CYBERNETICA

Co-funded by
the European Union

# Algorithmic bias risk management guideline

Guideline

Version 1.1

21 October 2025

D-16-583

PUBLIC

| | |
|---|---|
| Project Managers: | Liina Kamm (Cybernetica AS) |
| | Henrik Trasberg (Ministry of Justice and Digital Affairs) |
| | Sofia Paes (Ministry of Justice and Digital Affairs) |
| Authors: | Dan Bogdanov |
| | Paula Etti |
| | Hiroki Kaminaga |
| | Liina Kamm |
| | Tanel Mällo |
| | Tanel Pern |
| | Fedor Stomakhin |
| | Anto Veldre |

Cybernetica AS, Mäealuse 2/1, 12618 Tallinn, Estonia.

E-mail: info@cyber.ee, Web site: https://www.cyber.ee, Phone: +372 639 7991.

Co-funded by the EU. The views and opinions expressed are those of the authors alone and do not necessarily reflect the views or opinions of the European Union. The European Union is not responsible for them.

| Date | Version | Description |
|---|---|---|
| 23.07.2025 | 1.0 | Translation and adaption of D-16-571 v1.0 to English |
| 21.10.2025 | 1.1 | Updated translation of D-16-571 v1.1 to English |

# Table of Contents

# 1 Introduction

## 1.1 Goal

AI is a fast-evolving family of technologies that contributes to a wide array of economic, environmental, and societal benefits across the entire spectrum of industries and social activities ([1] rec 4). When properly applied, AI technology can provide a competitive advantage and support both the society and the environment ([1] pp 4). At the same time, depending on the circumstances regarding its specific application, use, and level of technological development, AI can create risks and cause harm to public interests and fundamental rights. Such harm might be material or immaterial, including physical, psychological, societal or economic harm ([1] rec 5).

The algorithmic bias of AI systems is a risk that needs to be addressed separately, as it can cause imperceptible distortions in decision-making and assessment processes and lead to unjust, inaccurate, or inefficient results. Ignoring bias will increase inequality, undermine trust, and can result in errors. Addressing bias is necessary for ensuring a sense of justice, accuracy, and reliability of decisions in all contexts (see examples in 3).

AI systems, including general-purpose artificial intelligence (GPAI), are increasingly being deployed in the EU public sector to enhance governance and services. Both official and informal use is spreading – many officials use GPAI tools individually even before official AI strategies are established. While AI is widely used to improve services and increase internal efficiency, GPAI brings complex issues of governance, ethics and legal compliance. [2]

Recent research shows that the introduction of GPAI is driven by technological progress, management support, and citizen expectations. Successful integration, however, also depends on ethical awareness, the adaptability of the organisation, and the development of internal skills [2].

The EU and Member States are developing guidance on the safe, transparent and lawful use of GPAI, focusing on accountability, data protection, and human control. The Interoperable Europe Regulation [3] and the AI Europe/World Action Plan [4, 5] (The AI Continent Action Plan) support this direction by encouraging cross-border cooperation and the deployment of reliable AI. [2] The Action Plan aims to transform Europe's strong industries and talent base into a powerful engine for AI innovation. [6]

The public sector plays a leading role in improving the quality of services in different sectors such as health, education, law and administration. Skilled use of artificial intelligence can prevent discrimination and increase accessibility for example for people with special needs. The use of diverse and high-quality databases is essential to reducing bias and ensuring fairness in AI applications. [2]

## 1.2 Scope

The algorithmic bias risk management tool focuses on managing the risks of bias in algorithmic and AI systems. The tool comprises three parts: this Guideline, which describes the nature and background of artificial intelligence systems and the biases present in these and reviews the possibilities of identifying and mitigating biases; a methodology document, which provides detailed instructions for setting up and carrying out the risk management process; and a workbook designed to simplify the documentation of information required for risk management. The risk management tool is primarily targeted at systems used by organisations. The requirements

for algorithmic and AI systems that private persons can use for their own needs are somewhat less strict. We are treating the risk management process as a generic one because systems can greatly vary in their design and the sources of bias can also be different. Depending on the implementation and deployment details of the system, the organisation may not be able to assess their risks using purely technological means, which calls for a more general approach.

## 1.3 Terms and abbreviations

### 1.3.1 Terms

**AI system**
: a machine-based system that processes input data to provide answers (e.g. forecasts, content, recommendations, or decisions) to queries which can affect the physical or virtual environment. The main feature of AI systems is their ability to produce inferences and derive relationships from intput data through the use of machine learning models

**algorithm**
: a step-by-step procedure to perform some action

**bias**
: a systematic difference in processing certain objects, persons, and groups compared to others, where the processing can be whatever action including perception, observation, representation, prediction, or decision

**explainability**
: the potentiality to describe an AI model's internal workings or outcomes in transparent and understandable terms; it is meant to answer the question 'Why?' without trying to claim the chosen course of actions is necessarily optimal

**guardrails**
: mechanisms and frameworks acting as a security checkpoint, evaluating the user input and generated answers based on the defined safety rules. The purpose of the guardrails is to prevent the generation of harmful, inappropriate, or off-topic content in cases where the safety-tuning of the model itself is insufficient

**machine learning model (ML model)**
: a specific algorithm or a set of such to predict outputs based on inputs

### 1.3.2 Abbreviations

**AI** – artificial intelligence

**AI Act** – regulation (EU) 2024/1689 on AI

**AI system** – artificial intelligence system

**EU** – European Union

**GDPR** – General Data Protection Regulation (EU) 2016/679

**GPAI** – general-purpose artificial intelligence

**LLM** – large language model

**TEU** – Treaty on European Union

**TFEU** – Treaty on the Functioning of the European Union

**UN** – the United Nations

**Estonian legislative acts**:

**AvTS** – Avaliku teabe seadus (Estonian) *Public Information Act*)

**IKS** – Isikuandmete kaitse seadus (Estonian *Personal Data Protection Act*)

**KüTS** – küberturvalisuse seadus (Estonian *Cybersecurity Act*)

**TNVS** – Toote nõuetele vastavuse seadus (Estonian *Product Conformity Act*)

**Technical abbreviations** used citing legal texts:

**Rec** – recital (an informative introductory section in EU legislative acts)

**Art** – article (in EU legislative acts)

**Sec** – section (in EU legislation, the second hierarchy level within Articles)

## 1.4 Structure of the document

The algorithmic bias risk management guideline helps the user of the methodology to understand the bias related concepts and the actual threats. We recommend studying the structure of the Guideline before starting reading the Methodology or filling in the associated workbook. In several sections, the Guideline describes systems from all over the world that have been discovered to include algorithmic bias or biased AI components. These case reports help the reader understand and better identify and prevent threats.

In Section 2 we explain what an AI system is and how algorithms and ML models fit into it. We list the parties involved in artificial intelligence systems from both legal and IT perspectives and explain how models and systems are created. The section provides a (non-exhaustive) list of quality metrics and test kits for machine learning models and also explains the concept of guardrails. All this helps understand the concept of bias and the moments during the life cycle of a ML model or algorithm where bias can be introduced.

Section 3 explains why why bias should be avoided. For this purpose we describe what kind of damage algorithmic bias can cause to persons and organisations. This is followed by an overview of regulations that mandate the prevention of bias-related damage. We provide the reader with links to international standards and other materials on AI bias.

Section 4 illustrates how algorithmic bias risk management can be tied to other risk management activities. We revise the risk management process and give recommendations on how to describe the AI system. We explain what kinds of damage can be caused by bias present in an algorithmic system, as well as by over-reliance on AI in decision-making. The section concludes with recommendations on which stage of the system's life cycle provides the most opportune and cost-efficient moment for dealing with AI bias.

In Section 5 we assist user of the methodology in certain specific situations and describe practical tools. First, we help understand whether the AI system in question is a black box (no access to the algorithm or machine learning model and its training data) or a white box (the deployer can study or change the algorithm, machine learning model, or the training data). We will then explain what can be done for bias reduction in both types of systems.

# 2 AI systems and their bias

## 2.1 Inside of an AI system

An **AI system** is an adaptive machine-based system able to operate at different levels of autonomy. Based on the input data, it makes predictions and conclusions to create output data (prognoses, content, recommendations or decisions) that may influence the physical or virtual environment. The main feature of AI systems is their ability to produce inferences and derive relationships from input data. AI systems differ from purely algorithmic software systems in that the latter only operate based on predefined rules. Both AI systems and algorithmic systems are viewed here strictly as specific ML models (AI components) or sets of decision-making rules.

**Example. Threat assessment wizard for an Emergency Call Centre**

The threat assessment wizard is an AI-based application which, by means of transcribing the incoming call and taking into account the earlier cases, offers the rescue coordinator the most likely follow-up scenario blueprint for their actions. The system is meant to support a human (the rescue coordinator) during the threat assessment phase because gathering all the information by sequential questions in critical situations is too time consuming.

**ML model** is a specific algorithm or a set of such used to predict output data based on input data. An important property of the model is its trainability: the model can derive essential relations or patterns from data to improve its accuracy for future predictions. ML models are central components in AI systems enabling them to learn from data and make decisions or provide recommendations.

**Example. Business early notification prototype**

The Ministry of Economic Affairs and Communications together with Statistics Estonia have created a ML model capable of predicting whether a company's situation is deteriorating (based on the overall situation of that sector of economy). If additional analysis uncovers a threat, the computer system sends the business an early notification which does not lead to legal consequences but helps the business-owner get their business on track.

**Algorithm** is a step-by-step procedure to perform some action.

To understand the effect of bias it is essential to understand where in the system algorithms and ML models (see Figure 1) are located. An AI system gets its input from our living environment, this is further transformed into digital data and fed into algorithm(s) and ML model(s). The latter are following goals defined by humans and process input data until output data are produced. The impact the output data has on the world can be direct (if the AI system is provided with appropriate levers) or indirect (when a human changes the world based on the AI system's output data). In both cases, the AI system affects our living environment and if the AI system is biased, the effect on our environment will also be biased. The methodology presented here has been created to facilitate the recognition of this kind of bias and prevention of its consequences.

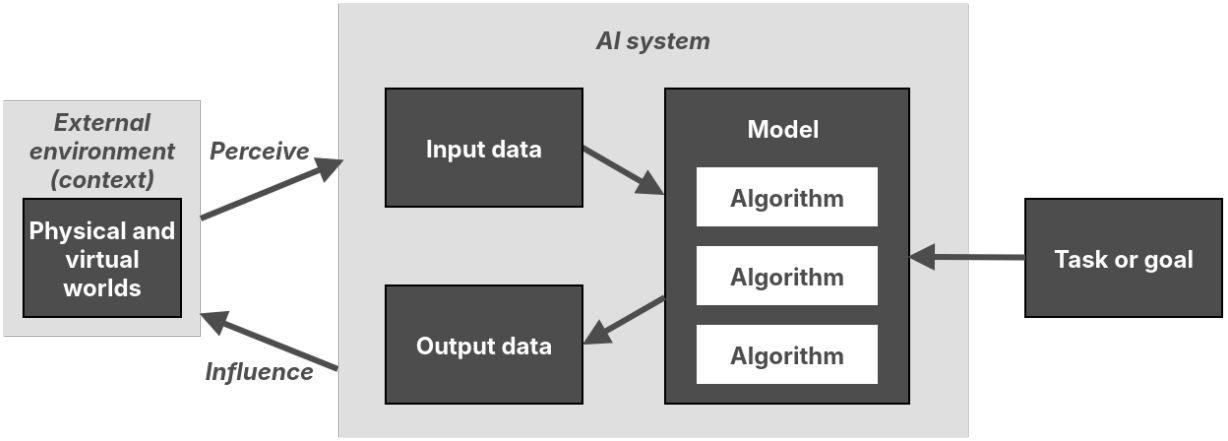


**Figure 1. Overall model of AI (adapted from [7]). The external environment, which can be either physical or virtual, is perceived by the AI system. This results in the generation of input data which is transmitted to the model. The model, which is trained with some (implicit or explicit) purpose in mind, generates output data, which, in turn, influences the external environment, e.g. via decisions made based on the data**

## 2.2 Creators and managers of AI systems

The EU AI Act defines a number of AI stakeholders (organisations and persons) as well as their roles in relation to AI systems. These roles are explained in the infobox on the next page. At the same time, AI systems are still software systems, meaning that they can be conceptualised in terms of ordinary IT systems.

Algorithmic and AI systems are created to simplify, accelerate, as well as substitute human labour. Therefore AI systems have an important relationship with the society and especially with people who can be using the AI systems or be affected by their output.

The methodology presented here is primarily targeted at systems used by organisations. The requirements for algorithmic and AI systems that private persons can use for their own needs are somewhat less strict.

Several organisational models are in use for the development and maintenance of IT systems; some of these models have been standardised. The following roles are vitally important in algorithmic and AI bias risk management.

**Service manager, project manager**. Each system has a person responsible for running the project of creating a system or for running an already created product or service. This role is well suited assume responsibility for the risk of the created system, as they have the capacity to direct the use of system-related resources. In a bigger team the actual risk management tasks can be delegated e.g. to the infosec manager, IT architect, or analyst.

**Product owner, service owner, analyst, architect**. The owners, analysts and architects know the system the best and know who to turn to for details related to the system. It is essential to involve these people into bias risk management.

**Specialists**. Programmers, data scientists, specialists in the fields of law, cyber security or data protection know the aspects of the system, product or service in depth. It is a good idea to involve them to clarify specific details of the particular algorithm, AI component, or usage scenario.

The representatives of these roles may be in-sourced from other organisations. For example, technical roles may actually work at a development partner, AI product vendor, or service provider.

The AI Act (EU) 2024/1698 lists the following AI stakeholders:

- **humans** – the society, social groups, individuals. The AI Act states that AI should be a human-centric technology, it should serve as a tool for people, with the ultimate aim of increasing human well-being (Rec 6).
- **provider** – a natural or legal person, public authority, agency, or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge (Art 3 (3));
- **authorised representative** – a natural or legal person located or established in the Union who has received and accepted a written mandate from a provider of an AI system or a general-purpose AI model to, respectively, perform and carry out on its behalf the obligations and procedures established by the AI Act (Art 2 (5))
- **deployer** – a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity (Art 3 (4)). 'Deployer' should be interpreted to mean any natural or legal person, including a public authority, agency or other body, using an AI system under its authority (Rec 13). The regulation applies to deployers of AI systems that have their place of establishment or are located within the Union (Art 2(1)(b)), as well as to providers and deployers of AI systems that have their place of establishment or are located in a third country, where the output produced by the AI system is used in the Union (Art 2(1)(c)).
- **importer** – a natural or legal person located or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established in a third country (Art 3 (6)).
- **distributor** – a natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market (Art 3 (7)).
- **operator** – a provider, product manufacturer, deployer, authorised representative, importer or distributor (Art 3 (8)).
- **downstream provider** – a provider of an AI system, including a general-purpose AI system, which integrates an AI model, regardless of whether the AI model is provided by themselves and vertically integrated or provided by another entity based on contractual relations (Art 3 (68)).
- **product manufacturer** – product manufacturers placing on the market or putting into service an AI system together with their product and under their own name or trademark (Art 2(1)(e)). Also see (Rec 87).

## 2.3 AI system life cycle

### 2.3.1 Creation of ML models

The development and implementation of an AI system comprises several stages; taken together, these make up the system's life cycle. These stages are essential in understanding and comprehending how bias comes into the system and distributes from there. The life cycle of an algorithm or ML model is not the same as the life cycle of a system implementing it.

Recommendations and decisions from an AI system have an effect on the external world; given that this in turn affects future data collection and decision-making, the system's life should be conceptualised as a cyclical process, not just linear progression from development to operation. The cyclical nature of an AI system's evolution is often the source of bias, especially if the machine learning system servers as a basis for newer versions of the system.

MIT scientists Suresh and Guttag [8] provide a comprehensive framework for the description of an ML model's life cycle stages. This model will also be used here as a basis for systematising the sources of AI and algorithmic bias. The stages making up an ML model's life cycle are as follows.

**Data collection.** This stage includes the identification of the development population, followed by the identification and measurement of relevant characteristics and notations. A development set is formed from the collected data. It is important to note that rather than starting with data collection from scratch, existing datasets are often re-used.

**Data preparation** – or pre-processing of collected data. This includes addressing missing data (aka imputation), simplification of characteristics, and normalisation of measurements. During the preparation phase, the dataset (development set) is usually divided into training data (for model development), validation data (for tuning the model during training), and test data (for the final evaluation of the model).

**Model training.** A ML model is created using training data (except test data). The model is trained to optimise a specific target function (e.g. minimising the average square error). At this stage, different model types, hyperparameters, and optimisation methods are tested and the best ones selected on the basis of validation data.

**Model evaluation.** After selecting the final model, the model's performance will be evaluated on test data that has not been previously used in the development of the model. In addition to test data, other reference data sets may be used to demonstrate or compare the model's robustness with other methods. Performance metrics corresponding to the characteristics of the task and data must be selected for the evaluation.

**Model postprocessing.** In certain cases, post-processing is necessary after training the model. For example, if a loan decision model produces a continuous score between 0 and 1, it may need to be converted into discrete categories (e.g. 'low risk', 'inconclusive', 'high risk') or a binary recommendation (e.g. 'grant/don't grant loan'). This can also be considered a part of the deployment process.

### 2.3.2 Creation of a system based on algorithms or ML models

**Decision to use an AI system.** First, the creator of the system produces a vision. To this end, they may analyse, for example, usage scenarios, commercial necessity, and the legal environment. It is important to consider alternative solutions and to evaluate different possibilities of creating the system.

**Model deployment or integration into AI system.** This includes a number of conventional IT system development activities such as design, architecture, design of operating models, system programming, user interface development. The development of systems using algorithms or ML models may involve further steps, such as compliance with explainability requirements or the introduction of a feedback mechanism to amend the model.

It is important to remember that the use population (data that the model sees and processes in the production environment) may not be identical to the development population, leading to a

situation where the model in fact is not familiar with the types of data found in the production environment. The implementation process is just the beginning of the operational life of an AI system and is followed by new life cycle stages.

**Operation and monitoring of the AI system.** After deployment, the system starts to operate in its natural environment and to process real data. This stage requires monitoring the system. First, the technical characteristics of the AI system and its components, e.g. the speed, performance and accuracy of the model(s) of the AI system, have to be monitored. Next, the stability of the statistical characteristics of real-world input data also needs to be monitored in order to detect data drifts that may reduce the accuracy of the system. Last but not least, the social impact and ethical risks of the system must be constantly monitored. This includes the quantitative assessment of fairness and bias and, where appropriate, automated real-time intervention to alleviate the legal, commercial and reputational risks arising from potential bias.

**Maintenance and updating of the AI system.** At this stage, monitoring data will be used to ensure the viability of the system. This includes both regular model upgrades (including retraining) and replacements, as well as wider changes in system architecture. Not limited to error correction, this is a continuous risk management process that assesses and mitigates new risks appeared during the operations (e.g. security risks, changes in business requirements). In extreme cases, it could lead to the system being decommissioned.

**Decommissioning the AI system.** The end of life of a system may come when the maintenance of a particular AI system is no longer feasible, or its commercial viability has been exhausted. At this stage, planning is needed to ensure a smooth transition. System-dependent processes must be analysed and it must be decided how to archive or securely delete data and models related to the system and proceed it in accordance with data protection rules.

## 2.4 Measurement of the quality and safety of ML models

The quality of ML models is a complex concept. Model quality assessment must take into account the nature of the task, the commercial objectives of the system implementing the model, as well domain-specific risks. No universal quality metrics suitable for all situations exist. Quality indicators can be divided into two: performance metrics for classical machine learning, and benchmarks for LLMs.

### 2.4.1 Classical performance metrics

A number of standard metrics are known for classification tasks where a model has to assign a category to data points (such as spam/non-spam).

**Accuracy** is the ratio of all correct classifications, whether positive or negative. This indicator is suitable when classes are balanced during model training, but is misleading for unbalanced datasets. Provided 99% of emails are not spam, the model that classifies all emails as non-spam is 99% accurate but still useless.

**Precision** is the ratio of all positive classifications made by the model that are actually positive. Precision is important when wrong results are costly (e.g. moving an important email to junk folder).

**Recall** shows the ratio of actual inputs predicted to fall into a specific class (e.g. 'positive result') that the model was able to identify. High recall is important if a false negative result is costly (e.g. failure to diagnose a disease).

**F1-score** is calculated as the harmonic mean of Precision and Recall, ensuring that both metrics are given equal importance..

For regression tasks where a constant value (e.g. property price) is predicted, other metrics such as mean squared error (MSE) or root mean squared error (RMSE) are used.

### 2.4.2 LLM performance evaluation

The evaluation of large language models (LLM) is more difficult due to the fact that their output data takes the form of free text. Simple classification metrics such as accuracy or mean squared error (MSE) do not work here. Therefore, the standard approach in the field is to use complex benchmarks that measure different capacities of the model. In addition to LLMs, the use of benchmarks is also common in models for imaging, speech recognition and synthesis, and other more specific domains where simple regression or classification is insufficient.

Benchmarks such as **MATH** and **GPQA** measure a model's capabilities in complex domains, such as mathematics, biology, physics, and chemistry. Others, such as **EvalPlus** and **Livecodebench**, focus on the ability to write functional code. More recently, benchmarks assessing the performance of agents have also appeared (**$\tau$-Bench**, **$C^3$-Bench**). These assess the model's ability to plan and execute tasks and use external tools. In this way, the ability of the model to work autonomously can be assessed. Below, you will find a brief outline of the most important and most referenced benchmarks used with LLMs.

**MMLU (*Massive Multitask Language Understanding*)** is a benchmark for measuring broad general knowledge. MMLU covers 57 different subject areas, including humanities, social sciences and natural sciences, from primary school level to expert level. A high score indicates that the model has a broad knowledge base about the world.

**GSM8k (*Grade School Math 8k*)** is a benchmark for the assessment of mathematical reasoning. It consists of elementary school level textual tasks with multi-step solutions. The emphasis is not so much on complex mathematics as on the ability of the model to divide the problem into parts, extract the correct information and then perform the required sequence of operations.

**MATH (*Mathematical Problem Solving*)** is a mathematical capability test significantly more difficult than GSM8k. It includes olympiad-level tasks in algebra, geometry, and mathematical analysis. A good result in the MATH test shows the model's high level of symbolic thinking.

**GPQA (*Graduate-Level Google-Proof Q&A*)** is designed to test deep, expert-level reasoning in science. Questions are deliberately difficult even for humans and cannot be answered via a simple web search. This benchmark evaluates the model's ability to debate based on first principles, rather than simply reproducing information found or acquired.

**EvalPlus / MultiPL-E** are the leading benchmarks for evaluating the ability to generate code. They test the model's ability to write a functionally correct code based on a natural-language description (*docstring*). High scores indicate the practical value of the model to programmers.

**$\tau$-Bench (*Tool-Agent-User Interaction Benchmark*)** evaluates the capabilities of agents in a realistic, changing environment. It tests an agent's ability to have a multi-step conversation with a simulated user, use predetermined tools (application interfaces), and follow domain-specific rules (e.g. for customer service). This helps to evaluate how reliably an agent based on the test model can communicate with humans and systems to complete complex tasks.

### 2.4.3 LLM safety evaluation

In addition to performance, the safety of the model is also of great importance in the assessment of an AI model. Whereas in the past, the focus was mainly on filtering obviously harmful content (e.g. instructions for illegal activities), modern safety benchmarks are more complex. They test the model's resistance to manipulation attempts, its tendency to give dangerous instructions, and its behaviour in the form of autonomous agents. Safety assessments are essential to ensure the responsible development and implementation of the model.

**SafetyBench** is a broad-based benchmark that is considered the standard for assessing the safety alignment of the model. It consists of thousands of multiple-choice questions in seven risk categories, including offensive content, illegal advice, self-harm and prejudice. SafetyBench gives an overview of the model's ability to avoid generating clearly harmful responses in different scenarios.

**AgentHarm** focuses, unlike fixed questionnaires, on the risk of agent-based misuse. It measures the model's behaviour in multi-stage tasks that can be considered harmful. Two aspects are evaluated: firstly, refusal rate, i.e. whether the agent refuses to perform a dangerous task, and secondly, jailbreaking resistance, i.e. how well the model can resist attempts to circumvent its trained security restrictions.

**VLBiasBench (*Vision-Language Bias Benchmark*)** is a benchmark specifically designed for large-scale visual-language models (VLMs) focused on identifying social biases and stereotypes. It is based on a synthetically-generated image dataset designed to prevent data leaks where the test data has already been present in the model training set. VLBiasBench covers nine categories of bias (incl. age, gender, race, profession) as well as intersecting bias, asking both open and multiple choice questions to models. The aim is to determine whether the relationships inferred from visual and textual information give reason to consider the model biased or promoting inequality.

### 2.4.4 Guardrails for language models

In addition to the internal safety setup of the models themselves, real-life implementations of artificial intelligence also include so-called guardrail systems. These are external frameworks or models that serve as a security checkpoint, evaluating user input and model-generated responses based on pre-defined safety rules. The purpose of guardrails is to prevent the generation of harmful, inappropriate, or off-topic content in real time in situations where the model itself failed to do so. The nature of the implementation of guardrails depends on the deployment model of the particular AI system.

**Guardrails integrated into APIs from large vendors** Companies such as Google, OpenAI and Anthropic build safety systems directly into their APIs. Often these are so-called black box intermediate software that the userd cannot configure themselves. They automatically filter the content according to the service provider's own safety rules. They can filter both the user's input and the model's output. Many users only need this kind of non-configurable default safety level that is constantly updated by the service provider.

**Open source guardrails**. The best-known example of this category is Meta's Llama Guard. This is a language model that is fine-tuned to mark inputs and responses based on a safety taxonomy (e.g. violence, hate speech, recommendations to perform illegal acts). The main advantage of this approach is transparency and adaptability. Developers can implement such a model as part of a guardrail framework or system, modify settings, study its behaviour, and even fine-tune it to

meet their own application-specific rules and risk tolerance.

**Specialized frameworks and cloud service solutions.** Specialised frameworks have emerged for more complex or specific needs. One well-known example is NVIDIA's open-source NeMo Guardrails. In addition to easy content moderation, NVIDIA's solution also offers programmable guardrails. For example, a model can be prevented from discussing certain topics, hallucinations can be managed by knowledge-based fact-checking, and dialogue can be directed along predetermined paths. A growing number of cloud service companies, such as Guardrails AI and Giskard, also offer guardrails as a service. These platforms provide sophisticated and managed safety solutions for business customers, such as data leak prevention, ensuring the coherence of the brand's voice and image, and safety analytics.

## 2.5 The essence of bias in AI systems

### 2.5.1 Understanding bias

Under the bias of an algorithmic or artificial intelligence system, we mean a situation where a system using an algorithm or an artificial intelligence component (such as a machine learning model) provides decisions or assessments that display preference for specific groups or characteristics regardless of their relevance to the task at hand. It is important to note that algorithmic bias differs from statistical bias. In quantitative terms, bias is a systematic deviation of a value from a base value, whereas algorithmic bias relates to decisions and opinions. Also note that bias as a phenomenon is not always derogatory or directly harmful for end users – bias can also make the AI system exceptionally generous.

**Example. Exam grades prediction system in the UK**

In 2020, in the UK during the coronavirus (SARS-CoV-2) event, A-level final exams, required for university admission, were cancelled due to the pandemic. Ofqual, the State Agency for Qualifications and Examinations, thus awarded the students grades generated by an algorithm, for which the most weighted element was the previous performance of both the student and their school (!) over three preceding years. It appeared that 40% of school-assessed grades across England, Wales and Ireland got adjusted downwards by the algorithm. Students from more disadvantaged backgrounds were scored worse and those from private schools were scored far better, with twice as many top grades being awarded to the latter over students in public sector schools. Many students thus had a lower-than-expected grade. The system was angrily criticized for historical and locational bias, as even alumni of the same school from different years can produce significantly different results [9], [10].

In the case of the AI system depicted in Figure 1 in Section 2.1, bias means that the outputs of the system are favouring certain persons, organisations or objects in the external environment over others.

**Example. Crime prevention system PRECOBS in Baden-Württemberg (Germany)**

The PRECOBS system was created to support the police work and predict crime rates in the region. The system was tested for 4 years and then discontinued due to its weak predictive power. One reason was that the system was inclined towards predicting certain types of professional burglaries and paid less attention to several other types of crime. Thus, the effectiveness of the system depended heavily on the area where it was used and the types of crimes prevalent there [9].

### 2.5.2 Mechanisms of the emergence of bias

Bias can emerge in an AI system at any stage starting from data collection and ending with the application of the model. Each stage involves choices made by humans – what to measure, who to include, how to process data, how to design the system. These choices determine which patterns of bias are preserved in the system or manifest in its behaviour.

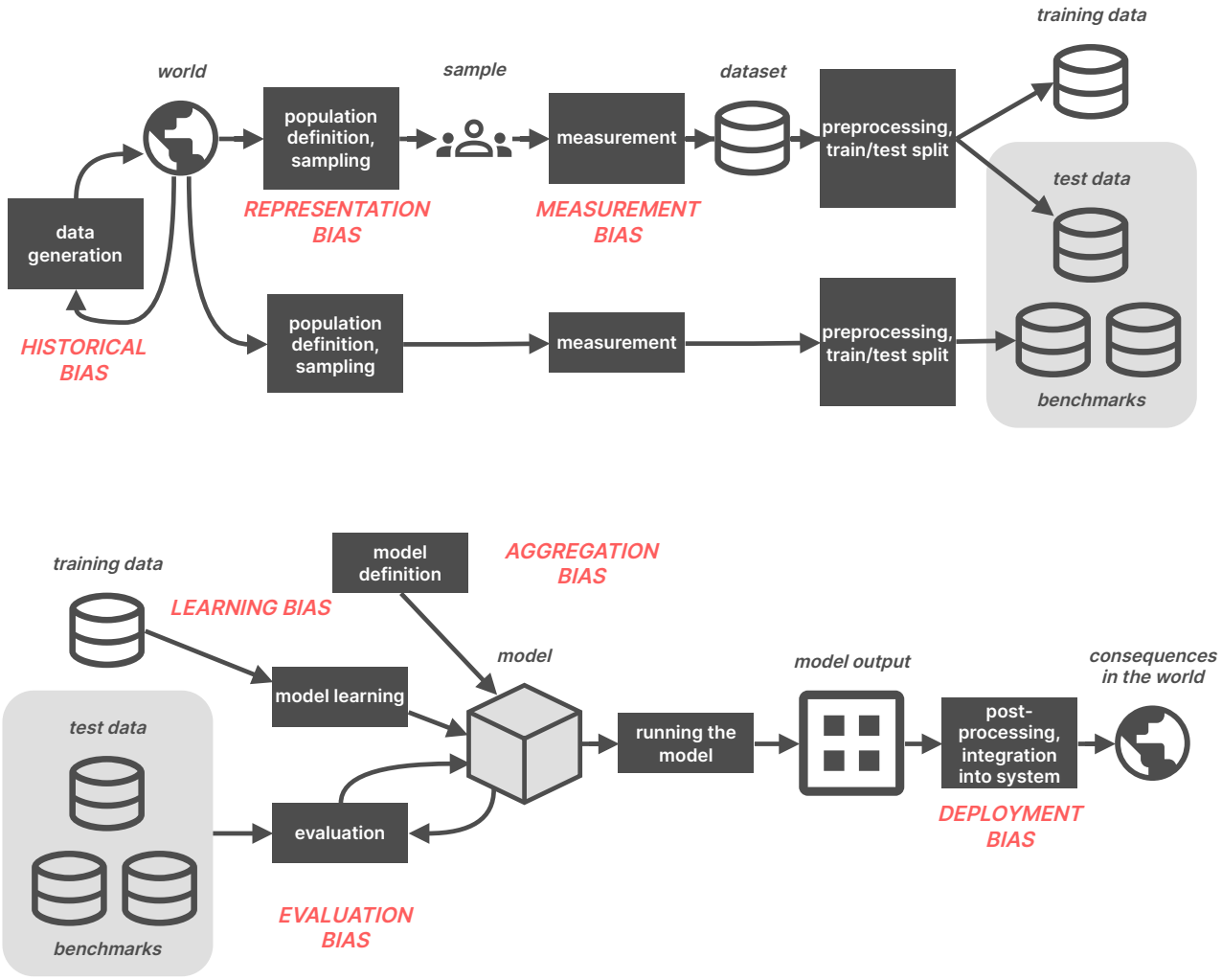


**Figure 2. Sources of bias in machine learning, adapted from [8]**

Figure 2 depicts the typical data and model cycle along with the main points of emergence of bias. This perspective helps understand bias as a systemic phenomenon, not just a characteristic of

the model, and creates the basis for connecting risks and alleviation measures to specific points of origin.

- **Data generation and population definition** — historical and representative bias. *Example*: Training data mostly includes past male employees → model infers that a suitable candidate must be male.
- **Measurement and labelling** — measurement bias. *Example*: Negative examples in image recognition are low-quality or labelled incorrectly → model learns patterns that do not describe the target phenomenon.
- **Data preprocessing and split into training and test data** — learning bias. *Example*: Data from under-represented groups are filtered out as noise → model does not learn to handle them at all.
- **Model definition and mechanisms of generalisation** — aggregation bias. *Example*: The same rigid model is used in different context even though inter-group relationship patterns do not coincide.
- **Evaluation using unsuitable benchmarks** — evaluation bias. *Example*: Model accuracy is evaluated only on a single population → the model seems good but fails in a new environment.
- **application in an unsuitable environment or work flow** — deployment bias. *Example*: Law violation risk assessment is used for determining the length of a prison sentence even though the model was not created for this.

Identification of these points of origin helps evaluate (see next sections) which risk alleviation measures cover which sources of bias and there may be any gaps in their coverage.

### 2.5.3 Forms and manifestations of bias

Reinforcement learning (RL) is a type of machine learning process that focuses on decision making by autonomous agents. ML algorithms used in RL can be either model-based or model-free, most frequently the latter. Algorithms take into account parameters obtained from the processing of previously available data or training a model[1]. Model-free algorithms have specific business rules in place and decisions are made in revealed form. These systems are also called rule-based systems[2]. Not many algorithms fall into this category, but the bias there has to be considered a little differently.

The view that predictions of a model-based AI system depend solely on the content of the data used for training is simplistic. Training data is not a static 'given' and data generation is not a neutral process in which the resulting deviations are mere quantitative shifts. Data generation involves a variety of human choices at each stage. The same stands for the ML process as a whole.

In the next sections we will explain in more detail the main forms of bias depicted n Figure 2 and the relevant terms, such as non-representativeness of the dataset, historically discriminatory patterns, as well as the use of the dataset in an inappropriate situation or environment. More fine-grained bias classifications can also be found in scientific literature; however, for a practical methodology, 40 different bias categories are just too many.

[1] See AI Risks report [11], Sections 2.2.2, 2.2.3 and 2.2.4.
[2] See AI Risks Report [11], Section 2.2.1.

**Historical bias** is the most intuitive form of data bias. Although the reflections of the world in input data may result in a formally accurate model, it still can harm certain population groups. This is because the model is also reflecting a historically discriminatory pattern in the dataset. For example, Google's machine translation can still struggle with translations from ungendered languages (such as Estonian) to gendered ones (such as English), returning results showing e.g. that only males are doctors and all nurses are female.

Understanding historical bias

Historical bias arises from natural historical changes in the population. Different rules and customs have prevailed in different periods: for example, certain social groups have been preferred in certain occupations. Institutional racism and institutional sexism are the most common examples of this. Finally, the social bias towards certain social groups comes from the context of a particular era. These factors have implications for the data we have historically.

**Representation bias** arises when some part of the population[3] is under-represented or over-represented in the data sample used to develop the model. As a result, the model cannot be extended to some parts of the population. This form of bias may result e.g. from the incorrect selection of the target population (data describing people living in Abu Dhabi may not be suitable to analyse inhabitants of Canberra), the under-representation of certain groups in the target population, or the limitations or unevenness of the sampling methods.

Understanding representation bias

When collecting datasets for the purpose of creating a model, bias may be caused by the exclusion of certain groups in the collection of data, uneven (non-random) selection of representatives in the sample, over-representation of the most common group in the population in the sample, or over-representation of enthusiastic data donors in the sample. The latter is a concern when encouraging data altruism: a dataset based only on data provided by activists cannot be guaranteed to be representative and balanced.

**Measurement bias** can occur in the selection, collection, or inference of attributes and labels for a predictive model. These attributes and labels are meant to replace or mediate any construct directly unobservable or impossible to encode like 'loan eligibility'. However, such an intermediate attribute may oversimplify the complex construct because aggregated measurement results may be obtained by different methods and the accuracy of the measurements may vary for different groups.

[3] Population in this context means any set of entities that should be described or represented by the ML model being created.

Understanding measurement bias

The most common source of measurement bias is uneven data quality, incomplete or carelessly applied classifications, ontologies and formats. When textual, speech or video data are collected, ensuring consistency in diffuse data collection is particularly difficult. In addition to the source data, errors may also arise from labelling – if the person labelling the data undervalues or overvalues a character, effect, or signal in the data, it also affects the quality of the model and may cause the emergence of bias. And, of course, in quantitative data, measurement errors (and biases) are one of the causes of measurement bias.

**Aggregation bias** can occur when a rigidly generalising model is applied to a dataset that includes groups or types of examples assuming different treatment. The root cause of aggregation bias is the assumption that relationships between input data and labels are the same across all subgroups of the dataset, which however may not be true. When such a model is deployed it can turn out not to work well for any part of the population, or (e.g. in the case of a biased sample) only work for the dominant population.

Understanding aggregation bias

The training datasets of the machine learning model can contain different groups of persons, organisations or objects, and a single pattern, structure or rulebook may not suite to describe all of these. In this case, treating them as equivalents during the training may cause the machine learning model to produce incorrect output data for some groups.

**Learning bias** manifests itself in a situation where the differences in model performance across different datasets are amplified by choices related to system objectives and requirements. For example, if training makes use of differential privacy, this may limit the impact of under-represented data in the model. As a result, the performance of the model suffers compared to a model not optimised for privacy.

Understanding learning bias

When training predictive machine learning models, groups that are overrepresented in the source data can also be amplified in the model because the associated probabilities are higher and uncertainty lower. An example of learning bias is using the data generated by a ML system to train the next ML system – the new model may inherit the previous bias. At the same time, synthetic data may well serve as a tool to reduce bias, especially when extra amount of data is generated for under-represented groups.

**Evaluation bias** occurs when benchmarks not corresponding to the actual population are used to evaluate the performance of a particular model. The root cause to that is the desire to compare models to one another: executing the model on a variety of external datasets offers a good opportunity for juxtaposition and the results tend to be equated with the model's 'goodness'. However, this can result in an excessive focus on fitting the model to one specific benchmark. This is particularly problematic in a situation where the test itself is already characterised by historical, representative or measurement bias.

**Understanding evaluation bias**

When evaluating models, AI creators can get stuck in the results of a specific test or quality comparison and forget that tests and comparisons may also be biased. If a machine learning model produces good results in some tests, it does not automatically mean that the model will be effective, high-quality and non-biased in a production system. The bias of the evaluation test itself should also be assessed.

**Deployment bias** emerges when the intended use of the model differs from its standard usage. In some cases the users themselves can deviate from the intended task: e.g. to utilise a system designed to assess the likelihood of committing crimes to determine the duration of a prison sentence. In other cases the system may be developed as abstract and autonomous, but in reality it must operate in and be interfaced to a complex socio-technological context not even reflected in the abstract model.

**Understanding deployment bias**

When an algorithm or ML model is applied outside its originally planned scope, a conceptual drift may occur, and the output data produced in the new context can turn out to incorrect and biased. The bias can also occur in interactions between the user and the AI system. Certain user interface elements or data presentation modes (colours, sounds) are known to cause biases in user behaviour even if the model itself is bias-free. If the user does have a clear picture of the quality or reliability of the AI system's output decision or whether it requires human validation, this will result in a decrease in the user's human agency and amplification of theAI system's bias. The user tasked with decision-making might also simply ignore assessments made by the AI system if these are in conflict with their own beliefs.

Modelless algorithms, such as rule-based systems, differ from model-based algorithms in that learning bias is ruled out, as there is no model learning to fit a function. In this cases, we can instead often speak of **design bias** arising from the potential biases of the system's designers. Apart from this, model-free algorithms are subject to all of the forms of bias discussed above.

**Understanding design bias**

Prejudiced or incorrect design choices made during algorithm development and utilisation of the algorithm in an inappropriate situation are both forms of **design bias**. Design bias can also occur in model-based systems where it can be introduced before the selection of a model or datasets (partial overlap with evaluation bias).

# 3 Avoiding bias

## 3.1 Reasons to avoid AI bias

As highlighted in the report of the European Commission's Independent High Level Expert Group on Artificial Intelligence [12], avoiding bias is one of the seven key requirements for achieving reliable AI, critical to protecting European values and the rule of law. Artificial intelligence must adhere to the principle of fairness, and AI systems must respect all moral values of humans in the same way; bias is in conflict with the principles of reliable artificial intelligence. In its existence, AI systems do not function objectively – they reflect or even amplify former prejudices and inequalities of people and society.

When AI systems shape decisions that affect a person's access to services, rights or opportunities, bias can have serious consequences – especially for those in an already vulnerable situation. Often, the bias of AI is hidden, as algorithms are considered neutral. In fact, easily scalable AI systems can spread harmful patterns much faster and more systemically than human and societal biases at their own. We have an ethical and legal responsibility to ensure that such technologies are not used as means of discrimination, of deepening social inequalities or of concentrating power in the hands of the few.

Broadly speaking, the reasons to avoid AI bias can be divided into three categories depending on the object being affected: AI bias can affect individuals, organisations, and the society as a whole.

### 3.1.1 Individual level: Harm to fundamental rights and freedoms

Avoiding bias in AI systems is necessary to protect fundamental rights and freedoms. Biased decision-making algorithmic systems violate the fundamental principles of democracy and individual freedoms that form the foundations of fair societies, undermining trust in both technology and institutions. Such systems are often characterised by a lack of transparency, both in terms of the overall application of AI and in terms of its precise impact. People have no choice as to its impact. Vulnerable groups are particularly at risk: children, people with disabilities, ethnic minorities, women.

Bias in AI systems can lead to inequal access to critical services, products, and technologies. This endangers the following principles:

- the right to equal opportunities and impartiality within the meaning of socio-economic, gender, age, ethnic, religious or sexual preference;
- the right to a fair justice system and an administrative system in general terms (for example, when the principle of presumption of innocence is violated);
- the right to privacy and the protection of personal property.

### 3.1.2 Organisational level: Risks of economic and reputational damage and legal compliance

AI bias can cause significant economic and reputational damage to the organisation. It can, for example, enable unfair competition by exploiting consumer bias or market opacity and anticom-

petitive cooperation. On the other hand, the disclosure of AI bias may result in reputational damage to specific organisations, institutions, companies (and thereby loss of existing or potential customers and market opportunities.

Other significant risks can be related to legal compliance: the use of biased AI systems can lead to legal consequences – supervision measures and penalties. Conflicts within THE organisation where employees perceive a contradiction between their employer's AI practices and their personal values and work ethic must be highlighted as a category requiring special attention both in situations where the AI is used by persons outside the company as well as in cases where it is used inside the organisation (e.g. for recruitment).

### 3.1.3 Social level: Undermining trust in institutions and technology

Avoiding bias in automated solutions implemented is necessary to ensure trust and confidence in such systems and in the institutions implementing them. Biased AI systems undermine trust in both technologies and institutions and prevents the society from realising its full potential through inclusive participation.

Systemic injustice damages the principles of democracy and rule of law. Injustices that already exist in the areas of social policies (housing, health), education, insurance, finance and trade (credit and other financial services; different pricing and availability of goods and services) make citizens vulnerable to technology adoption, which can undermine social cohesion. Bias can normalise, increase, and amplify existing social inequality, marginalisation, and prejudices and thereby endanger democratic processes and equal opportunities, especially if it hinders access to education, goods, services, and technologies.

General social loss of trust in technology can slow down the rate of innovation and reduce overall economic efficiency and economic growth.

Example. Systemic Risk Indicator (SyRI) of the Dutch Ministry of Social Affairs and Employment (2008–2020)

SyRI (System Risico Inventarisatie) was a large data analysis system used by the Dutch Ministry of Social Affairs and Employment between 2008 and 2020 to identify tax fraud and abuse of social benefits. The system combined various personal data – identity, employment, property, education, pension, business, income, and debt data – and used algorithmic models to analyze them to detect irregularities or potential fraud. SyRI generated risk reports on addresses in areas at higher risk of fraud. People were registered with the system and could potentially become subject to investigations. The system was used nationwide. The main problems were the opacity of the system, the fact that all citizens became suspects, and that the system disproportionately targeted areas of marginalised communities. In February 2020, the European Court of Human Rights declared the system unlawful, finding it to be in breach of Article 8 of the European Convention on Human Rights. As a result of the complaint of the Coalition of Civil Society Organisations, the use of the system was discontinued and the Government decided not to appeal this decision, recognising the inefficiency of the system.

**At the level of the individual**, SyRI violated a number of fundamental rights and freedoms and resulted in inequal access to critical services. The system was based on the principle that all Dutch people are suspects, contradicting the presumption of innocence. This violated the right of citizens to equal opportunities, as the system systematically targeted marginalised regions and ethnic minorities. Privacy law was also violated by extensively combining personal data without the consent or knowledge of the citizens. According to the judge, social interest and privacy were not in balance.

**At the national level**, SyRI caused significant reputational damage to the Dutch state. A ruling by the European Court of Human Rights in 2020 declared the system illegal, forcing the government to end the project and resign. International attention was negative, mobilisation of civil society effective. The government acknowledged at the ministerial level that the system was not effective nor efficient, but the reputational damage was already done.

**At the level of the society**, SyRI undermined trust in institutions and damaged the fundamental principles of democracy through the creation of systemic injustice. This exacerbated existing social inequalities by targeting those communities that were already in a vulnerable position, which in turn undermined trust in the state and the institutions – citizens did not know they were being monitored and could not challenge the decisions. Social opposition intensified when it became clear that certain areas had been designated as suspects. This prevented equal participation and created mistrust in society.

## 3.2 Regulations and guidelines that require addressing AI bias

### 3.2.1 Fundamental rights and freedoms

The creation of a fairer and more human-centred digital state must focus on the humans and their rights and freedoms, the starting points of which are provided, inter alia, by the Constitution of the Republic of Estonia [13] (see Chapter II) and the Treaty on European Union (EU) [14]. The TEU together with the Treaty on the Functioning of the European Union (TFEU) [15] form the basis of EU law, defining the common principles [16].

Since AI systems, including LLMs, can amplify societal biases, e.g. gender, racial and age stereotypes [2], it is critical that existing AI systems are not unfairly biased. The example below concerns the use of real-time face detection technology by the London Police which was found to be discriminatory towards a dark-skinned person.

Example. *S. Thomson* v. *Commissioner of Police of the Metropolis* (Use of facial recognition technology by the Metropolitan Police) [17]

The Metropolitan Police's plan to expand live facial recognition technology was challenged by the Equality and Human Rights Commission (EHRC), which argued that the deployment could violate human rights laws, particularly concerning privacy and discrimination. The EHRC supported a judicial review initiated by Shaun Thompson, a Black man who was wrongly identified and detained due to the technology. The EHRC challenged the Metropolitan Police's planned expansion of live facial recognition as potentially unlawful under European human-rights and equality standards, raising discrimination and privacy concerns for mass surveillance deployments.

AI provides significant opportunities for increasing the efficiency of law enforcement processes – from investigations and border defence to criminal law and asylum proceedings. At the same time, prejudiced AI can result in significant ethical issues, such as discrimination, misidentification, and loss of public trust. The Europol report AI Bias in Law Enforcement: A Practical Guide [18] emphasises the socio-technological nature of prejudice and highlights the importance of human supervision, impact assessments, and diverse, interdisciplinary teams. The report also discusses the difficulties in the management of compromises in defining and measuring justice, noting that no indicator of justice is suited for every single context and involvement of human judgement is thus unavoidable. The report provides practical recommendations based on the AI Act for the responsible, transparent, and rights-preserving use of AI systems at law enforcement institutions.

Studies show that large language models traditionally associate women with home and family life, while men are associated with careers and business. It has also been found that image generators are less likely to depict women and minorities in professional roles, especially in senior positions [2]. As we move towards a more equal and just society, we must try to avoid such situations in every possible way. All this is illustrated by the examples below.

Example. *Mobley v. Workday, Inc.* – Case no. 23-cv-00770-RFL [19, 20]

The case of Mobley v. Workday, Inc. is one of the first large-scale legal challenges to the use of AI in recruitment. A federal court in California recently granted conditional certification of claims under the Age Discrimination in Employment Act, potentially creating one of the largest collectives ever certified, covering millions of rejected applicants. The plaintiff argues that Workday's AI-based screening tools embed employer biases and rely on skewed training data, resulting in systemic exclusion based on age, race, and disability. The ruling underscores the legal risks employers face when deploying algorithmic hiring tools and signals that courts are willing to scrutinise their disparate impact. Federal litigation alleges that Workday's automated applicant-screening system has a disparate impact in hiring and that vendor liability theories can proceed in court, making this a key test of how civil-rights law applies to employer AI tools [19].

**Example.** *Harper v. Sirius XM Radio, LLC* – Case no. 2:25-cv-12403-TGB-APP [21, 22]

In *Harper v. Sirius XM Radio, LLC*, a job applicant has filed suit in the U.S. District Court for the Eastern District of Michigan, alleging that the company's AI-powered hiring tool discriminated against him on the basis of race. The plaintiff claims the iCIMS Applicant Tracking System embedded historical biases in its evaluation process, resulting in his rejection from around 150 IT positions despite his qualifications. Harper asserts both disparate treatment and disparate impact theories under Title VII and Section 1981, and seeks to expand the suit into a class action for similarly affected applicants. He is pursuing compensatory and punitive damages, as well as injunctive relief to modify or discontinue the AI tool. These are allegations at this stage; Sirius XM has not yet responded, and no findings of fact or law have been made by the court.

**Example.** The case of the UK Uber Eats courier [23]

An Uber Eats courier, Pa Edrissa Manjang, filed a discrimination claim against the company, alleging that the facial recognition system used to verify his identity was less effective for Black individuals, leading to repeated login failures. The case was settled with a financial payout to Manjang, and the Equality and Human Rights Commission (EHRC) highlighted the need for AI systems to be tested for bias and fairness.

The right to equality and the prohibition of discrimination are today considered fundamental human rights [24]. § 12 of the Constitution of the Republic of Estonia states that everyone is equal before the law. No one must be discriminated against on grounds of nationality, race, colour, sex, language, origin, religion, political or other belief, property or social status or other circumstances. Incitement to hatred, violence and discrimination of a national, racial, religious or political nature is prohibited and punishable by law. It is also prohibited and punishable by law to incite hatred, violence and discrimination between social strata[1].

The Treaty on European Union [14] emphasises human dignity, freedom, equality, and respect for human rights, highlighting non-discrimination, fairness and equality as values. According to the Treaty, the fundamental rights guaranteed by the European Convention for the Protection of Human Rights and Fundamental Freedoms are general principles of EU law [14]. The Charter of Fundamental Rights of the European Union [26] confirms, inter alia, the rights arising from the constitutional practices and international obligations of the EU Member States, including the European Convention for the Protection of Human Rights and Fundamental Freedoms, the social charters adopted by the EU and the Council of Europe, as well as the case-law of the Court of Justice of the European Union and the European Court of Human Rights [27]. Various directives have also been adopted to ensure equal treatment, such as Council Directive 2000/43/EC on racial and ethnic equality [28] and Council Directive 2000/78/EC on equal treatment in employment and occupation [29].

The equality of persons before the law and protection against discrimination are also recognised in a number of international instruments such as:

- The Universal Declaration of Human Rights [30] (1948) adopted by the United Nations (UN) – the first international document to stress that human rights apply equally to all and are indivisible, inalienable and universal [31].
- various UN conventions on the elimination of discrimination (e.g. protection of women's

[1] Regarding § 12 of the Constitution, see also the explanations given in the commented version of it [25]

rights [32], ensuring racial equality [33]).

- UN pacts, e.g. on civil and political rights [34], as well as on economic, social and cultural rights [35].

The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law [36] establishes principles for the development and use of AI systems to ensure that both systems and related activities are consistent with fundamental rights throughout the life cycle of the AI system and comply with the principles of democracy and the rule of law (Art 1 (1) ). In the EU, the Convention applies to AI systems governed by Artificial Intelligence Regulation (EU) 2024/1689 (AI Act).

The AI Act [1] states that AI should be a human-centric technology. It should serve as a tool for people, with the ultimate aim of increasing human well-being (Rec 6). The AI Act is a risk-based framework (see also Art. 9), which imposes stricter requirements on higher-risk AI systems (see Chapter III of the Regulation) and bans the use of certain AI systems (see Art. 5 of the Regulation). The requirements are raised in accordance with AI-related roles (see subsection 2.2) (see also explanations for the obligations of the importers ([37] page 503ff) and distributors ([37] page 516ff) ).

According to Regulation (EU) 2024/1689 on AI:

**Diversity**, **non-discrimination** and **fairness** mean that AI systems are 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 bias (Rec 27).
As an example, the AI Act mentions that technical inaccuracy of AI systems for biometric remote identification of natural persons can lead to bias of results and lead to discrimination, with the risk of biased results and discriminatory effects being particularly high in relation to age, ethnicity, race, gender or disability (Rec 32, 54).

Article 27 of the AI Act sets out a comprehensive framework for fundamental rights impact assessments (FRIA) (see also [38] and [37] p. 553 et seq.). Under certain circumstances, the obligation to do so will arise in the case of high-risk AI systems before the system is introduced. FRIA focuses on risk management by identifying risks affecting people, assessing the likelihood and severity of their occurrence, and proposing mitigation measures and developing an effective risk management plan [39]. ISO/IEC 42005 [40] (see Table 1) is also helpful in assessing the effects of the AI system. Special attention should be paid to the introduction and implementation of AI systems where the target group is more vulnerable groups, e.g. minors, the elderly, persons with special needs.

There is also a need for regulatory sandboxes (also for the public sector) to support safe testing of high-risk AI systems before the requirements of the regulation on AI systems are fully applicable [41]. The AI Act requires Member States to ensure that at least one artificial intelligence regulatory sandbox is established at national level and must be operational no later than 2.08.2026 (Art 57 (1)). For example, the French Data Protection Authority (CNIL) [42] provides such an opportunity for legal and technical advice in Europe. According to the Estonian AI Action Plan 2024-2026 [43] AccelerateEstonia operates in Estonia, which enables to test a new solution in the market together with prototyping and design the necessary legal changes.

According to the Estonian White Paper on Data and Artificial Intelligence 2024-2030 [44], Estonia wishes to be involved at international level in the development of policies, legislation and standards that enable us to promote our interests and ensure the applicability and consistency of AI

with the principles of reliable AI. It also sees great importance in cooperation with key partner countries in the areas of the implementation and monitoring of reliable AI.

The Estonian Digital Society Development Plan 2030 [45] sets the goal that public services must be of high quality, predictable and accessible in every region, ensuring the fundamental rights of people. Where such services contain AI systems, the relevant legal requirements must be taken into account and the guidelines of the various expert groups and competent organisations must be taken into account. According to the White Paper on Data and Artificial Intelligence 2024-2030 [44] it is also important to identify and mitigate the risks associated with algorithmic discrimination in Estonian public sector systems.

According to the European Commission's High Level Expert Group (AI HLEG) Trusted AI Guidelines [12], AI must be legal, ethical and robust. The main requirements for AI are human-centered control, technical reliability and safety, privacy and data management, transparency, diversity and fairness, social and environmental well-being and responsibility. These principles support the protection of fundamental rights and the benefits of society.

The Organisation for Economic Co-operation and Development (OECD) has developed the Artificial Intelligence Principles [46], which are:

- inclusive growth and sustainable development,
- human rights and democratic values,
- transparency and explainability,
- security and reliability,
- responsibility.

The OECD recommendations to policy makers highlight the need to invest in AI research and development, create an inclusive and supportive ecosystem, develop an interoperable policy environment, develop human skills and prepare for changes on workforce market, and promote international cooperation for reliable AI [46].

The European Declaration on Digital Rights and Principles for the Digital Decade [47] considers AI as a tool with the ultimate goal of enhancing human well-being. According to this, everyone should have the right to use the benefits of AI systems and make informed choices in the digital environment, while being protected from risks and harm that affect health, safety and fundamental rights. According to the declaration, it is also necessary to ensure a safe and healthy working environment and that the use of artificial intelligence in the workplace is transparent and that its use takes a risk-based approach. It is also necessary to ensure that important decisions affecting workers are made under human supervision and that workers are informed when they come into contact with AI systems.

**Example. Deterioration of equality**

Applications of artificial intelligence can further exacerbate inequalities, e.g. in recruitment, where algorithms favour men, or in healthcare, where women may be misdiagnosed, as the models are mainly based on data collected from men. In education, algorithms can underestimate girls' opportunities in real-time, increasing the risk of falling out and limiting access to further learning programs. [2]

The European Parliament's 2024 study highlights cases that show the extent of AI misuse and opens up complex policy, regulatory and diplomatic challenges related to the past. [48] According to the study:

- the misuse of artificial intelligence in authoritarian regimes is becoming an increasing problem, reinforcing the spread of misinformation and the suppression of dissidents;
- the EU must invest in research and development of ethical AI in order to promote transparent and non-misuse technologies;
- effective sanctions and accountability mechanisms are necessary to punish human rights violations and address supply chain vulnerabilities;
- the balance between innovation and strict ethical compliance is critical to prevent EU technologies from being used for repression;
- close partnerships with academia, the private sector and civil society are essential for early warning and innovative solutions;
- the EU should establish global AI governance standards through a strong legal framework and active international cooperation. The EU needs to maintain high standards of human rights and transparency internally in order to strengthen its global credibility;
- dual use of artificial intelligence requires refined export controls to prevent misuse.

The European Commission's 2025 report [2] addresses the changing role of General Purpose Artificial Intelligence (GPAI) in the EU. The report highlights the potential of the GPAI to promote innovation, productivity and social development, but also highlights threats such as misinformation dissemination, bias, job transformation and privacy risks. The technological possibilities, economic impacts and societal consequences of GPAI are discussed. In addition, EU regulatory frameworks (e.g. AI Degulation and data laws) will be examined, which should ensure reliable and transparent development. The report underlines the need for balanced policies to maximise the benefits of the GPAI and reduce risks while respecting democratic values and the EU legal framework.

**Example. General Purpose AI (GPAI)**

GPAI can reinforce existing bias and stereotypes, especially when models are trained on data that reflect inequality. For example, GPAI may exacerbate gender bias when assessing credit risk. [2]

The above underlines the need for a fair and transparent decision-making framework and to reduce unfair bias and ensure accountability. The development of reliable AI therefore requires a diverse set of training data and fairness-focused algorithms. It is important to regularly audit systems and increase the diversity of AI system developers. For example, at the EU level, the Digital Services Regulation [49] requires annual risk assessments from major platforms to identify potential discrimination and threats to fundamental rights. [2]

Diversity, fairness and the avoidance of discrimination are the basis for the creation of reliable AI. Eliminating unfair bias is essential in order to avoid marginalisation of vulnerable groups and exacerbating prejudice and discrimination. The fairness of AI is assessed on the basis of protected characteristics (e.g. race, gender, age, disability) in accordance with the Charter. AI systems should be evaluated across different protected groups, especially in high-risk cases, and efforts should be made throughout the life cycle to avoid discriminatory or biased outcomes in order to ensure fairness and comply with EU anti-discrimination legislation. [2]

By the end of 2025, the European Commission is planning to publish two new strategy documents: the Apply AI Strategy [50], which focuses on the use of the potential of AI systems in target sectors and in the provision of public services, and the European Strategy for AI in Science [51], which promotes the responsible use of AI in science and innovation. In Q1 2026, the

Commission is also expected to publish a draft of the Cloud and AI Development Act [52], the goal of which is to facilitate investment in cloud and edge computing [53].

### 3.2.2 Data protection and data management

Data protection issues are comprehensively regulated in Europe. Requirements for the protection of personal data can be based on both national and EU legislation. Depending on the situation, it may be necessary to rely on, for example, GDPR [54], the Data Protection Directive of law enforcement agencies [55] or the Regulation on the processing of personal data by EU institutions, bodies and agencies [56]. The guidelines of data protection organisations and their proposed interpretations, as well as various guidance materials or legal literature on the processing of personal data in AI systems, e.g. [57, 58, 59], may also be helpful.

The Estonian Personal Data Protection Act (IKS) [60] specifies and supplements the provisions of the General Data Protection Regulation (EU) 2016/679 [54] (GDPR) and establishes rules for the transposition of Directive (EU) 2016/680. The IKS provides in several sections (e.g. §4, §5, §6 (3) 3), §10 (2) (3)) that data processing must not excessive damages to the rights of any data subjects. In addition, the IKS provides that it is prohibited to make a decision based on only automated processing if this brings negative consequences for the data subject. A decision based on profiling which discriminates against natural persons on the basis of specific types of personal data is also prohibited, unless the making of the decision is permitted by law which provides for appropriate measures for the protection of the rights and freedoms and legitimate interests of the data subject (see subsection 21 (1) of the IKS).

Recital 10 of the AI Act also states that requirements related to the processing of personal data also be taken into account in the design, development, or use of AI systems involves the processing of personal data, More specifically, data subjects must retain all rights and guarantees under EU law, including those concerning automated individual decisions and profiling (AI Act, Rec 10). The example below provides a vivid illustration of the challenges related to data analysis.

**Example. Automatic data analysis – Case no. 1 BvR 1547/19, 1 BvR 2634/20 [61, 62]**

In February 2023, the German Federal Constitutional Court ruled on the constitutionality of automated data analysis by police authorities. The Court held that processing stored personal data through automated analysis constitutes a distinct interference with the right to informational self-determination under the Basic Law. Such interference may be more severe than the initial collection of data and therefore requires additional legal justification, guided by principles of proportionality and purpose limitation. The judgment emphasises that only the legislature may define the fundamental rules on what data may be used and which methods of analysis are permissible, while administrative authorities may specify technical details under strict oversight. Where automated data analysis has a serious impact on fundamental rights, it is only permissible for protecting particularly weighty legal interests and subject to stringent safeguards.

The application of GDPR to AI systems may pose difficulties as the characteristics of artificial intelligence (e.g. opacity and extensive data use) may conflict with the general principles set out in Article 5 of GDPR – e.g. legality, transparency, purposefulness, data minimization and accountability [59]. GDPR is also stricter with regard to the processing of special types of data, which may create legal uncertainty [63]. At the same time, it has been found [64] that the principles of purposefulness and data minimisation can also be applied flexibly to support artificial intel-

ligence within the framework of GDPR. The purposefulness allows for the data re-use provided it is consistent with the original purpose of data collection. Minimisation of data may mean, in particular, a reduction in detectability (e.g. through pseudonymisation) rather than necessarily limiting data volumes.

An AI system's compliance with GDPR ensures that the personal data used in the system are adequate, relevant, and necessary for achieving a clearly defined purpose. Data controllers must therefore carefully consider and justify the necessity of each data element; this must be analysed separately for each stage in the system's life cycle (see Section 2.3). A good overview of data, in turn, also facilitates compliance with the requirements of the AI Act, e.g. identification and mitigation of bias. The AI Act sets out rules for the processing of sensitive data and active bias management, thus ensuring compliance with the principles of fairness, transparency, and minimum necessary data in AI systems [65].

The Estonian Data Protection Inspectorate has shared tips on how to protect people's privacy and ensure data protection in a situation where the AI system processes personal data. They require that data be collected and processed only for clearly defined and justified purposes and that transparency and security be ensured. It is also important to understand whether the input given to AI may contain personal data, information intended for internal use or trade secrets.[2] People need to be informed about how their data are used and to be able to view and correct their data.

In the example below, the court ordered Clearview AI to inform data subjects of the possibility of deleting their data from the company's database.

Example. *ACLU v. Clearview AI* – Biometric data [67, 68]

In the case ACLU v. Clearview AI, the American Civil Liberties Union (ACLU) and its Illinois affiliate challenged Clearview AI for violating the Illinois Biometric Information Privacy Act (BIPA) through the unauthorised collection and use of facial images. As part of the settlement, Clearview is permanently prohibited from granting most businesses and private entities access to its faceprint database nationwide and cannot provide state or local entities in Illinois access for five years. Illinois residents can now opt out of Clearview's database, and the company must advertise this opt-out mechanism publicly. The settlement highlights the importance of robust biometric privacy laws and provides protection for vulnerable communities, including survivors of domestic violence, undocumented immigrants, and sex workers. The case demonstrates that strong privacy regulations can impose meaningful limits on large-scale biometric surveillance.

In the next example, the personal data of Facebook users were collected without the informed consent of most of the subjects. The data originated from both users of the Facebook app and their friends, and were used for the generation of political profiles for targeted advertising.

[2] The Public Information Act [66] aims to ensure that public access to information intended for general use is possible. § $3^1$ (3) of the Act states that 'the granting of information for general use must guarantee the privacy of a person, the protection of copyright, the protection of national security, the protection of trade secrets and other information subject to restrictions on access." Section 34 (1) of the Act provides that information to which access is restricted pursuant to the procedure established by law is deemed to be restricted. Entering the restricted information into a public AI system is not permitted as it is not possible to exclude the leakage of such information to the service provider or implementer of the model, i.e. there is no control over the further processing of the information.

**Example. The Cambridge Analytica scandal [69, 70]**

The Cambridge Analytica scandal illustrates how algorithmic inference can propagate bias even without direct data collection from every individual. By using a personality quiz and the social connections of participants, Cambridge Analytica created psychometric profiles for millions of users, which were then applied to influence political messaging. This process amplified existing social and political biases, as the algorithm relied heavily on demographic correlations and assumptions rather than individual behaviors. The case highlights that surreptitious inference and behavioral targeting can inadvertently reinforce discriminatory patterns or ideological echo chambers. Consequently, it underscores the need for regulatory oversight of AI and algorithmic systems to identify and mitigate bias in predictive models and data-driven decision-making.

The large volume of personal data processed by AI systems may pose a high risk to fundamental rights, in particular privacy and non-discrimination [71]. The 2nd Cybersecurity Directive (NIS2) [72] also states that new technologies such as artificial intelligence must comply with EU data protection requirements, including principles such as data accuracy, fairness, transparency and confidentiality (encryption) and collection of as few data as possible, and that the protection-by-design and default data protection requirements set out in GDPR must be fully respected.

In order to assess the probability and severity of the threat, taking into account the nature, scope, context and objectives of the processing, a data protection impact assessment should be carried out before processing (GDPR Art. 35, pp. 90). Impact assessments of AI systems, including ethical and social aspects, are a good complement to the data protection impact assessment [71].

**Example. *Toeslagenaffaire* in The Netherlands [71]**

The Dutch *Toeslagenaffaire* (childcare benefits scandal) concerned the use of an algorithm by the Dutch Tax and Customs Administration to detect childcare benefit fraud. The system created erroneous risk profiles, often targeting people with dual nationality or migrant backgrounds, and wrongly classified thousands of families as fraudsters. As a result, many lost access to benefits, faced severe financial hardship, and suffered long-term social consequences. In 2021, the scandal triggered the resignation of the entire Dutch government and remains a central example of how algorithmic bias can cause systemic injustice.

Article 10 (1) of the AI Act requires that training, validation and testing datasets for high-risk AI systems be subject to appropriate data management and management practices. The requirements referred to in paragraph 2 of the same Article must be taken into account, including the examination of the referred datasets to exclude bias (Art 10 (2) (f) ) and the obligation to take appropriate measures to detect, prevent and mitigate bias. Article 10 (2) (3) of the AI Act requires that training, validation and test data sets must be relevant, sufficiently representative, error-free and complete to the extent possible, taking into account the intended purpose of the data, and relevant statistical characteristics for the target population. In addition, paragraph 4 specifies that the datasets must take into account, to the extent necessary for their intended purpose, the characteristics or elements that characterise the specific geographical, contextual, behavioural or functional environment in which the high-risk AI system is intended to be used. Katerina Yordanova has commented extensively on Article 10 of the AI Act in her comments on the AI Act (see pages 259–283 [37]). See also AI Act pp. 66, 67, 68, 69, 70.

**Article 10 (5) of Regulation (EU) 2024/1689 on AI**

To the extent strictly necessary to ensure that bias is detected and corrected for high-risk AI systems in accordance with points (f) and (g) of paragraph 2 of this Article, providers of such systems may exceptionally process specific types of personal data, subject to appropriate safeguards to protect the fundamental rights and freedoms of natural persons. In addition to the provisions of Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680, such processing must comply with all of the following conditions:

a) the identification and correction of bias cannot be achieved effectively by processing other data, including artificial data or anonymised data;

b) regarding the special categories of personal data, technical restrictions are in place regarding the re-use of personal data, also state-of-the-art security and privacy measures are in place, including pseudonymisation;

c) special categories of personal data must be subject to measures to ensure the security, protection and appropriate safeguards of the personal data processed, including rigorous access control and documentation, in order to prevent misuse and to ensure that only authorised persons with appropriate confidentiality commitment have access to such personal data;

d) the special categories of personal data may not be transmitted or transferred, they may not be otherwise accessible to other persons;

e) the special categories of personal data should be deleted after the bias has been corrected or when the retention period of the personal data is over, whichever comes first;

f) the registration of processing operations pursuant to Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680 includes the reasons why the processing of specific categories of personal data was strictly necessary to detect and correct the bias and why this objective could not be achieved by the processing of other data.

The AI act requires high-risk AI systems to be able to identify and mitigate bias. This is the primary subject of Article 10 (5) of the regulation. Bias management is seen as a continual obligation than must be followed throughout the entire life cycle of the AI system, taking into account the principles of accuracy, transparency, and fairness. At the same time, the requirement to collect a sufficient amount of representative data to prevent data bias must be balanced with the de minimis principle set out in the GDPR, which requires careful planning. [73]

In addition to the above, audits also play a significant role. The Digital Regulation Cooperation Forum (DRCF) [74] explores how regulators can address the challenges posed by the increasing use of algorithms across different sectors. It maps the current landscape of algorithm auditing, identifies gaps, and considers how regulators might coordinate in overseeing algorithmic systems. A key focus is on accountability, transparency, and ensuring that auditing practices evolve alongside technological developments. Importantly, the report stresses that algorithmic bias is a major risk, since automated systems can replicate or amplify existing inequalities if trained on imbalanced data. Therefore, systematic audits should explicitly test for discriminatory outcomes in high-impact areas such as finance, healthcare, and online content moderation.

The AI act sets out that high-risk AI systems must be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret the system's output and use it appropriately (Article 13 (1) ). This requirement is also relevant for the public sector when purchasing an AI system from an external provider. A public sector organisation, as an implementer, which is about to implement an AI system must have sufficient

understanding of the logic of the AI system in order to make reasonable use of it.

### 3.2.3 Cybersecurity and product safety

Effective and modern cyber security measures are essential to ensure the protection of fundamental rights and personal data. The general principle is that consumer products must be safe [75] and a product that is not safe may not be placed on the market or put into service (Estonian Product Conformity Act (TNVS) [76] § 5 (1)).

The AI Act focuses on the safety of AI systems and provides that they must be accurate, stable and secure (see Article 15). High-risk AI systems must be designed and developed to achieve appropriate levels of accuracy, stability and cybersecurity throughout the life cycle (Art. 15 (1) AI Act). The referred accuracy levels and relevant accuracy parameters must be documented in the user manual accompanying the system (Art. 15 (3) AI Act).

Article 15 (4) of the AI Act requires high-risk AI systems must be as resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates. High-risk AI systems that continue to learn after being placed on the market or put into service must be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures. These AI systems must withstand unauthorized attempts by third parties to alter the system's output or performance by exploiting system weaknesses (Art. 15 (5) AI Act).

These requirements are not only important for high-risk AI systems, but also for other AI systems used in the public sector, as the use of public services depends on people's trust in solutions offered in the country and also in the State and administration. Thus, according to Article 15 (5) of the AI Act, the technical solutions used to address the various weaknesses must include effective measures to prevent, detect, counter, tackle and control attacks.

The purpose of the risk management system should be to identify and mitigate the risks to health, safety and fundamental rights of AI systems (AI Act pp. 65). Cyber security is also ensured by complying with the relevant requirements of the following legislation - Cybersecurity Directive [72], Cyber Resilience Act [77], Estonian Cyber Security Act [78].

The Cybersecurity 2nd Directive (NIS2) [72] sets out measures aimed at achieving a uniformly high level of cybersecurity across the EU (Art 1 (1) ). The requirements of the NIS2 Directive are transposed into Estonian law by The Cybersecurity Act (KüTS). KüTS sets out the requirements for the operation of network and information systems of the public sector, the requirements for liability and supervision and for the prevention and resolution of cyber incidents (Subsection 1 (1)).

The assessment of the cyber risks of a product containing digital elements, which is also a high-risk AI system, must be based on the different stages of the product and take into account the risks to the cyber complexity of the AI system in relation to unauthorised attempts by third parties to change the use, behaviour or performance of the system. Weaknesses inherent in artificial intelligence, such as data poisoning or counter-attacks, must be taken into account. Where appropriate, the risks to fundamental rights in accordance with the AI Act (Cyber Resilience Act rec 51; see also Article 12) should also be taken into account.

Compliance with cybersecurity and product safety requirements is crucial for AI systems. The security of AI systems is not only limited to technical safeguards, but also involves a wider responsibility to people and society as a whole. At the same time, it is imperative that AI meets the

strictest product safety requirements, avoiding risks and ensuring user protection throughout the life cycle of the AI system. By integrating cybersecurity and product safety principles in the initial stages of the design and development process, it is possible to design reliable, durable and human-centred AI solutions that support European and Estonian values and principles and increase society's trust in technology.

## 3.3 Standards that require addressing AI bias

The following international standards include provisions related to the mitigation of bias in AI systems. Standards that handle bias mitigation more indirectly are listed in Annex A.

**Table 1. Standards related to mitigation of AI bias**

| Number of the standard | Name of the standard | Explanations |
|---|---|---|
| ISO/IEC 42005 [40] | AI system impact assessment | The assessment focuses on understanding how the proposed AI systems and applications containing AI systems can affect people, groups of society or society as a whole. The standard supports transparency, accountability and trust in AI systems, enabling the organisation to identify, evaluate and document potential impacts throughout the life cycle of the AI system. |
| ISO/IEC 24027:2021 [79] | Bias in AI systems and AI aided decision making) | The Standard deals with bias in relation to AI systems, particularly when making AI-based decisions. The techniques and methods for measuring bias are described, with a view to identifying and reducing bias-related vulnerabilities. All stages of the life cycle of the AI system are covered, including data collection, training, continuous learning, design, testing, evaluation and use. |
| IEEE 7003-2024 [80] | IEEE Standard for Algorithmic Bias Considerations | The Standard describes processes and methodologies that help algorithm creators identify and reduce bias, providing guidance on the selection of validation datasets, definition of application limits, and management of user expectations to prevent unintentional misuse and misunderstanding of algorithms. |
| NIST AI 100-1:2023 [81] | AI Risk Management Framework | Fairness in AI involves addressing harmful bias and discrimination, though definitions of fairness vary across cultures and contexts. Mitigating bias doesn't guarantee fairness, as systems may still exclude or disadvantage certain groups. NIST identifies three main types of bias in AI – systemic, computational/statistical, and human-cognitive – all of which can exist without intent and can amplify harm if unaddressed. Also see [82]. |

| Number of the standard | Name of the standard | Explanations |
|---|---|---|
| NIST AI 600-1:2024 [83] | Generative AI Profile | Harmful bias in AI can amplify historical and systemic inequalities, cause performance gaps across groups due to non-representative data, or lead to uniform outputs that distort results and decisions. Additionally, human-AI interaction may result in issues like over-reliance, automation bias, or emotional attachment, affecting how people perceive and use AI systems. |

# 4 Addressing bias in AI systems risk management

## 4.1 How to approach bias within risk management

A generic risk management process consists of three steps: establishing the context, risk assessment and risk treatment. While establishing the context of operations, first the system and its stakeholders are described, then the organisation's risk appetite and risk acceptance criteria are to be defined and documented for the particular context. The risk assessment stage that follows, provides the identification of threats, marking some threats realistic and thus further handling these as risks and, evaluation of the risks – during the assessment usually also the risk owners are destined. Further, the risk treatment stage follows deciding on how to handle each of the risks – to avoid, mitigate, share/transfer or retain.

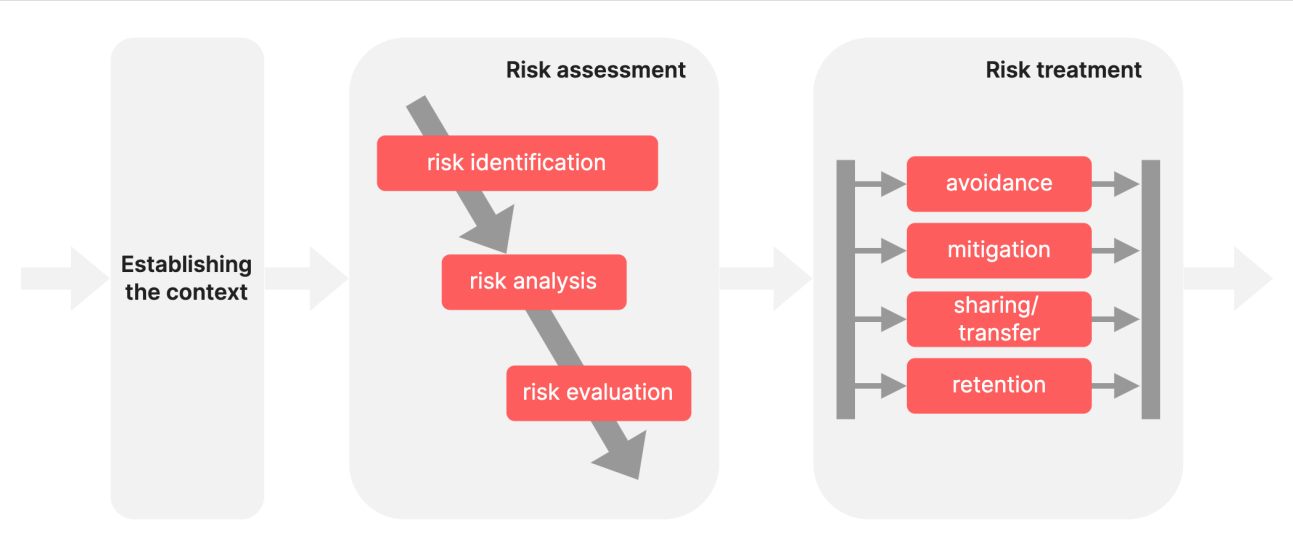


**Figure 3. Risk management process, adapted from [84]**

While risk management as a process can be somewhat unified, the objects the risk management deals with as well as the properties of these objects can be different. This is the reason why the risk based approach in loan business and insurance inherently differs from that in, e.g., infosec or nuclear safety. In this context, the risks of AI systems are somewhat close to the IT security risks while still requiring independent taxonomies, such as described in Section 2.5.

Although under the EU AI Act the risk management process only is mandatory for high-risk systems, it is strongly recommended to carry out risk analysis and assessment also for lower-risk systems. That does not necessarily mean setting up a formal risk management process but rather identifying and mitigating risks at the early stages of the system's life cycle that helps to prevent threats from realization or at least assists in mitigation of their effects. The identification and evaluation done even outside of a formal risk assessment process are beneficial as it becomes obvious what are the threats around the system. Once threats become evident and plausible, it becomes not so easy to overlook the necessity to treat these even in the full absence of a formal treatment process.

**Risk management standards and frameworks**

Risk management as a discipline is described in the ISO standard ISO 31000 [85] and the risk management framework (RMF) in NIST publication [86]. The specifics of risk management in information security is described in ISO/IEC 27005 [87] and that of cybersecurity in NIST cyber security framework (CSF) [88]. Standard ISO/IEC 23984 [89] describes how to extend the risk management process in compliance with ISO 31000 to an organisation that uses, develops or deploys AI systems. Finally, the Estonian Information Security Standard (E-ITS) [90] aligns with the ISO/IEC 27001 series, thus E-ITS implementers can make use of a risk-based approach. The risk management methodology for AI systems described in the [11] aligns with the ISO 31000, ISO/IEC 27005 and E-ITS workflows and is also in accordance with the NIST RMF and CSF frameworks.

## 4.2 Context description: Important aspects

### 4.2.1 System passport

As a very first step, it is necessary to identify and document the system itself and the stakeholders involved. System bias may arise due to biased input data or e.g. by wrong design decisions made by the system creators. If the organisation has access to the data used to train the models, it is possible to identify a possible data bias by just analysing the data set. When there is no access to the training data, bias assessments must rely on a limited information. Once the creators of the system and its rules have been properly surveyed, they can be contacted easily later and questioned over the potential bias to clarify the risks.

The stakeholders affected directly are:

- service users,
- persons, organisations or social groups related to the result of the service (whom the decisions made by the system affect),
- implementer of the system (reputational damage);

indirectly affected stakeholders are:

- data subjects whose data are used to train the model or create the business rules,
- system creator (reputational damage).

When it comes to a system built for a specific purpose, it is easy to record the direct and indirect purpose of the system. Where specific metrics exist to measure the achievement of the system goals, the organisation must document these. While it comes to a problem-adjusted or a narrowed GPAI model, it is yet possible to write down the direct purposes of the system, but it becomes more difficult to discover the rest (often hidden part) of the purposes of the GPAI. When using GPAI, it is very important to find out as much as possible about the underlying model and the data used to train it. Unfortunately, we often have to accept that these aspects remain opaque.

### 4.2.2 System usage scenarios

Next the usage scenarios[1] of the system are to be surveyed. These may partially overlap with the system goals, but they are more specific. Consider an example where the purpose of the system is to assess the likelihood of the potential benefits fraud. Then, depending on the design of the application, the usage scenario may be, for example, to compile a list of suspected fraudsters or instead, to calculate a fraud risk score for any client. These applications differ in that one issues a list of suspects, the other issues assessment results for all clients. Although the first routine can internally use the functionality of the second, these systems have very different outputs and thus different risks (both legal and ethical).

The system may contain one or more AI components, so one system may implement multiple usage scenarios. All usage scenarios must be described separately. It is important that the details of each usage scenario are documented, then it will be easier to make an later adequate overview of the threats in the system. Since several usages may share same risks, then if possible, these will be grouped and dealt with jointly at a later stage. However, it may happen that the same threat is assessed repeatedly, once per each usage story. This could happen when the risk scenario or controls applied differ per scenario.

For each usage scenario, write down the details of the scenario, describe the input data and related AI components (if a component supports multiple scenarios, that part can be copied), and information about the expected result of the use story with the indication how it will be used. The source of the information used in the compilation need to be correctly referenced so that it can be quickly retrieved and updated at a later date.

Details of the usage scenario must indicate who and when the usage scenario starts, what the system gets as input data, what is calculated or evaluated, and what is the result produced by the system. The subsequent use of the result must also be described.

For input data, any presence of personal data, its special categories as well as trade secrets must be made explicit. The collection of this data and the measures in place to protect the integrity and confidentiality of the data must also be described.

For the algorithm or AI component in use, its version and the year of creation, the type (algorithm or model) and parameters, the creator or origin, the deployment model and the component's interfaces with the rest of the system must be documented. Where possible, the accuracy of the system must be recorded, measures to reduce bias already in place must be described and existing test results and assessments, such as audit results, ethics assessments, tests, legal assessments, energy consumption assessments and environmental impact assessments, must be listed.

It should be documented when the result is personalised or contains special categories of personal data or well as trade secrets. A description is to be provided how the results will be used and what kind of decisions will be made based on the results. It should be inspected whether integrity and confidentiality of the result are protected.

[1] We are forced to use the term **usage scenario** for the reason terms '**user story**' and '**use case**' are reserved for IT development purposes and have a pre-defined structure and semantics unreasonable to adhere to for risk purposes.

## 4.3 Bias-related threats

### 4.3.1 Categorisation of threat scenarios in public sector AI systems

Public sector institutions utilise AI systems in extremely diverse contexts and for different purposes. To understand the risks of bias, it is first important to categorise the functional role of the AI system. The system's role is an important factor in the nature of both the threat scenarios and potential damage. It must be noted that the same technological solution can simultaneously have multiple functions and therefore be related to multiple threat scenarios. For example, a chatbot can both provide services and collect data for shaping policy.

**Policy-making and strategic management assistance tools** provide inputs for developing political decisions or long-term strategies through the analysis of large datasets or prediction of future trends. Bias risk is related to inaccurate or distorted analyses which can lead to misguided policy decisions affecting the entire society.

**Internal process support systems** include recruitment, resource distribution, organisation of work or other organisational decisions. Bias risk primarily manifests as unfair treatment or unfair distribution of resources within the organisation which can harm employees as well as operational efficiency.

**Administrative proceedings automation systems** make decisions or help make decisions on the rights, obligations, or benefits of citizens. This includes, e.g. permit issuance systems, benefit calculation algorithms, violation detection mechanisms, or tools for identifying circumstances relevant to the proceedings. Bias risk can materialise in the form of incorrect or inaccurate decisions or other outputs related to the use of the system. An incorrect decision can essentially mean the violation of a person's rights, restriction of their freedoms, or establishment of unlawful obligations.

**Systems providing public services and interacting with citizens** include chatbots, information systems, service provision platforms. Bias risk is related to differences in the quality of service provision or discriminatory interaction, which can negatively impact equal access to public services.

### 4.3.2 How the materialisation of threat scenarios can lead to damage

Should any of the threat scenarios described here materialise, the outcomes can be diverse and serious. Below, we will describe the main categories of damage that can result from bias in AI systems.

### 4.3.3 Bias can cause physical or economic damage

**Origins of threat.** AI models trained on historical data can learn and reproduce statistical patterns found in these data through their outputs. If these patterns reflect societal differences, the model may begin to associate certain demographic traits (such as gender, race, age) with certain outcomes. This may lead to discriminatory behaviour by the AI system implementing the model.

**Why is this a problem?** Such systems can systematically make adverse decisions against certain groups of people, regardless of individual differences. This leads to unfair access to jobs, loans, healthcare or law enforcement. Models designed to optimise processes can thus embed and automate earlier differences, making it more difficult to identify and address bias-related problems.

**What can go wrong? What kinds of damage can occur in what kinds of systems?**

Example. Healthcare

An algorithm widely used in US hospitals directed black patients to special care programs less frequently compared to white patients with a similar condition. This caused unnecessary health-related suffering to a group. The problem appeared because the algorithm evaluated current health condition as a function of the cost of medical procedures provided previously and leaving out of consideration the fact that historically, less money has been spent on the treatment of black patients. Thus, the model learned that lower costs meant less need for assistance, which was an incorrect relation. [91].

Example. Recruitment

Amazon's recruitment algorithm began to systematically give lower scores to female candidates. The system was trained using the biographies submitted over 10 years, most of which came from men. The model learned that 'male gender' is a predictor of success, and began to punish résumés that included the word 'women' (e.g. 'captain of the women's chess club'). This caused economic damage to a population group [92].

Example. Public services

Algorithms for assessing the risk of fraud in social benefits in the Netherlands systematically put citizens with immigrant backgrounds at a disadvantage, which led to discrimination in the processing of benefits and later to a widespread scandal [93]. Another example – it has been shown that predictive policing algorithms can disproportionately target police resources to certain areas, based on historical arrest data that already reflect previous patterns of policing and possible bias [94].

Example. Linguistic inequality affects quality

Large language models are predominantly trained on English language data. Therefore, their ability and accuracy in languages with a lower number of speakers is systemically lower. For example, the **MMLU-ProX** benchmark has shown differences of up to 24.3% in the ability of language models between languages [95]. However, the problem is not limited to lower quality (e.g. factual errors or unnatural phrasing), but also concerns safety. Experimental results show that due to the shortage of corpuses to fine-tune safety, all LLMs produce significantly more unsafe responses for non-English queries than English ones. The adverse effect for smaller nation groups is that their AI systems tend to be of lower quality or unsafe. [96].

**Example. Language inequality affects culture**

Specifically in scenarios where no information is in the language of the query, LLMs prefer documents in high-resource languages during generation, potentially reinforcing the dominant views. Such bias exists for both factual and opinion-based queries [97]. A digital divide may appear putting the quality and safety of the service into dependence of the user's language and thus cultures with smaller populations become weaker.

### 4.3.4 Erosion of human agency, dissipation of responsibility

**Origins of threat.** The threat arises from the human tendency to place excessive trust in the outputs of automated systems (*automation bias*). This effect is particularly strong when using so-called black-box processing where decision-making processes are not transparent or require additional human effort to validate them. This leads to a dissipation of responsibility: if a harmful decision is made by an algorithm, then it may be unclear who is legally or morally responsible – the developer, implementer or human operator.

**Why is this a problem?** In this situation, any error (including bias) in the algorithm or AI system will amplify damage. A system that guides a human in decision-making will erode professional expertise and critical thinking, as humans delegate decision-making to a machine. In case of public services, it reduces the impact of important human factors in decision-making, such as empathy, ethics and contextual understanding. Systems are becoming more inflexible and less human [98]. What is more important, however, is the deterioration of responsibility as a fundamental principle of democratic and legal law. If the chain of responsibility is severed, the affected parties may lose the right or even possibility to challenge the decisions or obtain compensation [99].

**What can go wrong? What kinds of damage can occur in what kinds of systems?**

- **Administrative proceedings:** Official's professional discretion is replaced by an inflexible algorithm that cannot account for the particulars of the specific case or human aspects. In practice, many public sector automatic decision-making systems have been cancelled due to the fact that they replace human discretion with an inflexible rule set or ignore principles such as the presumption of innocence. Such automation can cause public resistance which can lead to the termination of the entire project [9].
- **Policy-making:** Critical analysis is replaced by 'data-based' decision-making which can mask the fact that significant questions have been left unanswered. For instance, public discourse and the media can paint of a picture of an 'expected AI' which is often much more autonomous and capable than actually existing technology. This cultural image facilitates the emergence of the belief that an AI system is independent and infallible, overlooking the fact that the system is a result of human development and operational processes [100].
- **Internal processes:** The skills and experience of leaders and HR specialists remain untapped if decision-making is delegated to algorithms. This erodes the organisation's internal expertise and can result in decisions which may be technically correct but overlook the organisation's culture and human aspects.
- **Public services:** Service providers lose contact with citizens and their real needs. Systems become less flexible and less human, reducing the impact of important human factors, such

as empathy, ethics and contextual understanding in decision-making processes [98]. T

### 4.3.5 Organisational damage, reputational damage

**Origins of threat.** This is a risk from the perspective of the organisation deploying the AI system. If a biased or otherwise faulty system causes public damage, the subsequent negative attention and setback may prove to be very damaging to the organisation itself and its objectives. The risk lies not only in the unfair outcome, but also in the direct negative consequences for the creator or implementer of the system.

**Why is this a problem?** The problem is expressed in direct financial damage (failed projects, legal costs), loss of public confidence, and increased regulatory pressure. This could make the implementation of future AI projects considerably more difficult as the public, customers and employees become more sceptical of AI. For public authorities, this means failing to serve their citizens and erodes confidence in the state and the government in general.

**What can go wrong? What kinds of damage can occur in what kinds of systems?**

- **Empirical evidence:** The **RealHarm** database, which collects examples of AI failures, found that reputational damage is the most common type of damage for organisations. The study also showed that existing guardrails could not have prevented many of these incidents, indicating a gap in the practice of deploying AI protection systems [101].
- **Cancelled projects:** The survey, which analysed the failures of automated public sector systems, identified 61 cases where projects were terminated. This was due to a combination of technical deficiencies, budget overruns, proven bias and public pressure. All these factors contribute directly to organisational damage [9].
- **Business failure:** The previously mentioned case of Amazon's recruitment tool is a classic example of business failure. Wasted time and large financial investment resulted in an unusable product, plus negative media coverage and reputational damage, which forced the company to suspend the project [92].

**What can go wrong and what kind of damage occur in the public sector?**

- **Administrative proceedings:** The violation of citizens' rights will lead to legal disputes, media scandals and political pressure.
- **Internal processes:** Discrimination of employees can lead to labour disputes and loss of talent.
- **Public services:** Low-quality service or discrimination will damage the institution's reputation and public trust.
- **Policy-making:** Policy decisions made based on biased analyses can mean widespread criticism and political responsibility.

**Example. Administrative proceedings: Public services**

Dutch social benefit fraud risk evaluation algorithms systematically placed citizens with an immigrant background in a disadvantaged position, leading to their discrimination in benefit awarding procedures and subsequently resulted in a major scandal [93] (see also the SyRI case discussed in Section 3.1.3). It has also been demonstrated that predictive policing algorithms can unproportionally direct police resources into specific areas based on historical arrest data which already reflect past police work patterns and potential partiality [94]. Both examples demonstrate how systems designed to automate administrative proceedings can unlawfully limit citizens' rights and subject them to unfair obligations, thereby eroding the public's trust in the state and its services and causing reputational damage.

### 4.3.6 Damage to democracy, the rule of law, and social cohesion

**Origins of threat.** The bias in AI systems threatens the very foundations of a democratic society – the trust of citizens in the state and public institutions and the rule of law. Biased algorithmic systems violate the fundamental principles of democracy and individual freedoms which form the foundations of just societies. Such systems are often characterised by a lack of transparency, both in general as regards the implementation of AI and its precise impact. In addition, a person may lack the choice about the impact of AI on their life. At the same time, the bias of AI systems can normalise, increase and amplify existing social inequalities, marginalisation and prejudice. The injustices that already exist in the areas of social policy (housing, health), education, insurance, finance and trade make citizens vulnerable to restructuring. This creates divisions in society and hinders the equal participation of all citizens in social life.

**How do different types of systems create such risks?**

- **Administrative proceedings:** Unfair automatic decisions regarding citizens' rights, obligations or benefits will undermine the perception of the principles of rule of law and equality. Systemic discrimination in the distribution of benefits or determination of obligations will also deepen existing inequalities.
- **Policy-making:** Biased data analysis or forecasts can lead to discriminatory policies which will systematically damage certain groups, e.g. by ignoring their needs or amplifying their marginalisation.
- **Public services:** Differences in service quality between different groups will damage the principle of equality and public trust. This can increase the social (digital) divide and stratification.
- **Internal processes:** Discrimination in recruitment or resource distribution within the public sector can reduce the diversity and quality of public services which will indirectly affect the entire society.

**Why is this a problem?** Loss of trust in technology and institutions will undermine democratic processes and can lead to a crisis of legitimacy. AI bias will endanger the fundamental principles of rule of law: right to equal opportunities and impartial treatment in socioeconomic, age, ethnic, religious or sexual perspectives; the right to a just judiciary system and a just administrative system in general; and the right to the protection of privacy and personal assets. For instance,

the presumption of innocence will be turned upside down if all applicants for social benefits are checked for potential violations rather than qualification criteria. The deepening of bias together with the automation can undermine social cohesion and prevent society from realising its full potential through inclusivity. If certain groups are systematically deprived of equal opportunities in education, the labour market, health care or other areas of life, then the society cannot take advantage of the talents and contributions of all its members. This leads not only to individual loss but also to collective loss. Society is missing out on innovation, economic growth and social development, which could be brought about by full and equal participation of all citizens. If constitutional rights and democratic values are rendered meaningless by automatic systems, this will damage the legitimacy of the state and erode public trust, which will in turn deepen social divides and hinder the development of the society.

## 4.4 Evaluation of bias risks

### 4.4.1 Assessment of the severity of bias-related threats

Each organisation has to define its risk appetite – i.e. readiness to deal with the consequences of a threat event. If the consequences of a threat are evaluated to be so serious that the organisation prefer not to face these, a strategy has to be prepared to prevent and avoidance these. In Section 4.3 we presented examples of possible consequences of bias of algorithms and AI systems (material loss, health damage, damage on national and social levels). There are a number of aspects which significantly increase the risks of an AI system thus needing the attention from the deployer of the system.

**The output of the AI system affects life, health or financial situation.** If the output of the AI system directly or indirectly affects the provision of money, work, assistance, medical treatment, physical or mental violence against another person, or the unjustified restriction of his or her rights, the consequences of bias in the system are serious. Damage to the autonomy, privacy or other freedoms of a person is at least as important depending on the organisation of society, but risk analysts are often more clear about specific material or physical damage. Therefore, when describing the risks, it is worth considering whether, for example, data leaks can lead to later direct property or physical damage, even if the damage has not yet been demonstrably arrived.

**The AI system is an immediate decision-maker and/or implementer.** If the AI system's output is automatically transformed into a decision, the consequences of the bias will be serious. An AI-based system of automatic penalty procedures (e.g. for parking irregularities or speeding) has an immediate and specific impact. If such an AI system is biased and without proper monitoring and supervision, it can lead to measurable losses, especially if the person did actually behave correctly. We note that even when a human has been added to the system to oversight the decisions, they may develop a dependence on the opinion of the machine over time, and the effectiveness of the oversight decreases.

**The AI system changes beliefs, values or behavioural patterns.**

If the AI system affects the behaviour of a large number of people over a longer period of time, the consequences of bias are severe. For example, AI systems created for educational, psychological counselling or entertainment (but biased) can affect people's behaviour over a longer period of time. If they also perform their primary task (promoting skills, solving complex relationship problems or entertainment), then in addition, they can create harmful patterns of behaviour (addiction to the machine in all activities, unwillingness to communicate with other people). These risks are also present when services are provided by humans, however the amount

of human guiding and associated harms remain limited, unlike for AI systems which, while following exactly the same guidelines can affect hundreds of millions of people.

### 4.4.2 The art of setting thresholds

The risk management methodology for the bias of algorithmic and AI systems inherently consists of multiple steps. The more serious the threat, the more thoroughly should the mitigation measures be weighted.

**Number of persons affected** The easiest way to evaluate a threat is the number or percentage of affected persons in a given period of time. Examples of evaluating damage are as follows:

- During the system lifetime, one person will die due to the system's bias[2].
- One person per year gets health damage due to the system's bias.
- Every year, the system's bias leads to the end of the activities of ten viable organisations[3].
- $0.1\%$ of the users of an AI system suffer unreasonable financial losses due to its bias.
- $0.5\%$ of the population per year lose their jobs due to bias.

**Example. Training provider had to compensate the losses to rejected applicants caused by an age discriminatory algorithm**

The EEOC (Equal Employment Opportunity Commission), a federal agency in the United States, sued iTutorGroup, a language-teaching company, because their recruitment information system automatically rejected female applicants over 55 and male over 60. The company had to pay $365,000 in compensation, implement anti-discrimination policies, and provide training for its recruiters [102].

**Time to first damage** In addition to thresholds related to victims, it may be expedient to estimate how long it will take for the damage event to occur. This is particularly important for threats where the bias of an algorithm or an AI system affects people over a long period of time. A longer-term perspective must also be considered in a situation where the risk assessment is carried out before the system is built and the implementation lasts several years, for example:

- threats associated with bias may occur within 10 years from the assessment; or
- threats associated with bias may occur within 7 years after commissioning the system.

## 4.5 Risk treatment for bias threats

### 4.5.1 Possible outcomes

Evaluating and treating the risk of bias in the algorithm or AI system will likely have one of three possible outcomes.

1. The system is commissioned without changes and all residual risks will be accepted (possibly after implementing additional measures).
2. The system is deemed biased to the extent that its development or use must be suspended because treating the risks is either not possible or not economically rational. While this is a

[2] We assume here the AI system is of civilian use.
[3] Note that it could be also quite possible and even desirable to create an AI system to terminate the activities of certain organisations (e.g. criminal ones).

sure way to avoid the threats of bias, the expected benefits of an algorithmic or AI system will also be lost.

3. The system is commissioned on the condition that appropriate additional organisational and/or technical protection measures can be implemented. This is probably the most frequent result.

### 4.5.2 Complexities and trade-offs in bias mitigation

While there are many technical methods for bias mitigation, it is important to understand that this is not a simple technical bug-fix. The mitigation of bias is a complex sociotechnical problem without a generic and definitive solution. Each intervention comes with its own compromises and philosophical choices. The first and most fundamental challenge is that fairness lacks a single, universally accepted definition. An article introducing Google's What-If tool illustrates this, citing the example of five experts who all define gender fairness in loan decisions differently [103].

- **Group unaware**: Gender must not be taken into account at all, even if it means that no woman ever can get a loan.
- **Group thresholds**: Thresholds for creditworthiness should be set separately for women and men in order to compensate for the historical data disadvantage.
- **Demographic parity:** The gender distribution of approved borrowers must correspond to the gender distribution of applicants.
- **Equal opportunities:** Qualifying men and women must be equally likely to receive approval for a loan application.
- **Equal accuracy:** The accuracy of the model's predictions (for both positive and negative decisions) must be equal across both genders.

All these definitions contradict each other. Each option is a compromise that requires a decision according to the specific design and purpose of the AI system, the risk tolerance of the organisation, business requirements, and the legal landscape.

**Differences in outputs do not always equate to discrimination.** If it assumed that any statistical differences in the results for different groups is necessarily the result of unfair discrimination, there is a risk of falling into the trap of poor epistemology. It is possible that the model has identified real, non-discriminatory, underlying variables that correlate with demographic groups. In such a case, changes can lead to new injustices, for example by favouring less-qualified candidates to meet a quota. It is therefore critical to thoroughly analyse the actual causes of the differences before implementing mitigation measures.

**Trade-off between fairness and accuracy.** Most bias mitigation techniques act as limitations on the model optimisation process. A system with limitations is almost always suboptimal in performing its original, unfettered task. In practice, this often means that fairness is increased at the expense of the overall accuracy or performance of the model. As the Fairlearn example showed (see the description in Section 5.4), calibration of decision thresholds did reduce gender inequality, but also led to a slight decrease in the overall accuracy of the model. This is a compromise that organisations must take into account in a system's development.

### 4.5.3 Cost-efficiency of bias mitigation measures at different life cycle stages

In Section 2.3, we described the life cycles of creating a machine learning model and deploying an algorithm or model in an application. Bias avoidance at each stage of the life cycle comes at different prices [104, 105].

**The best time to prevent bias is during the system's design process.** The time after the vision and objectives of the system are formulated is the best moment to think about both the beneficiaries of the system and donors of the data in it. We recommend doing the following.

1. Define the population groups on whose data the algorithm or model has to work correctly.
   - Determine whether it has to behave correctly on all the people on planet Earth, on the citizens of a particular country, or on a smaller group.
   - Define the distribution of the age groups – how old and how young persons the system has to work with?
   - Define the expected gender distribution.
   - Define the natural language the system has to work with.
2. Define the rules about how to behave when choices and compromises need to be made in a situation where an ideal dataset, algorithm or model is not available, based on Section 4.5.2.
3. Add this information to the system's procurement or development requirements as an important requirement.

**The best (but not necessarily the cheapest) moment to avoid bias is while the algorithm or a machine learning model is being designed.** The changes of mitigation of bias risks are the best if the creation of the machine learning model is under the control of the client ordering the AI system (the model is created in accordance with their requirements). We give an overview of these possibilities in Section 4.5.4.

**In case the system has already been built or procured, dealing with bias is more expensive and perhaps even impossible.**

If unacceptable risks materialise in a system procured as a complete package, then the system must be either modified or replaced. If it is possible to replace an algorithm or machine learning model in the AI system (standard interfaces have been used), then it may be possible to develop or acquire a better model that works better on the target population. Sometimes the system can be rebuilt or additional technical or organisational measures can be added to prevent the replacement of the whole product, algorithm or model. Possible solutions are described in Section 4.5.5.

If the behaviour of the system cannot be changed, it must be discarded and, if necessary, a new system must be obtained. Relevant cases are covered in Section 4.5.6.

### 4.5.4 Mitigation of bias in ML model training

The economic reality of the IT industry makes in-house (on-premises) development of machine learning models expensive so it is often necessary to make do with models trained by others. If the developer of the AI system builder has control over how a machine learning model is created, then relevant requirements can be established in both the data collection and preparation, and model training stages.

**Data collection and preparation stage**

Bias can result from:

- under- or overrepresentation of specific groups within the model (representational bias) [104];
- measurement errors or high-noise data which reinforce historical and/or representational bias in the model being developed [106].

Possible measures for bias mitigation are as follows.

- In case the collection of training data is under the control of the system's creator (e.g. is a part of the project) then we recommend organise additional data collection to ensure the diversity and representativeness of the data (finding additional data on underrepresented groups) or leaving out data on overrepresented groups (while monitoring quality indicators)[4] [104].
- If the statistical distribution of data on underrepresented groups is known (or it is analytically identifiable) we recommend considering synthetic generation of additional data on these groups.
- We recommend considering the use of bias detection and data auditing tools, with an emphasis on tools that are being improved and updated [106].
- In case acquiring additional data is not possible, we recommend the application of balancing techniques in model training [106].

**Model training stage**

Bias can result from:

- unfairness of algorithms due to bias in training data or unsuitable optimisation criteria [107];
- prejudiced inferences in training data (e.g. connection of names to stereotypes) [108];
- overlearning based on data on dominant groups [107];
- inputting non-relevant personal data and sensitive data without any need or legal basis [108].

Possible measures for bias mitigation are as follows.

- If certain features are known not to increase the quality of the machine learning model but amplify the model's bias we recommend leaving these features out of the training data.
- We recommend using smaller and more easily controlled models with fewer 'surprises' in their behaviour [108].
- We recommend carrying out quality control during the preparation of training data, which should include the deletion of duplicates and filtration of the training dataset [108].
- If fairness-based algorithms are available for the functionality of the system being developed, we recommend considering their use [107].
- We recommend regular validation of the system's outputs using inputs from different groups, especially when deploying a new version of the model [107].
- We recommend carrying out bias impact assessments throughout the whole development process [107].
- If personally identifiable data are used for training or fine-tuning we recommend following the principles of data minimisation and data protection by design, as well as implementing additional protection measures for the protection of such data.

[4] Before using these (and other) mitigation measures we recommend reading Section 4.5.2. Also recall that using data synthesis to train ML models can amplify learning bias.

**Example. ML models that rely on a biased samples are also inefficient**

Several studies have attempted to replicate previous experiments in machine learning model training, and failed – the models in recurrent studies are not as successful as the original study models. The reason for this has often been sample bias – the effect appears only on certain training data, and the result cannot be repeated on larger and more representative datasets. This means that biased models can also just not work. Inefficiency of models trained with biased datasets is common, for example, in research on autism [109], radiological images [110], and obstetrics [111].

### 4.5.5 Reducing bias when interfacing or deploying an algorithm or model

Bias will already be present in AI system or algorithm at the interfacing stage. It is important to understand that a provider of a marketed and functioning AI system has significantly more legal obligations than in the course of its development (see Section 3). It is also worth considering that it may be more expensive to mitigate bias at a later stage than at the beginning of the creation of the system. Below, we will list some potential mitigation measures that can be employed at different life cycle stages.

**Fine-tuning the ML model for an AI system**

Bias can result from:

- reinforcement of statistical bias during fine-tuning;
- use of sensitive data in fine-tuning without lawful basis [108];
- reinforcement learning or automatic feedback cycles which amplify the ML model's bias [108].

Potential bias mitigation measures are as follow.

- In a situation where the base model cannot be changed but can be fine-tuned, we recommend setting bias reduction as an additional fine-tuning goal.
- We recommend ensuring the transparency of the fine-tuning process, documenting all changes made to reduce bias.
- Similar to training, we also recommend the principles of data minimisation and data protection by design and employing additional protection measures in fine-tuning [108].

**Model implementation, i.e. interfacing the model into an AI system**

Bias can result from:

- in high-impact contexts, discriminatory or hallucinated responses can lead to sustained bias in the AI system [104, 108];
- non-explainability of the model's decisions makes it difficult to identify the sources of bias and hinders their mitigation [112];
- biased or profiling information present in the shared context (e.g. context window of a chat) can affect the model's responses;
- the lack of protection measures against prompt injections enables bad-faith users to give biased instructions to the model [108].

Potential bias mitigation measures are as follow.

- In case the model cannot be fixed, we recommend considering swapping the algorithm or ML model for a less-biased option.
- Independent of the model's features, we recommend including a human in the process as a final decision-maker.
- Human supervision of the system must be strengthened [107].
- We recommend training the decision-makers and supervisors to understand the verification and explainability of the AI-system's outputs.
- In case the final decision will not be made by a human, the system must include output explainability and quality monitoring features. For this purpose, we recommend adding the option of requesting an output review which will forward the decision to a human for verification [108].
- Responsibility schemes and transparency reports must be implemented in the system [107].
- We recommend also treating bias related to the integration of the ML model to datasets by employing bias-reducing and privacy-preserving architectures (e.g. RAG with guardrails).
- Regular audits of the AI architecture and assessments of explainability and decision quality are required [108].

**Operation and monitoring of the AI system**

Bias can result from:

- drift in model bias over time due to changes in social expectations [105, 108];
- the ML model's output accidentally reveals personally identifiable data which allows a human decision-maker to make a biased decision which the machine would not have done.

Potential bias mitigation measures are as follow.

- We recommend organising continous assessment and real-time monitoring of the quality of the model's outputs [105].
- Feedback on the system's behaviour must be regularly gathered from stakeholders and users [108].

### 4.5.6 Situations necessitating termination of the system's use

If risk assessment reveals that a system using an algorithm or a machine learning model does not comply with the requirements of existing laws, it cannot be commissioned. If a law prohibiting the use is adopted during the operation of the system, the use must be terminated. If a law establishes additional conditions which the system does not comply with, the use of the system must be suspended until compliance is ensured. In certain cases, a similar ban can be initiated e.g by a relevant ethics committee.

In some situations, an otherwise lawful system can produce an unlawful output. In such cases there is no reason to immediately terminate the use of the system; instead, the system can be fixed, e.g. following the recommendations presented here. This also illustrates why the reduction of risks related to the system to near-zero is not always required – this may not be realistic, and a threshold established in such a manner renders the development of new services prohibitively difficult.

In the course of risk assessment, some risks can remain undetected and the treatment of other risks might not mitigate them completely. Mitigating a risk also does not bring the risk to zero. The remainder is called the residual risk. If the residual risks of an AI system exceed the organi-

sation's risk appetite, it is rational to either terminate the use of the system or to not deploy it at all.

A residual risk can exceed the risk appetite e.g. when:

- it cannot be ruled out that a biased AI system causes bodily harm or death to a person;
- it cannot be ruled out that the bias of the AI system directs a person to cause bodily harm or death to themselves or another person;
- financial damage arising from the deployment of the system exceed the possibilities of the organisation to cover them.

# 5 Tools for dealing with bias

## 5.1 Black box vs white box system

The terms 'black box' and 'white box' systems are often invoked in the context of AI systems. A black box is a system whose internal operation is opaque, while for a white box it is transparent. In practice, however, this distinction is not so clear-cut. It is more useful to think of the openness of the AI system and model as a spectrum that ranges from the most closed to the most open.

This spectrum is defined by several principal factors: how much information has been published about the development of the model (its data and training process), what access is available to the model itself, and how it is implemented. At the same time, the internal complexity of the models themselves also adds to the lack of transparency. Even a system that is fully transparent in its structure can functionally be a black box if we are unable to interpret its decisions.

### 5.1.1 Spectrum of openness of AI systems

**Dimensions of openness.** We list the main dimensions used to assess the openness of the system.

- **Training data.** It is the foundation of the model's knowledge, abilities and bias.
  - Most closed (opaque): no information available regarding the data.
  - Partly transparent: the main data sources (e.g. Common Crawl, Wikipedia) have been disclosed, but the final, processed data and blending relationships are not shared.
  - Most open (transparent): full and pre-processed training data is available in machine-processable form.
- **Data processing. Training blend.** The way the data is filtered and mixed affects the model's ability as much as the data itself.
  - Most closed: no details are published about the process.
  - Partly transparent: the overall methodology is described in a technical report (e.g. 'we used aggressive de-duplication').
  - Most open: Full source code is published for for cleaning, filtering and blending the data.
- **Architecture and weights of the model.** This determines whether the model can be used and studied by by others.
  - Most closed: no information on architecture and number of parameters.
  - Partly transparent (architecture only): architecture is described, but training parameters (weights) are closed. The user understands the essence of the model, but cannot run it independently.
  - Most open (with open weights): model weights are publicly downloadable. Generally, this is meant when a model is called an open source AI model.
- **Training details and 'recipe'.** This enables other scientists to replicate the results.
  - Most closed. No information available
  - Semi-transparent: main hyperparameters (e.g. learning step) and hardware information are published.
  - Most open (training recipe is open): Full training code, configuration files and logs are published.

- **Access methods and the extent of control.** This determines how much control the developer has over the behaviour of the model.
  - Most closed (integrated into the product): AI is only a specific functionality of specific applications (e.g. an image processing program).
  - API[1]-based access: Standardised for commercial models. The host system sends a query and receives a response. The APIs themselves are also on the spectrum: from simple input-output to more accurate control (e.g. access to log scores).
  - Most open (direct access): The user downloads the model scales and runs it on their own hardware, with full control.
- **License.** This determines what is legally permitted to do with the model.
  - Most closed: use is limited by terms of use; the model cannot (may not) be downloaded or modified.
  - Open licence with restrictions: the licence allows certain uses (e.g. research) but imposes restrictions on commercial use or requires disclosure of changes.
  - Most open (permissive licence, e.g. Apache 2.0, MIT): The licence allows the model to be used, modified and distributed almost without restriction, including for commercial purposes.

Table 2 illustrates the openness spectrum by comparing several well-known AI model manufacturers and service providers. The openness of the AI system or model is not a single attribute, but a set of independent choices made by developers and service providers. We note that the openness of the models is described at the time of writing this Guide and it may change over time. The table describes a number of different market strategies.

Closed API Providers (OpenAI, Google Gemini, Anthropic). Their models are black boxes by design, but they offer powerful APIs for developers. Their business model is providing a state-of-the-art service regulated by the terms of service.

**Open Weight Providers (Meta, Mistral, Qwen, DeepSeek, Google Gemma).** The most common form of open-source AI. Data and training processes often remain proprietary, which makes them semi-opaque for scientific reproducibility.

**Diversity of licenses.** Even among open models, there are big differences in licenses. Whereas Allen AI uses the permissive Apache 2.0 license, Meta (Llama) and Google (Gemma) use their private licenses, which can place restrictions on e.g. very large companies. For example, Cohere's licence is non-commercial, restricting immediate commercial use.

**Ecosystem-based openness (NVIDIA).** NVIDIA Nemotron series models are open-weights, but their license is strategically limited: it allows free testing, but for commercial use in production it requires joining NVIDIA's paid AI Enterprise platform. This is a try-before-buy strategy designed to link users to their hardware and software ecosystems.

**AI integrated into the product (Apple, Midjourney).** These systems are not designed as developer tools but are integrated into end-user products (e.g. operating system or chat application). Users communicate with AI through the product, not directly.

**Fully Transparent Model (Allen AI).** The OLMo series is exceptional because it aims at scientific reproducibility. Everything is public: data, code and model parameters (weights).

[1] API – Application Programming Interface

**Table 2. Openness of common LLMs**

| Model / Developer | Training Data | Pre-processing of training data | Architecture and Weights | Training details ("recipe") | Access and Control | License |
|---|---|---|---|---|---|---|
| **OpenAI** (Top models) | Closed | Closed | Semi-open | Closed | API | Proprietary |
| **Google** (Gemini family) | Closed | Closed | Semi-open | Semi-open | API | Proprietary |
| **Anthropic** (Claude family) | Closed | Closed | Semi-open | Semi-open | API | Proprietary |
| **Meta** (Llama family) | Semi-open | Closed | **Open** | Semi-open | **Weights** and partners APIs | Llama licence (restricted) |
| **Mistral** (Open models) | Semi-open | Closed | **Open** | Semi-open | **Weights** and API | **Apache 2.0 (permissive)** |
| **Alibaba** (Qwen3 family) | Semi-open | Closed | **Open** | Semi-open | **Weights** and partners APIs | **Apache 2.0 (permissive)** |
| **DeepSeek** (DeepSeek-R1) | Semi-open | Semi-open | **Open** | Semi-open | **Weights** and API | **MIT (permissive)** |
| **Google** (Gemma family) | Semi-open | Closed | **Open** | Semi-open | **Weights** | Gemma license (restricted) |
| **NVIDIA** (Nemotron family) | Semi-open | Closed | **Open** | Semi-open | **Weights** (restricted) | Private license (NVIDIA, paid in production) |
| **OpenAI** (Whisper family) | Semi-open | Closed | **Open** | Semi-open | **Weights** and API | **MIT (permissive)** |
| **OpenAI** (o1-mini) | Closed | Closed | **Open** | Closed | **Weights** | **Apache 2.0 (permissive)** |
| **xAI** (Grok family) | Semi-open | Closed | Closed (Grok-1) | Semi-open | **Weights** and API | **Apache 2.0 (permissive, Grok-1)** |
| **Cohere** (Command R+) | Semi-open | Closed | **Open** | Semi-open | **Weights** and API | CC BY-NC-SA (non-commercial) |
| **Apple Intelligence** | Closed | Closed | Closed | Closed | Integrated with product | Proprietary |
| **Midjourney** | Closed | Closed | Closed | Closed | Integrated with product | Proprietary |
| **Black Forest Labs** (FLUX family) | Closed | Closed | Semi-open | Closed | **Weights** and API | Flux License (non-commercial) |
| **Allen AI** (OLMo 2 family) | **Open** | **Open** | **Open** | **Open** | **Weights** | **Apache 2.0 (permissive)** |

### 5.1.2 Explainability of the model

It is important to understand that even a fully white box model with all components (data, code, scales) in the public domain is still a functionally black box without specific explainability and interpretability techniques. The internal operation of the model is just too complex to understand directly, even if we have its weights and their activations for a specific input. Therefore, separate methods are needed to investigate why the model makes certain decisions.

However, there is a difference between whether just the model or the whole system is a black

box. If an external API service is used, the entire system can be considered a black box. The integrator of such an API is not in control of the infrastructure of the model, any filters prior or post to queries, and is unable to guarantee the stability of the service. On the other hand, when hosting an open-weights model in-house, the system becomes more transparent because the implementer then has control over the entire chain, from hardware to processing the model's outputs. This distinction is often more important in practice than the full transparency of the model itself. To simplify the issue, we further distinguish the black box AI systems where the implementer has no access to the model and its operating processes and the white box systems where the model is operated by the deployer.

## 5.2 Measures for mtigating the bias of a black box system

For black box systems, such as using an external API service, there is no access to model training data and source code. It is therefore not possible to change the model itself. Bias mitigation measures have to be external to the system and focus on how the model is tested, implemented and what protection measures are built around it.

**Systematic testing and auditing.** The nature and extent of the bias present in the system must be understood before mitigation measures can be implemented. Because of the intrinsics of the model remain invisible, the model has to be tested behaviourally, i.e. through input-output analysis. This is similar to software penetration testing.

- **Benchmark tests.** The system can be tested using standard bias measurement tools (such as SafetyBench or AgentHarm) to compare its results with other models.
- **Robustness tests (red teaming).** This includes providing deliberately provocative or sensitive inputs to the system to detect potentially biased or harmful responses. The aim is to find weaknesses in the system.
- **Counterfactual analysis.** The system is provided with inputs that have been minimally modified, for example by changing only one demographic characteristic (e.g. name, gender, age), and must monitor whether the output of the model changes significantly. This can help to identify discriminatory behaviour.

**Using external guardrails.** Since the model itself cannot be changed, a protective layer has to be created around it, which filters both inputs and outputs.

- **Input validation.** User queries are validated before they are sent to the model. If the query is provocative in nature or may cause bias, it can be blocked or recast.
- **Output validation.** The model's response is validated before displaying it to the user. If the answer contains biased language, harmful stereotypes or factual errors, it can be replaced by a neutral message ('I cannot answer this question') or directed to a human operator.

**Organisational measures.** Often the most effective measures are not technical but organisational.

- **Restriction of usage scenarios.** If testing shows that the system is unreliable or biased in a given area (e.g. medical recommendations or candidate assessments), the most robust measure is to prohibit the use of the system in this context.
- **Ensuring human oversight.** For high-risk decisions, AI output should never be the final

decision, but only an informational input to a human expert. This ensures that the ultimate responsibility is borne by a human.

- **Changing Service Provider.** If the bias of the system is excessive and the service provider does not provide adequate solutions, the last resort is to choose another, more reliable product or service provider. This creates market pressure motivating developers to create fairer systems.

## 5.3 Measures for mitigating the bias of a white box system

In a white box scenario, there is access to the internal components of the model: model weights, and sometimes training data. This situation usually occurs when the system is developed within an organisation or in cooperation with a partner that ensures transparency, or when the organisation implements the external system in-house. This offers wider possibilities for mitigating bias, as it is possible to address the root causes of the problem at every stage of the life cycle of the AI system.

**Auditing the dataset and model.** Model diagnostics tools such as Google's What-If Tool or Microsoft's Responsible AI Toolbox can be used to find the root causes of the bias. They enable interactive analysis of the behaviour of the trained model on different data segments, visualisation of differences in results between demographic groups, and the generation of counterfactuals. Such an audit will provide input for the selection of the following mitigation measures.

**Data-level measures.** If an audit reveals problems with the dataset, the following measures can be applied before training the model:

- **Data balancing.** The impact of underrepresented data groups can be reduced, for example, by re-weighing data in a training process. Greater weight is given to the underrepresented data points. [113].
- **Data augmentation.** Additional synthetic data can be generated for underrepresented groups to improve their representation.

**Model-level measures.** These measures are implemented during model training to guide it towards fairer results.

- **Fairness-optimised training.** Algorithms can be supplemented with fairness-related restrictions, such as a penalty function that prevents the model from using sensitive features or promotes equal results between different groups.
- **Model fine-tuning.** The pre-trained model can be fine-tuned using a high-quality and balanced dataset to 're-educate' unwanted inferences and teach fairer behaviour.
- **Adversarial debiasing.** A technique where one model tries to make a prediction and the other tries to guess a sensitive trait based on the prediction. The first model is trained so that the second fails in its task [114].
- **Prompt engineering.** The behaviour of the model can also be guided by carefully crafted instructions, or prompts. It is possible to reduce discriminatory behaviour by adding reminders of the importance of fairness to the query or asking the model to analyse potential bias before responding [115].
- **Feature steering.** This is a technique in which the internal neural network activations of the

model are directly altered. By identifying specific characteristics related to bias within the model, their effects can be suppressed or altered, thereby reducing unwanted behaviour [116].

**Output related measures.** These measures are implemented after the model's prediction in order to correct its output.

- **Calibration of decision thresholds.** The decision thresholds of the model (e.g. from which score the application is considered approved) can be calibrated separately for different groups in order to achieve the desired measure of fairness. Tools such as Fairlearn offer this functionality [117].

## 5.4 Implementation examples of bias mitigation measures

Finding successful, real-life examples of the mitigation of AI bias is more difficult than finding failures. This is because failures are scandals and therefore newsworthy. A system that works in an unnoticed, expected and fair manner is not as interesting to the public.

Thus, this section focuses not so much on the success stories of real systems in production as on the demonstration of technical methods and approaches. These examples come mainly from academic studies and tutorials on toolkits and illustrate how bias can be mitigated at the technical level.

**Example. Data re-weighing in credit scoring (AIF360).**

The tutorial for the AIF360 toolkit illustrates the principle of data-level intervention using the example of age bias in the German credit dataset.

- **The problem.** The untrained model showed a clear preference for the older age group.
- **Method applied.** A preprocessing technique named re-weighing was used, which gives greater weight to data points of under-represented groups during training.
- **The result.** In the demonstration, the difference between the groups in positive loan decisions went down to zero. This shows how the pre-processing of data can reduce bias under ideal conditions [113].

**Example. Optimisation of decision thresholds for loan decisions (Fairlearn).**

Microsoft's Fairlearn tutorial demonstrates a post-processing technique that corrects the decisions of an already-trained model.

- **The problem.** The model for loan decisions showed gender bias.
- **Method applied.** The ThresholdOptimizer algorithm was used, which finds an optimal decision-making threshold for each demographic group (e.g. the score from which the loan is approved) to ensure that the fairness metrics are met.
- **The result.** This method significantly reduced the gap between groups, leading to a slight decrease in the overall accuracy of the model, which is a typical compromise between fairness and accuracy [117].

**Example. Reducing discrimination by prompting (Anthropic).**

In the case of LLMs, an effective measure is to guide the behaviour of the LLM through carefully crafted prompts.

- **The context.** It was found that the language model of the Claude family systematically made different decisions in high-risk scenarios, depending on the demographic characteristics mentioned in the query.
- **Measures applied.** Before issuing the actual task, special prompts were added to the model input, for example, asking the model to analyse the potential bias before answering, and to focus only on the candidate's qualifications.
- **The result.** Such prompt-based interventions significantly reduced the model's tendency towards discriminatory decisions, while maintaining the model's overall judgment in other respects [115].

**Example. Correction of social bias by feature steering (Anthropic).**

This is a technically more complex approach that directly interferes with the internal functioning of the model.

- **The context.** In LLMs, specific patterns (indicators,features) can be identified in neural network activations associated with certain concepts, including social bias.
- **Measures applied.** The feature steering technique was used. When a model generates a response, these neural network activations associated with bias are modified in real time.
- **The result.** Moderate intervention was able to reduce social bias in model responses without significantly impairing the overall performance of the model in other tasks [116].

# Bibliography

[1] *Euroopa Parlamendi ja Nõukogu määrus (EL) 2024/1689, 13. juuni 2024, millega nähakse ette tehisintellekti käsitlevad ühtlustatud õigusnormid ning muudetakse määruseid (EÜ) nr 300/2008, (EL) nr 167/2013, (EL) nr 168/2013, (EL) 2018/858, (EL) 2018/1139 ja (EL) 2019/2144 ning direktiive 2014/90/EL, (EL) 2016/797 ja (EL) 2020/1828 (Tehisintellekti määrus)*. https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=OJ:L_202401689. Euroopa Liidu Teataja L 2024/1689, 12.7.2024. Kasutatud: 2025-06-03. 2024.

[2] K. Abendroth Dias et al. *Generative AI Outlook Report – Exploring the Intersection of Technology, Society and Policy*. Ed. by E. Navajas Cawood et al. Publications Office of the European Union, Luxembourg. EUR 40337, JRC142598. 2025. DOI: 10.2760/1109679. URL: https://publications.jrc.ec.europa.eu/repository/handle/JRC142598.

[3] Euroopa Parlament ja nõukogu. *Euroopa Parlamendi ja nõukogu määrus (EL) 2024/903, 13. märts 2024, koostalitlusvõime tugevdamise kohta avalikus sektoris kogu liidus (Interoperable Europe Act)*. Euroopa Liidu Teataja L 2024/903, 11.04.2024. 2024. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=OJ:L_202400903.

[4] European Commission. *AI Continent Action Plan*. https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan. COM(2025) 165 final. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels, Apr. 2025.

[5] European Commission. *AI Continent Action Plan: Accelerating trustworthy AI in key sectors including public services*. European Commission – Digital Strategy Library. COM(2025)165 final; Accessed: 2025-07-09. 2025. URL: https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan.

[6] European Commission. *Commission sets course for Europe's AI leadership with ambitious AI Continent Action Plan*. European Commission – Digital Strategy. Accessed: 2025-07-09. Apr. 2025. URL: https://digital-strategy.ec.europa.eu/en/news/commission-sets-course-europes-ai-leadership-ambitious-ai-continent-action-plan.

[7] Stuart Russell, Karine Perset, and Marko Grobelnik. *Updates to the OECD's definition of an AI system explained*. Nov. 29, 2023. URL: https://oecd.ai/en/wonk/ai-system-definition-update.

[8] Harini Suresh and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2021.

[9] Joanna Redden et al. *Automating Public Services: Learning from Cancelled Systems*. Supported publication by the Data Justice Lab. Sept. 2022. URL: https://www.carnegieuktrust.org.uk/publications/automating-public-services-learning-from-cancelled-systems/.

[10] UNIO. *Artificial intelligence: 2020 A-level grades in the UK as an example of the challenges and risks*. 2020. URL: https://officialblogofunio.com/2020/10/26/artificial-intelligence-2020-a-level-grades-in-the-uk-as-an-example-of-the-challenges-and-risks/.

[11] *Risks and controls for artificial intelligence and machine learning systems*. Tech. rep. Riigi Infosüsteemi Amet, Cybernetica AS, 2024. URL: https://www.ria.ee/en/cyber-security/national-coordination-center-ncc-ee/cybersecurity-future-technologies.

[12] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[13] "The Constitution of the Republic of Estonia". In: (2020). URL: https://www.riigiteataja.ee/en/eli/530122020003/consolide.

[14] *Euroopa Liidu lepingu ja Euroopa Liidu toimimise lepingu konsolideeritud versioonid Euroopa Liidu lepingu konsolideeritud versioon Euroopa Liidu toimimise lepingu konsolideeritud versioon Protokollid Euroopa Liidu toimimise lepingu lisad Deklaratsioonid, mis on lisatud 13. detsembril 2007 alla kirjutatud Lissaboni lepingu vastu võtnud valitsustevahelise konverentsi lõppaktile.* ELT C 202, 7.6.2016, p. 1–388. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/?uri=celex%3A12016ME%2FTXT.

[15] *Euroopa Liidu toimimise lepingu (ELi toimimise leping) konsolideeritud versioon*. C 326/49. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/PDF/?uri=CELEX:12012E/TXT.

[16] European Union. *Summary: Treaty on the Functioning of the European Union*. URL: https://eur-lex.europa.eu/ET/legal-content/summary/treaty-on-the-functioning-of-the-european-union.html.

[17] Equality and UK Human Rights Commission. *Met Police's use of facial recognition tech must comply with human rights law, says regulator*. Permission granted to intervene in judicial review regarding Met Police live facial recognition policy. Aug. 2025. URL: https://www.equalityhumanrights.com/met-polices-use-facial-recognition-tech-must-comply-human-rights-law-says-regulator.

[18] Europol Innovation Lab. *AI Bias in Law Enforcement: A Practical Guide*. Observatory Report from the Europol Innovation Lab. Luxembourg, 2025. DOI: 10.2813/8081090. URL: https://www.europol.europa.eu/publications-events/publications/ai-bias-in-law-enforcement.

[19] USA District Court for the Northern District of California. *Order Granting Preliminary Collective Certification – Mobley v. Workday, Inc.* Case No. 23-cv-00770-RFL. May 2025. URL: https://www.govinfo.gov/content/pkg/USCOURTS-cand-3_23-cv-00770/pdf/USCOURTS-cand-3_23-cv-00770-1.pdf.

[20] Various legal & news outlets. *Mobley v. Workday, Inc. – AI hiring discrimination litigation (claims proceed)*. Federal case alleging disparate impact by Workday's applicant-screening AI; claims allowed to proceed and conditional,certification steps in 2025. June 2025. URL: https://www.jdsupra.com/legalnews/ai-bias-lawsuit-against-workday-reaches-5119500/.

[21] Arshon Harper. *Complaint in Harper v. Sirius XM Radio, LLC*. Case No. 2:25-cv-12403-TGB-APP, U.S. District Court for the Eastern District of Michigan. Aug. 2025. URL: https://www.fisherphillips.com/a/web/9MgBFhRPkToes9HKGYtqKF/aAkZr3/harper-v-sirius-xm-radio-ed-mi-225cv12403.pdf.

[22] Fisher Phillips. *Another Employer Faces AI Hiring Bias Lawsuit: 10 Actions You Can Take to Prevent AI Litigation*. Complaint filed 8/2025 alleging AI hiring tool produced racially biased results; case at complaint stage. Aug. 2025. URL: https://www.fisherphillips.com/en/news-insights/another-employer-faces-ai-hiring-bias-lawsuit.html.

[23] Equality and Human Rights Commission. *Uber Eats courier wins payout with help of equality watchdog, after facing problematic AI checks*. Mar. 2024. URL: https://www.equalityhumanrights.com/media-centre/news/uber-eats-courier-wins-payout-help-equality-watchdog-after-facing-problematic-ai.

[24] Ülle Madise et al., eds. *Eesti Vabariigi Põhiseadus – Kommenteeritud väljaanne*. https://pohiseadus.ee. Accessed: 2025-06-03. 2020.

[25] Ülle Madise et al., eds. *Eesti Vabariigi Põhiseadus – Kommenteeritud väljaanne: § 12*. https://pohiseadus.ee/sisu/3483/paragrahv_12. Accessed: 2025-06-03. 2020.

[26] *Euroopa Liidu põhiõiguste harta*. 2012/C 326/02. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:12012P/TXT.

[27] Publications Office of the European Union. *Summary: Charter of Fundamental Rights of the European Union*. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/?uri=legissum:l33501.

[28] *Nõukogu direktiiv 2000/43/EÜ, 29. juuni 2000, millega rakendatakse võrdse kohtlemise põhimõte sõltumata isikute rassilisest või etnilisest päritolust*. EÜT L 180, 19/07/2000, p. 22–26. URL: http://data.europa.eu/eli/dir/2000/43/oj.

[29] *Nõukogu direktiiv 2000/78/EÜ, 27. november 2000, millega kehtestatakse üldine raamistik võrdseks kohtlemiseks töö saamisel ja kutsealale pääsemisel*. EÜT L 303, 2.12.2000, p. 16–22. URL: http://data.europa.eu/eli/dir/2000/78/oj.

[30] Ühinenud Rahvaste Organisatsiooni Peaassamblee. *Inimõiguste ülddeklaratsioon*. ÜRO peaassamblee resolutsioon 217 A. Dec. 1948. URL: https://inimoigusedeestis.ee/hartadpaktid-deklaratsioonid-kohtud-jne/inimoiguste-ulddeklaratsioon/.

[31] ÜKP fondide võrdõiguslikkuse kompetentsikeskus. *Inimõiguste ülddeklaratsioon*. URL: https://kompetentsikeskus.sm.ee/et/vordsed-voimalused/vordne-kohtlemine/oigusaktid/uro/inimoiguste-ulddeklaratsioon.

[32] Ühinenud Rahvaste Organisatsioon. *Konventsioon naiste diskrimineerimise kõigi vormide likvideerimise kohta*. URL: https://www.riigiteataja.ee/akt/23988.

[33] Ühinenud Rahvaste Organisatsioon. *Rahvusvaheline konventsioon rassilise diskrimineerimise kõigi vormide likvideerimise kohta*. URL: https://www.riigiteataja.ee/akt/23980.

[34] Ühinenud Rahvaste Organisatsioon. *Kodaniku- ja poliitiliste õiguste rahvusvaheline pakt. Mitteametlik tõlge*. URL: https://www.riigiteataja.ee/akt/23982.

[35] Ühinenud Rahvaste Organisatsioon. *Majanduslike, sotsiaalsete ja kultuurialaste õiguste rahvusvaheline pakt. Mitteametlik tõlge*. URL: https://www.riigiteataja.ee/akt/23981.

[36] Council of Europe. *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*. Council of Europe Treaty Series 225. Vilnius: Council of Europe, Sept. 2024. URL: https://rm.coe.int/1680afae3c.

[37] Ceyhun Necati Pehlivan, Nikolaus Forgó, and Peggy Valcke, eds. *The EU Artificial Intelligence (AI) Act: A Commentary*. Netherlands: Kluwer Law International, 2024. ISBN: 9789403532271. URL: https://law-store.wolterskluwer.com/s/product/the-eu-artificial-intelligence-ai-act-a-commentary/01tPg000007gkK9IAI.

[38] CEDPO AI and Data Working Group. *Fundamental Rights Impact Assessments: What are they? How do they work?* https://cedpo.eu/wp-content/uploads/CEDPO-micro-insight-paper-fundamental-rights-impact-assessments.pdf. Micro-Insights Series; Authors: Thomas Ajoodha and Jared Browne; published [10/1/2025]. Jan. 2025.

[39] CEDPO. *CEDPO Micro Insight Paper: Fundamental Rights Impact Assessments*. 2025. URL: https://cedpo.eu/wp-content/uploads/CEDPO-micro-insight-paper-fundamental-rights-impact-assessments.pdf (visited on 07/08/2025).

[40] ISO/IEC. *ISO/IEC 42005:2025 Artificial Intelligence — Management system for artificial intelligence — Requirements and guidance*. 2025. URL: https://www.iso.org/standard/42005 (visited on 07/08/2025).

[41] Council of the European Union. *Outcomes of the discussions on simplification activities in the digital field*. Document ST-9383-2025-INIT. https://data.consilium.europa.eu/doc/document/ST-9383-2025-INIT/en/pdf. June 2025.

[42] Commission Nationale de l'Informatique et des Libertés (CNIL). *Artificial Intelligence and Public Services: CNIL Publishes Results of Its Sandbox*. 2025. URL: https://www.cnil.fr/en/artificial-intelligence-and-public-services-cnil-publishes-results-its-sandbox (visited on 07/08/2025).

[43] *Tehisintellekti tegevuskava 2024-2026*. URL: https://www.kratid.ee/_files/ugd/7df26f_21000a2dd36c4a66a30eea97563370a3.pdf.

[44] *Andmete ja tehisintellekti valge raamat 2024-2030*. URL: https://www.kratid.ee/_files/ugd/7df26f_9a6d060409214b9da2774e5b5eabf717.pdf.

[45] *Eesti digiühiskonna arengukava 2030*. URL: https://www.mkm.ee/digiriik-ja-uhenduvus/digiuhiskonna-arengukava-2030.

[46] High-Level Expert Group on Artificial Intelligence. *OECD AI Principles overview*. 2019. URL: https://oecd.ai/en/ai-principles.

[47] *Euroopa deklaratsioon digiõiguste ja -põhimõtete kohta digikümnendiks*. OJ C 23, 23.1.2023, p. 1–7. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32023C0123(01).

[48] European Parliament: Directorate-General for External Policies of the Union and A. Ünver. *Artificial intelligence (AI) and human rights – Using AI as a weapon of repression and its impact on human rights – In-depth analysis*. Publications Office of the European Union, 2024. URL: https://data.europa.eu/doi/10.2861/52162.

[49] European Parliament and Council of the European Union. *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on European Critical Raw Materials*. EUR-Lex. Accessed: 2025-07-09. 2022. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32022R2065.

[50] *CCommission launches public consultation and call for evidence on the Apply AI Strategy*. European Commission – Digital Strategy. Apr. 2025. URL: https://digital-strategy.ec.europa.eu/en/consultations/commission-launches-public-consultation-and-call-evidence-apply-ai-strategy.

[51] *Commission seeks feedback on the future Strategy for Artificial Intelligence in Science*. European Commission – Research & Innovation. Apr. 2025. URL: https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/commission-seeks-feedback-future-strategy-artificial-intelligence-science-2025-04-10_en.

[52] *AI Continent – new cloud and AI development act*. European Commission. 2025. URL: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14628-AI-Continent-new-cloud-and-AI-development-act_en.

[53] Arcangelo Leone de Castris. *AI governance around the world: European Union*. Report (version v3). Published August 8, 2025; accessed 2025-09-23. Aug. 2025. DOI: 10.5281/zenodo.17054419. URL: https://zenodo.org/records/17054419.

[54] Euroopa Liit. "Euroopa Parlamendi ja nõukogu määrus (EL) 2016/679, 27. aprill 2016, füüsiliste isikute kaitse kohta isikuandmete töötlemisel ja selliste andmete vaba liikumise ning direktiivi 95/46/EÜ kehtetuks tunnistamise kohta (isikuandmete kaitse üldmäärus) (EMPs kohaldatav tekst)". In: *Euroopa Liidu TEataja L119* 59 (May 4, 2016), pp. 1–88.

[55] *Euroopa Parlamendi ja nõukogu direktiiv (EL) 2016/680, 27. aprill 2016, mis käsitleb füüsiliste isikute kaitset seoses pädevates asutustes isikuandmete töötlemisega süütegude tõkestamise, uurimise, avastamise ja nende eest vastutusele võtmise või kriminaalkaristuste täitmisele pööramise eesmärgil ning selliste andmete vaba liikumist ning millega tunnistatakse kehtetuks nõukogu raamotsus 2008/977/JSK*. ELT L 119, 4.5.2016, p. 89–131. URL: http://data.europa.eu/eli/dir/2016/680/oj.

[56] *Euroopa Parlamendi ja nõukogu määrus (EL) 2018/1725, 23. oktoober 2018, mis käsitleb füüsiliste isikute kaitset isikuandmete töötlemisel liidu institutsioonides, organites ja asutustes ning isikuandmete vaba liikumist, ning millega tunnistatakse kehtetuks määrus (EÜ) nr 45/2001 ja otsus nr 1247/2002/EÜ (EMPs kohaldatav tekst.)* ELT L 295, 21.11.2018, p. 39–98. URL: http://data.europa.eu/eli/reg/2018/1725/oj.

[57] European Data Protection Board. *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*. Tech. rep. Adopted on 17 December 2024. European Data Protection Board, Dec. 2024. URL: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf.

[58] Kris Shrishak. *AI-Complex Algorithms and Effective Data Protection Supervision - Bias Evaluation*. Tech. rep. European Data Protection Board, Support Pool of Experts Programme, Jan. 2025. URL: https://www.edpb.europa.eu/system/files/2025-01/d1-ai-bias-evaluation_en.pdf.

[59] P. Dewitte. "AI Meets the GDPR: Navigating the Impact of Data Protection on AI Systems". In: *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Ed. by N. A. Smuha. Cambridge Law Handbooks. Cambridge University Press, 2025, pp. 133–157.

[60] *Isikuandmete kaitse seadus*. RT I, 04.01.2019, 11. URL: https://www.riigiteataja.ee/akt/104012019011?leiaKehtiv.

[61] German Federal Constitutional Court. *Judgment of 16 February 2023 – Automated data analysis*. Feb. 2023. URL: https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/2023/02/rs20230216_1bvr154719en.html.

[62] German Federal Constitutional Court (Bundesverfassungsgericht). *Legislation in Hesse and Hamburg regarding automated data analysis for the prevention of criminal acts is unconstitutional*. Press release (English) summarising 1 BvR 1547/19 and 1 BvR 2634/20 judgments. Feb. 2023. URL: https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2023/bvg23-018.html?nn=148454.

[63] Euroopa Parlamendi mõttekoda. *Algorithmic discrimination under the AI Act and the GDPR*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2025)769509. 2025.

[64] Giovanni Sartor, Francesca Lagioia, et al. “The impact of the General Data Protection Regulation (GDPR) on artificial intelligence”. In: (2020).

[65] European Data Protection Board. *EDPB Opinion 2024/28 on certain data protection aspects related to the processing of personal data in the context of AI models*. 2024. URL: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf.

[66] *Avaliku teabe seadus*. RT I 2000, 92, 597. URL: https://www.riigiteataja.ee/akt/130122024005?leiaKehtiv.

[67] American Civil Liberties Union. *ACLU v. Clearview AI*. 2022. URL: https://www.aclu.org/cases/aclu-v-clearview-ai.

[68] American Civil Liberties Union and ACLU of Illinois. *Big Win, Settlement Ensures Clearview AI Complies With Groundbreaking Illinois Biometric Privacy Law*. May 2022. URL: https://www.aclu.org/press-releases/big-win-settlement-ensures-clearview-ai-complies-with-groundbreaking-illinois.

[69] Jane Bambauer. “Cambridge Analytica and the Meaning of Privacy Harm”. In: (). URL: https://pep.gmu.edu/wp-content/uploads/sites/28/2019/01/Bambauer_PEP_White_Paper_Cambridge_Analytica.pdf.

[70] *Facebook hit with $645,000 fine in UK over Cambridge Analytica scandal*. CNET. Oct. 2018. URL: https://www.cnet.com/tech/tech-industry/uk-information-commissioners-office-hits-facebook-with-645000-fine/.

[71] Michael Veale and Frederik Zuiderveen Borgesius. “AI Meets the GDPR”. In: *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Ed. by Markus Dubber, Frank Pasquale, and Sunit Das. Cambridge University Press, 2022, pp. 576–596. URL: https://www.cambridge.org/core/books/cambridge-handbook-of-the-law-ethics-and-policy-of-artificial-intelligence/ai-meets-the-gdpr/94476F95CE264B80C00B46BA8506F474.

[72] European Parliament and Council of the European Union. *Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 concerning measures for a high common level of cybersecurity across the Union, and amending Regulation (EU) 2019/881*. EUR-Lex. Dec. 2022. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32022L2555.

[73] Hajo Michael Holtz and Jonas Ledendal. *AI Data Governance – Overlaps Between the AI Act and the GDPR*. Preprint, Uppsala University / Lund University. Forthcoming in "Law, Innovation and Technology" (Spring 2026). Sept. 2025. URL: https://uu.diva-portal.org/smash/get/diva2:1996843/FULLTEXT01.pdf.

[74] GOV.UK. *Auditing algorithms: the existing landscape, role of regulators and future outlook*. Digital Regulation Cooperation Forum. Findings from the DRCF Algorithmic Processing workstream – Spring 2022. Sept. 2022. URL: https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook.

[75] Euroopa Parlament ja Euroopa Liidu Nõukogu. *Määrus (EL) 2023/988 Euroopa Parlamendi ja nõukogu 10. mai 2023. aasta määrus üldise tooteohutuse kohta, millega muudetakse Euroopa Parlamendi ja nõukogu määrust (EL) nr 1025/2012 ja direktiivi (EL) 2020/1828 ning tühistatakse Euroopa Parlamendi ja nõukogu direktiiv 2001/95/EÜ ja nõukogu direktiiv 87/357/EMÜ*. EUR-Lex. 2023. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32023R0988.

[76] *Toote nõuetele vastavuse seadus*. URL: https://www.riigiteataja.ee/akt/111032025003?leiaKehtiv.

[77] Euroopa Parlament ja Euroopa Liidu Nõukogu. *Määrus (EL) 2024/2847, mis käsitleb digielemente sisaldavate toodete küberturvalisuse horisontaalseid nõudeid ja millega muudetakse määrusi (EL) nr 168/2013 ja (EL) 2019/1020 ning direktiivi (EL) 2020/1828*. EUR-Lex. 2024. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=OJ:L_202402847.

[78] *Küberturvalisuse seadus*. RT I, 22.05.2018, 1. URL: https://www.riigiteataja.ee/akt/121062024015?leiaKehtiv.

[79] International Organization for Standardization. *Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making*. ISO/IEC TR 24027:2021, Edition 1. Nov. 2021. URL: https://www.iso.org/standard/77607.html.

[80] IEEE Standards Association. *IEEE P7003 - Algorithmic Bias Considerations*. Tech. rep. Published: 2025-01-24. 2023. DOI: 10.1109/IEEESTD.2025.10851955. URL: https://ieeexplore.ieee.org/servlet/opac?punumber=10851953.

[81] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. NIST AI 100-1. Available at: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf. U.S. Department of Commerce, Jan. 2023. URL: https://doi.org/10.6028/NIST.AI.100-1.

[82] R. Schwartz et al. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. Available online at https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence. Special Publication (NIST SP) - 1270, National Institute of Standards and Technology. 2022. URL: https://doi.org/10.6028/NIST.SP.1270.

[83] National Institute of Standards and Laurie E. Locascio Technology Gina M. Raimondo. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1, U.S. Department of Commerce. July 2024. URL: https://doi.org/10.6028/NIST.AI.600-1.

[84] Riigi Infosüsteemi Amet. *Eesti Infoturbestandard. Riskihaldusjuhend*. 2024. URL: https://eits.ria.ee/et/abimaterjalid/riskihaldusjuhend.

[85] *Risk management — Guidelines*. en. Standard ISO 31000:2018. International Organization for Standardization, 2018. URL: https://www.iso.org/standard/65694.html.

[86] *Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy*. en. Standard NIST SP 800-37 Rev. 2. US National Institute of Standards and Technology, 2018. URL: https://csrc.nist.gov/pubs/sp/800/37/r2/final.

[87] *Information technology — Information security, cybersecurity and privacy protection — Guidance on managing information security risks*. en. Standard ISO/IEC 27005:2022. International Organization for Standardization, 2022. URL: https://www.iso.org/standard/80585.html.

[88] *NIST Cybersecurity Framework 1.1*. en. Standard NIST CSF v. 1.1. US National Institute of Standards and Technology, 2018. URL: https://www.nist.gov/cyberframework/framework.

[89] *Information technology — Artificial intelligence — Guidance on risk management*. en. Standard ISO/IEC 23984:2023. International Organization for Standardization, 2023. URL: https://www.iso.org/standard/77304.html.

[90] Riigi Infosüsteemi Amet. *Eesti infoturbestandard (E-ITS)*. 2023. URL: https://eits.ria.ee/.

[91] Heidi Ledford. "Millions affected by racial bias in health-care algorithm". In: *Nature* 574.31 (2019), p. 2.

[92] Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women". In: *reuters.com* (2018). URL: https://www.reuters.com/article/uk-amazon-com-jobs-automation-insight-idUKKCN1MK08G.

[93] Gerwin van Schie, Laura Candidatu, and Diletta Huyskes. "Motherhood in the Datafied Welfare State:: Investigating the Gendered and Racialized Enactment of Citizenship in Dutch Algorithmic Governance". In: May 2025, pp. 209–226. ISBN: 9789048562718. DOI: 10.2307/jj.28874939.18.

[94] Yuxuan Dai and Zhaohui Wang. "Predictive Policing and Algorithmic Fairness". In: *Synthese* 201.3 (2023), pp. 1–29.

[95] Weihao Xuan et al. *MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation*. 2025. arXiv: 2503.10497 [cs.CL]. URL: https://arxiv.org/abs/2503.10497.

[96] Wenxuan Wang et al. "All Languages Matter: On the Multilingual Safety of LLMs". In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 5865–5877. DOI: 10.18653/v1/2024.findings-acl.349. URL: https://aclanthology.org/2024.findings-acl.349/.

[97] Nikhil Sharma, Kenton Murray, and Ziang Xiao. "Faux Polyglot: A Study on Information Disparity in Multilingual Large Language Models". In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 8090–8107. ISBN: 979-8-89176-189-6. DOI: 10.18653/v1/2025.naacl-long.411. URL: https://aclanthology.org/2025.naacl-long.411/.

[98] Julia Unwin. *Kindness, Emotions and Human Relationships: The Blind Spot in Public Policy*. Report. Carnegie UK Trust, 2018. URL: https://www.carnegieuktrust.org.uk/publications/kindness-emotions-and-human-relationships-the-blind-spot-in-public-policy/.

[99] Karen Yeung. *A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility Within a Human Rights Framework*. MSI-AUT Report MSI-AUT (2018) 05. Posted: 10 Dec 2018. The University of Birmingham, Nov. 2018, p. 94. URL: https://ssrn.com/abstract=3286027.

[100] Auli Viidalepp. "The expected AI as a sociocultural construct and its impact on the discourse on technology". PhD thesis. University of Tartu, 2023.

[101] Pierre Le Jeune et al. *RealHarm: A Collection of Real-World Language Model Application Failures*. 2025. arXiv: 2504.10277 [cs.CY]. URL: https://arxiv.org/abs/2504.10277.

[102] EEOC. *Press Release: iTutorGroup to Pay $365,000 to Settle EEOC Discriminatory Hiring Suit.* 2023. URL: https://www.eeoc.gov/newsroom/itutorgroup-pay-365000-settle-eeoc-discriminatory-hiring-suit.

[103] David Weinberger. *Playing with AI Fairness*. https://pair-code.github.io/what-if-tool/ai-fairness.html. Google PAIR blog post. 2018.

[104] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *arXiv preprint arXiv:1908.09635* (2019). URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11880872/.

[105] S. Ghosh, S. Sanyal, and S. Mukhopadhyay. "A Classification Framework for AI Bias Impacts and Effective Mitigation Strategies". In: *Social Sciences* 6.1 (2024), p. 3. DOI: 10.3390/socsci6010003. URL: https://www.mdpi.com/2413-4155/6/1/3.

[106] S. Ghosh, S. Sanyal, and S. Mukhopadhyay. *Artificial Intelligence (AI) Bias Impacts Classification Framework for Effective Mitigation*. 2023. URL: https://pure.jgu.edu.in/id/eprint/6745/1/Artificial%5C%20intelligence%5C%20%5C%28AI%5C%29%5C%20bias%5C%20impacts%5C%20classification%5C%20framework%5C%20for%5C%20effective%5C%20mitigation.pdf.

[107] X. H. Phan, T. H. Nguyen, H. M. Le, et al. "Fairness-aware Artificial Intelligence: A Systematic Review". In: *Neural Networks* (2024). DOI: 10.1016/j.neunet.2024.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0893395224002667.

[108] Isabel Barberá and Murielle Popa-Fabre. *Expert Report on Privacy and Data Protection Risks in Large Language Models (LLMs)*. Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Convention 108. Presented at the 48th Plenary Meeting, 17 June 2025. 2025. URL: https://media.licdn.com/dms/document/media/v2/D4E1FAQFUWjwoEARZpQ/feedshare-document-pdf-analyzed/B4EZdAFWr0HgAY-/0/1749126851206.

[109] Daniel Bone et al. "Applying machine learning to facilitate autism diagnostics: pitfalls and promises." eng. In: *J Autism Dev Disord* 45.5 (May 2015), pp. 1121–1136. ISSN: 1573-3432 (Electronic); 0162-3257 (Print); 0162-3257 (Linking). DOI: 10.1007/s10803-014-2268-6.

[110] Michael Roberts et al. "Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans". In: *Nature Machine Intelligence* 3 (Mar. 2021).

[111] Gilles Vandewiele et al. "Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling". In: *Artificial Intelligence in Medicine* 111 (Jan. 2021), p. 101987. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2020.101987. URL: http://dx.doi.org/10.1016/j.artmed.2020.101987.

[112] Ching-Hua Chuan et al. "EXplainable Artificial Intelligence (XAI) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives". In: *Telematics and Informatics* 91 (2024), p. 102135. ISSN: 0736-5853. DOI: https://doi.org/10.1016/j.tele.2024.102135. URL: https://www.sciencedirect.com/science/article/pii/S073658532400039X.

[113] Rachel K. E. Bellamy et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Oct. 2018. URL: https://arxiv.org/abs/1810.01943.

[114] Jenny Yang et al. “An adversarial training framework for mitigating algorithmic biases in clinical machine learning”. In: *NPJ Digital Medicine* 6.1 (2023). Published online 2023-03-29, p. 55. DOI: 10.1038/s41746-023-00805-y. URL: https://www.nature.com/articles/s41746-023-00805-y.

[115] Alex Tamkin et al. *Evaluating and Mitigating Discrimination in Language Model Decisions*. 2023. arXiv: 2312.03689 [cs.CL]. URL: https://arxiv.org/abs/2312.03689.

[116] Esin Durmus et al. *Evaluating Feature Steering: A Case Study in Mitigating Social Biases*. Oct. 25, 2024. URL: https://anthropic.com/research/evaluating-feature-steering.

[117] Hilde Weerts et al. “Fairlearn: Assessing and Improving Fairness of AI Systems”. In: *Journal of Machine Learning Research* 24 (2023). URL: http://jmlr.org/papers/v24/23-0389.html.

[118] *ISO/IEC 11179-1:2023. Information technology — Metadata registries (MDR)*. URL: https://www.iso.org/standard/78914.html.

[119] *ISO/IEC/IEEE 15288:2023.Systems and software engineering — System life cycle processes*. URL: https://www.iso.org/standard/81702.html.

[120] *ISO/IEC/IEEE 15939:2017. Systems and software engineering — Measurement process*. URL: https://www.iso.org/standard/71197.html.

[121] *ISO/IEC 19501:2005. Information technology — Open Distributed Processing — Unified Modeling Language (UML) Version 1.4.2*. URL: https://www.iso.org/standard/32620.html.

[122] *ISO/IEC 25002:2024. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model overview and usage*. URL: https://www.iso.org/standard/78175.html.

[123] *ISO/IEC 25012:2008. Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model*. URL: https://www.iso.org/standard/35736.html.

[124] *ISO/IEC 25024:2015. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality*. URL: https://www.iso.org/standard/35749.html.

[125] *ISO/IEC 25023:2016. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of system and software product quality*. URL: https://www.iso.org/standard/35747.html.

# Appendix A Standards indirectly related to mitigation of bias in AI systems

International standards directly related to bias in AI systems are set out in 3.3. Here we highlight the standards that indirectly help to mitigate bias in AI systems.

**Table 3. Standards indirectly related to mitigation of AI system bias**

| Number of the standard | Name of the standard | Explanations |
|---|---|---|
| ISO/IEC 11179-1:2023 [118] | Metadata registries (MDR) (multiple parts) | The Standard provides a basis for understanding metadata and MDRs by addressing metadata management and data descriptions. |
| ISO/IEC 15288:2023 [119] | Systems and software engineering – **System life cycle processes** | The standard describes system life cycle processes that support the development, procurement and cooperation of systems between different parties. |
| ISO/IEC 15939:2017 [120] | Systems and software engineering — **Measurement process** | The standard describes the measurement process system, helping to define, implement and refine metrics according to specific needs. |
| ISO/IEC 19501:2005 [121] | Information technology – Open Distributed Processing – Unified Modeling Language (UML) | The standard describes the UML language, which allows visual modelling and documentation of software systems in a uniform and standardised way. |
| ISO/IEC 25002:2024 [122] | Systems and software Quality Requirements and Evaluation (SQuaRE) – **Quality model overview and usage** | The standard defines the framework for quality models, describing their structure, meaning and relationships with measurement, requirements and evaluation. |
| ISO/IEC 25012:2008 [123] | Software product Quality Requirements and Evaluation (SQuaRE) – **Data quality model** | The standard describes a general data quality model that can be used to define, measure and evaluate data quality requirements for structured data. |

| Number of the standard | Name of the standard | Explanations |
|---|---|---|
| ISO/IEC 25024:2015 [124] | Systems and software Quality Requirements and Evaluation (SQuaRE) – **Measurement of data quality** | The standard defines data quality metrics that enable quantitative assessment of data quality in accordance with the characteristics set out in ISO/IEC 25012. The Standard provides guidance on how to apply these metrics throughout the data-life cycle and is designed for use in a variety of information systems and organisational roles, from developers to quality managers. |
| ISO/IEC 25023:2016 [125] | Systems and software Quality Requirements and Evaluation (SQuaRE) – **Measurement of system and software product quality** | The standard defines quantitative metrics for evaluating the quality of the system and software to be used in conjunction with ISO/IEC 25010. The standard supports the definition and evaluation of quality requirements throughout the development cycle and is suitable for quality assurance, management, delivery, procurement and maintenance. |