



**Responsible AI Assessments**

**Identify and assess potential harms and biases in AI systems**

**with a focus on use cases in Sub-Saharan Africa and Asia Pacific**

**Pacific**

**Part B: Qualitative Guide**

As a federally owned enterprise, GIZ supports the German Government in achieving its objectives in the field of international cooperation for sustainable development.

**Published by:**Deutsche Gesellschaft für   
Internationale Zusammenarbeit (GIZ) GmbH

Registered offices

Bonn and Eschborn, Germany

Global programme Digital Transformation

FAIR Forward – Artificial Intelligence for All

Friedrich-Ebert-Allee 32 + 36

53113 Bonn, Germany

T +49 228 44 60-0

F +49 228 44 60-17 66

E fairforward@giz.de

I [www.giz.de](http://www.giz.de)

**Responsible**

GIZ - FAIR Forward – Artificial Intelligence for All

Nadine Dammaschk – AI Advisor ([nadine.dammaschk@giz.de](mailto:nadine.dammaschk@giz.de))

Jonas Gramse – AI Advisor ([jonas.gramse@giz.de](mailto:jonas.gramse@giz.de))

**Authors**

Eticas: Mariano Martín Zamorano, Luis Rodrigo González Vizuet, Gemma Galdon Clavell

FAIR Forward: Nadine Dammaschk, Jonas Gramse

**Content review**

FAIR Forward: Nadine Dammaschk, Jonas Gramse, Sheila Kibughi, Deshni Govender, Kathleen Ziemann, Balthas Seibold.

Community of AI inclusion experts: Favour Borokini, Josia Paska Darmawan, Mohamed Kimbugwe, Mercy King'ori, Meena Lysko, Raashi Saxena, Kofi Yeboah.

**Special acknowledgements**

We are additionally grateful to:

* Namritha Murali, Mitchel Ondili, Raphael Leuner and Francesca Trevisan for their contributions in the initial phase;
* Sheila Kibughi and Isabela Miranda for their support in steering the activity;
* Eva Keller for her contributions to the documents in their final stages;
* the general FAIR Forward team and involved partners for their continuous feedback and contributions.



This document is available in Open Access under the [Attribution-ShareAlike 4.0 International](https://creativecommons.org/licenses/by-sa/4.0/) license CC BY-SA 4.0 DEED.

**Attribution**    
You must give [appropriate credit](https://creativecommons.org/licenses/by-sa/4.0/#ref-appropriate-credit), provide a link to the license, and [indicate if changes were made](https://creativecommons.org/licenses/by-sa/4.0/#ref-indicate-changes).

Please cite as follows: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) - FAIR Forward (2024). Responsible AI Assessments. Part B: Qualitative Assessment Guide. Licensed under CC BY-SA 4.0 DEED.

**ShareAlike**    
If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](https://creativecommons.org/licenses/by-sa/4.0/#ref-same-license) as the original.

**Re-use/Adaptation/Translation**

Any derivative work should include the following visible disclaimer “The present work is not an official GIZ publication and shall not be considered as such.” Use of the logos of GIZ and FAIR Forward or the imprint on the back matter of the publication is not permitted on derivative works.

GIZ is not liable for any alteration of the original content as used in the derivative work.

For any derivative work, we would also appreciate greatly if you can notify us briefly via [fairforward@giz.de](mailto:fairforward@giz.de) and beyond 2025 reach out to [vanessa.dreier@giz.de](mailto:vanessa.dreier@giz.de).

**Disclaimer**

The data in the publication has been collected, analysed and compiled with due care; and has been prepared in good faith based on information available at the date of publication without any independent verification. However, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH does not guarantee the accuracy, reliability, completeness or currency of the information in this publication. GIZ shall not be held liable for any loss, damage, cost or expense incurred or arising by reason of any person using or relying on information in this publication.

Bonn, Germany, June 2024

**Table of contents**

[1. INTRODUCTION 1](#_Toc169162687)

[1.1 About the Responsible AI Assessments 1](#_Toc169162688)

[1.2 About this Qualitative Guide 2](#_Toc169162689)

[1.3 Structure of the Qualitative Guide 2](#_Toc169162690)

[1.4 How can you use this Guide? 3](#_Toc169162691)

[Overview on the algorithm lifecycle and associated ethics principles 5](#_Toc169162692)

[2. Pre-Processing: Problem formulation & data collection 7](#_Toc169162693)

[**2.1 Problem formulation** 7](#_Toc169162694)

[2.1.1 Proportionality, do not harm, fairness and non-discrimination 9](#_Toc169162695)

[2.1.2 Multi-stakeholder, adaptive governance and collaboration 12](#_Toc169162696)

[2.1.3 Sustainability 13](#_Toc169162697)

[2.1.4 Problem formulation score card 14](#_Toc169162698)

[**2.2 Data collection** 16](#_Toc169162699)

[2.2.1 Proportionality and do not harm 18](#_Toc169162700)

[2.2.2 Transparency and explainability 20](#_Toc169162701)

[2.2.3 Right to privacy and data protection 20](#_Toc169162702)

[2.2.4 Fairness and non-discrimination 21](#_Toc169162703)

[2.2.5 Data collection score card 22](#_Toc169162704)

[3. In-Processing: Model selection, training and evaluation 25](#_Toc169162705)

[3.1 Proportionality and do not harm 27](#_Toc169162706)

[3.2 Transparency and explainability 28](#_Toc169162707)

[3.3 Fairness and non-discrimination 29](#_Toc169162708)

[3.4 Safety and security 30](#_Toc169162709)

[3.5 Multi-stakeholder, adaptive governance, and collaboration 30](#_Toc169162710)

[3.6 In-processing and score card 30](#_Toc169162711)

[4. Post-Processing: Deployment and monitoring 33](#_Toc169162712)

[4.1 Human oversight and determination 35](#_Toc169162713)

[4.2 Transparency and explainability 35](#_Toc169162714)

[4.3 Responsibility and accountability 36](#_Toc169162715)

[4.4 Awareness and literacy 37](#_Toc169162716)

[4.5 Fairness and non-discrimination 38](#_Toc169162717)

[4.6 Post-processing score card 39](#_Toc169162718)

[5. REFERENCES 42](#_Toc169162719)

[6. Biographies: Community of AI Inclusion Experts 44](#_Toc169162720)

# **INTRODUCTION**

### **1.1 About the Responsible AI Assessments**

As AI technologies evolve, so does the imperative to ensure that their **development and deployment align with human rights principles** and avoid causing harm or perpetuating social inequalities. The **Responsible AI Assessments** are a proactive response to address these challenges head-on, co-created by the GIZ-project “[FAIR Forward – Artificial Intelligence for All](https://www.bmz-digital.global/en/overview-of-initiatives/fair-forward/)” [[1]](#footnote-2), [Eticas](https://eticas.ai/) and a diverse [community of AI inclusion experts](#_7._Biographies:_) from Sub-Saharan Africa and Asia Pacific.

The Responsible AI Assessments are a method to identify, assess and mitigate potential harms and biases in AI. As an AI risks and ethics assessment tool, they guide AI stakeholders (e.g. as an assessor, developer, project managers or deployer of AI) in critically analyzing their AI resources, emphasizing human rights and ethical considerations throughout the AI lifecycle.

The **Responsible AI Assessments**[[2]](#footnote-3)consist of the following parts:

* **Step-by-Step Guide** (Part A):

It orientates on how to apply the Qualitative and Quantitative Assessment Guides, enriched with best practices and lessons learned.

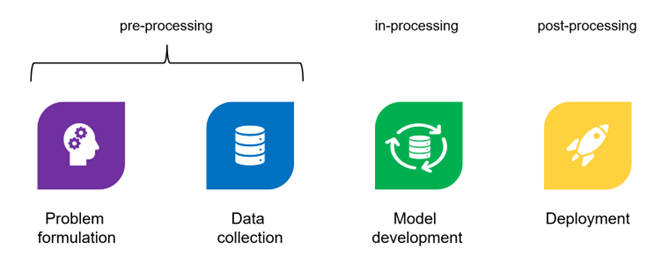
* **Qualitative Guide** (Part B):

It provides critical questions for each stage of the AI lifecycle to assess societal implications, potential biases, fairness, and effects on diverse stakeholders.

* **Quantitative Guide** (Part C):

It focuses on quantitative methods and metrics for critical analysis of data as well as AI models and systems. It builds on the insights from the Qualitative Guide.

Each parts serves a unique purpose. Their main aim is to guide AI stakeholders to **critically analyze their AI resources**, draft actionable insights and mitigate risks. They guide **reflection during** the following stages of an **AI lifecycle**:



Drawing on experiences with real-world AI assessments, the Responsible AI Assessments are a living framework adaptable to the evolving AI landscape. In 2023, they were tested on 7 AI activities from 6 countries on the African and Asian continent.

The original version of the Responsible AI Assessments is available under the [FAIR Forward website](https://www.bmz-digital.global/en/overview-of-initiatives/fair-forward/).

### **1.2 About this Qualitative Guide**

This **qualitative AI risk assessment guide** allows **to evaluate the potential human rights risks and ethical risks of AI resources** like training data for AI, AI models and systems.[[3]](#footnote-4) It aims to guide AI stakeholders (e.g. assessor, developer, project manager or deployer of AI) on what to consider when striving for algorithmic fairness and how to enhance AI fairness by:

* incentivizing **critical reflection**;
* identifying and creating awareness **of human rights risks and ethical concerns**;
* formulating appropriate **mitigation measures**, considering all elements through the AI design and development journey life cycle.

Additionally, this document seeks to:

* improve **transparency of AI resources** throughout their lifecycle by providing a framework for documenting the decision-making process, data sources and evaluation metrics used in the AI systems. This makes it easier to track and audit the performance of the systems;
* **facilitate collaboration** **and knowledge sharing** among stakeholders involved in developing and deploying AI systems by providing a common language and framework for evaluating AI systems.

The Qualitative Guide has been developed and tested with real-live AI use cases that have been supported by FAIR Forward. These use cases primarily involved AI models and systems that are **based on supervised learning**.All of these AI use cases were in different stages of their development, most of them in the pre-processing stage. Therefore, the Qualitative Guide is naturally tailored to these activities and their contexts. Nonetheless, it should be applicable for assessing AI systems in similar or related contexts.

### **1.3 Structure of the Qualitative Guide**

The Qualitative Guide is **structured around two axes**: The **AI lifecycle**, on the one hand, and **UNESCO’s** [**Recommendations on the Ethics of AI**](https://unesdoc.unesco.org/ark:/48223/pf0000381137)(2022), on the other. Principles developed by this recommendation entail an up-to-date reflection on AI human rights implications addressing critical aspects from a holistic perspective. Furthermore, they are the first global standard-setting instrument on the subject, which was adopted jointly by UNESCO’s member states in 2021[[4]](#footnote-5).

For each stage within an AI lifecycle[[5]](#footnote-6), the guide provides:

* a brief definition of the stage
* related AI principles of the UNESCO recommendation
* info box on questions an auditor might have
* moments of bias
* exemplary questions
* that correspond to principles for AI as suggested by **UNESCO’s** [**Recommendations on the Ethics of AI**](https://unesdoc.unesco.org/ark:/48223/pf0000381137)

The Qualitative Guide focuses on the principles most salient for each stage of the AI lifecycle to maintain applicability and gain internal **coherence and consistency**.[[6]](#footnote-7)

### **1.4 How can you use this Guide?**

For further information on **how to apply the guide as part of a Responsible AI Assessment** and whom to include in which phase, **please check out the** [**Step-by-Step Guide**](https://gizonline.sharepoint.com/teams/AI-Labwithguests/Freigegebene%20Dokumente/HF-REGIO%20Übergreifend%20und%20Vernetzung%20der%20Länder%20un/26%20Activity_Q_of_Global_Regional%20-%20Risk%20board/Final%20documents/Responsible%20AI%20Assessments%20-%20Part%20A%20-%20Step-by-Step%20Guide.docx?web=1) **first** before continuing.

Generally, while this guide appears to move through AI lifecycle stages sequentially, it is crucial to understand that these stages are interconnected and flexible. **Insights gained at any point can lead to a reassessment of previous stages.** Each lifecycle stage can be explored independently, even if an AI system is already in use. It is valuable to retrospectively assess past decisions or consider data collection aspects during the problem formulation stage. Think of the guide as an **iterative exercise.** We recommend that you make the results of your analysis publicly available.

For a holistic analysis, we recommend going through the complete guide. However, as this might not be feasible for everyone, we suggest focusing on questions that are particularly relevant to the AI use case in question, and that have been selected with an independent assessor/ AI auditor.

For a comprehensive reflection, we recommend using the guide with a **diverse set of stakeholders.** This might include AI developers, project managers, gender and inclusion experts, affected stakeholders[[7]](#footnote-8), or external auditing professionals. This approach ensures a holistic and nuanced perspective.

|  |
| --- |
| **Important disclaimers**  The Responsible AI Assessments are a method developed to conduct a holistic AI risk and ethics assessment. It can be used by any individual (applying it themselves), but it is highly recommended that the method is utilised with the expertise of external assessors or auditors. A Responsible AI Assessment **does not qualify as a formal audit** (in any form), nor does it replace an audit process. Use of the Responsible AI Assessments alone **does not guarantee compliance** with local and/or international laws, regulations or standards. Please engage independent auditors and/or legal advisors to ensure compliance of your product or service with local and/or international laws.  **This guide does not attempt to be a ‘holy grail’** – and there will probably never be a perfect template for ensuring AI Ethics for all AI use cases. This guide simply strives to make the opaque field of AI Ethics more operationalized and tangible and to provide exemplary guidance for AI stakeholders on how to incorporate considerations of AI Ethics throughout the algorithm lifecycle.  We acknowledge that content may be perceived as subjective, particularly in the association of questions with specific AI ethics principles or stages of the algorithm lifecycle. Furthermore, we recognize that some readers may find essential questions missing.  **If you have any feedback** on some of the points above or other points, **we would like to hear from you via** [fairforward@giz.de](mailto:fairforward@giz.de)to further improve the Responsible AI Assessments and their underlying guides.  *Please also note that this is the* ***fill-out version*** *of the Responsible AI Assessment. This version is meant for those who want to apply the assessment guides. To avoid unwanted changes, in this version* ***only the following parts can be edited****:*   * ***Answer blocks*** *below questions* * ***Tables*** *that are meant to be filled out.*   *If you encounter issues, please reach out to* [*fairforward@giz.de*](mailto:fairforward@giz.de)*.* |

### **Overview on the algorithm lifecycle and associated ethics principles**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm processing phases** | **Subphases** | **Entry points for ethical risks** | **Associated ethics principles** |
| [**Pre-processing**](#_2._Pre-Processing:_Problem) | 1. [Problem and system definition](#_2.1_Problem_formulation) | 1. Society / Real life → Data | 1. Proportionality and do not harm 2. Fairness and non-discrimination 3. Multi-stakeholder, adaptive governance, transparency and collaboration 4. Sustainability |
| 1. [Data collection & data quality](#_2.2_Data_collection) | 1. Society / Real life → Data 2. Data → Population 3. Population → Sample 4. Sample → Variables + Values | 1. Proportionality and do not harm 2. Transparency and explainability 3. Right to Privacy and Data Protection 4. Fairness and non-discrimination |
| [**In-processing**](#_3._In-Processing:_Model) | 1. Model selection, training and evaluation | 1. Variables + Values → Patterns 2. Patterns → Predictions | 1. Proportionality and do not harm 2. Transparency and explainability 3. Fairness and non-discrimination 4. Safety and security 5. Multi-stakeholder, adaptive governance, and collaboration |
| [**Post-processing**](#_4._Post-Processing:_Deployment) | 1. Deployment 2. Scaling 3. Implementation 4. Interpretability of the results 5. Monitoring | 1. Predictions → Decisions by humans 2. Decisions by humans → World | 1. Human Oversight and determination 2. Transparency and explainability 3. Responsibility and accountability 4. Awareness and literacy 5. Fairness and non-discrimination |

|  |
| --- |
| **Unpacking definitions of fairness and discrimination**  The principle of fairness and non-discrimination in the UNESCO guidelines says, “AI actors should promote social justice, fairness, and non-discrimination while taking an inclusive approach to ensure AI’s benefits are accessible to all.”  Given the highly interdisciplinary nature of responsible AI, definitions of fairness and discrimination sometimes vary and can lead to confusion. This box aims to create awareness on different language around responsible AI.  In the Qualitative guide of the Responsible AI Assessments, we mostly refer to the social interpretation of discrimination and fairness. In the Quantitative Guide, the statistical definition of fairness is more at the forefront.  **Fairness as social concept**  Fairness is closely intertwined with justice. In a narrower sense, a fair AI system can mean that it does not perpetuate existing social inequalities and injustices.  Fairness in the light of social justice is more aspirational. In this case, an AI system might have the goal to contribute to a more equitable, inclusive distribution of resources or access to opportunities.  **Fairness as statistical concept**  From a technical view, fairness in AI often refers to mathematical fairness concepts (see chapter 5.6 or [Verma, Rubin](https://fairware.cs.umass.edu/papers/Verma.pdf) 2018). Please note that the statistical concept of fairness is limited by the fact that the so-called “ground truth” might be inherently biased. Whether mathematical calculations alone are sufficient to solve the problem of fairness remains controversial and is subject of ongoing research.  **Discrimination as social concept**  Discrimination refers to treating people differently and causing them punctual or systemic harm based on sensitive aspects like race, colour, sex, language, religion, disability, age, sexual orientation, and gender identity, among others. This list is non-exhaustive and might change according to the (local) context. Discrimination can also occur at the intersection of these aspects.  **Discrimination as statistical concept**  In statistics, discrimination is rather used as a non-normative term to describe statistical concepts of dividing predictions into groups, e.g. “discriminating” between blue and red flowers.  As a tool, the Responsible AI Assessments aim to support you in creating fairer AI systems. Mainly they should support you in not perpetuating existing inequalities, but they also aspire to guide you in creating more inclusive AI systems. |

# **2. Pre-Processing: Problem formulation & data collection**

The pre-processing stage covers concrete steps from problem formulation to data collection and processing, influencing the later performance of an AI system.

Assessing the pre-processing stage serves two primary objectives:

* Attain a **precise definition** of the problem. The AI system should align with real-world scenarios, considering various variables around the main problem and evaluating the problem-solving approach.
* Ensure the data used for training the AI system is of **high quality**, aiming for the most accurate representation of reality and its complexity.

Biases associated with this development phase are mainly cognitive or statistical.

## **2.1 Problem formulation**

The first moment in which bias can enter the algorithm lifecycle is when you formulate the problem that an AI system should solve. The problem formulation stage requires **making assumptions about the problem and the data required to solve it**.

For example, these assumptions can be biased if they are:

* based on **incomplete or inaccurate information** about the phenomenon at stake
* or **influenced by prejudice** or personal biases of the people involved in the project.

For the problem formulation stage, this guide provides exemplary questions for the following AI principles of UNESCO:

* [Proportionality, do not harm, fairness and non-discrimination](#_2.1.1_Proportionality,_do) (2.1.1)
* [Multi-stakeholder, adaptive governance, and collaboration](#_2.1.2_Multi-stakeholder,_adaptive) (2.1.2)
* [Sustainability](#_2.1.3_Sustainability) (2.1.3)

|  |
| --- |
| **Tips for tackling algorithm audits**  An auditor might ask you for:   * The intended and unintended purposes of the AI system * Potential intended and unintended consequences * A map of stakeholders (in-)directly impacted and how they might be affected by the system * A map of roles and responsibilities |

#### **Moments of bias in problem formulation**

Each moment of the lifecycle of an algorithmic system is subject to different sources of bias. These moments of bias are not mutually exclusive nor relevant in all cases. Identifying which ones apply to an AI use case makes them easier to tackle.

Below you will find an overview of the main sources of bias for problem formulation:

|  |  |  |  |
| --- | --- | --- | --- |
| **Bias Type** | **Definition** | **Example** | **Exemplary testing** |
| **Techno-solutionist bias** | **Over-reliance on high-tech solutions** without considering simpler alternatives or potential social and environmental impacts. | An AI system for crop disease detection might overlook simpler, low-tech solutions like farmer education. | 1. Analyze alternatives. 2. Assess expectations of developers vs. performance in other relevant use cases. 3. Check desirability and compliance. 4. Evaluate environmental impact. 5. Study social impacts. |
| **Data availability/ scarcity** | The selected **datasets** are **not relevant** to the problem, or there is insufficient data to represent (crucial parts of) the real-world problem. | An AI system for predicting crop yields might lack sufficient data on crops grown by women farmers, leading to inaccurate predictions for women. | 1. Check available datasets for relevance. 2. Assess labels and features. 3. Ensure data minimization compliance. 4. Consider and analyze statistical differences between groups. |
| **Historical bias** | The data used for training the AI system reflects **existing social and societal biases**, leading to potentially harmful outcomes. | An AI system for allocating farm resources might favour male farmers based on historical data, perpetuating existing gender inequalities. | 1. Perform contextual analysis and identify protected groups. 2. For Natural Language Processing (NLP): validate word embeddings. 3. Analyze statistical differences between groups. |
| **Population bias** | The **target population** during the design phase does not represent the actual user population, leading to non-representative results. | An AI system designed for large-scale farms might not work well for small-scale farmers, who make up the majority of the user population. | 1. Identify affected stakeholders and protected groups. 2. Check distribution differences of groups. 3. Perform randomization testing. |

### **2.1.1 Proportionality, do not harm, fairness and non-discrimination**

The following questions explore the problem that you wish to solve with an AI system, whether an AI system would be an appropriate solution and what kind of harm it might exhibit for affected stakeholders and protected groups[[8]](#footnote-9).

**Exploring the problem and solution space:**

* What is the **problem that needs to be solved** by the system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Imagine the AI system is deployed: What **positive change** do you envision it to create?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Why do you think this problem should be solved through AI? Could the problem be **solved through a non-AI-related solution** (e.g. other low- or non-tech solution)?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Besides the typically protected groups, have you also considered the system's potential impact on other groups related to emerging marginalization derived from technology implementation?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

**Exploring affected stakeholders of the AI system**

Fill out the table below to explore how envisioned stakeholders might be affected by the AI system:[[9]](#footnote-10)

|  |  |  |  |
| --- | --- | --- | --- |
| **Who is (in)directly affected by the AI system?** | **Do they belong to a *protected group*?** | **How would the AI system affect this group positively?** | **How could the AI system affect this group negatively?**  (e.g. via false positives/  false negatives) |
| Stakeholder 1 |  |  |  |
| Stakeholder 2 |  |  |  |
| Stakeholder x |  |  |  |
| … |  |  |  |

**Exploring human rights implications**

From the risks that you identified above, fill out the table below to explore which [human rights](https://www.un.org/sites/un2.un.org/files/2021/03/udhr.pdf) might be affected by the AI system. Try to categorize how likely it is that they will occur and how severe they might be should they occur:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Risk (exemplary)** | **Likelihood for risk to occur.**  (inexistent, minor, moderate, significant) | **Why?** | **Severity if risk occurs** (inexistent, minor, moderate, significant) | **Why?** | **Mitigation** (to be filled out with your assessor[[10]](#footnote-11)) |
| Risks to life, sustenance or security of a person |  |  |  |  |  |
| Risks to unfair distribution of services for certain groups |  |  |  |  |  |
| Risks to privacy or reputation |  |  |  |  |  |
| Risks to freedom of opinion and expression |  |  |  |  |  |
| … |  |  |  |  |  |

**Exploring dynamics of the team working on the AI system:**

Diversity within the teams leading design and deployment of algorithms is crucial to identify and mitigate discrimination. Exemplary aspects to reflect on:

* How diverse is the **team working on the AI system,** in terms of disciplinary (professional) backgrounds, age, gender, culture and nationalities?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How engaged is the **team with the communities for which** they are designing the AI system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How is relevant **domain expertise** included within your team’s composition?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What are the **hierarchical roles** and **decision-making capacity** of these team members? (*Please, fill in the table below*)

|  |  |  |  |
| --- | --- | --- | --- |
| **Team member** | **Diversity perspectives** | **Role** | **Responsibility over  model design** |
| Member X | * Tag members by (age, gender, gender identity, disability status, etc. | * Developer * CEO * PM * ... |  |
| Member Y | * … | * …. |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Based on table information:

* What is the **balance between team internal diversity and power relations** among team members? Does the present structure enable space for becoming aware of and openly voicing existing biases, voicing those biases and considering them in developing the AI system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How does the team aim to **avoid potential biases** despite the lack of representation from a specific group when making critical design decisions (e.g. how are perspectives from non-binary people respected if they are not represented within the team or any focus group activities)?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **2.1.2 Multi-stakeholder, adaptive governance and collaboration**

The following explores how different stakeholders would be able to participate throughout the AI lifecycle to ensure the AI system’s inclusivity:

* What strategies will you use to engage **end-users or beneficiaries** during the design process, and how will this engagement be maintained throughout the AI lifecycle?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + What specific methods or platforms will you use to engage stakeholders?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + On what aspects will you engage them?   
    Will their involvement be limited to user experience, or will it extend to deeper aspects like system goals and ethical values behind the system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + How do you intend to integrate conclusions from these collaborations and evaluations into the system design?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Has there been a **public debate or case law/current jurisprudence** on similar technologies?[[11]](#footnote-12) If so, how have the main conclusions of such debates been considered for your AI system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* If the technology will be used in a public setting, how do you intend to **consult the public** about plans for the AI system?
  + Through what channels?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + Has the language and the message of the consulting process been adapted to the different communities and audiences?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + How do you intend to integrate their input into the system design, especially from low-income communities?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **2.1.3 Sustainability**

The following invites you to consciously reflect on the Sustainable Development Goals (SDGs) agreed as common focus of the international state community and how they relate to your planned AI system.

Which of the following [**targets of the SDGs**](https://sdgs.un.org/goals) does the AI system promote? (see table below)



Source: https://sdgs.un.org/goals

|  |  |  |
| --- | --- | --- |
| **SDG Goal** | **SDG Target** | **How the AI system promotes it** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

### **2.1.4 Problem formulation score card**

Before moving to the next stage, the scorecard below is a **guide for evaluating and summarizing the AI system’s ethical findings** concerning problem formulation. It provides a concise and comprehensive overview of the AI system's ethical considerations and any actions taken in response to the assessments.

The score card can be used to communicate the findings of an AI system with team members, decision-makers, auditors and affected communities.

|  |
| --- |
| **How to use the score card**  **Assign an ordinal rating** for each variable based on the perceived severity of the issues identified. Use the following scale. The scales are explained below with an example from an AI crop mapping use case:  **0: No issues** e.g.: All regions covered by an AI-based crop mapping system are equally represented in the training data. The AI model is expected to consistently provide accurate crop-type predictions across various geographical areas without any bias.  **1-2: Minor concerns** e.g.: There is a slight imbalance in the representation of certain crops in the training data, but the impact on predictions is expected to be minimal. Therefore, the team considers that the system will generally perform well across different crop types.  **3-4: Moderate issues**  e.g.: The AI model is suspected to be moderately biased in predicting crop types, with certain regions experiencing lower accuracy rates. This may be due to identified variations in data quality or insufficient representation of specific geographical features.  **5: Significant issues**  e.g.: Identified unbalanced data representation indicates significant bias in predicting crop types, with certain regions expected to consistently receive inaccurate predictions. Following the preliminary assessment, this may lead to unfair resource allocation or policy decisions, impacting the agricultural practices or livelihoods of affected communities.  **Provide specific comments** for each variable, explaining the rationale behind the assigned rating and include observations, concerns, or noteworthy aspects related to each variable. |

* **Name of the AI system:**
* **Evaluator name:**
* **Model date, version and type:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Principle** | **Sub-aspects per principle** | Assign **severity rating** for issues per sub-aspect(0-5) | **Why did you assign this score?** If you identified issue(s), what kind? | **How can you mitigate this issue?** (ideally filled out with your assessor[[12]](#footnote-13)) |
| Proportionality and do not harm  & fairness and non-discrimination | Problem & solution space |  |  |  |
| Affected stakeholders |  |  |  |
| Human rights implications |  |  |  |
| Team dynamics |  |  |  |
| Multistakeholder & Governance | Involving end-users / beneficiaries |  |  |  |
| Current public debates |  |  |  |
| Consultation of public |  |  |  |

## **2.2 Data collection**

The second pre-processing moment in which bias can enter the algorithm lifecycle is when **reality is translated into data**. That way, **social inequalities**, **historical imbalances and discriminatory patterns** mayfeed into thedata collection process and create harmful downstream patterns.

For the data collection in the pre-processing stage, this guide provides exemplary questions for the following AI principles of UNESCO:

* [Proportionality and do not harm (2.2.1)](#_2.2.1_Proportionality_and)
* [Transparency and explainability (2.2.2)](#_2.2.2_Transparency_and)
* [Right to privacy and data protection (2.2.3)](#_2.2.3_Right_to)
* [Fairness and non-discrimination (2.2.4)](#_2.2.4_Fairness_and)

|  |
| --- |
| **Tips for tackling algorithm audits**  An auditor might ask you for:   * If you collected the data, how did you conduct this process technically and legally? (e.g. did you collect the data via web scraping; how did you ensure informed consent) * A description of the types of variables used to train, test or operate the model * How complete or representative the data is vis-à-vis the purpose of the model and intended deployment contexts * The URL for the datasets you have used * Compliance with data protection regulations, incl. how data will be processed and stored |

#### **Moments of bias in data collection**

Each moment of the lifecycle of an algorithmic system is subject to different sources of bias. These moments of bias are not mutually exclusive nor relevant in all cases. Identifying which ones apply to an AI use case makes them easier to tackle.

Below you will find an overview of the main sources of bias for data collection:

|  |  |  |  |
| --- | --- | --- | --- |
| **Bias Type** | **Definition** | **Example** | **Mitigation Techniques** |
| **Labeling bias** | Inaccurate data labeling due to subjective perceptions. | In satellite images, a **wheat field** might be **misclassified** as a corn field due to similar color patterns. | * Use/ create labeling guidelines. * Have multiple and diverse experts label data and reach consensus. * Use majority vote for labeling if many experts are available. * Check how labels (categories) are distributed in your data. * Review and correct labeling errors. * Consider using open-source datasets and labels that allow for crowd-verification of labelling. * Publish your own datasets & labels under open-source licenses to allow for third-party checks, corrections, and crowdsourcing of improvements. |
| **Sampling/ generaliza-tion bias** | Non-random or uneven sampling leading to **unrepresentative data.** | An AI system trained on data from farms run by men mightfail to predict yields for farms run by women due to different farming practices. | * Test classifier performance on under-sampled and control groups. * Assess oversimplification instances. * Calculate confidence intervals separately for each group. * Perform statistical significance tests across different groups. * Ensure balanced and representative training data. * Consider open-source datasets, see above. |
| **Survey bias** | Inaccurate, incomplete, or inconsistent data from surveys or interviews. | A survey asking farmers about their fertilizer use might lead them to overstate the amount used due to social **desirability bias**. | * Address issues in data collection method, e.g. for questions where people are hesitant. * Compare and contrast data from different sources. * Useprobabilistic matching and individual reference identifiers for dataset harmonization. * Minimize likelihood for recall bias through survey design. * Distribute surveys evenly, account for under-sampled groups. |
| **Survivorship bias** | Only considers **data that ‘survived’** or remained till the end, ignoring the ones that did not.  This could lead to overestimating the ‘survived’ data. | An AI system analyzing the success of different crop varieties might only consider data from farms that successfully harvested, ignoring those that failed due to harsh weather conditions. | * Collect and record all relevant information. * Work with confidence intervals in your data analysis. * Check how labels (categories) are distributed in your data. * Consider open-source datasets, see above. |

|  |
| --- |
| **On the importance of datasheets**  **If you are collecting data yourself, it is highly recommended that you create a datasheet**. After going through this section of the Qualitative Guide, you can use many of the insights to draft a datasheet. Here you can find a famous [template for datasheets](https://arxiv.org/pdf/1803.09010.pdf) (Gebru et al. 2018).  **When using datasets** for model training **check for an existing Datasheet** (regardless of whether it is [open, shared or closed data](https://theodi.org/insights/tools/the-data-spectrum/)).  Datasheets are a best practice to create transparency about a dataset and help you to understand the collection, processing and quality of the data.  If there is no available datasheet, search at least for basic available information as a rough estimator for the quality of the data, e.g.:   * Why was the data collected * When was it collected where and how * For what can it be used * How it was further processed   If you do not have enough information available to assess the quality of a certain dataset, reconsider whether and why you want to use this data source or resort to other data sources. |

### **2.2.1 Proportionality and do not harm**

The following questions aim to evaluate the appropriateness and setup of your data. They are designed to assess the potential impact of inaccurate predictions or misinterpretations on algorithmic bias.

**Exploring the configuration of your data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **What variables do you intend to collect?** | **Is this variable directly related or a proxy?** | **Is this variable related to a protected group?** | **What is the datatype (e.g., string, int, float) of this variable?** | **Why do you think this variable is suitable to predict the target variable?** | **What might be the shortcomings of this variable in predicting the target variable?** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Once the above aspects are identified, data-quality aspects should also be examined:

* Via which **procedures** do you intend **to collect the data**? For example:
  + Generally: hardware apparatus, web scraping, web surveys
  + NLP: communal voice collection
  + Agriculture: manual human curation (via phones)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How do you ensure that your data is **representative of the target population**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Which **missing data points** might impact the system’s performance?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What kind of data quality and representation challenges might arise from **manual data collection**, with a specific focus on issues related to gender and its proxies?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Example**: AI-driven precision farming may provide farmers with tailored agronomic advice using both satellite imagery data and ground truth data from sensors.  Potential biases arise if historical data, collected manually and reflecting traditional gender roles, skews representation. Mitigation measures involve a revised data collection, bias mitigation in algorithms, and regular evaluation to ensure fair and effective outcomes for the entire farming community. |  |  |

* For agriculture use cases:
  + If you are collecting data with sensors: Have you tested whether the sensors are broken or miscalibrated?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + If you are collecting data with smartphones, drones or other technologies: How do you consider how accessible the technology is for farmers, particularly for the later deployment stage? (e.g. disparity in technology access for female in comparison to male farmers or sparse connectivity in some areas)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + Have you addressed expected challenges such as limited regional/local data, indirect personal data sharing, gender bias, and outliers in your system training data?"

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* For NLP use cases:

Have you reviewed your datasets for potential incomplete or inaccurate data dictionaries to ensure the robustness and accuracy of the algorithm's training data?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* For satellite imagery cases:   
  Have you examined data granularity and representative geographical and temporal representation of your data?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **2.2.2 Transparency and explainability**

The following questions explore how the data has been collected, labelled and pre-processed. Thus, they provide insights into the data used to train an AI system. This makes it easier to comprehend the system’s later decisions and outcomes.

* Which **decisions** **or trade-offs** did you make **during the data collection** process that impact how representative the sample is of the larger population?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Which **decisions or trade-offs** did you make on **how to label your data** and whether those decisions impact the later output and functioning of your AI system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How do you intend to **pre-process** the data? (e.g. cleaning, normalizing, transforming data)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* If you use **external data**: Have you checked the **rules of use**, including any licensing terms or usage restrictions, set by the data providers?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

|  |
| --- |
| As a minimum, it is recommended to **document these answers** and make them publicly accessible, for example in a [datasheet](https://arxiv.org/pdf/1803.09010.pdf). |

### **2.2.3 Right to privacy and data protection**

The following questions explore the need for balancing high-quality datasets for AI training while ensuring that this data is collected in a way that respects privacy rights of individuals:

* What legal requirements on data protection, AI or similar do you have to consider when developing the system in your region?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + Is the data initially collected on a **lawful basis**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + How will you ensure that your data collection and the potential publication of the data is based on **informed consent**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* If your dataset contains personal data:
  + What techniques do you use to ensure that the dataset is properly **pseudonymized/ anonymized**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + Could your dataset expose individuals/commercials to **the risk of being re-identified (if the data was anonymized)**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

|  |  |
| --- | --- |
|  | **Pseudonymized data** refers to personal information that has been changed to make it more challenging, but not impossible, to identify the person it belongs to. It is still considered personal data under data protection laws because a separate key can link it back to the individual.  **Anonymized data** has had all personally identifiable information (PII) changed or removed to the extent that individuals cannot be reasonably re-identified. It ensures more privacy because there is no way to link it back to specific individuals. In comparison to pseudonymization, the process cannot be reversed. |

### **2.2.4 Fairness and non-discrimination**

The following question should be considered to address and mitigate the above forms of bias embedded in **training data** regarding the data that you have or intend to collect and to foster fairness in broader context:

* For socio-demographic data: What **demographic categories** are represented in the data (e.g., age, gender, language)?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Within the categories of your data that are relevant for your AI system, which subgroups are **more likely to be (over-)represented**? For example:
  + For socio-demographic data:   
    imbalanced representation of certain social groups, e.g. in terms of gender (men vs women vs non-binary identifying people), age (young vs old) or location (urban vs rural)
  + For agricultural data:   
    imbalanced representation of certain farming practices (that may be rather associated with men than women) or crop images from specific regions
  + How might this potential **imbalanced representation** in the data affect the quality of later outputs?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Are there any relevant **intersectional groups** to be considered in relation to your system (e.g., a transgender black person)?
  + If so, how are these groups distributed in the data?
  + Are there any factors in the data that could introduce noise related to these groups? (e.g. underrepresentation of certain groups that leads to skewed results)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Can the data be (mis-)used to target prejudiced groups, harm or **unfairly restrict access** to public services of certain groups or individuals?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What broader implications on fairness will your decisions on accessibility have:
  + How openly will you share your training dataset and models with the community, for instance through **open-source licenses** allowing others to build on your work?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + Could you improve fairness by **open sourcing your dataset** and thereby mitigate risks on community and societal level? (for example, risks relating to power concentration, technological dependency, unequal economic opportunity, resource distribution/allocation and, under-/overuse of AI) [[13]](#footnote-14)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **2.2.5 Data collection score card**

Before moving to the next stage, the scorecard below is a **guide for evaluating and summarizing the AI system’s ethical findings** concerning data collection. It provides a concise and comprehensive overview of the AI system’s ethical considerations and any actions taken in response to the assessments.

The score card can be used to communicate the findings of an AI system with team members, decision-makers, auditors and affected communities.

|  |
| --- |
| **How to use the score card:**  **Assign an ordinal rating** for each variable based on the perceived severity of the issues identified. Use the following scale. The scales are explained below with an example from an AI landslide detection use case:  **0: No issues:** e.g.: An AI system for landslide detection, demonstrates effective and representative data collection during the pre-processing phase. Robust strategies ensure equitable benefits and ethical considerations, minimizing data-related fairness concerns.  **1-2: Minor concerns:** e.g.: While addressing core data needs, minor imbalances in data ingestion and installed cameras in the pre-processing phase exist. Continuous efforts in participatory design, and compliance may contribute to mitigating these concerns, ensuring a minor impact on fairness.  **3-4: Moderate issues:** e.g.: Pre-processing introduces moderate concerns related to biases in data collection. The system may exhibit issues due to sampling and data collection favoring specific regions, and imbalances in landslide/non-landslide samples. Mitigation strategies, including comprehensive bias analysis and adjustments seem to be crucial for maintaining fairness in data handling.  **5: Significant issues:** e.g.: The system faces significant data-related fairness challenges during pre-processing, including sampling bias, data collection bias, and imbalances. Incomplete representation from temporal and regional imbalance in data collection raises concerns about accuracy and fairness, impacting risk assessment accuracy. Rigorous bias mitigation and diverse data sources are essential for equitable pre-processing and subsequent deployment.  **Provide specific comments** for each variable, explaining the rationale behind the assigned rating and include observations, concerns, or noteworthy aspects related to each variable. |

* **Name of the AI system:**
* **Evaluator name:**
* **Model date, version and type:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Principle** | **Sub-aspects per principle** | Assign **severity rating** for issues per sub-aspect(0-5) | **Why did you assign this score?** If you identified issue(s), what kind? | **How can you mitigate this issue?** (ideally filled out with your assessor[[14]](#footnote-15)) |
| Proportionality and do not harm | Data configuration |  |  |  |
| Data collection methods |  |  |  |
| Transparency and explainability | Data labelling decisions |  |  |  |
| Use of external data |  |  |  |
| Right to Privacy and Data Protection | Legal basis |  |  |  |
| Informed consent (when applicable) |  |  |  |
| Data protection ensuring purpose limitation |  |  |  |
| Fairness and Non-Discrimination | Data points |  |  |  |
| Harmful effects on data subjects |  |  |  |
| Intersectionality |  |  |  |
| Balanced representation |  |  |  |
| Mitigation of risks on community and societal level |  |  |  |

# **3. In-Processing: Model selection, training and evaluation**

The in-processing stage involves selecting, training, and adapting the model to tailor it to a specific use case and context. This process also includes evaluating the algorithm's performance and fairness from a statistical standpoint.

The main goal of assessing the in-processing stage is to pinpoint and minimize the influence of various biases—such as technical, statistical, or cognitive biases—on the AI model.

**Important note:** This part of the qualitative assessment is closely connected to the quantitative assessment guide. Building on insights from the qualitative assessment so far, the quantitative evaluation will help you assess your AI model's fairness statistically.

For the in-processing stage, this guide provides exemplary questions for the following AI principles of UNESCO:

* [Proportionality and do not harm (3.1)](#_3.1_Proportionality_and)
* [Transparency and explainability (3.2)](#_3.2_Transparency_and)
* [Fairness and non-discrimination (3.3)](#_3.3_Fairness_and)
* [Safety and security (3.4)](#_3.4_Safety_and)
* [Multi-stakeholder, adaptive governance, and collaboration (3.5)](#_3.5_Multi-stakeholder,_adaptive)

|  |
| --- |
| **Tips for tackling algorithm audits**  An auditor might ask you for:   * Access to the code * Type of method used and the justification (e.g., Linear regression, logistic regression, decision tree, SVM algorithm, Naive Bayes algorithm, KNN algorithm, K-means, random forest) * The system architecture, e.g. algorithms, machine learning models, data pipelines, and infrastructure * The model performance such as: * performance metrics (e.g., accuracy, precision, recall, F1 scores) * metrics related to computation efficiency * fairness metrics * Information on Bias and Impact: * Error and fairness metrics * Profiling categories * Vulnerable categories (see Glossary in annex) * Selection rate per category and profile * Average score for individuals and per profile in each category * Impact ratio per category and profile * Reverse-engineering results |

#### **Moments of bias in in-processing**

Each moment of the lifecycle of an algorithmic system is subject to different sources of bias. These moments of bias are not mutually exclusive nor relevant in all cases. Identifying which ones apply to an AI use case makes them easier to tackle.

Below you will find an overview of the main sources of bias for in-processing:

|  |  |  |  |
| --- | --- | --- | --- |
| **Bias** | **Definition** | **Example** | **Mitigation strategy** |
| **Over- and underfitting** | **Overfitting** (or high variance): occurs when an AI model is excessively complex. It **fits the training data too well** but performs poorly on new data.  **Underfitting** (or high bias): occurs when an AI model is **overly simplistic**. It can neither model the training data nor generalize to new data. | A model predicts crop yield based on historical data so precisely that it fails to adapt to new climate patterns.  A model oversimplifies crop yield prediction, neglecting crucial environmental factors. | * Introduce penalties for complex models to prevent overfitting. * Obtain more diverse data or reduce reliance on insufficient data. * Adjust model complexity based on data patterns. * Optimize model constraints during training to balance bias and variance. |
| **Measurement bias** | Arises from a **mismatch between training** data types/tools **and** actual **target data**. | A drone-based crop monitoring AI is trained on high-resolution satellite images. But the real-world inputs have a lower resolution. This leads to discrepancies in the recognition of crop diseases or nutrient deficiencies. | * Ensure that testing conditions align with real-world scenarios, including the resolution of measurement tools. * Capture data more frequently to account for variations. Avoid relying on infrequent measurements. * Test in lab and real settings before launch. |
| **Hot hand / Gambler's fallacy** | Arises when assuming a model will continue to perform well based solely on past success, **without proper re-evaluation or testing**. | A model has been successfully predicting high yields for several consecutive seasons. The fallacy occurs if we assume that the model will continue to successfully predict high yields, without considering changing environmental conditions or other factors. | * Reassess ground truth, variables and measurements before launch of major system updates. * Check for recent event/dynamics that may impact model performance and impacts. |
| **Aggregation bias** | Occurs when a model is **suboptimal or biased toward a specific** (e.g. dominant) **subgroup** in the data. This leads to incorrect conclusions about the group more generally. | An AI system mostly trained on data from large, techno­logically advanced farms may not work well for smallholder farms due to differences in resources and practices. | * Check for differential performance across groups. * Calculate results/rates for each protected group identified in pre-processing. * Calculate and document impact ratio per group (results for protected vs results for most salient group) and individuals (results for individuals of a protected group vs. results for individuals in the most salient group). * Define/validate fairness definitions and metrics. |

### **3.1 Proportionality and do not harm**

The following questions explore how an AI model is developed and measured. Their aim is to provide an opportunity to scrutinize the algorithm and to enhance our comprehension of the system’s requirements and potential risks.

**Exploring underlying data:**

* Have you re-examined that training, test and validation datasets are each **representative of the same underlying population/ actual target group**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What variances are you observing between the **training data versus production (live) data**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How does the model handle **outliers, over- and underrepresentation, and bias** towards certain populations?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

**Exploring model evaluation:**

* With which **metrics** have you decided **to evaluate the performance and fairness** of the model? Why these metrics?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What are the **thresholds** used by the model (if any) and how have these been selected?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What approach do you have in place to **evaluate** the model’s performance and fairness **against new and/or unseen data** (e.g. via cross-validation or bootstrapping)?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Do you perceive the **results of the AI model** as sufficient? Why is this the case?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **3.2 Transparency and explainability**

The following questions inquire about how you intend to create transparency for the AI model. The model should be as easy to understand as possible for those who are affected by it.

* How do you intend to communicate the **model’s limitations and uncertainties** to the users, stakeholders and wider public?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How are you **documenting crucial decisions of the in-processing stage** (e.g. performance and fairness metrics, thresholds, evaluation outcomes and changes)?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What techniques are you using to analyze and test **how the model arrives at its decisions?**

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What steps are you taking to make the model’s **decision-making process easy to understand** for people without a technical background?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What **mechanisms** are in place for users **to question or appeal** the model’s decisions? If so, how is this process managed and communicated to the users?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

|  |
| --- |
| As a minimum, it is recommended to **document these answers** and make them publicly accessible. |

### **3.3 Fairness and non-discrimination**

The following questions explore how assessing algorithmic fairness on trained algorithms helps with conducting **human rights-related risk mitigation**. This step is crucial before deploying the system in the real world.

* How do you **define a ‘fair outcome’** for the users in terms of the AI model’s predictions or decisions?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Based on the metrics that you selected to evaluate the fairness of the model how does the model perform:
  + **Group performance:** across different groups, especially those protected by law?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + **Differential impact**: within a protected group, i.e. does the model’s performance change when considering another factor?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + **Intersectional impact**: when multiple factors intersect, especially for protected groups?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

|  |  |
| --- | --- |
|  | **Examples for each category**, based on an AI model predicting credit scores:  **Group performance**: does the model perform equally well for men, non-binary people, and women or not?  **Differential impact**: does the model perform differently for older and younger women? And how does this compare to the performance for non-binary people or men of similar ages?  **Intersectional impact**: does the model perform differently for older women with disabilities compared to younger, non-disabled women? And how does this compare to non-binary people or men with similar age and disability status? |

* How do certain **(protected) variables**, like gender, interact with or **influence other aspects of the data**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How are you assessing whether your model is prone to **aggregation bias** that favours certain dominant patterns within the data – and may contribute to inequalities?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **3.4 Safety and security**

The following questions explore rights to safety and security to protect those who are affected by the AI system:

* What and where are system weaknesses that could make it (or the data that it stores) vulnerable to **cyber-attacks**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What are the **organizational and technical measures** to ensure the robustness and security of your system against potential attacks throughout its life cycle?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* **Robustness**: How do you intend to test the AI model in contexts like its later deployment contexts (i.e. validation experiments)?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **3.5 Multi-stakeholder, adaptive governance, and collaboration**

The following questions explore how different stakeholders have been incorporated into the model development to represent their perspectives in the final model:

* How have citizens or those affected by the AI system been **included in the development & validation process** of the AI model?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How will you incorporate **their perspectives and issues they might perceive** about the AI system into the final model before deployment?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **3.6 In-processing and score card**

Before moving to the next stage, the scorecard below is a **guide for evaluating and summarizing the AI system’s ethical findings** concerning the in-processing stage. It provides a concise and comprehensive overview of the AI system's ethical considerations and any actions taken in response to the assessments.

The score card can be used to communicate the findings of an AI system with team members, decision-makers, auditors and affected communities.

|  |
| --- |
| **How to use the score card:**  **Assign an ordinal rating** for each variable based on the perceived severity of the issues identified. Use the following scale. The scales are explained below with an exemplary AI chatbot use case, designed to answer Data Protection queries:  **0: No issues** e.g.: The chatbot ensures accurate and timely information aligned with the Data Protection Act, covering all user categories and target-languages (e.g. Luganda and English)  **1-2: Minor concerns** e.g.: The chatbot usage may have slight diversity limitations in user categories (I.e. according to their expertise, lawyers, citizens, etc.). However, efforts to expand categories, ensure dataset diversity, and define fairness goals are expected to minimize impacts.  **3-4: Moderate issues**  e.g.: Concerns arise in measuring fairness during in-processing. The model may show moderate literacy bias, affecting the accuracy of responses. Robust user feedback systems and live data monitoring could tackle these concerns.  **5: Significant issues**  e.g.: The chatbot exhibits constant problems in responding to user queries in Luganda accurately. Continuous retraining and improvement in language detection are not guaranteed to enhance performance and negatively impact the use for Lugandan speakers.  **Provide specific comments** for each variable, explaining the rationale behind the assigned rating and include observations, concerns, or noteworthy aspects related to each variable. |

* **Name of the AI system:**
* **Evaluator name:**
* **Model date, version and type:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Principle** | **Sub-aspects per principle** | Assign **severity rating** for issues per sub-aspect(0-5) | **Why did you assign this score?** If you identified issue(s), what kind? | **How can you mitigate this issue?** (ideally filled out with your assessor[[15]](#footnote-16)) |
| Proportionality and do not harm | Underlying data |  |  |  |
| Model evaluation |  |  |  |
| Transparency and explainability | Model documentation |  |  |  |
| Model communication |  |  |  |
| Fairness and non-discrimination | Differential impact/treatment |  |  |  |
| Aggregation/intersectional bias |  |  |  |
| Safety and security | Cybersecurity |  |  |  |
| Technical and organizational resources |  |  |  |
| Multistakeholder and adaptative governance | Participatory governance |  |  |  |
| Feedback integration |  |  |  |

# **4. Post-Processing: Deployment and monitoring**

The post-processing stage consists of ensuring that the AI model works as intended in the real world. Several **factors need to be reconsidered when the algorithm runs** on integrated production data, software and hardware: Once deployed, algorithms may perform very differently on data subgroups. Also, previously identified correlations between input and output may change during implementation.

Assessing the deployment stage has the following main goals:

* Continuous evaluation of the performance of the AI model in the real-word scenario
* Identify biases or harms caused during implementation and tackle them

For the post-processing stage, this guide provides exemplary questions for the following AI principles of UNESCO:

* [Human oversight and determination (4.1)](#_4.1_Human_oversight)
* [Transparency and explainability (4.2)](#_4.2_Transparency_and)
* [Responsibility and accountability (4.3)](#_4.3_Responsibility_and)
* [Awareness and literacy (4.4)](#_4.4_Awareness_and)
* [Fairness and non-discrimination (4.5)](#_4.5_Fairness_and)

|  |
| --- |
| **Tips for tackling algorithm audits**  An auditor might ask you for detailed information on:   * how deployers and users are trained * informational material available to users * mechanisms in place to evaluate the system * how environmental aspects have been considered in AI systems[[16]](#footnote-17) (co2 footprint / energy consumption, etc.) |

#### **Moments of bias in post-processing**

Each moment of the lifecycle of an algorithmic system is subject to different sources of bias. These moments of bias are not mutually exclusive nor relevant in all cases. Identifying which ones apply to an AI use case makes them easier to tackle.

Below you will find an overview of the main sources of bias for post-processing:

|  |  |  |  |
| --- | --- | --- | --- |
| **Bias** | **Definition** | **Example** | **Mitigation strategies** |
| **Deployment bias** | Occurs when organizational, budgetary, technical, or training **issues impact outputs** of the AI system, potentially influencing those affected by the decisions. | A budget-limited farmer gets fertilizer recommendations from an AI system that favors cost over optimal nutrition, leading to lower crop yields.[[17]](#footnote-18) | * Assess model performance in the real-world context. * Incorporate human oversight and ethical considerations throughout the deployment process. * Update AI system based on feedback from continuous auditing. |
| **Benchmark test bias** | This bias occurs when a model’s quality is **judged solely on** itsperformance on specific **benchmark** datasets. | An AI model for crop disease classification is evaluated solely on well-known benchmark datasets, potentially overlooking biases that may exist in real-world farm conditions. | * Consider alternative metrics and evaluation criteria beyond accuracy. * Validate model performance across diverse datasets and conditions to ensure generalizability beyond specific benchmarks. |
| **Automation bias** | Human tendency to **favor suggestions from automated (or popular) systems**, possibly ignoring correct but contradictory non-automated information. | A farmer follows an automated system’s pesticide recommendation, despite knowing a more suitable alternative. | * Staff and user training and easy-to find information on possible limitations of system. * Clear attribution of roles and responsibilities for humans and AI systems. * Establish documented procedures that justify and validate any human intervention, ensuring equal accountability for accepting or rejecting algorithmic decisions. |
| **Visualization bias** | This bias is introduced through the Framing Effect, where the **way options are presented can influence decisions**.  Related biases include dark patterns or marketing, availability bias, anchoring bias, and signal error, which may result from how data is visualized and interpreted. | A visually appealing crop yield dashboard may lead farmers to focus disproportionately on certain crops, neglecting other important decision-making factors. | * Awareness and education: Ensure that individuals who oversee data are aware of visualization biases and cognitive effects. * At a more technical level: causal methods (like directed acyclic graphs) may provide a more nuanced understanding of data. |

### **4.1 Human oversight and determination**

The following questions explore the actual establishment of **AI governance and protocols** oriented towards ongoing mitigation of automated decisions and other AI-related risks:

* How do you intend to regularly **evaluate the system**, i.e. its outcomes and impacts? (e.g. via user feedback, production data, AI audits)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How regularly will the system be **revalidated for fairness** **and performance** and by which team or group of experts?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How will you check how people and the AI system interact so that the **system works well for users** and can be managed effectively (e.g. ensuring that the AI system functions properly)?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How will you ensure that **feedback** from any kind of ongoing monitoring and evaluation will be **fed back into training and re-measuring** the model performance and fairness?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **4.2 Transparency and explainability**

The following questions explore how those affected by an AI system are properly informed about its social implications, inputs and outcomes:

* How will the system disclose and make transparent that it is **based on AI**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How will the AI system provide clear and understandable information about its **purpose, use and limitations**?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How will the AI system provide feedback to users on **how it arrived at a certain output** (e.g. prediction or recommendation)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Which parts of the system are open for independent third-party checks (e.g. training data, weights, model, deployment code etc.), among others to check for not-yet-found biases?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Which parts of the system are available under open conditions (e.g. open-source licenses) that allow others to make changes and contribute to a more inclusive system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How do you intend to provide detailed and easy-to-understand **information on the AI system’s development** from the pre- to the post-processing stage? How will you make this information openly available?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How will you **assess the** **explainability** of the system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How will you ensure that **(future) changes and updates** made to the model are transparent and documented for the wider public or for regulatory review?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

|  |
| --- |
| As a minimum, it is recommended to **document these answers** and make them publicly accessible. |

### **4.3 Responsibility and accountability**

The following questions explore how to consider accountability in case someone should get harmed by an AI system. In this regard, understanding and defining accountability principles is crucial to ensure **meaningful trust in AI systems**.

* How can individuals **object to a decision** that is based on an AI system?
  + If so, is this process transparent, accessible and smooth?
  + Is there a body or organization that they can defer to for assistance with an objection?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How can the actors that develop/ deploy the AI system be **held accountable** for harm that users of such systems may suffer?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What mechanisms and measures have you put in place to enable those affected (negatively) by an AI system to **exercise their rights?**For instance: rights to data access, rectification or deletion of personal data, new obligations for providers established in the AI Act (including testing, documentation, transparency, and notification duties)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* **Human-in-the-loop accountability:** How do you ensure that those who operate the AI system stay responsible and can influence the outputs/results of the AI system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **4.4 Awareness and literacy**

The following questions explore how people are **informed and trained about the AI system**:

* How will you provide **training or instructions** to the end users so that they are aware of and can navigate its operation, capabilities and limitations?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How are those affected by the AI system educated on its operation, capabilities and limitations?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Will you provide specific **accessibility aid** for this information when it comes to protected groups? (e.g. audio information for people with hearing impairment)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* For those who will use the AI system: How are you ensuring that people are aware of how their **input data** will be stored or further used?
  + Do you provide this information in easy-to-understand language for a non-technical audience?
  + Is there an option to opt out of (personal) data being stored or processed?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **4.5 Fairness and non-discrimination**

The following questions explore how you can **consider and re-evaluate aspects of fairness in the deployment stage**, particularly unfair and/or disadvantageous treatment of protected and vulnerable groups.

* How are you evaluating whether the AI system provides **equal treatment** and **inclusive access** with respect to protected groups, in terms of multilingualism, gender, age, disability, cultural and other factors?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + If certain protected groups are disproportionally negatively affected or unfairly treated by the AI systems: Which ones are affected and how?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + How have you examined the perception of these groups concerning the system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + Would the system be accessible and **usable by vulnerable groups**, e.g.people with disabilities like visual or hearing impairment?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* How will you ensure that users can provide **easily and accessibly feedback** on the model’s performance and impact?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + How are you ensuring you collect feedback from protected groups, particularly to identify where the AI system might be mistreating certain groups?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + How is this feedback incorporated into the model’s development and refinement?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* Once deployed, can users **introduce biases into the algorithm** (for instance, by inputting biased data)?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + If so, to what extent might the altered model have a negative impact on users and beneficiaries of the system?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* What broader implications on fairness will your decisions on accessibility have:
  + Could you improve fairness by **open-sourcing the system** (datasets, labeling sets, model weights, deployment code)and thereby mitigate risks on community and societal level? (for example, risks relating to power concentration, technological dependency, unequal economic opportunity, resource distribution/allocation and, under-/overuse of AI) [[18]](#footnote-19)

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

* + What **types of operating models** (financial models, business models) could you use to maximize permanence of the system and its parts, ensuring continuous improvement and maximum societal and operational benefit?

|  |
| --- |
| Klicken oder tippen Sie hier, um Text einzugeben. |

### **4.6 Post-processing score card**

This scorecard below is a **guide for evaluating and summarizing the AI system’s ethical findings** concerning post-processing before moving to the next stage. It provides a concise and comprehensive overview of the AI system's ethical considerations and any actions taken in response to the assessments.

The score card can be used to communicate the findings of an AI system with team members, decision-makers, and affected communities.

|  |
| --- |
| **How to use the score card:**  **Assign an ordinal rating** for each variable based on the perceived severity of the issues identified. Use the following scale. The scales are explained below with an exemplary AI landslide detection system:  **0: No issues** e.g.: All regions covered by an AI-based crop mapping system are equally represented in the training data. The AI model is expected to consistently provide accurate crop-type predictions across various geographical areas without any bias.  **1-2: Minor concerns** e.g.: There is a slight imbalance in the representation of certain crops in the training data, but the impact on predictions is expected to be minimal. Therefore, the team considers that the system will generally perform well across different crop types.  **3-4: Moderate issues**  e.g.: The AI model is suspected to be moderately biased in predicting crop types, with certain regions experiencing lower accuracy rates. This may be due to identified variations in data quality or insufficient representation of specific geographical features.  **5: Significant issues**  e.g.: Identified unbalanced data representation indicates significant bias in predicting crop types, with certain regions expected to consistently receive inaccurate predictions. Following the preliminary assessment, this may lead to unfair resource allocation or policy decisions, impacting the agricultural practices and livelihoods of affected communities.  **Provide specific comments** for each variable, explaining the rationale behind the assigned rating and include observations, concerns, or noteworthy aspects related to each variable. |

* **Name of the AI system:**
* **Evaluator name:**
* **Model date, version and type:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Principle** | **Sub-aspects per principle** | Assign **severity rating** for issues per sub-aspect(0-5) | **Why did you assign this score?** If you identified issue(s), what kind? | **How can you mitigate this issue?** (ideally filled out with your assessor[[19]](#footnote-20)) |
| **Human oversight and determination** | System regular evaluation |  |  |  |
| Feedback loop integration |  |  |  |
| Human management |  |  |  |
| **Transparency and explainability** | Data disclosure and communication |  |  |  |
| External assessment of explainability |  |  |  |
| Regular and updated communication |  |  |  |
| **Responsibility and accountability** | Users rights exercise tools |  |  |  |
| Object decisions protocols |  |  |  |
| Accountability mechanisms |  |  |  |
| **Awareness and literacy** | Accessibility |  |  |  |
| End-user training |  |  |  |
| Open information on data management |  |  |  |
| **Fairness and non-discrimination** | Equal treatment assessment |  |  |  |
| Data release on system performance |  |  |  |
| Regular evaluation of data input and bias |  |  |  |
| Maximizing benefits and minimizing risks on community and societal level |  |  |  |

# **5. REFERENCES**

Barocas, S., & Selbst, A. D. (2016). Big data’s disparate impact. *California Law Review,* 104(3), 671-732.

Castillo, C. (2018). Assessing the impact of machine intelligence on human behaviour: an interdisciplinary endeavor. In *Proceedings of HUMAINT Workshop*. <https://arxiv.org/pdf/1806.03192.pdf>.

Caton, S., & Haas, C. (2020). Fairness in machine learning: a survey. *ACM Computing Surveys.* <https://dl.acm.org/doi/full/10.1145/3616865?af=R>

Criado, N., Ferrer, X., & Such, J. M. (2020). *A normative approach to attest digital discrimination*. ArXiv.org. <https://arxiv.org/abs/2007.07092>

Eticas, & Castillo, C. (2021). *Guide to algorithmic auditin*g. Eticas. <https://eticas.ai/wp-content/uploads/2024/04/Guide-to-Algorithmic-Auditing-EN.pdf>

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K (2020). Datasheets for datasets. C*ommunications of the ACM,* 64, 86-92. <https://arxiv.org/abs/1803.09010>.

Haijan, S., Bonchi, F., Castillo, C. (2016). Algorithmic bias: from discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Digital Library. <https://doi.org/10.1145/2939672.2945386>.

High-level Advisory Body on Artificial Intelligence (2023). *Interim report: governing AI for humanity.* United Nations. <https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf>.

Ho, D. A., & Beyan, O. (2020). *Biases in data science lifecycle*. arXiv.org. <https://arxiv.org/abs/2009.09795>.

Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T. (2019). *Quantifying the carbon emissions of machine learning*. ArXiv.org. <https://arxiv.org/abs/1910.09700>.

Laato S., Tiainen M., Najmul A., Mäntymäki M. (2022). *How to explain AI systems to end users: a systemic literature review and research agenda*. Emerald Insights, 32(7), 1-31. <https://doi.org/10.1108/intr-08-2021-0600>

Lippert-Rasmussen, K. (2013). *Born free and equal? A philosophical inquiry into the nature of discrimination*. Oxford University Press.

Mantelero, A., Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review*, 41. <https://doi.org/10.1016/j.clsr.2021.105561>

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., Gebru, T. (2018). Model cards for model reporting. In *FAT\* '19: Conference on Fairness, Accountability, and Transparency.* <https://arxiv.org/abs/1810.03993>.

Mitchell, S., Potash, E., Solon, B., D’Amour, A., Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Applicatio*n, 8(1), 141-163. <https://doi.org/10.1146/annurev-statistics-042720-125902>

O’Neill, C., Gunn, H. (2020). Near-term artificial intelligence and the ethical matrix. In S. Matthew Liao (Eds.), *Ethics of artificial intelligence,* 237-270. Oxford University Press. <https://doi.org/10.1093/oso/9780190905033.003.0009>

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 560-568. <https://doi.org.10.1145/1401890.1401959>.

Prabhakaran, V., Mitchell, M., Gebru, T., Gabriel, I. (2022). *A human rights-based approach to responsible AI*. ArXiv.org. <https://arxiv.org/abs/2210.02667>.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 33-44. Association for Computing Machinery.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: definition and background. In *Mission AI, Research for policy,*15-41. Springer, Cham. <https://doi.org/10.1007/978-3-031-21448-6_2>

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*.

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings* of *Equity and access in algorithms, mechanisms, and optimization*. <https://doi.org/10.1145/3465416.3483305>

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>

United Nations. (1948). *Universal declaration of human rights*. <https://www.un.org/en/about-us/universal-declaration-of-human-rights>

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 8919-8928.

# **6. Biographies: Community of AI Inclusion Experts**

[**Favour Borokini**](https://ng.linkedin.com/in/favourborokini)

Favour Borokini is an expert on data and digital rights with experience in gender and responsible AI. In her vast professional experience, Favour has worked in multiple organizations like Pollicy, The Future Society, Cumberland Lodge, Tech Hive, and more. Favour is focused on the relationship of humans with technology, and how technology can be creative, inclusive and can improve the quality of human life.

[**Josia Paska Darmawan**](https://www.linkedin.com/in/paska/?originalSubdomain=id)

Josia Paska is a research activist and social data analyst. Their main interests include diversity, equity, inclusion, digital justice, marginalized community empowerment, and more. Josia has experience in international institutions like the Center for Digital Society, Fairwork Foundation, and more.

[**Mohamed Kimbugwe**](https://ug.linkedin.com/in/mohamed-kimbugwe-56741044)

Mohamed is an international development professional with expertise in human-centred digital transformation, and the intersectionality of digital systems and gender, human rights and social impact. His professional experience has been in multiple institutions as an advisor in GIZ, consultant in Light for the World, co-founder of Silent World Foundation, and more. Mohammed is interested in the research on gender, bias in automated decision-making processes, and more.

[**Mercy King'ori**](https://ke.linkedin.com/in/mercy-king-ori-00a411101)

Mercy King’ori is a technology and policy researcher working as a Policy Analyst for Africa at the Future of Privacy Forum, an organization that advances responsible data practices in support of emerging technologies.

[**Meena Lysko**](https://mbc.lysko.com/index.htm)

Meena Lysko has experience in the application of remote sensing techniques, ethics, and AI applications in agriculture. During her vast professional experience, Meena has worked as an international consultant, as director of Move Beyond Consulting, as a researcher at the Council for Scientific and Industrial Research, and more.

[**Raashi Saxena**](https://www.linkedin.com/in/raashi-saxena-18a90684/?originalSubdomain=in)

Raashi Saxena is a public interest technologist based out of Bangalore. Her expertise lies in digital accessibility, digital governance and its implications on human rights, and their application on artificial intelligence systems. During her professional experience, Raashi has collaborated with multiple institutions including Superbloom, Accessibility Lab, The Sentinel Project, Studio intO and more.

[**Kofi Yeboah**](https://gh.linkedin.com/in/kofiyeboah)

Kofi Yeboah is an experienced consultor and researcher in multiple international organizations, startups, and social enterprises like HOPin Academy, Mozilla, Paradigm Initiative, University of Alberta, and more. His main interest is focused on digital rights, governance and development of Sub-Saharan Africa.



Deutsche Gesellschaft für

Internationale Zusammenarbeit (GIZ) GmbH

Registered offices

Bonn und Eschborn

Friedrich-Ebert-Allee 32 + 36 Dag-Hammarskjöld-Weg 1-5

53113 Bonn, Deutschland 65760 Eschborn, Deutschland

T +49 228 44 60-0 T +49 61 96 79-0

F +49 228 44 60-17 66 F +49 61 96 79-11 15

E fairforward@giz.de

I www.giz.de

1. Implemented on behalf of the BMZ, FAIR Forward strives for a more open, inclusive and sustainable approach to AI on a local and global level. To achieve this, FAIR Forward is working together with seven partner countries (Ghana, India, Indonesia, Kenya, Rwanda, South Africa and Uganda). [↑](#footnote-ref-2)
2. Please note that the Responsible AI Assessment is an assessment method. It is suitable for those who want to conduct a holistic AI risk and ethics assessment by themselves or with an external assessor (recommended option). It is not an audit and cannot guarantee compliance with upcoming standards (e.g. AI EU Act). Please engage certified auditors if you want to ensure compliance. [↑](#footnote-ref-3)
3. An AI system can be understood as any software that is developed with: one or more AI-related techniques, like machine learning, deep learning, statistical models, Bayesian estimation or any other method that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with. See also [Sheikh et al.](https://link.springer.com/chapter/10.1007/978-3-031-21448-6_2) (2023) and [Laato et al.](https://www.emerald.com/insight/content/doi/10.1108/INTR-08-2021-0600/full/html) (2022). [↑](#footnote-ref-4)
4. The Recommendations on the Ethics of AI covers the following principles:

   * Proportionality and do-not-harm
   * Safety and security
   * Fairness and non-discrimination
   * Sustainability
   * Right to privacy and data protection
   * Human Oversight and Determination
   * Transparency and explainability
   * Responsibility & accountability
   * Awareness and literacy
   * Multi-stakeholder, adaptive governance and collaboration

   [↑](#footnote-ref-5)
5. I.e. pre-processing (encompassing problem formulation and data collection), in-processing (model selection and development), and post-processing stage (deployment). [↑](#footnote-ref-6)
6. It should be noted, though, that there are overlaps in how the principles are operationalized. In general, the principles cannot always be clearly separated, they also blur into each other. [↑](#footnote-ref-7)
7. This includes directly affected stakeholders (e.g. end-users, deployers or beneficiaries), but also indirectly affected stakeholders. [↑](#footnote-ref-8)
8. Protected individuals belong to groups that have historically faced systemic discrimination or marginalization. These groups are identified by characteristics such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth, disability, age, sexual orientation, and gender identity, among others. This list is non-exhaustive and might change according to the (local) context.

   The identification of protected groups is guided by international and national legislations, among others the [Universal Declaration of Human Rights](https://www.un.org/en/about-us/universal-declaration-of-human-rights) (1948). In Machine Learning, it is particularly important to consciously assess the effects of AI that might exacerbate these harms and discrimination, in order to mitigate their perpetuation. [↑](#footnote-ref-9)
9. The following table is inspired by [Cathy O’Neil & Hanna Gunn](https://academic.oup.com/book/33540/chapter-abstract/287905687?redirectedFrom=fulltext&login=false) (2021). If you would fill this table out for an agricultural AI use case that recommends measures for diseased crops, you could consider to following stakeholders, among others: Female smallholder farmers (directly affected), male smallholder farmers (directly affected), customers from local food markets (indirectly affected), environment (directly affected). [↑](#footnote-ref-10)
10. The person who is responsible to assess the AI use case. This person should have expertise in assessing AI systems for harms, ethical risks and bias (e.g. as a certified AI auditor). [↑](#footnote-ref-11)
11. For example, if relevant for your AI use case: you could look into how the GDPR’s ban on profiling, which limits how AI systems’ can process certain sensitive data, has been interpreted in court. This legal insight could then be factored into the system’s design. [↑](#footnote-ref-12)
12. The person who is responsible to assess the AI use case. This person should have expertise in assessing AI systems for harms, ethical risks and bias (e.g. as a certified AI auditor). [↑](#footnote-ref-13)
13. See [High-level Advisory Body on Artificial Intelligence](https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf) (2023, p. 10f) [↑](#footnote-ref-14)
14. The person who is responsible to assess the AI use case. This person should have expertise in assessing AI systems for harms, ethical risks and bias (e.g. as a certified AI auditor). [↑](#footnote-ref-15)
15. The person who is responsible to assess the AI use case. This person should have expertise in assessing AI systems for harms, ethical risks and bias (e.g. as a certified AI auditor). [↑](#footnote-ref-16)
16. Two core questions must be considered in this regard:

    How can the algorithm be optimized to reduce energy consumption and minimize its carbon footprint?

    What measures are in place to ensure the sustainability of the data centers hosting the AI infrastructure? [↑](#footnote-ref-17)
17. If a model would be developed to optimise outputs for cost over nutrition, such rationales should be communicated to farmers a) in a language and manner they understand and b) explicitly when using the system (e.g. not hidden in legal jargon). [↑](#footnote-ref-18)
18. See [High-level Advisory Body on Artificial Intelligence](https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf) (2023, p. 10f) [↑](#footnote-ref-19)
19. The person who is responsible to assess the AI use case. This person should have expertise in assessing AI systems for harms, ethical risks and bias (e.g. as a certified AI auditor). [↑](#footnote-ref-20)