CYBERNETICA

Co-funded by
the European Union

# Algorithmic bias risk management

Methodology

Version 1.1

21 October 2025

D-16-584

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-16-572 v1.0 to English |
| 21.10.2025 | 1.1 | Updated translation of D-16-572 v1.1 to English |
| | | |

# Table of Contents

# 1 Introduction

## 1.1 Purpose

[Paula] Anto kommenteeris: Üldist - lingid EU seadustele vajavad ingliskeelses tekstis vahetamist ingliskeelsete dokude vastu. 122 viite hulgas on ka Eesti seadusi, seal tuleb viidata ingliskeelsele tekstile, jne. Hetkel tegemata. Paula, palun vaata üle ja asenda. Sul vast kõige lihtsam seda teha. Võid teha sama viite nime ja panna lõppu -eng (n cite: gdpr-eng). Ja siis vahetada viite tekstis ka ära eng peale.

Whereas the ALGORITHMIC BIAS RISK MANAGEMENT GUIDELINE ('the Guideline') describes the continuum of concepts related to algorithmic bias, this methodology document teaches the reader how to put the Guideline's postulates into practice.

The methodology provides detailed steps and actions that help to analyse algorithmic bias and manage risks in systems utilising AI and algorithmic decision-making.

The methodology is flexible and can be used together with risk management standards implemented by the organisation. The methodology is complemented by a workbook that can be filled in using a spreadsheet application.

## 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: the Guideline, which presents an account of the nature of AI systems and biases inherent in such systems and the possibilities of identification and mitigation of such biases; this 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 Concepts and 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 input 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 oth-

ers, 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

# 2 Using the methodology

Figure 1 shows the overall AI and algorithmic system bias risk analysis workflow. The figure also shows the relationships between workflow steps and the sections of this document.

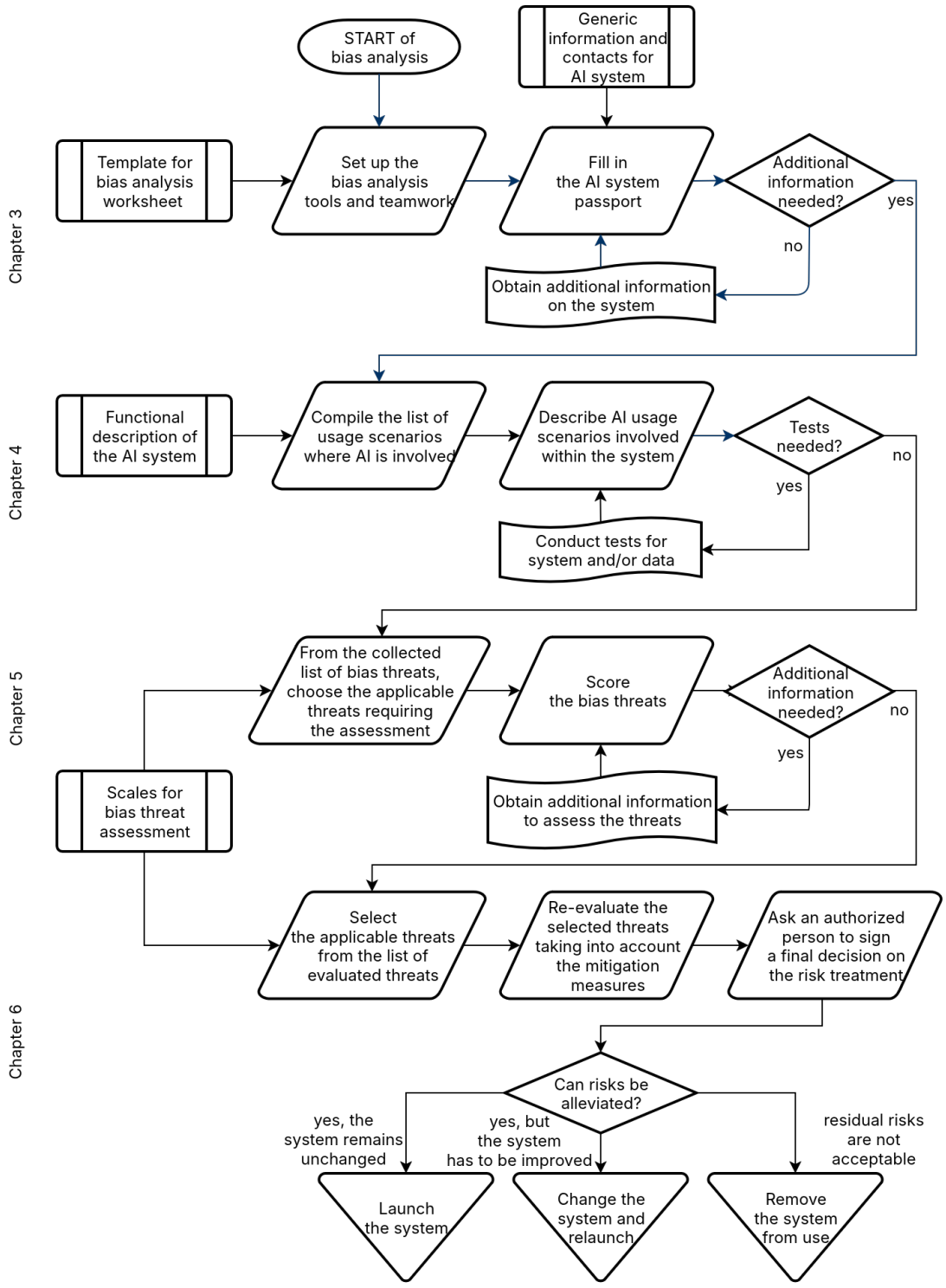


**Figure 1. Bias analysis workflow**

During the testing of the methodology, a number of questions emerged that are also likely to

be asked by other future users. Below, you will find answers to some of these frequently asked questions.

**What should I do if I have just begun the development of a system and many of its features are yet to be determined?**

Start early because changing a system still under development will be less expensive. Do not hesitate to leave some of the boxes on any of the worksheets empty, use information found in designs and plans, and change your designs and plans in accordance with the risk assessment!

**How should I use the workbook if I am considering the addition of a decision-making algorithm or AI to an existing system?**

In case no previous decisions need to be reconsidered, feel free to cover only the functions to be changed or added; this will make the work less taxing.

**How many usage scenarios should I write up?**

One to four usage scenarios should generally cover the main functions of the AI. However, your goal should be to write up one usage scenario for each significant AI model or algorithm.

**Is it necessary to consider threats in qualitative terms (e.g. percentages or number of endangered people) or can I also evaluate them qualitatively?**

Qualitative evaluation is always an option, but in that case it is recommended to involve more evaluators and decide how to compare the significance of the threats.

**What should I do if my organisation does not wish to accept residual risks?**

Everything that people do is open to risks. Risk-free life is either fundamentally impossible or prohibitively expensive. We recommend sharing experience with others on the subject of assessing practical risks.

# 3 Creation of an AI system passport

## 3.1 Who, when, why?

**Responsible person.** The owner of the AI system is responsible for filling in the AI system passport. For operational efficiency, this may be delegated to a project manager, technical expert, security or legal team either within the organisation or at a contracted development partner.

**Time.** The AI system passport may be filled in during the creation, design or development phase, at the start of risk analysis, or during an audit. The earlier the start, the less technical information is available about the AI system, but at the same time there are also more opportunities to guide the design and development of the system, establish requirements and mitigation measures.

**Purpose.** The AI system passport documents the system that is being evaluated and analysed. There may be different versions of the system and assessments made at different times. The passport will record both the current state of the system and the persons involved at the time of the assessment.

## 3.2 Preparations

**Step 1: Prepare the workbook.** Do the following.

1. Find the latest version of the AI bias analysis workbook file.
2. Create a new copy of the workbook and give the file a name associated with the analysed system.
3. Store the created file in a system document repository accessible to partners.

**Step 2: Decide on the manner of filling the workbook and assemble a working group.** Do the following.

1. Identify who knows more about the technical details, legal framework, or risks associated with the assessed AI system (e.g. data scientists, data protection expert, information security manager, quality manager, risk manager, other responsible managers).
2. Find out who needs to be included in the bias analysis. To this end, study the organisation's quality system, information security policy, data protection policy, social impact policy, risk management procedures, and general responsibility chains.
3. Decide whether you are filling in a workbook together in the form of a meeting or each one at a time that suits them. Set up cooperation tools for filling in the workbook accordingly.
4. Set a schedule of sessions and contributions in accordance with the form of cooperation (session times, deadlines for personal contributions). Share tasks by e-mail or through the work management system.

## 3.3 Filling in the workbook

**Step 1: Describe the AI system for later reference.** Fill yellow cells 1.1–1.X on the AI SYSTEM PASSPORT (1-5) worksheet. Follow the recommendations below.

1. When assigning a version, you can use a specific version number or name, year number, code repository identifier, etc. to avoid later confusion about which version of the system

was evaluated.

2. When determining the deploying organisation, consider who is the service owner, contracting authority or developer – who has the responsibility.
3. References can be added to the cell 1.x. In the case of a web service, the address of the AI system (URI/URL) or a website introducing it can be added. In the case of an app, a reference to the app store, a wiki page, a document repository, a source code repository can be used. Add lines if needed.

**Step 2: Specify the parties related to the AI system so that those responsible and those affected are clearly defined.** Fill yellow cells 2.1–2.X on the AI SYSTEM PASSPORT (1-5) worksheet. Follow the recommendations below.

1. The EU AI Act outlines the following roles: service provider, manufacturer, implementer, importer, distributor. Consult Section 2.2 of the Guideline.
2. The AI system may have several development partners. Add the name, area of responsibility, and the years when the partner was active.
3. As the main affected persons, list the main user group of the system (e.g. people handling applications at an organisation,decision makers), the target group affected by the decisions (e.g. citizen participating in administrative proceedings). If possible, add a broader social group affected by the system (e.g. health care providers, parents, minors, persons moving in the surveillance camera observation area in public spaces or buildings, users of communication systems).

**Step 3: Document the goals of the AI system so that the system's goals are clearly defined** Fill in yellow cells 3.1–3.X on the AI SYSTEM PASSPORT (1-5) worksheet. Follow the recommendations below.

1. Write down the expected functionality as a direct goal. What does the system need to do using the algorithm or AI component? What decisions does it need to support? What kind of human task does it need to carry out? See also Section 4.2 of the Guideline.
2. As indirect goals, write down why it was necessary to carry out this task. What kind of changes does the society want to achieve? Which processes does it want to make more efficient?
3. The metrics do not have to be invented by the working group, they can come from the owner of the system.

**Step 4: Write down contact information for subsequent inquiries.** Fill in yellow cells 4.1–4.7 on the AI SYSTEM PASSPORT (1-5) worksheet. Follow the recommendations below.

1. Not all contacts mentioned on the worksheet have to be documented. The aim is to simplify further work so that all contacts can be easily found.

**Step 5: Add references to technical documents so that when making assessments, everyone knows how the system is designed to work.** Fill in yellow cells 5.1–5.9 on the AI SYSTEM PASSPORT (1-5) worksheet. Follow the recommendations below.

1. Different documents can be available depending on the AI system's life cycle stage. If the system is still in the vision phase, a vision document may be available. If the system is mature, more documents will be available.
2. Document can be of any file type, format, or standard. The main point is that the content described on the worksheet is available in some format.

## 3.4 Follow-up tasks

**Step 1: Check the quality of the collected information in cooperation with the working group and partners.** Do the following.

1. If not all members of the working group were able to participate in the work, a short-term coordination round is recommended after the completion of the worksheet. The aim is to fill the gaps in the available documents and create a single information space for everyone for the next task (identification of the system's AI-based functionalities).

# 4 Description of AI usage scenarios

## 4.1 Who, when, why?

**Responsible person.** The owner of the AI system has to be aware of the usage scenarios related to AI. To clarify the system's functionality and technical details, it is reasonable to involve service owner, system analyst, architect and data scientists. To simplify future work, a legal expert and/or risk manager should be involved, either immediately in the assessment process or in follow-up activities, to validate the results.

**Time.** The **usage scenarios** comprise the main subject of investigation of this methodology. They are a very good starting point because usage scenarios and the related business cases provide the best overview of data streams and persons associated with these streams. An insufficient amount of details may be available during the vision and pre-analysis phases of the system development – in this case we recommend filling in as much of the usage scenario questionnaire as possible, leaving out the technical details of the AI component (which will be clarified later). In case the system being assessed for bias has already been implemented, more details will be available, making the threat assessment easier accordingly.

**Purpose.** Recording the main usage scenarios of AI components helps to clearly identify the persons affected by the bias present in the AI system. Additional technical details will help to estimate the extent of the bias.

## 4.2 Preparations

**Step 1:** If necessary, **familiarise yourself with AI and data terminology. Do the following.**

1. Go over the definitions of AI, machine learning models and decision support algorithms to make sure you can recognise them in a system. Sections 2.1 and 2.3 of the Guide will assist you in this.
2. Go over the definitions data-related terms (personalised data, re-identifiable data, special categories of personal data, data related to trade secrets). Section 3.2 of the guide will assist you in this.
3. If possible, share your findings and suitable support materials with the working group so that they too will be prepared for the task.

**Step 2: Familiarise yourself with sources describing the functionality of the AI system. Do the following.**

1. Familiarise yourself with the materials found in section 5 (Technical details) of the bias assessment workbook.
2. Find the functionality descriptions (usage scenario templates, business cases and schemes, scenarios and other materials). Read these thoroughly and assess whether you can identify involved parties and data elements in the information available to you. Prepare a draft of AI-related usage scenarios in your system. To reduce the bias assessment workload, group together scenarios involving similar parties or using the AI component/algorithm in a similar manner. Examples to aide in the identification of usage stories can be found in Section 4.2.2 of the Guideline.
3. Go over the materials to find information on which functions of the analysed system utilise

AI or decision support algorithms. Find descriptions of the algorithms in the referenced documentation (if it exists). See Sections 2.1 and 2.3 of the Guideline for assistance in identifying the AI components.

4. If necessary, share your findings and suitable support materials with the working group so that they too will be prepared for the task.

## 4.3 Filling in the workbook

**Step 1: Compile a list of usage scenarios making use of AI or decision support algorithms.** Fill in yellow cells 6.1–6.9 (add or remove the sections as required) on the AI USAGE SCENARIOS (6) worksheet. Follow the instructions on the worksheet and the recommendations below.

1. In case a working group is involved in the assessment, first present the usage scenarios identified during the preparations phase which involve the system's AI component or decision support algorithm. Check whether your other people familiar with the system agree with this list of scenarios involving AI. Discuss whether anything should be added, removed, or which scenarios could be grouped together to save time (same users or affected persons, same AI component or algorithm).
2. In an ideal case, the system should have 1–7 AI-related usage scenarios, but if you find more, let it be so. Just plan more time for the work.
3. After filling in Section 6 of the AI USAGE SCENARIOS (6) worksheet, copy the worksheet AI USAGE SCENARIO A (COPY ME!) as many times as needed, so that a separate worksheet is created for each scenario identified. Enumerate the sheets using sequential capital letters of the alphabet (A, B, C, etc.). You should now have worksheets AI USAGE SCENARIO A, AI USAGE SCENARIO B, AI USAGE SCENARIO C etc.

**Step 2: Provide generic descriptions for the identified AI and decision support-related usage scenarios.** Fill in the yellow cells on each of the worksheets AI USAGE SCENARIO {XYZ} starting from scenario A (A.1–A.X, B.1–B.X, etc.), following the recommendations below.

1. The initiator of the particular scenario (A.2) can be either a user (a human person) or an automatic component of the system reacting to input data (like current time, received message or camera picture). In both cases it is important to document the event (A.3) triggering the usage scenario and starting up the AI.
2. The absolute minimum to document in the context of an AI or decision supporting algorithm is the general character of the input data (A.4), the purpose of the data processing (A.5) ,and the output data (A.6).
3. For bias analysis, it is extremely important to understand how the output data of the AI or algorithm will be used. Do any of the system's users utilise this information in decision-making? If yes, then to what extent – does the user simply confirm the machine's recommendation or is there any active post-processing taking place in the user's head to reach a final decision? Or does the machine make the decision autonomously and implement it automatically (e.g. in coordination with other systems)?
4. If any particular sources of information are used in addition to these cited earlier, these should be recorded in cell (A.X).

**Step 3: To the extent possible, specify the character of the input data.** Fill the yellow cells AA.1–AA.X on the worksheet AI USAGE SCENARIO A and the respective cells (BB.1–BB.X, etc.) on its copies, following the recommendations below.

1. The identification of composition and types of data in cells AA.1–AA.4 helps to understand which data can be either directly (personalised) or indirectly (personalisable) connected to specific physical or legal persons and potentially lead to their discrimination in an automated or machine-supported decision process.
2. The way data was gathered/obtained (cell AA.5) characterises the quality and accuracy of the data. There is a big difference input data acquired from a state digital database on a digitally identified person and data acquired from a much noisier channel (text messages, audio, video or image analysis). In case the data are aggregated from several sources, the sources should be listed to enable subsequent assessment of the quality of the data and its suitability for data analysis.
3. The measures for data integrity and confidentiality (cell AA.5) are expected to support the correctness of the input data for AI or decision algorithm as well as to provide protection during the processing. Data covered by confidentiality requirements either for legal or for practical reasons should be processed using additional protection measures to avoid their leakage during the administrative process as well as an unintended discrimination outside the official usage scenarios.

**Step 4: Elaborate the technical details of the AI or algorithmic processing to the extent possible.** Fill in the yellow cells AAA.1–AAA.X on the worksheet AI USAGE SCENARIO A, then similar cells BBB.1–BBB.X on copy B and so on, following the recommendations below.

1. The first four cells (AAA.1–AAA.4) help to identify the technical implementation of the AI component. Depending on the system, these may contain the name of the system and the ML model, as well as the name of the algorithm found in a book or scientific publication. Version info (number, name, year) helps to find information on the particular component. Multiple models or names can be recorded here, especially if they are run together.
2. The interface of the algorithm or model with the system (AAA.5) helps to understand where the model or algorithm gets its input data. Make sure to also note under whose control the algorithm or model is operated (under the system operator's control at their data centre, as a cloud service via an API, or anything in between). Knowing this is essential as it helps to evaluate the risk of the model or algorithm being changed, resulting in changes in the system's behaviour, without the operator's knowledge.
3. In case there is sufficient information available on the training and test data used in the creation of the model or algorithm, this can be noted down in cell (AAA.6). NB! This information is not mandatory, as it might not be available. When complemented by the available quality indicators (AAA.7), this will later enable the assessment of cases where data quality or presence of errors in the data can affect the system's output.
4. In an ideal case, the creator of the AI component has already described in their documentation which bias reduction or control measures (cell AAA.8) have already been implemented and whether these have been tested in any way (AAA.9). It is not infeasible, however, that documentation on such measures is lacking even though test results have been provided. NB! It is important to understand the organiser of the risk analysis can also order a survey or testing of the model or algorithm in order to gain the information necessary for completing these two cells.

**Step 5: Elaborate the implementation of the result to the extent possible.** Fill in the yellow cells AAAA.1–AAAA.X on the worksheet AI USAGE SCENARIO A, then similar cells BBBB.1–BBBB.X on worsksheet B and so on all copies of the worksheet, following the recommendations below.

1. The output data composition cell (AAAA.1) must clearly state what data produced by the ma-

chine learning model or algorithm the human or machine decision will be based on. Cells AAAA.2–AAAA.4 must specify which parts of these data can help identify a person, special categories of personal data, or the organisation's trade secrets or classified information.

2. We recommend adding extra details to cell AAAA.5 compared to cell A.7 and focus on the explainability and transparency of the data. The information provided in the cell has to help the reader to understand to what extent the affected persons get information about the process leading to the specific decision made in relation to them, including the models and algorithms used and the outputs of these models and algorithm.
3. The measures ensuring the correctness of the result should be described in cell AAAA.6. If the machine learning system in use is capable of hallucinations, this cell is of extra importance. The measures for the protection of data covered by the algorithm or ML model's confidentiality requirements should also be recorded (be these related to personal data, trade secrets, or classified information).

## 4.4 Follow-up activities

**Step 1: Check the quality of the collected information in collaboration with the working group and partners.** Do the following.

1. In case the technical information available to you is insufficient but there is reason to believe that the information in question should exist (because, e.g., the system has been built), we recommend approaching the contacts listed in chapter 4 and asking them to fill in the missing cells.
2. In case the system has not been tested but testing seems viable, we recommend consulting Sections 5.2 and 5.3 of the Guideline and organising a bias test using either a black box or a white box method.
3. If not all members of the working group were able to participate in the work, we recommend holding a short coordination round after filling in the workbook. The purpose of the coordination round is to fill in any remaining gaps in the materials and to create a common information space for all participants in preparation for the next task (description and analysis of threats).

# 5 Evaluation of bias-related threats

## 5.1 Who, when, why?

**Responsible person.** The owner of the system is responsible for describing the bias risks of the AI system in cooperatation with a legal expert, a risk manager and/or a security expert. To describe the protection measures, persons familiar with the technical implementation and operation of the system should be involved.

**Time.** The best time for a bias threat analysis is at the time period when the initial technical solution is already in place but not yet completely implemented. During this period it is still cost-effective to implement changes based on the threats found. According to the software lifecycle model, the optimum arrives when either the architecture or prototype has been elaborated but the development has not yet begun. In the case of a purchasing a ready made product, the situation is different – the risk analysis should be performed after selecting the product or product candidates but before heavily investing into the product.

**Purpose.** The description and evaluation of threats will help clarify which AI bias-related risks are applicable and thus need further attention. The assessments made in this stage enable continuing the process with risk treatment.

## 5.2 Preparations

**Step 1: Familiarise yourself with threats related to AI bias.** Do the following

1. Study the examples presented in the Guideline to illustrate the concept of AI bias and the associated threats (Section 4.3).
2. If possible, study available past bias analyses (regardless of the methodology used) and the policies and documents describing the risk tolerance of your organisation.
3. If necessary, share your findings and suitable supporting materials with the working group to help them prepare for the next task.

**Step 2: Define the criteria that designate the lower boundary for threat analysis.** Do the following.

1. For the sake of the efficiency of the risk analysis, determine a reasonable threshold for threats to be assessed and treated.
2. The assessment should cover threats that could foreseeably materialise during the system's lifetime and result in real harm. Check the section CRITERIA FOR A THREAT... on the worksheet THREATS CONSIDERED (T) regarding the conditions and thresholds described there. Choose the ones suitable for your organisation and record these in the light gray cells under the heading 'Threshold chosen'.
3. NB! In case bias risks have been previously analysed in the system owner's organisation (regardless of the methodology), consider re-using the same thresholds.

**Step 3: Create scales to express the impact and likelihood scores.** Do the following.

1. To ensure uniformity in the evaluation of risks, it is necessary to establish evaluation scales which would help comparing the risks under consideration.

2. Fill in the light gray cells on the THREAT ASSESSMENT MATRIX worksheet to generate the distinct gradations on your qualitative scales of impact and likelihood. Make use of the examples provided.
3. NB! In case bias risks have been previously analysed in the system owner's organisation (regardless of the methodology), consider re-using the same evaluation scales.
4. NB! In case the scales have been modified before the re-assessment of a system, all the threats have to be re-evaluated according to the new scales.

## 5.3 Filling in the workbook

**Step 1: Compose a list of possible threat scenarios.** Fill in the cells 7.1, 7.2, etc. on the worksheet THREATS CONSIDERED (7).

1. Review all the usage scenarios recorded in the table. Leave out scenarios where the AI component's output does not affect decisions related to persons, organisations, or groups of the same.
2. For the remaining usage scenarios, evaluate whether they can be related to a threat scenario. Examples of threat scenarios are provided in Section 4.3 of the Guideline.
3. If you see any threat scenarios, record these on the worksheet. Add extra lines if needed.
4. It is critical that the description of the threat scenario is short and states the specific kind of harm caused. We recommend using the following pattern: 'This event or a short chain of events causes this kind of harm'. See examples on the worksheet THREATS CONSIDERED (7).
5. Disregard the specific significance of the threats while recording threat candidates. This will be done in the next step.

**Step 2: Choose which threats will be assessed.** On the worksheet THREATS ASSESSED (T), fill in the yellow cells O1.1, O1.2, O2.1, O2.2, O3.1, O3.2 for each threat scenario to be assessed. Follow the recommendations below.

1. For each threat scenario candidate on the worksheet THREATS CONSIDERED (7), conduct a preassessment to decide whether or not to keep the scenario. Make use of the criteria previously defined on the worksheet THREATS CONSIDERED (7).
2. In case a particular threat scenario matches some condition or threshold, copy the threat scenario to the worksheet THREATS ASSESSED (T) and fill in the first three cells (e.g. O1.1, O1.2, O1.3 with the extended description of particular threat scenario.

**Step 3: Collect information needed to evaluate the likelihood of the threat scenario.** On the worksheet THREATS ASSESSED (T), fill in the yellow cells (e.g. O1.3, O1.4, O1.5, O1.6, O2.3, O2.4, O2.5, O2.6) for each of the identified scenarios under assessment, following the recommendations below.

1. Prerequisites for a threat scenario to materialise (cell O1.3) should contain a minimum number of steps that, in parallel or in succession, cause the materialisation of the threat scenario and harm.
2. It is easiest to evaluate the plausibility of a threat scenario (cell O1.4) by looking around for cases where a similar threat scenario has already materialised (see the Guideline for examples). If no straight comparison is available, base your assessment on expert opinions.
3. in the case of some threat events, the materialisation of the event can also potentially result in the violation of a law or other norm. Record some of the most serious of these violations

in cell O1.5.

4. In case the system already has protection measures in place to avoid this particular threat scenario, describe these in cell O1.6. Note that the economic rationality of the measures is unimportant here. Just write down what, if at all, can be done to alleviate the risks.
5. If no protection measures have been implemented yet but the assessment working group can propose some, then it is reasonable to record these in cell O1.7.
6. Continue with the next threat related to the usage scenario or move on to another scenario, adding the new threats to the worksheet. If possible, group similar threats. Try to keep the number of threats down. For the sake of efficiency, it is reasonable to describe 10–30 threat scenarios per system.

**Step 4: Proceed with the assessment of the threat scenario.** For each identified threat scenario, fill in the yellow cells O1.7, O1.8, O2.7, O2.8 on the worksheet THREATS ASSESSED (T), following the recommendations below.

1. Each working group member will score the impact and likelihood of each threat based on the threat description on worksheet THREATS ASSESSED (T), as well as the description of the related usage scenario.
2. We advise copying and filling in the worksheet THREAT ASSESSMENT MATRIX for each particular threat. Evaluation scales are handily provided on the worksheet, as well as the formulas to calculate the averages and modes of the scores given. A graph with the distribution of the scores is also provided.
3. NB! The same evaluation scales have to be used for the entire system.
4. If possible, each working group member should a sentence or two to justify their evaluation.
5. In case there are multiple evaluators, we advise that the scores first be given privately to prevent other evaluators from being affected by the first person's scores. In cases when confidential scoring is not possible nor efficient, the collection of each score should be started from a different person; note, though, that this, too, can lead to biases in the risk analysis.
6. In case there are multiple evaluators, a final score has to be agreed among the participants and recorded in the threat assessment cell. The methodology for determining the final score depends on the employed evaluation scale. In the case of a linear scale, a simple arithmetic mean will do. For non-linear scales, the most popular score can be used.

## 5.4 Follow-up activities

**Step 1: If necessary, organise quality control for the scores.** Do the following.

1. In case there was insufficient information to evaluate the impact and likelihood of a threat and some threats were not scored as a result, seek assistance from persons familiar with the details of the particular threat scenario. They can be data scientists, lawyers, or infosec specialists. Their knowledge should help clarify the input data for the evaluation, after which a new evaluation can be carried out to reflect the new knowledge.

# 6 Bias risk treatment

## 6.1 Who, when, why?

**Responsible person.** Risk treatment should be carried out by the same group of people who described and assessed risk scenarios. They have the best background knowledge for risk treatment. Experts who can tell how threats can be practically reduced must also be involved. It is particularly useful at this stage to have the head of the organisation operating the AI system be involved in the evaluation. They will be authorised to accept risks on behalf of the organisation. If this person cannot be involved in risk treatment, he or she must be involved in the approval of the decision.

**Time.** Risk treatment should take place as soon as possible after risk assessment.

**Purpose.** Risk treatment is the final stage of the AI system bias assessment, at the end of which decisions are reached.

## 6.2 Preparations

**Step 1: Determine the signs that make a bias risk significant.** Do the following:

1. It may not be economically feasible or possible to mitigate all the assessed risks. Only significant risks should be considered.
2. Review the section CRITERIA on the worksheet RISK TREATMENT (R) and the conditions and thresholds below. Under the threshold examples, select the appropriate ones for your organisation and record these in the light gray cell under the selected threshold.
3. Attention! If the organisation that owns the AI system has previously analysed bias risks (according to this or any other methodology), consider reusing the previous thresholds.

**Step 2: Determine possible risk mitigation measures.** Do the following:

1. Find out what are the options for changing or replacing the system (AI component, training dataset or general implementation) and whether there are funds or other agreements for these activities.
2. If it is not possible to change the system, it is necessary to clarify the extent to which the AI system can be configured and how these changes may affect the bias in its outputs.
3. Finally, clarify the extent to which it is possible to change the process of using the outputs of the AI system, e.g., by adding additional controls.

## 6.3 Filling in the worksheets

**Step 1: Select significant threat scenarios.** For each significant threat scenario, fill the yellow cells R1.1, R1.2, R2.1, R2.2, R3.1, R3.2, etc. (add or remove them on the worksheet as appropriate) on the worksheet RISK TREATMENT (R). Follow the recommendations below.

1. Check all the threat scenarios listed on worksheet THREATS ASSESSED (T). Apply the significance criteria from the CRITERIA table on worksheet RISK TREATMENT (R) to each threat assessed and copy the names of the threats above the threshold to the Risk Treatment Table (box R1.1).
2. Add an explanation about which threshold was exceeded to cell R1.2.

**Step 2: Find possible mitigation measures for bias risks.** For each significant threat scenario, fill the yellow cells R1.3, R1.4, R2.3, R2.4, etc on the worksheet Risk Treatment (R). Follow the recommendations below.

1. If possible bias risk mitigation measures have already been documented together with the threat description (worksheet Threats Assessed (T)), select measures which are feasible with available resources or for which the necessary resources or working hours are likely to be available.
2. Technical measures may require the replacement of the entire system or its AI component, replacement of the ML model, or training with additional data. See Sections 4.5 and 5.4 of the Guideline for inspiration.
3. Organisational measures generally mean the addition of additional human controls, but may also mean restrictions on system users (e.g., risk groups are not allowed to use the system) or the introduction of additional transparency measures. See more in Section 4.5 of the Guideline.

**Step 3: Re-evaluate the threat scenario as if mitigation measures have been implemented.** For each significant threat scenario, fill the yellow cells R1.5, R1.6, R1.7, R2.5, R2.6, R2.7, etc. on the worksheet Risk Treatment (R). Follow the recommendations below.

1. Act in the same way as in the previous assessment, but now imagine yourself in a situation where the mitigation measures planned have already been implemented.
2. It is important to understand that realistic mitigation measures can, but do not necessarily, change the risk scores. It is possible that, for example, the impact of the threat will be reduced, but its likelihood will remain the same. It is also possible that neither will decrease. Follow the recommendations for the organisation of the assessment to reduce the bias of the evaluators.
3. In cell R1.7, record the threat level remaining even after mitigation measures (the most likely scenario for a bias-related threat to materialise which remains after the system is upgraded).
4. Be sure to use the same description of metrics and the same methodology for aggregating multiple scores.

**Step 4: Make a risk treatment decision for each significant threat.** For each significant threat scenario, fill the yellow cells R1.8, R2.8, etc. on the worksheet Risk Treatment (R). Follow the recommendations below.

1. Decide whether the residual bias risk after the implementation of mitigation measures is acceptable, taking into account the risk policy of the service owner.
2. If the risk is acceptable, record the risk as 'Acceptable' in cell R1.8 and explain how the owner of the service can deal with the possible harm.
3. If the risk is not acceptable, record the risk as 'Unacceptable' in cell R1.8 and explain why the organisation chose to refuse this risk.

## 6.4 Follow-up activities

**Step 1: Formalise the final bias risk assessment decision with a person authorised by the AI system owner.** Do the following.

1. Find the person whose mandate includes taking risks on behalf of the organisation.
2. Introduce them to the significant risks and the risk treatment decisions.

3. Ask for a decision on the general deployment of the system (to be deployed, to be deployed with modifications, to be removed from use). Document the decision in the cell S.1 on worksheet CONCLUSION (C).
4. In cells S.2–S.6, document the changes without which the system may not work. Get a confirmation from the authorised person of the means necessary for making the changes.
5. In cells S.7–S.9, document the changes that are recommended but can be postponed if needed.
6. In cell S.X, add a reference to the decision document, if possible.
7. Let the authorised person confirm the decision (e.g., with a digital signature).

**Step 2: Implement the bias risk analysis decision.** Do the following.

1. Communicate the decision to all persons involved in the risk analysis.
2. If the AI system remains in operation but with changes, initiate the implementation of the agreed mandatory and, if possible, recommended changes with the means provided for this purpose.
3. If the AI system is removed from use, initiate related activities.