Algorithm Audit

Stakeholder panels for AI bias testing

Towards inclusive, deliberative and transparent AI bias testing standards in JTC21 CEN-CENELEC, SC42 and ISO

January 2024

Overview

1 What is a stakeholder panel?

2 Why are stakeholder panels needed for AI bias testing?

3 Bias and fundamental rights in the AI Act

4 Relevant JTC21/SC42 and ISO activities

Activities NGO Algorithm Audit

Normative advice commissions	Advising on ethical issues emerging in concrete algorithmic practices through deliberation, resulting in *algoprudence*
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

Financially supported by

SIDN fonds

European Artificial Intelligence & Society Fund

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

1. **What is a stakeholder panel?**

2. Why are stakeholder panels needed for AI bias testing?

3. Bias and fundamental rights in the AI Act

4. Relevant JTC21/SC42 and ISO activities

A diverse group of people having a deliberative conversation on ethical issues emerging in AI

Stakeholder panel

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

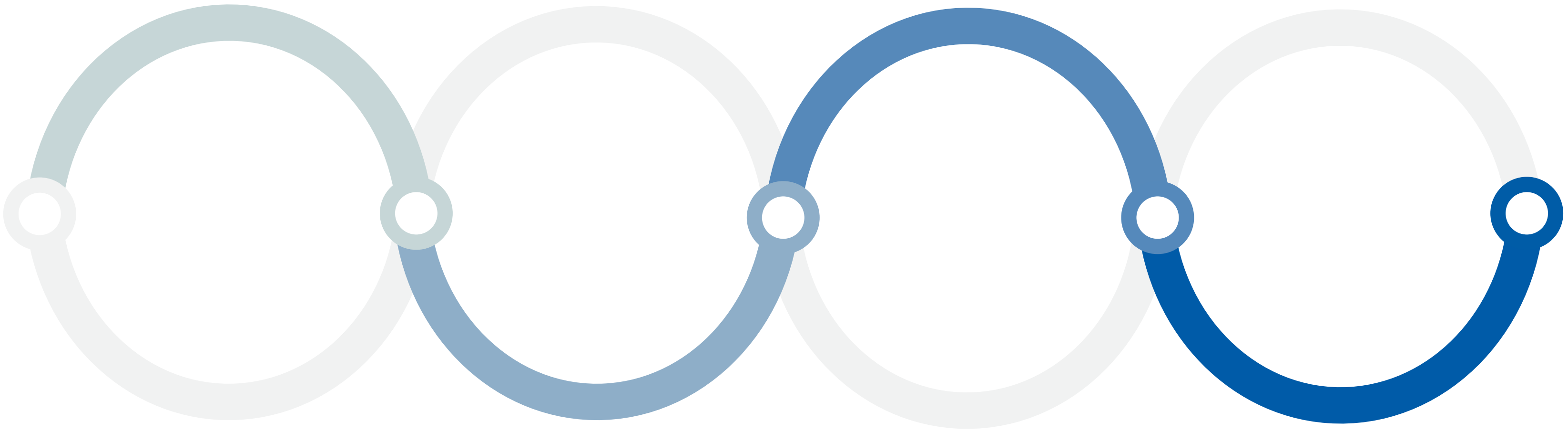
diverse

inclusive

deliberative

transparent

Working plan to convene stakeholder panels to assess bias in AI input and output



Step 1

Identify issue

Identify a concrete ethical concern in a real algorithm or data-analysis tool

Step 2

Problem statement

Describe ethical issue and hear stakeholders and affected groups

Step 3

Stakeholder panel

Deliberative conversation on ethical issue by diverse and inclusive stakeholder panel

Step 4

Public advice

Advice of panel is published together with problem statement

Composition of stakeholder panels vary per case, but share common dividers

Overview of stakeholders (not exhaustive)

- Model owner
- People subjected to the algorithm
- Legal, statistical, ethical experts
- Representatives of affected groups
- Subject matter experts

There is no universally optimal method for incorporating people subjected to an algorithm in a normative advice commission. Experiment with various working formats is therefore encourages, among others:

> Include a person subjected to the algorithm as part of the normative advice commission;

> Include people subjected to the algorithm in defining the problem statement prior to the panel gathering;

> Include people subjected to the algorithm by hosting focus sessions in parallel to the panel gathering.

The above options are not mutually exclusive. Please reach out if you think other options should be taken into account.

1. What is a stakeholder panel?
2. **Why are stakeholder panels needed for AI bias testing?**
3. Bias and fundamental rights in the AI Act
4. Relevant JTC21/SC42 and ISO activities

Testing AI bias cannot be automated. Context-dependent answers are needed

Technical component of AI – objectively verifiable

- Focus on pre-defined technical standards, set by (inter)national agencies, e.g., plug design, fixed wavelength for all microwaves

- AI product safety measures is aim of EU AI Act
- AI auditing as check-list exercise:

□ Risk management: Documentation, monitoring and evaluation practices in place?
□ Record-keeping: Logging capabilities conform (CE-) standards?
□ Conformity assessment: Procedures for internal and/or external validation?

Standardizable | Objective truth | Private auditors

Normative component of AI – open for interpratation

- AI Governance provides a framework, but does not answer difficult questions:
  - What is good? What is bad?
  - When is training data biased?
  - Is all disparate impact harmful?
- Political questions, difficult to square with mandate of commercial parties
- Normative component in AI auditing is different from existing auditing practices:
  - Financial sector: Asset-liability risk management is less value-driven;
  - Drug approval: Consensus over objective safety measures and medical trial procedures

Context-dependent | Open for debate | Public task

Algorithm Audit advocates inclusion of stakeholder panels in CEN-CENELEC/JTC21 standards

How to strike a balance between precision and recall?

- Judicial system: minimize false positives (sentencing innocent)
- Medical system: minimize false negatives (leave diseases undetected)

Similar approach in other regulatory instruments:

- AI Act: EU office for foundation models, i.e., multi-stakeholder composition
- GDPR art. 39(5): When a Data Privacy Impact Assessment (DPIA) is mandatory, stakeholders should be heard

Key take-away for AI bias testing standards:

- Guidelines can be developed how processes for bias testing can be made inclusive, deliberative, and transparent
- Standardized way to resolve non-standardizable issues

Algorithm Audit

1. What is a stakeholder panel?

2. Why are stakeholder panels needed for AI bias testing?

3. Bias and fundamental rights in the AI Act

4. Relevant JTC21/SC42 and ISO activities

Overview of AI Act articles relating to bias and fundamental rights

Art. 4 – Amendments to Annex I

‘diversity, non-discrimination and fairness’ means that AI systems shall be developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law;

Art. 9 – Risk management system

- Assess whether risk management system is in place
- Document and maintain risk management obligations for algorithm documentation, monitoring and evaluation

Art. 10 – Data and data governance

- Assess existing data collection, data processing and data quality checks
- If these exist, assess documentation of relevant design choices and assumptions, including bias detection and mitigation measures

Recital 18

Technical inaccuracies of AI systems intended for the remote biometric identification of natural persons can lead to biased results and entail discriminatory effects. This is particularly relevant when it comes to age, ethnicity, sex or disabilities.

Art. 69 – Codes of conduct

including where they are drawn up in order to demonstrate how AI systems respect the principles set out in Article 4a and can thereby be considered trustworthy

Art. 15 – Accuracy, robustness, cyber security

after being placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as an input for future operations (‘feedback loops’) are duly addressed with appropriate mitigation measures.

Recital 44

Training, validation and testing data sets … with specific attention to the mitigation of possible biases in the datasets, that might lead to risks to fundamental rights or discriminatory outcomes for the persons affected by the high-risk AI system.

Art. 43 – Conformity assessment

- Comply to CE certification and available non-CE certifiable content
- Carry out examination, test and validation procedure before, during and after development of AI system
- Pre-market assessment and post-market monitoring

Art. 28 – **Obligations of the provider of a foundation model**

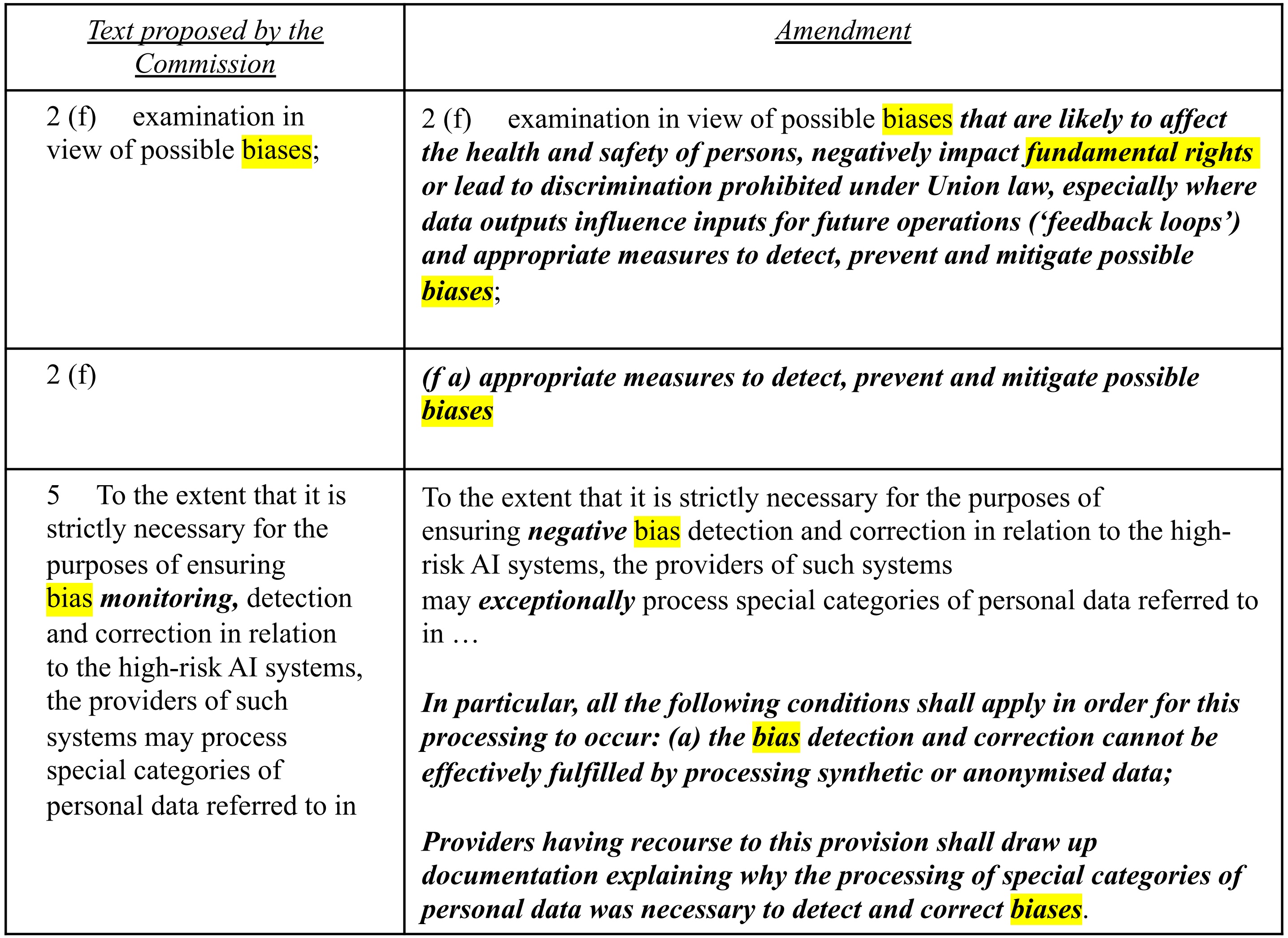
- Process and incorporate only datasets that are subject to appropriate data governance measures for foundation models, in particular measures to examine the suitability of the data sources and possible biases and appropriate mitigation

Example #1 on bias testing – Art. 10 Data and data governance

Art. 10 – Data and data governance

2. Application of appropriate techniques for data governance and data management
   f. Examination in view of possible biases;

5. To the extent that it is strictly necessary for the purpose of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems … appropriate safeguards for the fundamental rights of natural persons



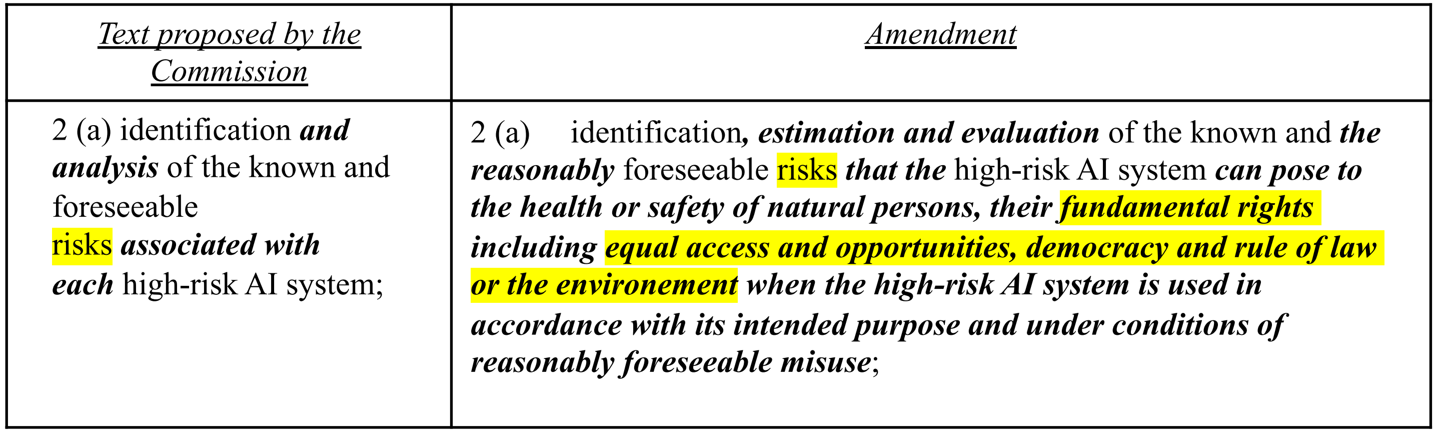
Text proposed by the Commission	*Amendment*
2 (f) examination in view of possible biases;	2 (f) examination in view of possible biases ***that are likely to affect the health and safety of persons, negatively impact fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations ('feedback loops') and appropriate measures to detect, prevent and mitigate possible biases***;
2 (f)	***(f a) appropriate measures to detect, prevent and mitigate possible biases***
5 To the extent that it is strictly necessary for the purposes of ensuring bias ***monitoring,*** detection and correction in relation to the high-risk AI systems, the providers of such systems may process special categories of personal data referred to in	To the extent that it is strictly necessary for the purposes of ensuring ***negative*** bias detection and correction in relation to the high-risk AI systems, the providers of such systems may ***exceptionally*** process special categories of personal data referred to in … ***In particular, all the following conditions shall apply in order for this processing to occur: (a) the bias detection and correction cannot be effectively fulfilled by processing synthetic or anonymised data;*** ***Providers having recourse to this provision shall draw up documentation explaining why the processing of special categories of personal data was necessary to detect and correct biases***.

Example #2 on risks of violating fundamental rights – Art. 9 Risk management system

Art. 9 – Risk management

1. A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems.

a) identification ***and analysis*** of the known and foreseeable risks ***associated with each*** high-risk AI system;



Text proposed by the Commission	*Amendment*
2 (a) identification ***and analysis*** of the known and foreseeable risks ***associated with each*** high-risk AI system;	2 (a) identification***, estimation and evaluation*** of the known and ***the reasonably*** foreseeable risks ***that the*** high-risk AI system ***can pose to the health or safety of natural persons, their fundamental rights including equal access and opportunities, democracy and rule of law or the environement when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse***;

1. What is a stakeholder panel?
2. Why are stakeholder panels needed for AI bias testing?
3. Bias and fundamental rights in the AI Act
4. **Relevant JTC21/SC42 and ISO activities**

Overview of AI bias testing related standards

Europe

AI Act

Product safety regulation

↓ Articles

CEN CENELEC

↓

JTC21/SC42

- WG2 – Risk management
- WG3 – Data engineering
- WG4 – Trustworthiness

- N472 – AI Risk management
- N204 – Bias requirements for managing bias in AI systems
- N468 – Trustworthiness framework

International

ISO IEC ITU

- ISO/IEC 23894 AI bias terms
- ISO/IEC 12791 Treatment of unwanted bias in classification and regression machine learning tasks
- ISO/IEC 42001 AI System Management
- ISO/IEC 42005 Impact assessment

International values

⚡

Vienna agreement

European values

WG4 – Fundamental Rights Impact Assessment and stakeholder panels as part of Trustworthy AI

Benchmark analysis of FRIAs, with Dutch FRIA as benchmark

Ministry of the Interior and Kingdom Relations

March 2022

Impact Assessment

Fundamental rights and algorithms

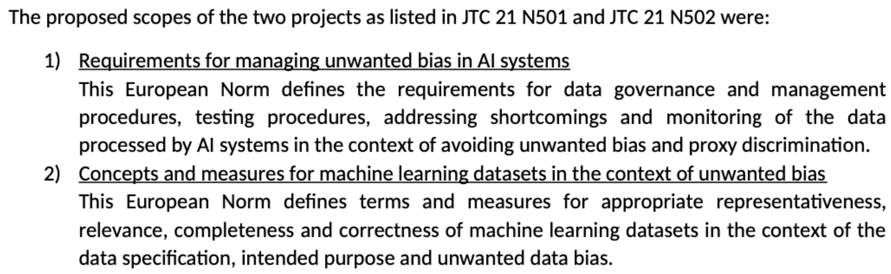
+

Stakeholder panels

Algorithm Audit

WG3 – Stakeholder panels as part of AI bias testing procedure (engineering aspects)

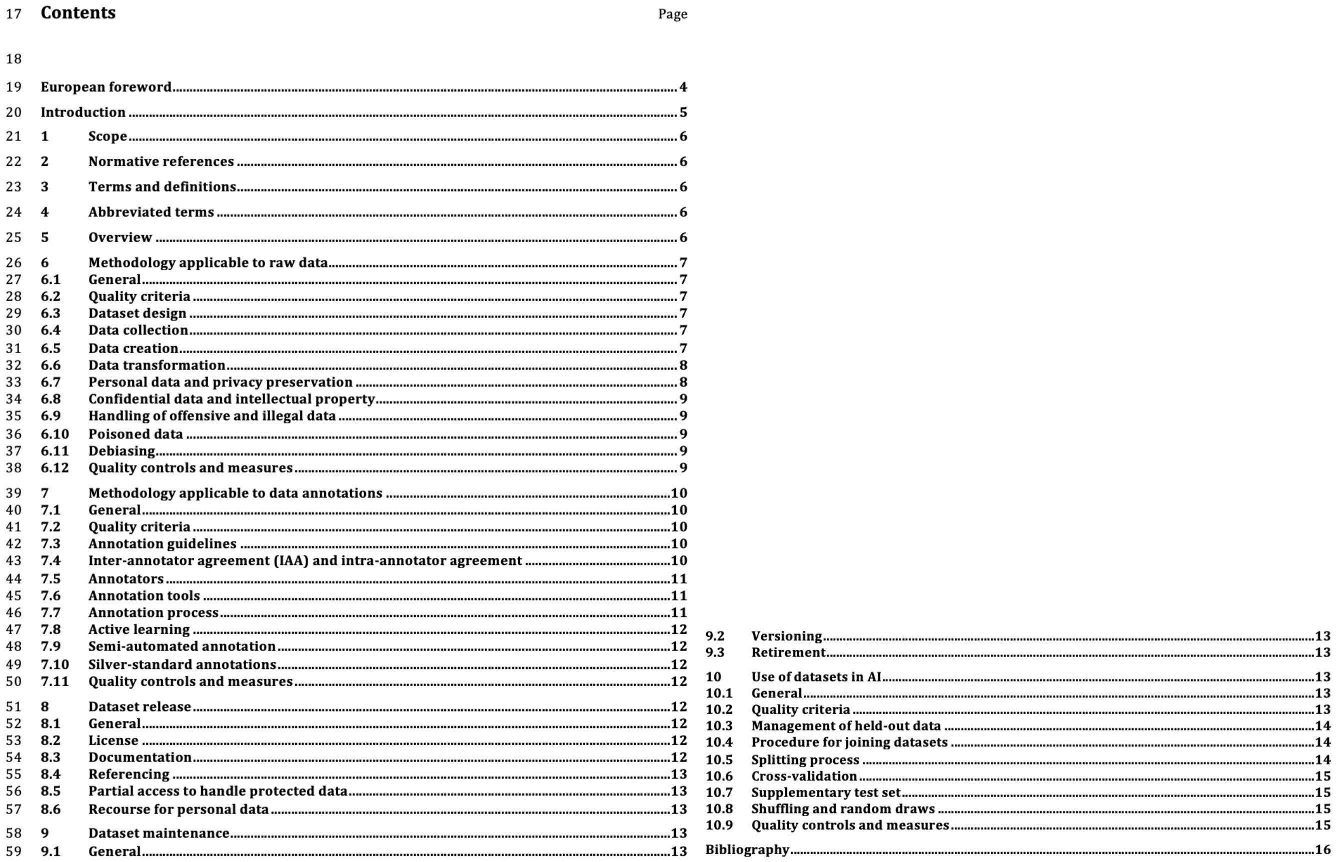
Scope preliminary work item (PWI) on bias standards



The proposed scopes of the two projects as listed in JTC 21 N501 and JTC 21 N502 were:

1) Requirements for managing unwanted bias in AI systems
This European Norm defines the requirements for data governance and management procedures, testing procedures, addressing shortcomings and monitoring of the data processed by AI systems in the context of avoiding unwanted bias and proxy discrimination.
2) Concepts and measures for machine learning datasets in the context of unwanted bias
This European Norm defines terms and measures for appropriate representativeness, relevance, completeness and correctness of machine learning datasets in the context of the data specification, intended purpose and unwanted data bias.

NWIP Data outline v0



Contents Page

WG2 – Risk management system

Scope of NWIP encourages contributions regarding fundamental rights:

Risks covered include both risks to health and safety and risks to fundamental rights which can arise from AI systems, with impact for individuals, organisations, market and society. This document also defines methods that can be used to determine if a package of risk management measures associated with an AI system will be able to ensure that certain risks arising from that product or system are identified, monitored, and managed, leading to an acceptable level of risk.

Algorithm Audit

We build public knowledge for ethical algorithms

🌐 www.algorithmaudit.eu

✉ info@algorithmaudit.eu

in 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