools, Dutch Data Protection Authority (2024).
10 Advice on automated selection techniques, Pels Rijcken (2024).
11 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01), European Data Protection Board.
12 Legal protection against risk profiling based on the GDPR, the ECHR, and the Charter of Fundamental Rights, F. Çapkurt, Dutch journal for legal professionals (2025).
13 The Right to an Explanation in Practice: Insights from Case Law for the GDPR and the AI Act, L. Metikos en J. Ausloos, Law, Innovation and Technology (2025).

Box 1

Advice Dutch Data Protection Authority on article 22 GDPR and automated selection tools

In light of the Schufa ruling and the subsequent advice from the Dutch state's own lawyers[10] the Dutch Data Protection Authority (AP) has issued guidance on how Article 22 GDPR should be interpreted in the context of risk profiling algorithms.[9] In this guidance, the AP states that risk profiling can be applied without specific legal provisions, provided the following five conditions are met:[14]

i. Investigate discriminatory processing and implement mitigating measures if necessary;
ii. Periodically assess whether discrimination occurs;
iii. Ensure that the consequences of risk profiling only take effect after meaningful human intervention, see Steps 3-5 of this standard;
iv. Prevent risk profiling from having other significant consequences, see Step 1.1 of this standard;
v. Make the use of risk profiling known to the individuals concerned.

This public standard provides practical guidance for conditions **iii-iv** of the AP's advice. For steps **i-ii**, the public standard on Profiling algorithms by Algorithm Audit can be used.[15]

The above advice from the AP is not undisputed. It deviates on crucial points from the legal analysis of the Dutch state's own lawyers, particularly in the interpretation of what constitutes a decision with significant consequences. Other experts point to a one-sided reading of the GDPR, especially in relation to other European legislation, such as the European Convention on Human Rights (ECHR) and the Charter of Fundamental Rights of the European Union (Charter).[12] Nevertheless, the government has fully adopted the AP's advice.[16] Algorithm Audit acknowledges both the criticism on the AP's advice and the initiative to provide concrete actionable guidance for algorithm-driven decision-making processes. By publishing the public standards on Profiling algorithms and Meaningful human intervention for risk profiling algorithms, Algorithm Audit hopes to contribute to how risk profiling algorithms can be responsibly used in practice.

It should also be noted that prioritizing 'blind' evaluation by decision-makers in Step 3 of this standard deviates from the recommendation of the Dutch Data Protection Authority (AP). On p.6 of its recommendation, the AP states: "*To be meaningful and to actually prevent or modify outcomes if necessary, the person who intervenes must be able to adequately assess whether the selection in a particular case is justified. Therefore, the handling officer must know how the automated process (selection rule/algorithm/technology) works and must understand how and in what way this (the formation of) the final decision is shaped and influenced.*" Algorithm Audit places greater weight on the risk of *automation bias* in this approach (the tendency of evaluators to adopt the recommendations of the risk profiling algorithm without critical thought) than on the importance of evaluators understanding how risk profiling works. This is because such understanding should not be the primary responsibility of first-line decision-makers but rather that of second-line algorithm specialists.

[14] Supra note 9, p.14-17.
[15] Supra note 1.
[16] Dutch parliamentary papers 2024/25 2024D47487.

Step-by-step guide

Performing Steps 1-5 helps preventing prohibited automated decision-making when using risk profiling algorithms, but it does not provide a guarantee, as it depends on how the steps are carried out and the choices made during the process.

Step 1 – Create overview decision-making process

1.1 Create a schematic overview of the entire decision-making process. Determine whether the outcomes of the risk profiling algorithm will be stored for the long term or shared internally or externally, potentially leading to 'significant consequences'.[17]

1.2 Ensure that stakeholders in the decision-making process have the ability to appeal the decision.

1.3 Ensure that the use of a risk profiling algorithm is sufficiently disclosed to the individuals concerned, for example, by mentioning the input data for the algorithm through a letter or a pop-up notification in the app of the platform or service.

1.4 Determine at which moments in the decision-making process meaningful human intervention may be required.

NOTE: If Step 1.1 establishes that the outcomes of a risk profiling algorithm are stored long-term or shared internally or externally, significant consequences for stakeholders can easily arise. This is prohibited without specific legal provisions and corresponding safeguards.[18] When this is the case, it is no longer possible to prevent prohibited automated decision-making under Article 22(1) GDPR by following the subsequent steps in this standard. The current approach can only continue if one of the exceptions in Article 22(2) GDPR applies and adequate safeguards have been implemented.

Step 2 – Determine type of decision

Whether a risk profiling algorithm falls under the prohibition in Article 22 GDPR depends on the effect of the decision that, informed by the algorithm's outcome, is made. Assess whether the algorithm informs a decision that has a 'legal effect' on individuals or otherwise 'significantly affects' them. This is the case, for example, if one of the following types of decisions is made:[19]

i. A formal decision, such as imposing a tax assessment, granting or denying a benefit or allowance, making a decision following an appeal, or granting or denying a permit or subsidy;

ii. A decision with financial consequences, such as the ability to obtain a payment plan or qualify for credit;

iii. Entering into an agreement, such as an employment contract or a purchase agreement;

iv. Selection for an inspection, if the inspection is intrusive for the individual, such as a home visit;

v. A decision affecting someone's access to education, such as admission to a university or school assignments;

vi. Decisions affecting someone's employment opportunities, such as processing job applications or assigning projects to freelancers;

vii. Otherwise significantly impacting the individual.

The above is not an exhaustive list. A context-dependent assessment must always be made to determine whether a decision has significant

[17] note 9, p.6-8 and p.16. When the outcomes of an algorithm are stored or shared for other purposes, significant consequences for the individuals involved can quickly follow. For instance, stored or shared outcomes of a risk profiling algorithm can have a long-lasting or permanent effect on the individual. This could happen if a stored risk selection or risk score repeatedly leads to investigations, or through a self-reinforcing effect of successive investigations. By sharing outcomes, a risk selection or risk score can take on a life of its own, especially if third parties use the outcome in unforeseen ways, without proper safeguards, or by creating a "blacklist." For example, various municipalities used the outcomes of the "Preselect recidivism" algorithm to create lists to monitor young people. See article of Follow The Money.

[18] Supra note 9, p.16.

[19] Supra note 8, 10, 11 and 13.

consequences for an individual. This assessment should take into account potential outcomes (opportunities/risks). The consequences do not need to have already occurred, and they do not necessarily need to be the same for all individuals involved.[20]

Examples of decisions without legal or significant consequences include:

- **i.** Issuing a warning;[21]
- **ii.** Prioritization of applications, requests, or complaints, without affecting their processing;
- **iii.** Selection for inspection, when the inspection is not intrusive for the individual. The AP states that providing additional information for an inspection does not have significant consequences for the individual.[22]

There is no prohibited automated decision-making if the algorithm-driven decision-making process does not result in significant consequences for the individuals involved. Algorithm Audit recommends implementing measures for meaningful human intervention in this case (see Step 3), even though this is no longer a legal requirement.

NOTE: If the algorithm-driven decision-making process is implemented with limited justification and/or documentation, it can indirectly – through violations of fundamental rights (such as respect for privacy or equal treatment) – still significantly impact individuals and thereby fall within the scope of Article 22 GDPR. See also the public standard on Profiling algorithm by Algorithm Audit.[23]

Step 3 – Meaningful human intervention

3.1 Determine whether all cases are shared 'blindly' with evaluators. In a 'blind' selection, the evaluator does not know how the case was selected (by a risk profiling algorithm, randomly, or through another method). The evaluator also cannot infer this from the context. In blind evaluation, the evaluator is not influenced by the outcome of the algorithm.

3.2 Gain insight into the circumstances in which evaluators must make a decision. The following questions may help:[24]

- **i.** On the basis of which information should evaluators assess a decision or challenge the algorithm's outcome?
- **ii.** How much data do evaluators see when making a decision?
- **iii.** What requirements are placed on evaluators to make a decision?
- **iv.** How much time do evaluators typically have to assess the outcome of an algorithm? How does this relate to the nature of the decision to be made?

3.3 Determine whether there is meaningful human intervention by an evaluator. Check if the following questions can be answered affirmatively:[24]

- **i.** If there is no blind evaluation (see Step 3.1): Do evaluators understand how and on what data the profiling algorithm arrives at a result?
- **ii.** If there is no blind evaluation (see Step 3.1): Would evaluators be able to make the decision without the profiling algorithm?
- **iii.** Do evaluators have sufficient time to make an informed decision?
- **iv.** Can evaluators consider specific circumstances in their assessment that the algorithm does not take into account?

[20] Supra note 9, p.8.
[21] Note: When a formal decision is made here, it is indeed a decision with legal consequences.
[22] Supra note 9, p.8.
[23] Supra note 1.
[24] The 10 questions from steps 3.2 and 3.3 have been selected by Algorithm Audit as the most relevant questions out of the 93 questions included in the Consultation meaningful human intervention by the AP. The other question is included based on Algorithm Audit's practical experience; see note 25 below.

Ensure that profiling characteristics are not included both in the risk profiling algorithm and in the work instructions.[25]

v. Do evaluators have the opportunity to ask each other or a supervisor for assistance?

vi. Are quality reviews conducted on the work of evaluators?

vii. Is the algorithm adjusted based on feedback from evaluators, stakeholders, or monitoring?

When established in Step 3.1 that cases are shared blindly with evaluators, the evaluation is entirely independent of the risk profiling algorithm. The decision is then not solely based on profiling. In this case, there is no prohibited automated decision-making.

If the questions from Step 3.3 about the decision-making process can be answered affirmatively, it is likely that evaluators consider factors other than just the outcome of the risk profiling algorithm. The decision is then not solely based on profiling because meaningful human intervention is involved.[26]

Step 4 – Data-analysis effect of risk profiling algorithm

NOTE: If in Step 3.1 you have determined that cases are shared blindly with evaluators, then this step does not need to be carried out.

4.1 Divide the outcomes of the risk profiling algorithm into two categories before they are presented to an evaluator for decision-making, for example, a 'high risk' category and a 'less high risk' category. The 'high risk' category is referred to as positives, and the 'less high risk' category is referred to as negatives.[27]

4.2 Determine which portion of the cases presented to evaluators for decision-making fall into the 'high risk' category (positives) and the 'less high risk' category (negatives).

4.3 Calculate the true positive rate: divide the number of cases where the evaluator agrees with further action as recommended by the risk profiling algorithm (true positives) by the number of cases categorized as 'high risk' by the algorithm (positives).

4.4 Determine whether the follow-up actions taken after human intervention primarily consist of cases marked as 'high risk' by the risk profiling algorithm (high true positive rate). If the true positive rate is too high, the significance of human intervention should be questioned. In this case, repeat Step 3 and carry out Step 5.

Based on Steps 4.1-4.4, it can be determined how often evaluators disagree with the risk algorithm's prediction. A best practice is to not only present cases classified as 'high risk' by the risk profiling algorithm (blindly) to evaluators but to supplement them with a pre-established proportion of randomly selected cases, and possibly supplemented with other forms of selection, such as signal-driven cases.[28]

[25] For example, in the CUB process of DUO, both the risk profiling algorithm and the work instructions made a distinction based on the characteristic 'distance to parent(s)'.

[26] Article 29 Working Party, Guidelines on Automated Individual Decision-Making and Profiling for the Application of Regulation (EU) 2016/679, 2017, p. 24.

[27] In the CUB process of DUO, the outcome of the risk profiling algorithm (a risk score between 0 and 180) was divided into a 'high risk' category (score between 60-180) and a 'less high risk' category (score between 0-59).

[28] Sgnal-driven cases are selected based on signals from other parts of the organization. In businesses, for example, this could be the finance department monitoring overdue payments. In municipalities, this could be the Work and Income department informing the Enforcement and Supervision department about potentially suspicious situations.

Box 2

Example Step 4 – Determine true positive rate

Step 4 is illustrated with an example. In this example, it is assumed that all individuals in the target population are assigned a score by the risk profiling algorithm.

4.1 Of the 13 individuals involved, 5 are classified as 'high risk' (positive) by the risk profiling algorithm.
4.2 Not only the 5 positives, but also 2 negatives are blindly presented to an evaluator.
4.3 From the human intervention, the following results:

- 3 out of the 5 positives were correctly classified as positive (true positives);
- 2 out of the 5 positives were incorrectly classified as positive (false positives);
- 1 out of the 2 negatives was correctly classified as negative (true negative);

1 out of the 2 negatives was incorrectly classified as negative (false negative). Note that this ratio should be adjusted once the outcome of any potential objection procedure for an individual is known.

4.4 Determining the true positives ratio (3/5 = 60%) can assist in assessing whether there is meaningful human intervention. The higher the true positives ratio, the more likely it is that there is no meaningful human intervention. In that case, there may be (prohibited) fully automated decision-making.

Practical example regarding Step 4.2: In the CUB process of DUO in 2014, 2,400 out of 3,179 selected students for a home visit were assigned the 'high risk' label by the risk profiling algorithm (75%). 640 students selected by an evaluator for a home visit had been assigned the 'low risk' label, and 140 had been assigned the 'unknown risk' label by the risk profiling algorithm (20% and 5%, respectively).[29] Not only cases classified as 'high risk' were selected for further investigation.

Practical example regarding Step 4.4: An evaluator of a car-sharing platform decides to go against the recommendation of a risk profiling algorithm in 40-50% of cases and does not send a warning to a user for risky driving behavior. The evaluator is trained, has clear work instructions, and sufficient time to make a judgment. Additionally, users can appeal the decision and request the processed data from the platform. There is meaningful human intervention. In this step of the decision-making process, no prohibited automated decision-making occurs.[30]

[29] Supra note 5 p.40. Due to rounding to the nearest tens, the sum of the parts differs from the total.
[30] Supra note 6.

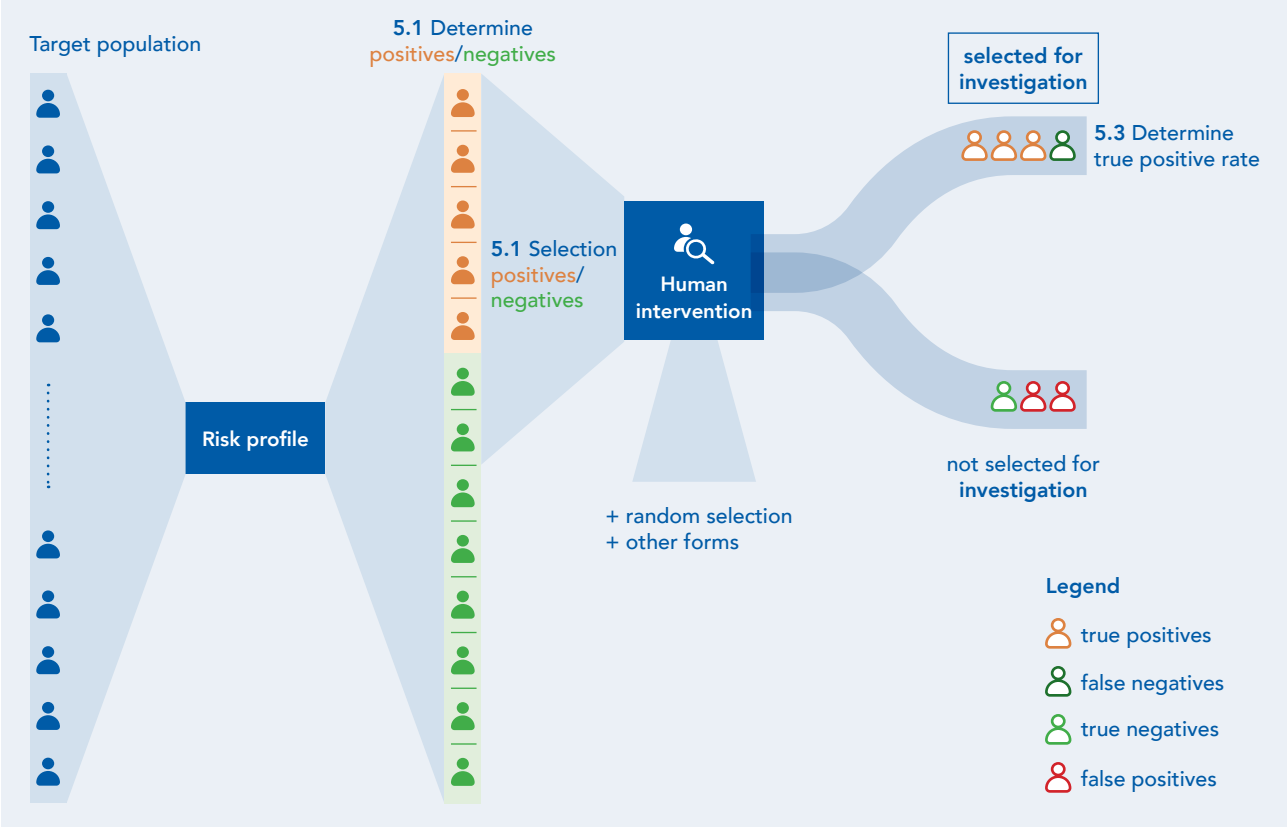


Figure 1 - Determining how often evaluators deviate from an action recommended by the algorithm can help in determining whether meaningful human intervention is involved.

Step 5 – Field experiment automation bias

NOTE: If Step 4 has established that evaluators frequently go against the prediction of the algorithm, then this step does not need to be carried out. If this is not the case, the following experiment can be conducted to determine whether an evaluator is influenced in their decision-making by seeing a label generated by a risk profiling algorithm.[31]

5.1 Formulate the following hypotheses:

- H_0: Visibility of a label generated by a risk profiling algorithm for a case influences the decision made by the assessor;
- H_A: Visibility of a label generated by a risk profiling algorithm for a case does not influence the decision made by the assessor.

5.2 Select a set of (e.g., 10) realistic cases. These can also be fictional. Ensure the outcomes are known, e.g., 'high risk' or 'less high risk' categories.

5.3 Divide the assessors into two groups (e.g., 10 assessors per group). Group A does not see the label generated by the risk profiling algorithm. Group B sees the label. Group A is the control group.

5.4 Determine the sample size: Decide how often the cases generated in Step 5.2 will be presented to the assessor groups divided in Step 5.3.[32]

5.5 Conduct the following experiment:

- Group A (Control group): This group sees the cases with all relevant information for assessment and the work instructions used in the regular process. They do not see

[31] The field experiment described in Step 5 is inspired by Dutch parliamentary papers 2023/24 2024D17779.

[32] Zie Random sample size for single and multiple hypothesis tests, Algorithm Audit (2024).

the label generated by the risk profiling algorithm. They assess the case based on this information;

- Group B: This group sees the cases with the information "This case has been rated as high risk by the risk profiling algorithm" or "This case has been rated as less high risk by the risk profiling algorithm." The risk category is randomly assigned: half of Group B gets a case labeled 'high risk,' and the other half gets a case labeled 'less high risk.' A real algorithm is not used to assign the risk categories.[33] This group also receives all relevant information for assessment and the work instructions used in the regular process. They assess the case based on this information, including the risk category.

5.6 Label decisions as correct (1) or incorrect (0) based on the decisions of the assessors for both Group A and Group B. Whether a case is correctly assessed depends on the actual outcomes established in Step 5.2. Calculate the percentage of correct decisions in Group A and Group B.

5.7 Apply a two-tailed Z-test to determine if there is a statistically significant difference between the correct assessments in Group A and Group B. In line with Step 5.1, the following hypotheses apply:

- H_0: the proportion of correctly assessed cases is different between Group A and Group B, i.e., $p_A \neq p_B$;
- H_A: The proportion of correctly assessed cases is the same between Group A and Group B, i.e., $p_A = p_B$.

5.8 Accept or reject H_0. Use a significance level of $p < 0.05$.

[33] By randomly assigning outcomes, this experiment is not dependent on the quality or functionality of an existing algorithm.

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. The goals of the NGO are three-fold:

Knowledge platform

Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, see for instance our AI Policy Observatory and position papers

Normative advice commissions

Forming diverse, independent normative advice commissions that advise on ethical issues emerging in real world use cases, resulting over time in algoprudence

Technical tools

Implementing and testing technical tools for bias detection and mitigation, e.g, bias detection tool, synthetic data generation

Project work

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

Structural partners of Algorithm Audit

SIDN Fund

The SIDN Fund stands for a strong internet for all. The Fund invests in bold projects with added societal value that contribute to a strong internet, strong internet users, or that focus on the internet's significance for public values and society.

European AI&Society Fund

The European AI&Society Fund supports organisations from entire Europe that shape human and society centered AI policy. The Fund is a collaboration of 14 European and American philantropic organisations.

Dutch Ministy of the Interior and Kingdom Relations

The Dutch Ministry of the Interior is committed to a solid democratic constitutional state, supported by decisive public management. The ministry promotes modern and tech-savvy digital public administrations and govermental organization that citizens can trust.

Building *AI auditing* capacity from a *not-for-profit* perspective

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Algorithm Audit

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