

**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 C: Quantitative 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 C: Quantitative 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](#_Toc168996435)

[1.1 About the Responsible AI Assessments 1](#_Toc168996436)

[1.2 About this Quantitative Guide 2](#_Toc168996437)

[1.3 Structure of the Qualitative Guide 2](#_Toc168996438)

[2. PREPARATORY WORK: Initial mapping and scoping 5](#_Toc168996439)

[Step 1: Fill out a Model Card (MC) 5](#_Toc168996440)

[Step 2: Based on Qualitative Guide, create initial bias and harm hypotheses 6](#_Toc168996441)

[3. INTRO TO MOMENTS OF BIAS 8](#_Toc168996442)

[4. QUANTITATIVE ASSESSMENT IN PRE-PROCESSING 9](#_Toc168996443)

[Step 1: Complete the Pre-Processing Stage of the Qualitative Guide 9](#_Toc168996444)

[Step 2: Complete MC sections for “data details” and “contextual information” 10](#_Toc168996445)

[Step 3: Evaluate your data 11](#_Toc168996446)

[Step 4: Detect biases 12](#_Toc168996447)

[A) Common moments of bias within the problem formulation stage 12](#_Toc168996448)

[B) Common moments of bias within the data collection stage 13](#_Toc168996449)

[5. QUANTITATIVE ASSESSMENT IN IN-PROCESSING 15](#_Toc168996450)

[Step 1: Go through the In-Processing Stage of the Qualitative Guide 15](#_Toc168996451)

[Step 2: Complete MC sections for “model details” and “bias and impact” 16](#_Toc168996452)

[Step 3: Calculate baseline performance metrics 16](#_Toc168996453)

[A) Confusion matrix 16](#_Toc168996454)

[B) Based on the confusion matrix, calculate basic performance metrics 17](#_Toc168996455)

[Step 4: Select and apply additional methods 20](#_Toc168996456)

[A) Evaluation metrics 20](#_Toc168996457)

[B) Algorithm training and hyperparameter tuning 21](#_Toc168996458)

[C) Parsimonious Principle 23](#_Toc168996459)

[Step 5: Fill out Score Card for additional methods 24](#_Toc168996460)

[Step 6: Select Fairness Metrics 25](#_Toc168996461)

[Step 7: Reassess 30](#_Toc168996462)

[Step 8: Detect biases 30](#_Toc168996463)

[6. QUANTITATIVE ASSESSMENT IN POST-PROCESSING 32](#_Toc168996464)

[Step 1: Go through the Post-Processing stage of the Qualitative Guide 32](#_Toc168996465)

[Step 2: Update MC and complete “redress measures” section 33](#_Toc168996466)

[Step 3: Generate documentation 33](#_Toc168996467)

[A) Training and outcome codebooks 33](#_Toc168996468)

[B) Commented coding script(s) 34](#_Toc168996469)

[C) Compilation of insights from applying the Qualitative and Quantitative Guides 34](#_Toc168996470)

[Step 4: Monitor dashboard (for internal use) 34](#_Toc168996471)

[Step 5: Visualization dashboards (for public use) 35](#_Toc168996472)

[Step 6: Reassess 36](#_Toc168996473)

[Step 7: Detect biases 36](#_Toc168996474)

[7. REFERENCES 38](#_Toc168996475)

[8. BIOGRAPHIES: Community of AI Inclusion Experts 42](#_Toc168996476)

# INTRODUCTION

## 1.1 About the Responsible AI Assessments

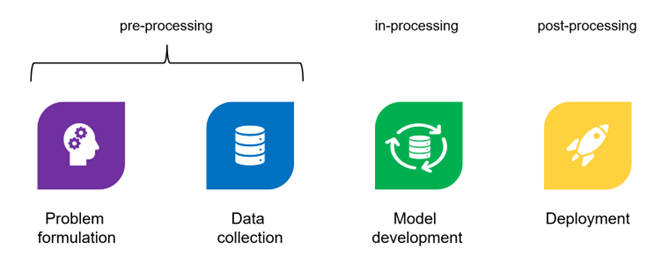
As AI technologies evolve, so does the imperative to ensure that their **development and deployment align with human rights principles** and avoid causing harm or perpetuating social inequalities. The **Responsible AI Assessments** are a proactive response to address these challenges head-on, co-created 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](#_8._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 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 Quantitative Guide

This Quantitative AI Risk Assessment Guide orientates you in evaluating biases and fairness-related issues in data for AI and AI models from a quantitative perspective. Its focus is on identifying differential impact, i.e. how an AI model might behave differently for different groups.[[3]](#footnote-4) This Quantitative Guide has been developed and tested based on AI use cases that have been supported by FAIR Forward in its partner countries Ghana, India, Indonesia, Kenya, Rwanda, South Africa, and Uganda. These use cases primarily involve AI models that are based on supervised learning.

It aims to also inspire other stakeholders on how to enhance AI fairness in technologies to critically reflect on aspects like:

* **Representation issues** in input data (e.g. gender bias, vulnerable groups).
* **Appropriateness** of the data for modelling, regarding the purpose and objective of the AI system.
* **Mathematical performance and fairness** of an AI model for individual subgroups.
* Methods for the **mitigation of technical issues**.

Because no standard exists yet for how to assess the different types of AI (e.g. narrow or generative AI) and algorithms (i.e., supervised, unsupervised, or reinforcement learning), this document aims to provide an **overview on common steps to quantify and measure** **differential impact** in AI systems.

## 1.3 Structure of the Qualitative Guide

The Quantitative Guide is **structured around** the **AI lifecycle** and organized into the following sections:

1. [Preparatory work: initial mapping and scoping](#_2._Preparatory_Work:)
2. Quantitative assessment in:
   1. [Pre-processing](#_4._Quantitative_Assessment) (problem formulation and data collection)
   2. [In-processing](#_5._Quantitative_Assessment) (model development)
   3. [Post-processing](#_6._Quantitative_Assessment) (deployment and monitoring)

For each stage within an AI lifecycle[[4]](#footnote-5), the guide provides:

* An info box representing exemplary questions to answer during each stage
* An overview of common methods and metrics.
* Practical examples to illustrate when to apply certain metrics or methods.

|  |
| --- |
| This Quantitative Guide simplifies quantitative terms to make them accessible for a broader audience. It aims for general understanding, not academic precision. Additionally, the guide introduces common quantitative audit methods and metrics, but it is not exhaustive. Consider it a starting point for further research. Wherever possible, work with an experienced assessor or auditor for a tailored quantitative assessment of your AI use case. |

Before using the Quantitative Guide, it is strongly recommended that users go through relevant lifecycle sections in the [Qualitative Assessment Guide](https://gizonline.sharepoint.com/teams/AI-Labwithguests/Freigegebene%20Dokumente/HF-REGIO%20%C3%9Cbergreifend%20und%20Vernetzung%20der%20L%C3%A4nder%20un/26%20Activity_Q_of_Global_Regional%20-%20Risk%20board/Final%20documents/Qualitative%20AI%20risk%20assessment%20tool--%20final%20draft2.docx?web=1). This allows them to better comprehend the following parts as a preparation for a quantitative assessment, for example:

* Problem definition
* Key vulnerable/ protected groups (directly or indirectly affected)
* Potential human rights risks
* Potential key issues within the pre-, in- or post-processing stage (see score cards)
* … and more.

|  |
| --- |
| **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.  **If you have any feedback**, **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 tables*** *meant to be filled out* ***can be edited****.*  *If you encounter issues, please reach out to* [*fairforward@giz.de*](mailto:fairforward@giz.de)*.* |

# PREPARATORY WORK: Initial mapping and scoping

To prepare the quantitative analysis, the first stage focuses on revisiting the rationale and theories behind an intended AI use case. For this, the following steps are recommended:

1. Fill out a Model Card (MC)
2. Based on Qualitative Guide, create initial bias and harm hypotheses

## Step 1: Fill out a Model Card (MC)

The Model Card (MC) provides a structured compilation of general information about an algorithmic system, its context and use. It can also be a helpful tool for any future auditor who might assess your AI system.

The MC can be filled with the outputs of the Qualitative Guide or completed from scratch. The information compiled in the MC will guide the assessment in the different lifecycle stages. Thus, it is recommended to revisit the model card regularly throughout the AI lifecycle to fill out empty sections.

The next table provides an exemplary template for an MC. The final structure and content of the MC should be tailored to the specific needs and context of the respective AI system and its users.

|  |  |
| --- | --- |
| **General information** | **Contextual information** |
| * Who is developing the prediction, classification, or causal model (internal team, external provider, external company, etc.)? * At what stage of development is the system? * What problem is the system designed to solve? * Does the system have any previous versions, upgrades, or developments? * What is the architecture of the system? | * What is the purpose of the algorithm? * What specific task or function does the algorithm perform in this project? * Which direct and indirect stakeholders would be affected by the algorithm and how? * What potential risks could arise from using this algorithm in this context? * What is the role of humans in the functioning and application of the algorithm? |
| **Data details** | **Model details** |
| * What are the data sources? * If using a proprietary method, what was the methodology used? * Is there a codebook with variable type, description, temporality, and variable name? * What groups of data/people are represented in the data? (particularly protected groups) * What methods are in place to ensure the quality of the data? * What type of database is used? * What data preprocessing steps have been performed before model training? | * What machine learning methods or algorithms are used in the model? * What kind of output does the model produce? * What variables were used for the model? * What parameters were used? * What specific thresholds or conditions were defined? |
| **Bias and impact** | **Redress measures** |
| * What mitigation measures have been undertaken to reduce system and/or model error? * Which population or sample groups were considered for the system and/or model? * What metrics were used to measure system performance? * Were specific metrics used to measure differences in output concerning population or sample groups? * Are the results expected? | * What measures are used to adjust and/or correct the system? * What measures are in place to monitor and update the system? |

## Step 2: Based on Qualitative Guide, create initial bias and harm hypotheses

Before continuing, we strongly recommend building on your insights from the score cards in the Qualitative Guide. The hypotheses of bias and harm that you define there for the pre-, in- and post-processing stages will respectively help you with completing the Quantitative Guide.

After completing the model card, you can come up with some initial hypotheses about potential biases involving specific groups and environments at the pre-processing, in-processing and post-processing phases.

|  |
| --- |
| **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. |

# INTRO TO MOMENTS OF BIAS

Each moment of the lifecycle of an algorithmic system (pre-processing, in-processing and post-processing) is subject to different sources of bias. Within this quantitative guide, bias mainly refers to any unfairness or discrimination of an AI system, which results in individual or collective harm, due to its underlying data or the model. This chapter introduces methods to identify moments that may introduce biases throughout the different lifecycle stages of an algorithmic system:

* Pre-processing (i.e. problem formulation and data collection)
* In-processing
* Post-processing

The following figure provides an overview of biases for each lifecycle stage of an algorithmic system:

A screenshot of a computer screen

Description automatically generated

Figure 1: © Eticas

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.

# QUANTITATIVE ASSESSMENT IN PRE-PROCESSING

The pre-processing involves various steps to collect, clean, transform, select, and prepare the raw data before it is used for model training, development, and testing**.**

During this phase, quantitative analysis and testing should answer the following questions:

|  |
| --- |
| * Are the **training data/datasets** used in the system **relevant to the issue** to be solved? * Is there a possibility that certain groups or scenarios are **underrepresented** or absent in the training data due to limited availability or lack of structure? * Was there evidence of **non-representative data favouring** certain salient groups over minority groups? * Was an assessment made to determine whether the training data's characteristics match those of the **intended deployment environment**? * Is there a process to ensure that conclusions drawn from the AI system's analysis are not disproportionately influenced by **outlier data points or omitted variables**? |

For the pre-processing stage, the following steps are recommended:

1. [Complete the Pre-Processing Stage of the Qualitative Guide](#_Step_1:_Complete)
2. [Complete MC sections for “AI process” and “training/validation data”](#_Step_2:_Complete)
3. [Evaluate your data](#_Step_3:_Evaluate)
4. [Detect of biases](#_Step_4:_Detect)

## Step 1: Complete the Pre-Processing Stage of the Qualitative Guide

To prepare the quantitative assessment, it is strongly recommended to **build on your score card for the pre-processing stages** of the Qualitative Guide. From the Qualitative Guide, it is particularly recommended to define the following aspects in relation to the local context for which you plan to develop an AI system. This will assist you in conducting the quantitative assessment:

* The **problem** that you want to solve.
* **Stakeholders affected** by your AI system.
* **Configuration of your data**, this includes:
  + **Collection and quality of data**. This includes considering the possible biases that can percolate in the data acquisition, origin, or methodology.
  + **Sub-groups** into which your data can be sorted[[5]](#footnote-6). Based on these groups, you will later assess how an algorithm may treat or impact these groups differently.
  + **Protected groups** eitherwithin your data or affected by the system that you build[[6]](#footnote-7).
* **Decisions** that have been made and have an impact on:
  + Data origin/collection and its representativeness of the target population.
  + Data labelling and potential labelling bias.
  + Data pre-processing, incl exploratory analysis

|  |
| --- |
| **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. |

## Step 2: Complete MC sections for “data details” and “contextual information”

If you have not yet filled out these details for the Model Card, please do so now.   
You can follow this [internal link to the Model Card](#_Step_1:_Fill).

## Step 3: Evaluate your data

This step involves various evaluation methods to assess whether your data is representative. Depending on your use case, different methods will apply. Still, each method provides a different perspective to spot potential problems in your data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Analysis** | **Definition** | **Relevant to your use case? Explain your answer** | **What actions can be taken to mitigate the challenges?** |
| **Sample-Population studies of represen­tativeness** | Used in any kind of model.  It determines the scope and quality of the data to ensure it represents real-world scenarios. It involves identi­fying different groups, spatial distribution, and subdivisions within the data. |  |  |
| **Anonymization of data** | Used in systems that require personal/sensitive data or capture it during the collection process.  It involves applying ­pseudonym­mization techniques like permutation, suppression, or anonymization to protect personal or sensitive information.[[7]](#footnote-8) Timely identification of personal or sensitive data can prevent future privacy, compliance, and data processing issues. |  |  |
| **Accuracy of labeling** | Used for identifying issues or imbalances in the labeling of datasets. [[8]](#footnote-9) It focuses on measuring the labeling accuracy between datasets |  |  |
| **Correlation matrix** | Used in any kind of model.  Visualizes the relation between variables. It helps identify variables that might induce endogeneity or violate the principle of parsimony (simplicity). It shows how variables move together and their interrelationships. |  |  |
| **Model endogeneity test** | Mostly used on econometric modelling.  It checks the assumption that the explanatory variables (causes) are independent and not correlated with the dependent variable (effect). |  |  |
| **Error modelling** | Used in any kind of model.  Aims to understand the distribution and relationship of errors within the model. It can help detect omitted variables, undesirable relationships, or incorrect model selection. |  |  |

## Step 4: Detect biases

Below you will find common biases within the pre-processing stage and exemplary strategies for testing and mitigating them. You can analyse which biases you perceive as relevant for your AI use case.

### A) Common moments of bias within the problem formulation stage

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Bias Type** | **Definition** | **Exemplary testing** | **Relevant for your use case? Please explain your answer** | **If relevant for your use case: How will you mitigate this bias?**  (Ideally filled out with your assessor[[9]](#footnote-10)) |
| **Techno-solutionist Bias** | **Over-reliance on high-tech solutions** without considering simpler alternatives or potential social and environmental impacts. | * Analyze alternatives. * Assess expectations of developers vs. performance in other relevant use cases. * Check desirability and compliance. * Evaluate environmental impact. * Study social impacts. | Klicken oder tippen Sie hier, um Text einzugeben. |  |
| **Data Availability/ Scarcity** | The selected **datasets** are **not relevant** to the problem, or there is insufficient data to represent the real-world problem. | * Check available datasets for relevance. * Assess labels and features. * Ensure data minimization compliance. * Analyze statistical differences between groups. |  |  |
| **Historical Bias** | The data used for training the AI system reflects **existing social biases**, leading to potentially harmful outcomes. | * Perform contextual analysis and identify protected groups. * For Natural Language Processing (NLP): validate word embeddings. * 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. | * Identify affected stakeholders and protected groups. * Check distribution differences of groups. * Perform randomization testing. |  |  |

### B) Common moments of bias within the data collection stage

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Bias Type** | **Definition** | **Exemplary testing** | **Relevant for your use case?** Please explain your answer | **If relevant for your use case: How will you mitigate this bias?**  (Ideally filled out with your assessor[[10]](#footnote-11)) |
| **Labeling Bias** | Inaccurate data labeling due to subjective perceptions. | * 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.** | * 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. | * Address issues in data collection method, e.g. for questions where people are hesitant. * Compare and contrast data from different sources. * Use probabilistic matching and individual reference identifiers for dataset harmonization. * Minimize likelihood for recall bias through survey design. |  |  |
| **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. | * 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. |  |  |

# QUANTITATIVE ASSESSMENT IN IN-PROCESSING

The **in-processing phase is related to selecting, training and evaluating the AI model**. Due to the amount of possible metrics and evaluation methods, this phase is one of the most complex ones to assess. During this phase, quantitative analyses and testing should aim to answer the following questions, based on the data that is used to train and test the AI model:

|  |
| --- |
| * Is there a systematic approach in place to evaluate the **model's performance** against new and unseen data? * Are the chosen **evaluation metrics** focused on not just training accuracy but also the model's ability to perform fair on new and diverse data? * Is there an assessment process to identify any unjustified inequalities or misleading trends that may arise due to **aggregation bias**? * Is the AI system evaluated for **fairness and performance metrics separately for each protected group** identified during pre-processing? * Is there a procedure in place to perform **explainability** analysis and testing to understand the reasons behind the model and the decision-making process?[[11]](#footnote-12) |

For the in-processing stage, the following steps are recommended:

1. [Go through the In-Processing Stage of the Qualitative Guide](#_Step_1:_Go)
2. [Complete MC sections for “AI model” and “bias and impact”](#_Step_2:_Complete_1)
3. [Calculate baseline performance metrics](#_Step_3:_Calculate)
4. Select and apply additional methods.
   1. [Evaluation metrics](#_A)_Evaluation_metrics)
   2. [Algorithm training and hyperparameter tuning](#_B)_Algorithm_training)
   3. [Parsimonious principle](#_D)_Parsimonious_Principle)
5. [Fill out Score Card for additional methods](#_Step_5:_Fill)
6. [Select and analyze fairness metrics](#_Step_6:_Select)
7. [Reassess](#_Step_7:_Reassess)
8. [Detect biases](#_Step_7:_Detect)

## Step 1: Go through the In-Processing Stage of the Qualitative Guide

During the in-processing stage, it is strongly recommended to go through the Qualitative Guide concurrently. Reflections from both guides particularly enrich each other in this stage. For the Quantitative Guide, it is recommended to define the following aspects concerning the local context for which you plan to develop an AI system. This will assist you in conducting the quantitative assessment:

* “**Fair outcome**” of the AI system.
* Training, test, and validation **datasets, each representative** of the same underlying population.
* Variances between your **data versus production (live) data.**

## Step 2: Complete MC sections for “model details” and “bias and impact”

If you have not yet filled out these details for the Model Card, please do so now where possible.

Insights from this stage might also prompt you to re-draft certain parts of the MC.   
You can follow this [internal link to the Model Card](#_Step_1:_Fill).

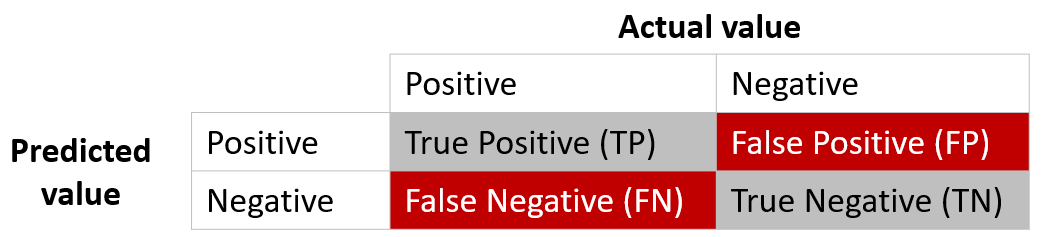
## Step 3: Calculate baseline performance metrics

The following baseline metrics are used to measure the overall performance of the algorithm. These metrics are the core references to understand the model and they can be recalculated as many times as necessary.

For example, the baseline metrics must be recalculated and evaluated if any of the metrics and methods of this chapter lead to modifying the training data, thresholds, or decisions. This constant process of recalculating the baseline metrics ensures the robustness of the system and helps in its explainability.

### A) Confusion matrix

The confusion matrix is a fundamental tool to assess the performance of an AI model and to conduct statistical analyses of bias in data. It is mainly used for supervised models that classify data into two or more categories[[12]](#footnote-13).



Below you will find the confusion matrix and its terms applied to a disease identification system for plants:

1. **True positive (TP)**: Number of cases when the predicted and actual outcomes are in a positive class. For example, in how many cases did the AI model correctly predict that a plant had a disease?
2. **False positive (FP)**: number of cases when the predicted outcome is in the positive class, but the actual outcome is in the negative class. For example, in how many cases did the AI model predict that a plant had a disease, though it was healthy?
3. **False negative (FN)**: number of cases when the predicted outcome is in the negative class, but the actual outcome is in the positive class. For example, in how many cases did the AI model predict that a plant was healthy, though it had a disease?
4. **True negative (TN)**: number of cases when the predicted and actual outcomes are in the negative class. For example, in how many cases did the AI model correctly predict that a plant was healthy?

### B) Based on the confusion matrix, calculate basic performance metrics

Once the confusion matrix is calculated with the general output and for each relevant sub-group, the following metrics provide different perspectives on the performance of a classification model.

**Each metric should be calculated for each relevant sub-group**. Please note that each metric emphasizes a different aspect of a classification problem. For example, precision is important when the cost of a false positive is high, while recall is crucial when the cost of a false negative is high. Therefore, you need to identify which metrics are relevant to your specific classification problem, also given that some metrics have mutually exclusive definitions.

The following table helps to define each performance metric and determine its relevance to the problem.

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Definition | How does this metric translate to your use case? | Do you think this metric is relevant for your use case? Explain your answer |
| Accuracy | Measures overall performance across all instances. It’s useful when the classes are balanced, and the cost of false positives and false negatives are roughly the same. |  |  |
| Precision | Measures the performance of a classifier when the cost of a false positive is high. It’s useful when we want to be very sure of our prediction. |  |  |
| Recall (Sensitivity) | Measures the performance of a classifier when the cost of a false negative is high. It’s useful when we want to capture as many positives as possible. |  |  |
| F1 Score | Harmonic mean of Precision and Recall. It’s useful when we want a balance between Precision and Recall and there is an uneven class distribution. |  |  |
| False Discovery Rate (FDR) | Measures the proportion of false positives among all positives. It’s useful when we want to minimize the number of false discoveries. |  |  |
| False Positive Rate (FPR) | Measures the proportion of actual negatives that are incorrectly identified as positives. It’s useful when we want to minimize the number of false alarms (false positives). |  |  |
| False Omission Rate (FOR) | Measures the proportion of actual positives that are incorrectly identified as negatives among all predicted negatives. It’s useful when we want to minimize the number of false negatives among all the instances that we predicted as negative. |  |  |
| False Negative Rate (FNR) | Measures the proportion of actual positives that are incorrectly identified as negatives among all actual positives. It’s useful when we want to minimize the number of missed positives (false negatives). |  |  |
| Positive Predictive Value (PPV) | Measures the proportion of actual positives that are correctly identified as positives among all predicted positives. It’s useful when we want to be very sure of our positive predictions. |  |  |
| Negative Predictive Value (NPV) | Measures the proportion of actual negatives that are correctly identified as negatives among all predicted negatives. It’s useful when we want to be very sure of our negative predictions. |  |  |
| True Negative Rate (TNR) | Measures the proportion of actual negatives that are correctly identified as negatives among all actual negatives. It’s useful when we want to minimize the number of false alarms (false positives) among all the instances that are actually negative. |  |  |
| AUC-ROC Curve | For classification models, mainly binary ones.  The AUC-ROC measures how well a model performs at various threshold settings. The curve is used to see the relationship between the TP and the FP rate. Useful when the costs of false positives and false negatives are significantly different. |  |  |

Please note: Applying the Quantitative Guide is an iterative process. Although the following steps indicate a sequence, it does not mean that you should follow it in this order. For example, you can jump from this chapter directly to the [fairness metrics](#_Step_6:_Select) and then look at the [additional methods](#_Step_4:_Select) afterwards.

Because the different steps, particularly in the in-processing stage have implications for each other, we also suggest to [reassess](#_Step_7:_Detect) your results.

|  |
| --- |
| **A light bulb with a black background  Description automatically generatedPractical example**  Imagine you want to detect and predict landslides in a certain rural region with machine learning. The data is collected through high-definition cameras installed in houses. Let’s compare how different metrics have different meanings in this context:  In this scenario, a False Positive (FP) would imply to predict that a landslide will occur, though this is not the case. If households were relocated based on this output, it can cost a lot of resources. Also, it can cause economic damage, particularly to those who are unnecessarily relocated. On the other hand, a False Negative (FN, saying no landslide will occur, but it happens nonetheless) can lead to late reactions of the government emergency systems, cost multiple lives and economic losses.  Though both cases lead to serious and unwanted consequences, particularly a False Negative can create severe harm. Now, let’s imagine that the local authority uses this landslide prediction model in their decision-making processes. Most likely, they would prefer a model with a lower rate of False Negatives, to keep the risk of a social crisis as low as possible.  When focusing on reducing the risk of False Negatives in a model, a team could analyze metrics like False Omission Rate (FOR) or Negative Predictive Value (NPV). This would allow us to assess the risks of the algorithm more in-depth from a technical perspective by measuring the sensitivity of the model in the negative prediction and reducing the error of the model by introducing better loss functions, thresholds, parameters, or new optimization methods.  This example shows how important it is to:   * understand the context and objective of an AI system (also in terms of the potential harm it can cause, should it not work as intended), * translate what each performance and fairness metrics mean in this context and for the potential (harmful) impact of an algorithm, and * determine which specific metrics have relevance for a certain context and objective of an AI system.   Example to fill the table:  False Omission Rate can be defined as the probability of having a landslide when the algorithm predicts **no** landslide in the region. In this sense, the system does not accomplish its objective and the socio-economic impact can be catastrophic. |

## Step 4: Select and apply additional methods

After having identified the relevant performance metrics for your use case (step 2), the next step is to select and conduct additional analyses that support you in making your algorithm more robust. As in the pre-processing phase, these analyses contribute to identifying and understanding causes of bias in your data.

You will have to determine which of the methods below are relevant to your use case. Insights from applying the Qualitative Guide and additional research beyond this guide will support you in prioritizing the methods below.

|  |
| --- |
| **Orange warning icon - Free orange warning iconsRelevant note**  Applying these methods is not a linear process. Rather think of it as an iterative process of continuous recalculation and interpretation. Detecting biases, data adjustments, change of data sources, and more, requires continuously checking accuracy, and errors, optimizing the model, and striving for simplicity.  During the in-processing phase, you will most likely need to continuously revisit and fix various issues with the data. This might also lead to revisiting the pre-processing stage to collect new data, adjust labels, etc., and then recalculate the metrics and methods for the in-processing stage.  Also, please note once again that the following metrics are not exhaustive but rather serve as an exemplary inspiration. |

### A) Evaluation metrics

This chapter introduces additional evaluation metrics. After identifying the metrics relevant to your use case, it’s crucial to understand how well your AI model’s predictions align with actual outcomes.

This is particularly important when there’s a clear target variable, such as crop yield or crop type. The focus of these metrics is not just on the model’s overall performance, but also on preventing erroneous predictions that could have significant repercussions, such as misdiagnosing a crop disease leading to a total crop loss. Hence, a thorough understanding and application of these metrics is vital for the success and reliability of your AI model.

You can document findings from this exercise and mitigation measures in the [**score card**](#_Step_5:_Fill) at the end of the chapter.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Definition** | **Do you think this metric is relevant for your use case? Explain your answer.** |
| **Cohen’s Kappa** | For classification models, mainly binary ones.  Useful to compare agreement between the predicted and the true value. Range: -1 to 1. |  |
| **Matthews Correlation Coefficient (MCC)** | For classification models, mainly binary ones.  Measures the strength of the prediction (e.g. disease or no disease), considering all four quadrants of the confusion matrix. |  |
| **Mean Absolute Error (MAE)** | For regression models.  Imagine you are throwing darts at a dartboard. MAE is the average distance of each dart from the bullseye. It’s used to understand the average magnitude of error. |  |
| **Root Mean Square Error (RMSE)** | For regression models.It is similar to MAE, but it’s more like a golf game where the penalties for missing the hole by a large margin are severe. RMSE gives higher weight to larger errors, making it useful for detecting outliers. |  |
| **Log Loss** | For classification models, mainly binary ones.  Used where we need to predict probabilities, e.g., how likely it is that a crop is infected with a disease. |  |
| **Intersection over Union (IoU)** | For object detection tasks.  Picture two overlapping photos. IoU is a number from 0 to 1 to measure the accuracy of the bounding box. |  |

### B) Algorithm training and hyperparameter tuning

Training an algorithm involves fine-tuning an AI model’s hyperparameters iteratively to optimize its performance. This crucial step encapsulates the intricate details of the data, including groupings, divisions, thresholds, and more, that have been identified in previous stages. Accordingly, the methods for model optimization below aim to address key challenges in AI such as over- and underfitting or minimizing error (Loucks P. & Beek E. 2017).

You can document findings from this exercise and mitigation measures in the [**score card**](#_Step_5:_Fill) at the end of the chapter.

**a) Train-test split of the data**

|  |  |  |
| --- | --- | --- |
| **Method** | **Definition** | **Do you think this metric is relevant for your use case? Explain your answer** |
| **KFold Cross-Validation** | Used in classification models.  It divides the data into ‘K’ subsets. The data is trained on ‘K-1’ subsets and tested on the remaining one. It helps to identify issues like overfitting. With these multiple subdivisions the algorithm can train and validate more times with the same training data. |  |
| **Bootstrapping** | Used for various tasks (e.g., regression, classification, etc.).  This method creates a random sampling with replacement to probe a metric, test, or training. It creates new samples from your data by randomly picking points - like shuffling a deck and drawing cards, then putting them back. This process is carried out to understand potential outcomes and estimate the sampling distribution of a statistic. |  |
| **Nested Cross-Validation** | For classification models assigning labels like final output.  This method optimizes the training of the model’s hyperparameters. It is a multiplication of the "k” validations in “n” models with “k” validations. In other words, each “k” fold is again divided and tested. |  |

**b) Hyperparameter tuning**

|  |  |  |
| --- | --- | --- |
| **Method for hyperparameter tuning** | **Definition** | **Do you think this method is relevant for your use case? Explain your answer** |
| **GridSearch** | Used for various tasks (e.g., regression, classification, etc.).  Optimizes hyperparameters for a specific performance metric (e.g. ROC AUC) on the training or validation set. It is like fine-tuning a recipe of a cake through trying exhaustively different combinations to find the best combinations of ingredients (hyperparameters) for the best cake (model). |  |
| **Random Searching** | Used for various tasks (e.g., regression, classification, etc.).  Randomly selects values for the model’s parameters. It chooses the combination that provides the best results and supports avoiding saddle points during optimization. It is like throwing darts randomly to find the best spot. Less computational expensive than GridSearch. |  |
| **Evolutionary Optimization** | Used for various tasks (e.g., regression, classification, etc.).  Simulates the process of natural selection where the “fittest individuals” are selected for reproduction to produce the offspring of the next generation. Simulates adapting and learning from past experiences, and thus “improving” at the task over time. |  |
| **Bayesian Modeling** | Used for various tasks (e.g., regression, classification, etc.).  Uses Bayes’ theorem to guide the search for optimal parameters. It learns from previous iterations to make “informed” decisions about where to look next. |  |
| **Regularization Functions** | Used for various tasks (e.g., regression, classification, etc.).  If a model is too complex, it may fit the training data too closely and perform poorly on unseen data. Regularization functions add a penalty term (cost function), helping to balance the trade-off between bias (oversimplification) and variance (over­fitting). |  |

### C) Parsimonious Principle

The parsimonious principle suggests that the simplest explanation is often the most acceptable. In AI, this principle can help reduce the variables in a model to minimize noise and correlation.

You can document findings from this exercise and mitigation measures in the [**score card**](#_Step_5:_Fill) at the end of the chapter.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Definition** | **Do you think this metric is relevant for your use case? Explain your answer** |
| **Akaike Information Criterion** | Used for various tasks (e.g., regression, classification, etc.)  AIC is used to compare different models based on their fit to the data and complexity. It favors models that achieve a good fit with fewer parameters. |  |
| **Bayesian Information Criterion** | Used for various tasks (e.g., regression, classification, etc.)  BIC is similar to the AIC metric, but BIC uses the probability function and is commonly used in time-series models. |  |
| **Bayes Factors** | Used for various tasks (e.g., regression, classification, etc.)  Bayes Factors are used to compare the evidence in favour of two competing hypotheses. The hypothesis that requires fewer assumptions (i.e., is simpler) is often preferred |  |
| **Mallow’s criterion** | For regression models.  Mallow’s criterion is like a balance scale that weighs the trade-off between the goodness of fit and the number of parameters in a model. |  |
| **Variance Inflation Factor (VIF)** | For regression models.  VIF is like a magnifying glass that reveals the degree of multicollinearity in a regression model. |  |

|  |
| --- |
| **A light bulb with a black background  Description automatically generatedPractical example**  Consider two AI use cases:   * + - 1. estimating the yield of coffee beans, by using sample pictures from individual branches and their coffee cherries, and       2. estimating poverty via satellite imagery.   For the coffee yield estimation, several issues can influence its functioning, for example, fluctuating photo quality or differences between branches and species. From this use case context, the following methods can prove particularly useful:   * understanding the accuracy and fitting of the model to have the best prediction possible, and * assessing errors to determine outlier branches with excessive cherries that are introducing bias into the model.   For the poverty estimation via satellite imagery, it would be rather relevant to emphasize the parsimonious principle. Poverty can be estimated by using a variety of strongly correlated variables, for example, development indexes, public services, or salaries. But the multiple-variable correlation does not make the model more predictive, it rather adds soundness and complexity to understanding the result. |

## Step 5: Fill out Score Card for additional methods

The following scorecard summarizes your quantitative findings before processing to the next stage. It allows for specific comments to understand technical concerns and areas needing attention. The scorecard can be used to communicate the findings of an AI model with team members, decision-makers, and affected communities.

**How to use it:**

Assign an ordinal rating for each variable based on the perceived severity of the issues identified. Use the following scale:

* **0: No issues**
* **1-2: Minor concerns**
* **3-4: Moderate issues**
* **5: Significant issues**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Insert metrics / method that you applied** | **Assign perceived impact of finding  (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[[13]](#footnote-14)) |
| Evaluation metrics |  |  |  |  |
|  |  |  |  |
| Training of the algorithm |  |  |  |  |
|  |  |  |  |
| Model optimization |  |  |  |  |
|  |  |  |  |
| Parsimonious principle |  |  |  |  |
|  |  |  |  |

|  |
| --- |
| **Orange warning icon - Free orange warning iconsRelevant note**  Before you continue, it is important to remember the continuous loop between stages and phases. During the training, optimization, and development of the initial model, many aspects can impact the pre-processing phase and cause a re-training and re-optimization of the model.  The next section, “Fairness metrics”, can modify the accuracy and general performance of the algorithm. Consequently, is crucial to record all the decisions on the tables suggested in the present document. The main objective of the next section is to make better decisions about data collection, processing, and prediction based on evidence and statistical tests. In short, the best model is not the one with the highest accuracy, but the one that best models the context where it will be applied. |

## Step 6: Select Fairness Metrics

After applying the methods above, it is crucial to assess the mathematical fairness of the AI model via fairness metrics. They allow us to **assess how the AI model performs across different groups** of data, including biases or disparities in outcomes (see [Rubin and Verma 2018](https://fairware.cs.umass.edu/papers/Verma.pdf) or [Carey and Wu 2022](https://link.springer.com/article/10.1007/s43681-022-00183-3)).

The selection of relevant fairness metrics is based on the type of algorithm, its objective, deployment context, the configuration of the data, groups in the data, and more. To apply the fairness metrics, the data needs to have different group categories. Fairness metrics are **relevant for datasets with and without sensitive demographic data**.[[14]](#footnote-15) Most often, they are applied for classification tasks and supervised learning.[[15]](#footnote-16) They are **based on basic performance metrics** ([step 3b](#_B)_Based_on)), so it is recommended to calculate them first.

Because there are **no one-size-fits-all fairness criteria**, contextual analysis supports you in determining the most suitable combination of fairness metrics for each use case. The Qualitative Guide can support you in understanding whether fairness metrics are relevant to your use case.

Fairness metrics are divided into four main groups. Respectively, they focus on mathematical fairness based on:

* predicted outcome,
* predicted and actual outcomes,
* the actual outcome, and
* similarity-based measures.

To explain the different groups, let’s imagine an AI model that is used to predict whether applicants will receive a loan for their farm.

**Predicted outcome:** These metrics look only at the model’s predictions. For example, the model might predict that 70% of male farmers and 60% of female farmers will receive a loan. If you use a fairness metric from this group, you will want both genders to have an equal chance of being predicted as receiving a loan. These metrics do not compare these predictions further to the actual outcomes that the validation or test data contain.

**Predicted and actual outcomes:** These metrics compare the model’s predictions to the actual outcomes of the data. Thus, they are considered more robust. For instance, if the model predicted that 70% of male farmers would receive a loan, but 80% actually did, these metrics would highlight this discrepancy.

**Actual outcome:** These metrics focus on the contrast between multiple characteristics of the data. They do so by considering the probabilities assigned to each prediction by the model. For example, a model might predict a 70% chance for a male farmer to receive a loan and a 60% chance for a female farmer. If the actual success rates were 80% for males and 75% for females, these metrics would capture both the discrepancy in probabilities and the discrepancy in actual success rates.

**Similarity-based measures:** These metrics evaluate model fairness beyond one attribute (e.g. gender). Even if relevant features (e.g. sensitive ones like gender) are properly identified as possible sources of discrimination, it is still possible that other attributes (e.g. insensitive[[16]](#footnote-17) ones like credit score) serve as proxies or latent variables. For instance, if two farmers have similar credit scores, farming experience, and farm sizes, but the male farmer has an 80% chance of getting the loan while the female farmer has only a 40% chance, a similarity-based measure would flag this as a potential fairness issue.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Definition** | **Do you think this metric is relevant for your use case? Explain your answer** |
| **Predicted Outcome** | | |
| Statistical Parity | Ensures equal probability for each group to be assigned to the positive class (TP and FP). *Example:* An AI model should equally classify the chance of approving loans for both male and female farmers. |  |
| Conditional Statistical Parity | An extension of statistical parity that considers additional legitimate attributes.  *Example:* The loan approval model should also consider factors like whether the male/female farmer is an individual smallholder or part of a cooperative. |  |
| **Predicted and actual outcomes** | | |
| Predictive Parity | Each group should have equal positive predictive value (PPV).  *Example:* The loan approval model should be equally accurate for both male and female farmers. |  |
| False Positive Error Rate Balance | Each group should have equal false positive rates (FPR).  *Example:* The loan approval model should not over-predict loan approvals for either male or female farmers. |  |
| False Negative Error Rate Balance | Each group should have equal false negative rates (FNR).  *Example:* The loan approval model should not under-predict loan approvals for either male or female farmers. |  |
| Conditional Use Accuracy Equality | Each group should have an equal positive predictive value (PPV) and negative predictive value (NPV).  *Example:* The loan approval model should accurately predict loan approvals and rejections for both male and female farmers. |  |
| Overall Accuracy Equality | Each group should have an equal prediction accuracy.  *Example:* The loan approval model should accurately assign approvals and rejections to both male and female farmers. |  |
| Treatment Equality | Each group should have an equal ratio of false negatives to false positives.  *Example:* The loan approval model should have the same ratio of under-predicted to over-predicted loan approvals for both male and female farmers. |  |
| **Actual outcome** | | |
| Test Fairness | For any confidence level the AI model has, entities from all groups should have an equal chance of being correctly predicted.  *Example:* Regardless of the confidence level, it should be equally accurate in predicting loan repayment for both male and female farmers. |  |
| Well-Calibration | For any confidence level the AI model has, the chance of a correct prediction should be equal to the confidence level for all subgroups (e.g. male, female or non-binary in gender).  *Example:* If the AI model is 90% sure a farmer will repay a loan, then it should be correct 90% of the time, for both male and female farmers. |  |
| Balance for Positive Class | Entities who are correctly predicted to belong to the positive class should have the same average confidence level from the AI model, regardless of their group.  *Example:* The AI model should be equally confident in its predictions for male and female farmers who will repay their loans. |  |
| Balance for Negative Class | Entities who are correctly predicted to belong to the negative class should have the same average confidence level from the AI model, regardless of their group.  *Example:* The AI model should be equally confident in its predictions for male and female farmers who will not repay their loans. |  |
| **Similarity-based measures** | | |
| Causal discrimination | For entities with the same characteristics, e.g. except for a sensitive attribute like gender, the AI system calculates the same output.  *Example*: If two farmers, one male and one female, have the same farming experience, credit score, and farm size, they should have the same chance of getting a loan. |  |
| Fairness through unawareness | This measure requires that the AI system does not use sensitive attributes like gender in its decision-making process. Note that this approach can be problematic as it overlooks the possibility of insensitive attributes acting as proxies for sensitive ones.  *Example:* If credit scores are generally lower for female farmers due to systemic biases, the AI system might end up denying more loans to female farmers even though it’s not directly considering gender. |  |
| Fairness through awareness | This metric evaluates fairness based on the principle that similar individuals should receive similar outcomes. Similarity is defined using a distance metric. The distance between the outcomes for two individuals should not exceed the distance between the individuals themselves.  *Example*: two farmers with similar farming experience, credit score, and farm size should have similar chances of getting a loan, regardless of their gender or other protected attributes. |  |

|  |
| --- |
| **A light bulb with a black background  Description automatically generatedPractical scenario**  Imagine an algorithm that classifies suitable sites to install solar mini-grids in rural areas. Among others, this classification is based on socio-economic variables, interconnection with other rural areas, productivity, or population growth. As this algorithm might influence decisions on which areas might get further electrified, it has a direct impact on people’s lives. Thus, it is crucial to mitigate possible biases or preferences for certain regions or social conditions.  For this use case, fairness metrics based on “predicted and actual outcomes” might be of particular interest. They help to spot whether the AI model’s predictions are evenly distributed among different regions. This would allow us to compare whether those differences in predictions are in line with the project goal (e.g., in the case of specific development policies) or not (e.g., if the project’s objective is equal electrification).  The third group of metrics “actual outcomes” might be particularly useful if we know of different success rates for electrification strategies in certain regions. Let’s say our AI model predicts a 70% chance of a successful solar mini-grid installation in Region A and a 60% chance in Region B. However, due to the better infrastructure, the actual success rate in Region A is 80%. In contrast, due to the challenging conditions in Region B, the actual success rate is only 50%.  In this case, the metric would highlight that the model’s predictions do not accurately reflect the unique conditions in each region. The model overestimates the success rate in Region B and underestimates it in Region A. |

## Step 7: Reassess

As has been emphasized throughout this guide, mitigating biases and harm is an ongoing, non-linear, and iterative process. For examples, insights from your fairness metrics might prompt you to collect additional data or have implications for how you conducted the previous steps of the Quantitative Guide so far.

Remember, concluding these steps and the in-processing chapter does not mean that your AI model is free of ethical harm. Mitigation is a constant journey, and this guide is designed to guide you along the way, highlighting areas where mitigation and further assessment can be relevant.

## Step 8: Detect biases

Below you will find common biases within the in-processing stage and exemplary strategies for testing and mitigating them. You can analyse which biases you perceive as relevant for your AI use case.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Bias Type** | **Definition** | **Exemplary testing** | **Relevant for your use case?** Please explain your answer | **If relevant for your use case: How will you mitigate this bias?**  (Ideally filled out with your assessor[[17]](#footnote-18)) |
| **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. | * 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**. | * 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**. | * 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. | * 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. |  |  |

# QUANTITATIVE ASSESSMENT IN POST-PROCESSING

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 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 when implementing AI system in their actual deployment context.

Thus, continuous monitoring and evaluation are required when and after implementing the AI system in the real world, also called ‘algorithmovigilance’. This allows to assess generalisability of the model, detect data shifts over time (when data is feed into the model real-time/continuously) or other harmful effects.

The main questions to answer during this stage are:

|  |
| --- |
| * Is the AI system **tested on multiple databases** **or subjected to adjustments** to ensure it does not overly rely on a single benchmark? * Does the AI system's data visualization process consider potential cognitive biases like the **Framing Effect and Availability Bias** that can influence decision-making? * Is there a process to ensure that data visualizations do not disproportionately rely on easily available data, potentially leading to **Availability Bias**? * Is the impact of human decision-making tracked and measured to ensure that **human intervention is meaningful**? * Does the AI system actively incorporate feedback and input from marginalized groups/outliers, to address **potential accessibility bias**? |

For the post-processing stage, the following steps are recommended:

1. [Go through the Post-Processing stage of the Qualitative Guide](#_Step_1:_Go_1)
2. [Update MD and complete sections for “Redress measures”](#_Step_2:_Update)
3. [Generate documentation](#_Step_3:_Generate)
4. [Monitor dashboards](#_Step_4:_Monitor)
5. [Visualize dashboards](#_Step_5:_Visualization)
6. [Re-assess](#_Step_6:_Reassess)
7. [Detect biases](#_Step_7:_Detect_1)

## Step 1: Go through the Post-Processing stage of the Qualitative Guide

During the post-processing stage, it is strongly recommended to go through the Qualitative Guide concurrently. Reflections from both guides enrich each other in this stage.

From the Qualitative Guide, it is recommended to define the following aspects of the local context for which you plan to develop an AI system. This will assist you in conducting the quantitative assessment:

* regular **evaluation** of the AI system’s **outcomes and impacts,**
* regular **revalidation for performance and fairness,**
* integrating **feedback from monitoring** into re-training and re-validation,
* **transparency** about the AI system’s **purpose, use and limitations,**
* **explainability** of output, and
* **human-in-the-loop accountability.**

## Step 2: Update MC and complete “redress measures” section

Update your Model Card where appropriate and particularly fill out the section on “Redress” if you have not yet filled out these details. You can follow this [internal link to the Model Card](https://euc-word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=de&rs=en%2DUS&wopisrc=https%3A%2F%2Feticastech.sharepoint.com%2Fsites%2FServices%2F_vti_bin%2Fwopi.ashx%2Ffiles%2F546e06dfa957405a89681c8724f7e568&wdpid=334dd0a2&wdenableroaming=1&mscc=0&hid=7731FAA0-C0A4-7000-D816-2DD402AA46A4&wdorigin=BrowserReload&jsapi=1&jsapiver=v1&newsession=1&corrid=8f71ccb5-ed97-4279-9eb0-3379d38690df&usid=8f71ccb5-ed97-4279-9eb0-3379d38690df&sftc=1&cac=1&mtf=1&sfp=1&instantedit=1&wopicomplete=1&wdredirectionreason=Unified_SingleFlush&rct=Normal&ctp=LeastProtected#_Step_1:_Fill).

## Step 3: Generate documentation

For the post-processing phase, it is recommended to create the following three main documents:

1. training and outcome codebooks,
2. commented coding script(s), and
3. a compilation of the tables in this Quantitative Guide.

These documents help others to understand the process of how the algorithm and its decision-making process have been developed.

### A) Training and outcome codebooks

Codebooks describe the main characteristics of the used data. These codebooks contribute to the replicability and transparency of the AI system.

The following table provides you with an exemplary schema for a codebook.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name of variable in dataset | Name of variable for public | Description of the variable | Type of variable (e.g. str, bool, int, list, etc.) | Unique values in the variable | Units of the variable | Source of the variable | Level of observation (e.g. person, household, city, state, country, etc.) |
| *Example:*  *pop* | Total population | Total censual population (2024) | Float | [0, …, n] | People | Census 2024 | Provinces |

|  |
| --- |
| **Orange warning icon - Free orange warning iconsRelevant note**  When creating codebooks, it’s beneficial to include the primary key(s) or unique identifiers used in the dataset (e.g. ID for each observation). These keys help uniquely identify each record and are crucial for data management and analysis.  If your AI system operates in a modular architecture, where different parts of the system interact with different subsets of the data, consider including an Entity Relationship Diagram (ERD).[[18]](#footnote-19) An ERD can help visualize the relationships between different data entities, enhancing understanding of the data structure and how different parts of the system interact with it. This can be particularly useful when trying to understand complex systems and ensure fairness and transparency in your AI model. |

### B) Commented coding script(s)

Prepare the coding scripts, model methodology, used thresholds, and/or system architecture diagrams with detailed comments and notes. The main objective is to create a compilation of the codes and explanations to facilitate replication of the AI system’s training and fitting process.

Clear and comprehensive comments are crucial for understanding the code and its functionality. For an example of well-commented scripts, you can check out the [IBM AI Fairness 360 GitHub repository](https://github.com/Trusted-AI/AIF360).

### C) Compilation of insights from applying the Qualitative and Quantitative Guides

**Compile all tables and information** from the Qualitative and Quantitative Guide **into a comprehensive document or database**. This document serves as an overview on decisions made and issues identified during the AI lifecycle. It is a crucial resource for future audits, risk assessments, or modifications to the original algorithmic system, ensuring transparency and accountability.

## Step 4: Monitor dashboard (for internal use)

A monitoring dashboard is essential for tracking the performance and updates of an AI model It is particularly relevant when algorithms are constantly training or learning. A dashboard, primarily designed for the technical team, can include various visualizations and metrics.

It requires ongoing data collection (based on data privacy and anonymization), including relevant variables that reflect the impact on different groups. For instance, if the model’s impact on gender is to be monitored, gender variables need to be collected.

Please find below exemplary key elements for a dashboard. For illustration, they are applied to an AI system that is identifying diseases in plants on smallholder farms owned by both women and men:

* **Performance and Fairness Metrics**: Charts showing the most relevant performance and fairness metrics of disease identification overall and broken down by gender of the farm owner.
* **Algorithm Activity**: Graphs showing the number of disease identifications made over time, response times of the system, number of users etc.
* **Time Series Analysis**: Plots showing how the model’s performance and fairness metrics have evolved over time.
* **End-User Feedback**: Feedback from the farmers about the quality of the disease identifications, presented in a way that’s easy to understand and act upon.
* **Model Description**: Information about the current model(s) being used, including the type of model, parameters, hyperparameters, and the date of the last update.
* **Training Data Updates**: Information about when the training data was last updated, and any significant changes in the data.
* **Special Monitoring**: Any specific fairness or statistical analysis results that the team wants to track. For example, a metric tracking the difference in false positive rates for disease identification between farms owned by women and men.

## Step 5: Visualization dashboards (for public use)

Visualization dashboards are designed with the end user in mind. They present key information about the AI system in an accessible and understandable format. It is important to collaborate with the end user or clients when designing a dashboard to ensure meaningful accessibility.

Please find below exemplary key elements for a publicly accessible dashboard. For illustration, they are applied to an AI system that is identifying diseases in plants on smallholder farms owned by both women and men:

* **System Outputs:** Visualizations of aggregated and anonymized data from the AI system’s disease identifications. For example, it could show the overall percentage of identified diseases or trends over time.
* **Performance Metrics**: Simple charts or graphs showing the most relevant performance and fairness metrics of the disease identifications in an easy-to-understand and non-technical way.
* **User Feedback Mechanism**: A simple and intuitive way for farmers to provide feedback on the accuracy of the disease identifications.
* **Explainability**: Simple, non-technical explanations of how the system works and makes its decisions. For example, it could explain that the system uses images of the plants to identify diseases or other variables that might be additionally taken into account.
* **Accountability**: Information about who to contact if there are issues with the system, and how decisions about the system are made.
* **Privacy**: Clear, easy-to-understand information about how the farmers’ data is protected.

## Step 6: Reassess

The final step is to reassess earlier stages. As has been emphasized throughout this guide, mitigating biases and harm is an ongoing, non-linear, and iterative process. Use the tables in the Quantitative Guide to collect critical observation and mitigation potentials. Summarize the main changes, adjustments, and considerations during the pre- and in-processing phases. Reflect critically on how and when different biases can be addressed at each stage. This step is crucial for planning future actions and ensuring the effectiveness of the whole risk assessment tool.

Remember, concluding this guide does not mean your AI system is now free of all harm. Mitigation is a constant journey, and this guide is designed to guide you along the way, highlighting areas where mitigation can be relevant.

## Step 7: Detect biases

Below you will find common biases within the post-processing stage and exemplary strategies for testing and mitigating them. You can analyse which biases you perceive as relevant for your AI use case.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Bias Type** | **Definition** | **Exemplary testing** | **Relevant for your use case? Please explain your answer** | **If relevant for your use case: How will you mitigate this bias?**  (Ideally filled out with your assessor[[19]](#footnote-20)) |
| **Deployment bias** | Occurs when organizational, budgetary, technical, or training **issues impact outputs** of the AI system, potentially influencing those affected by the decisions. | * 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. | * 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 systems**, possibly ignoring correct but contradictory non-automated information. | * 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, Availability Bias, Anchoring Bias, and Signal Error, which may result from how data is visualized and interpreted. | * 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. |  |  |

# REFERENCES

Barocas, S., Hardt, M., & Narayanan, A. (2018). Fairness in machine learning limitations and opportunities. The MIT Press. <https://api.semanticscholar.org/CorpusID:113402716>

Blatchford M, Mannaerts C. & Zeng Y. (2021). Determining representative sample size for validation of continuous, large continental remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 94. <https://doi.org/10.1016/j.jag.2020.102235>

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). *Toward trustworthy AI development: Mechanisms for supporting verifiable claims*. ArXiv.org. <https://arxiv.org/abs/2004.07213>

Burrel J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1)*.* <https://doi.org/10.1177/2053951715622512>

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. A*rXiv preprint.* [2010.04053 (arxiv.org)](https://arxiv.org/pdf/2010.04053)

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2), 72-80. <https://doi.org/10.1109/mts.2021.3056293>

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Hal, D. I., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. [Datasheets for datasets (acm.org)](https://dl.acm.org/doi/pdf/10.1145/3458723)

Gilbert, T. K., Dean, S., Zick, T., & Lambert, N. (2022). Choices, risks, and reward reports: Charting public policy for reinforcement learning systems. *CLTC White Paper Series.* UC Berkeley Center for Long-Term Security.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. Association for Computing Machinery New York, NY, United States. <https://doi.org/10.1145/2939672.2945386>

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). *The extent and consequences of p-hacking in science*. PLoS Biology, 13(3). <https://doi.org/10.1371/journal.pbio.1002106>

Ho, D. A., & Beyan, O. (2020). Biases in data science lifecycle. *arXiv preprint.* [2009.09795 (arxiv.org)](https://arxiv.org/pdf/2009.09795)

Jiang, H., & Nachum, O. (2020). Identifying and correcting label bias in machine learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR. <https://proceedings.mlr.press/v108/jiang20a.html>

Kaptchuk, T. J. (2003). *Effect of interpretive bias on research evidence*. British Medical Journal, 326(7404), 1453-1455. <https://doi.org/10.1136/bmj.326.7404.1453>

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018, July). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*, 2668-2677. PMLR.

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*. <https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf>

Lacoste, Y. (2012).Geography, geopolitics, and geographical reasoning. *Herodote*, 146-147(3-4), 14-44. <https://doi.org/10.3917/her.146.0014>

Li, J., & Chignell, M. (2022). FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. *AI And Ethics*, 2(4), 837–850. <https://doi.org/10.1007/s43681-022-00145-9>

Loucks, D. P., & Van Beek, E. (2017). *Water resource systems planning and management, an Introduction to optimization models and methods*. Springer eBooks. <https://doi.org/10.1007/978-3-319-44234-1>

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5). <https://doi.org/10.1016/j.ipm.2021.102642>

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287596>

Oca, A.M., Bahmanyar, R., Nistor, N., & Datcu, M. (2017). Earth Observation Image Semantic Bias: A Collaborative User Annotation Approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10, 2462-2477.

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. *Association for Computing Machinery*, 33–44. <https://doi.org/10.1145/3351095.3372873>

Rossini, M., Celesti, M., Bramati, G., Migliavacca, M., Cogliati, S., Rascher, U., & Colombo, R. (2022). Evaluation of the spatial representativeness of in situ SIF observations for the validation of Medium-Resolution Satellite SIF products. *Remote Sensing*, 14(20), 5107. <https://doi.org/10.3390/rs14205107>

Abd-El-Hafiz, S. K., AbdElHaleem, S. H., & Radwan, A. G. (2016). Novel permutation measures for image encryption algorithms. *Optics and Lasers in Engineering*, 85, 72–83. <https://doi.org/10.1016/j.optlaseng.2016.04.023>

Selbst, A. D., Boyd, D., Friedler, S., Venkatasubramanian, S., & Vertesi, J. (2018, August 23). *Fairness and abstraction in sociotechnical systems*. In *Proceedings of the conference on fairness, accountability, and transparency,* 59-68. Association for Computing Machinery. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3265913>

Sonobe, R., Yamaya, Y., Tani, H., Wang, X., Kobayashi, N., & Mochizuki, K. (2018). Crop classification from Sentinel-2-derived vegetation indices using ensemble learning. *Journal of Applied Remote Sensing*, 12(2).

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization,* 1-9. <https://dspace.mit.edu/handle/1721.1/143588>

Lv, T., Zhou, X., Tao, Z., Sun, X., Wang, J., Li, R., & Xie, F. (2021). Remote Sensing-Guided spatial sampling Strategy over heterogeneous surface ground for validation of vegetation indices products with medium and high spatial resolution. *Remote Sensing*, 13(14), 2674. <https://doi.org/10.3390/rs13142674>

Verma, S., & Rubin, J.S. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1-7. <https://fairware.cs.umass.edu/papers/Verma.pdf>

Wang, Z., Qinami, K., Karakozis, Y., Genova, K., Nair, P.Q., Hata, K., & Russakovsky, O. (2019). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8916-8925.

Widodo, Budiardjo, E.K., & Wibowo, W.C. (2019). Privacy preserving data publishing with multiple sensitive attributes based on overlapped slicing. *Inf.,* 10, 362. <https://doi.org/10.3390/info10120362>

Wong, D.W. (2004). The modifiable areal unit problem (MAUP). In *WorldMinds: Geographical perspectives on 100 problems,* 571-575. Springer, Dordrecht.

Wooldridge, J. (2013). *Introductory econometrics: A modern approach* (5th ed.). South-Western Cengage Learning.

Zhang, M., Singh, H., Chok, L., & Chunara, R. (2022). Segmenting across places: The need for fair transfer learning with satellite imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2915-2924.

# 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. These can be related to human-demographics or beyond (e.g. certain crops, regions, etc.). [↑](#footnote-ref-4)
4. I.e. pre-processing (encompassing problem formulation and data collection), in-processing (model selection and development), and post-processing stage (deployment). [↑](#footnote-ref-5)
5. These can be related to human-demographics (e.g. age, gender) or non-human demographics (e.g. certain crops, regions, etc.). [↑](#footnote-ref-6)
6. 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-7)
7. Pseudonymized datarefers 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. To briefly explain the mentioned techniques:

   Permutation refers to replacing personal identifiers with random.

   Suppression involves removing or replacing certain data elements to prevent the identification of individuals, especially in cases where the data points are unique or rare.

   In the case of permutation models, encryption models can be used to provide an extra level of security to sensitive data. Authors like Salwa et al. (2016) provide guidance on the selection of permutation and encryption models.

   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. [↑](#footnote-ref-8)
8. Jiang & Nachum (2019) developed a method that assigns appropriate weights to the classes of the training data, ensuring that the learned classifier will be approximately unbiased even with a finite sample size.

   Montes de Oca et al. (2017) suggested a collaborative user annotation approach for remote sensing data, which often has a common bias in the validation or classification of images. Their approach has been shown to improve the performance of classification tasks.

   Another example are reviews. They involve having multiple experts or data labelers independently label the data. Discrepancies are resolved through discussion or majority vote. [↑](#footnote-ref-9)
9. 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-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. Explainability is particularly relevant for AI models that do not rely on deep learning architectures. In deep learning networks, hidden layers complicate the process of understanding why a model arrived at a specific decision in a way that humans can grasp. Beyond explainability, transparency regarding training data, model architecture, weights, and other aspects is crucial to understand how and why a model was built in a certain way. [↑](#footnote-ref-12)
12. If you are working with a regression model, the confusion matrix can be applied in some cases (e.g., binomial regression). In this case, one option is to convert continuous variables into discrete ones. It is recommended to develop a complementary analysis for the regression model applied in AI systems in addition to the econometric conventional tests. Please note that for unsupervised and reinforcement learning defining and measuring fairness can be less straightforward and more complex. [↑](#footnote-ref-13)
13. 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-14)
14. For instance, if you’re developing a disease identification system for plants and the model consistently misclassifies diseases for a certain type of plant, this could lead to significant losses for farmers who grow that plant. Fairness metrics can help you identify and mitigate such imbalances. [↑](#footnote-ref-15)
15. Please note that for unsupervised and reinforcement learning defining and measuring fairness can be less straightforward and more complex. [↑](#footnote-ref-16)
16. Features that are theoretically not considered sensitive or as directly referring to protected groups, e.g. age, gender, disability status, etc. [↑](#footnote-ref-17)
17. 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-18)
18. For a more detailed explanation, see [here](https://www.lucidchart.com/pages/er-diagrams?usecase=erd#section_0). [↑](#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)