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

Part A: Step-by-Step Guide

german cooperation
DEUTSCHE ZUSAMMENARBEIT

implemented by:
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

FAIR FORWARD
Artificial Intelligence for all.

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

Responsible
GIZ - FAIR Forward – Artificial Intelligence for All
Nadine Dammaschk – AI Advisor (nadine.dammaschk@giz.de)
Jonas Gramse – AI Advisor (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.

Bonn, Germany, June 2024

Table of contents

Executive summary

The "Responsible AI Assessments" is a collaborative co-creation of "FAIR Forward – Artificial Intelligence for All," a programme of German Development Cooperation (GIZ) commissioned by the German Federal Ministry for Economic Cooperation (BMZ), Eticas and a distinguished community of AI inclusion experts from Sub-Saharan Africa and Asia Pacific.

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

They **Responsible AI Assessments** 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.

Drawing on experiences with real-world AI assessments, the Responsible AI Assessments are a living framework adaptable to the evolving AI landscape. The full and original version of the Responsible AI Assessments incl. all step-by-step guides (Part B, Part C) and editable versions is available under the FAIR Forward website (or under this direct link).

In 2023, they were tested on 7 AI activities from 6 countries on the African and Asian continent. These real-world assessments included a diverse set of AI activities such as a landslide detection project in Rwanda, cashew disease detection in Ghana, crop mapping in Telangana (India), site-identification for solar mini grids in Uganda or chatbot usage in Kenya.

We hope that its focus on practice and real-world contexts from Sub-Saharan Africa and Asia Pacific make it particularly useful for (future) assessors or developers of AI systems in such contexts as well as worldwide. This guide concludes with summarizing key insights and implications for the future development of the method, emphasizing the need for continuous refinement. In this spirit, we also open source all parts of the Responsible AI Assessments to enable wide use and further iterations by others.

Disclaimer

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.

Acknowledgement

As FAIR Forward and Eticas, we extend **our sincere gratitude to our involved partners and community of engaged AI inclusion experts for their invaluable contributions** to co-creating the Responsible AI Assessments.

The openness of our partners allowed us to apply and improve drafts of these Responsible AI Assessments with the ongoing AI-related activities that FAIR Forward supports. Thanks to our partners' commitment, we could gain all the valuable insights on how to design and improve the Responsible AI Assessments, as shown in this report.

Special thanks go to our distinguished pool of **AI inclusion experts**, involving Favour Borokini, Josia Paska Darmawan, Mohamed Kimbugwe, Mercy King'ori, Meena Lysko, Raashi Saxena, and Kofi Yeboah.

They departed with us on the journey to develop the method of the Responsible AI Assessments. Their critical perspectives guided us when drafting and iterating the Qualitative and Quantitative guides and provided invaluable insights when testing the Responsible AI Assessments on a wide range of AI-based activities.

This Step-by-Step Guide, as well as the Qualitative and Quantitative Assessment Guides, reflect the **collective expertise and commitment** of numerous actors, without whom it would not have possible to create the Responsible AI Assessments as they are. Thank you!

Biographies: Community of AI Inclusion Experts

Favour Borokini

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

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

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

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

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

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

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.

Introduction

In an era where Artificial Intelligence (AI) plays an increasingly pivotal role in shaping various aspects of communities and societies, the intersection of technology, ethics, and human rights has become a critical focal point. 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" [1], Eticas and a diverse community of AI inclusion experts from Sub-Saharan Africa and Asia Pacific.

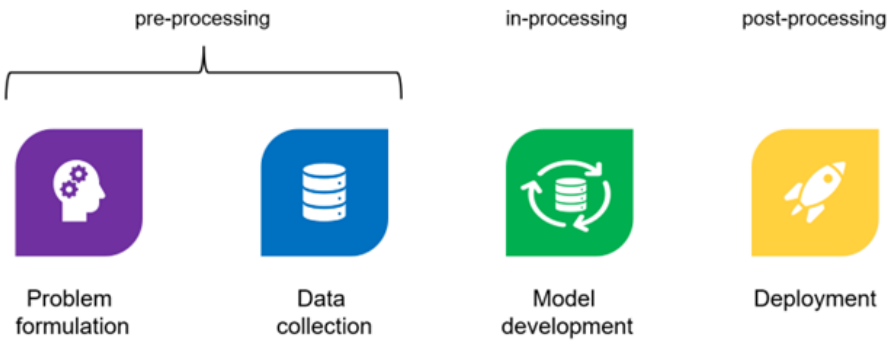
The Responsible AI Assessments are a method to identify, assess and mitigate potential harms and biases in AI. As an AI risks and ethics assessment tool, they guide AI stakeholders (e.g. as an assessor, developer or deployer of AI) in critically analyzing their AI resources, emphasizing human rights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

The **Responsible AI Assessments** 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**:

[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).



This Step-by-Step Guide is geared toward AI stakeholders (e.g. as an assessor, developer, project manager or deployer of AI) who seek to gain an understanding how AI systems can be assessed for potential harms, ethical risks and biases in AI.

In the following sections, we unravel the journey of developing the Responsible AI Assessments and provide actionable guidance how they can be applied:

- **Section 1** uncovers the **advantages** that the Responsible AI Assessments present and its **limitations**, setting the stage for a nuanced understanding of their scope.
- **Section 2 is a step-by-step guide** and equips users with the tools needed to implement the method and transition from theory to practice.
- **Section 3 concludes** this Step-by-Step Guide by summarizing critical takeaways and discussing implications for the future.
- **Annex:** Unveils the fundamental considerations for developing the method, including a panoramic view of how it was conceptualized and designed.

It should be noted that the Responsible AI Assessments are a **living framework** that will be further adjusted to apply to the dynamic nature of AI landscapes and FAIR Forward's learnings from applying the Responsible AI Assessments. They were based on a diverse set of AI activities on the African and the Asian continent such as a landslide detection project in Rwanda, cashew disease detection in Ghana, crop mapping in Telangana (India), site-identification for solar mini grids in Uganda or chatbot usage in Kenya.

All these AI use cases were in different stages of their development, with most of them in the pre-processing stage. Therefore, the Responsible AI Assessments are naturally tailored to these activities and their contexts. Nonetheless, they should be applicable for assessing AI systems and projects in similar or related contexts.

1. What the Responsible AI Assessments can do (and what not)

1.1 Advantages of the Responsible AI Assessments

You follow a globally accepted framework:
Drawing on the UNESCO principles from its Recommendation on the Ethics of AI (2021), the Responsible AI Assessments are based on a globally accepted framework of how to assess **human rights and ethical principles in AI**.[2]

You get a practical method to conduct a holistic risks assessment:
Offering a modular scheme, the method stands as an **opportunity to comprehensively address biases and human rights risks** by providing:

- a collection of questions (Qualitative Guide) that should be answered throughout an AI lifecycle and
- guidance how to navigate a quantitative fairness assessment (Quantitative Guide).

You get a method that has been tested in 6 different country-contexts:
The Responsible AI Assessments have been tested on 7 diverse AI use cases, ranging from landslide detection to chatbot usage up to disease identification in plants in 6 countries in Sub-Saharan Africa and Asia Pacific.

You can adapt the guides to your context:
The Qualitative and Quantitative Assessment Guides provide a **comprehensive and flexible template** to capture crucial biases and human rights risks across diverse contexts. Thanks to their mid-level specificity, they can be **adapted to the specifics of other domains and contexts**, ensuring relevance and effectiveness.

You get a method that helps you to prepare an audit:
Although the tool is not an audit itself (see limitations), the Responsible AI Assessments may be deemed a suitable preparation for any audit (see also “questions that an auditor might ask” in the respective guides).

You get a method that acknowledges the importance of active discussion:
The Responsible AI Assessments are **based on the premise of interaction**. They acknowledge that they can only provide the basis for critical reflections and discussions because the gaps of (local) context and sector specifics can only be filled through dialogue – in the best case with a diverse range of stakeholders and external assessors or auditors.

[2] The Recommendation has been adopted by UNESCO’s member states.

Your critical thinking on a wide range of AI Ethics-related issues is stimulated: In this regard, the guides encourage **critical reflection** when developing and deploying AI solutions. The sessions have provided a platform for AI stakeholders to reflect on critical and otherwise overlooked aspects. Thus, the Responsible AI Assessments create **awareness about** various sources for **human rights risks and biases** while opening the room for reflecting on mitigation strategies.

The method is based on a collaborative design and diverse perspectives: The guides incorporate by-design **inclusivity and collaboration**, facilitated by the active involvement of AI inclusion experts, AI developers and project managers, and end users from the respective countries during the Responsible AI Assessments. This collaborative approach ensures that the method is a living and practical tool shaped by diverse perspectives and practical insights.

1.2 Limitations:

While the Responsible AI Assessments hold promising opportunities, it is crucial to acknowledge their limitations transparently.

Recognize the limitations of applying the guides to non-supervised AI use cases:
One of the primary challenges lies in the **standardization level of these guides**. The guides and method have been mainly tested on supervised AI use cases, but their applicability to, for example, unsupervised or generative AI use cases needs to be evaluated.

Consider scalability and resource constraints:
Additionally, the reliance on end users, community representatives, and experts for input, while enriching the Assessments, and the dedication required to conduct the Assessments poses a potential **limitation for scalability**. The depth of engagement required for effective implementation may pose challenges when attempting to apply the same method across a broader spectrum, in contexts where resources are constrained or in dynamic environments where rapid interventions are required.

Ensure adequate training for effective utilization:
It **requires training and expertise in AI risk assessments, auditing** or related fields to utilize the assessment guides effectively, adequately iterate between qualitative and quantitative dimensions and propose suitable mitigation strategies on a socio-technical and quantitative level. Therefore, if you or any assessor only receives a limited introduction to the method, it may lead to suboptimal usage, reducing the Responsible AI Assessment's potential benefits.

Adopt an iterative approach for increased accountability:
The current method, as outlined below, presents a **short-term intervention**: If you use it only once within the lifecycle of an AI system development, you may get only a snapshot of potential issues within AI development and deployment. To be more effective and impactful, we recommend deepening the **method by applying it several times** throughout the lifecycle of an AI activity. This iterative assessment approach facilitates increased accountability to follow up on recommendations made in previous sessions.

Support smaller organizations with resource constraints:
Particularly, smaller organizations might struggle with participating in such assessments or implementing mitigation strategies due to **resource constraints**. It would be relevant to identify how to support such smaller organizations and partners effectively in risk assessments and their follow-ups to ensure that such assessments are also accessible and meaningful for stakeholders with resource constraints.

Understand the difference between assessment and audit:
It is important to note that the assessment Responsible AI Assessments **do not resemble an official AI audit**. With "audit", we refer in this context to an official examination of an AI algorithm that is conducted by an independent body (see Carrier & Brown 2021; Hasan, Brown, et al. 2022). Usually, such audits entail a set of tests to check adherence to predefined official standards or regulations. The assessment may however be a suitable preparation for any audit (see also “questions that an auditor might ask” in the respective guides).

Acknowledge the imperfect nature of AI risk mitigation:
Although the Responsible AI Assessments considers broader societal implications, encompassing ethical considerations, potential biases, and the algorithm's effects on diverse stakeholders, it should be noted that **neither an audit nor assessment** such as this **can guarantee that an AI solution will cause no harm**. Nonetheless, they represent important even if imperfect instruments to mitigate the possibility of harm as much as possible.

Despite these limitations, the Responsible AI Assessments represent an orientation **towards responsible AI practices**. The recognition of these challenges lays the groundwork for ongoing refinement and remodeling, ensuring that the method evolves in tandem with the dynamic landscape of AI technologies and their real-world applications.

2 Step-by-step guides for sessions and AI systems analysis

This section serves as a **comprehensive guide if you want to conduct a Responsible AI Assessment**. It outlines its different phases in detail (see figure 1):

Figure 1. Phases of the Responsible AI Assessments (as of 12/2023)

For each phase, this section provides:

- General information on the phase
  (including recommended time and participants)
- Step-by-step guide on important steps
- Best practices

As a general note:
To be more effective and impactful, we **highly recommend conducting the assessment** not only once, but **periodically** throughout the lifecycle of an AI system development. Applying it once might only allow a snapshot of potential issues evident at that moment in time within AI development and deployment. An iterative assessment approach facilitates increases accountability to follow up on recommendations made in previous sessions.

Who should do this assessment – note on engaging with assessment experts

This Step-by-Step Guide starts from the baseline assumption that you want to assess an AI use case for its potential risks. You can either be part of an AI development team yourself or an internal/ external assessor. Additionally, you should have identified a concrete AI activity to assess, and your organization should have agreed to such an assessment both on leadership level and on level of the respective AI project team.

If you are considering doing an internal assessment on an AI activity of your team or organization, we recommend you engage with external AI assessment experts. Why?

1. **Fresh and more critical perspectives:**
   The closer you are to your AI activity, the more difficult it is to ask the uncomfortable questions that Responsible AI provokes. External assessors should not have such constraints or familiarity bias. Their independence allows them to be more critical.
2. **Expertise and Experience:**
   External assessors often have specialized expertise and experience in AI risk assessment across various industries and use cases. This can lead to a more thorough and effective risk assessment.
3. **Resource Efficiency:**
   Hiring an external assessor can free up internal resources, allowing the in-house team to focus on their core responsibilities and/or towards mitigation of risks identified in the process.
4. **More diverse perspectives through additional experts:**
   It is recommended to involve **specific experts with local contextual or domain knowledge** (on AI product or the specific sector at hand, e.g. agriculture) **and AI inclusion experts**. This broadens the discussion with topics that might otherwise not be discussed. The depth of these experts can be **adapted to the resources available**. For example, they can either serve as an additional critical voice during the Deep Dive or be engaged throughout the whole process of the Responsible AI Assessments.

In addition to Eticas' auditing expertise, the input of the community of AI inclusion experts that supported us was a highly valuable and critical addition. Thus, we highly **recommend engaging suitable additional experts** in similar assessments.

2.1 Scoping Call

The main function of scoping calls is to narrow down the focus for Deep Dive sessions. They aim at identifying:

- **main issues** that the activity faces
  (e.g. bias in data collection, fairness issues of model)
- **relevant stages of AI lifecycle** for the assessment
  (pre-, in- or post-processing stage)
- **assessment type** to apply
  (qualitative vs quantitative vs mix of both)
- **topics of additional interest** to the project team

General information:

- Recommended time for the Scoping Call:
  - at least 30 min.
  - best case around 45 min.
- Required participants:

    - Facilitators & main assessors
    - Core team (e.g. AI developers, project managers)
    - Additional relevant team members (e.g. compliance officers or ethicists)
- Recommended:
    - (external) data ethics and inclusion experts
    - (external) domain experts

Step-by-step guide

Before Scoping Call

- **Collect initial information** from the AI project team:
    - Request and analyze existing documentation (e.g. reports, training data, model)
    - Ask the stakeholders to fill out a short overview of their system, a so-called model card[3] (as far as possible for them)
    - Conduct brief interviews with stakeholders (e.g. AI developers, project managers, beneficiaries) on challenges that they face within the activity. Even 10-15min interviews may be sufficient.
- Based on the material, **develop initial hypotheses** about**:**
    - Which stages of the AI lifecycle an assessment could focus on
    - Which issues and AI principles might be of relevance for this use case
    - Whether the assessment should focus on the qualitative or quantitative assessment guide – or a mix of both.
- Based on your hypotheses, **develop initial questions** for the scoping call, by using the qualitative and quantitative assessment guide as an inspiration. Adopt those questions to the use case at hand.
- In advance, **share questions** with the project team to receive their perspectives on what might be missing and refine the questions (if necessary).

During Scoping Call

- Facilitators provide the **platform for exchange** where mainly the project team speaks and engages in a dialogue on the prepared questions.
- **Define collaboratively** what the **Deep Dive** should focus on:
    - main issues the activity faces
    - relevant stages of AI lifecycle for assessment (pre-, in- or post-processing)
    - assessment type to apply (qualitative vs quantitative vs mix of both)
    - topics are of additional interest to the project team.
- Provide an **outlook for** the **Deep Dive** session.

[3] A Model Card provides a structured compilation of general information about an algorithmic system, its context and use. It can also be a helpful tool for any auditor who might assess your AI system.
For an exemplary template, please refer to the Quantitative Guide, chapter 2: *Preparatory Work*.

- Discuss with the AI project team already **whether and how the final report (or parts of it) can be made publicly accessible** – and how this might influence the openness of discussions.

After Scoping Call

- **Share agreement** with everyone for potential adjustments.
- **Prepare Deep Dive** (see section 4.2).

Best practices

- Use a model card[4] as a tool to gather information on the training data, model, intended use, evaluation, benchmarks, ethical considerations, and other factors.
- When sharing the initial questions via e-mail divide between:
  - Information that you still need for preparing the call that can be quickly answered by the project team (e.g. what data sources were used for model development),
  - Questions for the scoping call that are more open and focused on dialogue.
- To keep the call manageable, select 5-8 questions that facilitate open discussions.
- The scoping call should have a more informal character so that the project team can understand the process of the assessment and feel ready for the Deep Dive.
- To not delay the Scoping Call too much, a smaller circle of participants would be advisable.
- On inclusion of (external) experts:
  - It is recommended that identified (external) experts already participate in the Scoping Call. This way, they can gain more background information on the use case. During the Deep Dive, this will facilitate that they can provide more specific feedback.
  - From experience, the experts can serve as an additional sounding board to bring in additional questions and provide comments from their end.
  - Depending on how closely the external experts can and want to be involved: they can also be more strongly included in the preparation process (e.g. by reviewing the information on the use case, providing additional questions for the scoping call and co-developing initial hypotheses for bias).

[4] As mentioned above, for an exemplary template, please refer to the Quantitative Guide, chapter 2: *Preparatory Work*.

Example of questions for the Scoping Call

The questions below are **exemplary from an AI use case** that was assessed in the pilot phase of the Responsible AI Assessments. The use case is centered around coffee yield prediction. Information was available on the data collection and the model.

Datasets and algorithm:

- How many of the collected coffee plant photos are useful for building the model?
- Have you tried other algorithms apart from regression models?
- What is your approach towards open sourcing the datasets and the model?

Data privacy and awareness:

- How do you communicate the project objective with farmers and ensure their informed consent to the data collection and later processing?
- How is personal or sensitive data, if any, handled within these datasets?

Transparency and fairness:

- Is the model transparent and explainable in its decision-making process? How do you ensure the model's output can be understood and scrutinized?
- What are the main sources of possible errors or biases? (e.g. with respect to the quality of photos taken, the characteristics of the coffee farms, selection of farmers for collaboration, etc.)
- How do you address potential biases in the data? Are there measures in place to ensure fairness in the decision-making process?

Example of bias hypotheses

Below you will find exemplary bias hypotheses from the AI use case on coffee yield prediction.

Type of bias	Application to use case
Sampling bias	Since training data predominantly comes from specific regions, with specific farming practices (i.e. one with more advanced farming practices), this **might not accurately represent the diversity of farming conditions** in other regions, leading to biased recommendations for farmers outside that region.
Selection bias	If the system collects data primarily from farmers who have access to smartphones and can upload photos, it might exclude those with limited access to technology, **potentially biasing the system toward more tech-savvy farmers**.
Labeling bias	If labels assigned to coffee crop photos are influenced by preconceived notions about what a healthy coffee plant looks like (i.e., pruning, soil erosion control, sucker removal, weeding, pest and disease control), the AI **might learn biases in labeling, impacting the accuracy of yield estimation**.
Historical bias	If historical data on coffee crop health and practices reflects past gender inequalities, where **women's contributions to coffee farming** were **underrepresented**, the system may perpetuate these inequalities in its recommendations.

2.2 Deep Dive

The Deep Dive represents the core of the Responsible AI Assessments. Based on the Qualitative and Quantitative Assessment Guides, its main function is to:

- **Analyse risks and issues** as identified in the scoping call
- **Reflect on** the potential **impact** of those risks
- **Draft mitigation measures**, suited for the local context and domain

General information:

- Recommended time for the Deep Dive session:
  - at least 90 min.
  - best case 120 min.
- Required participants:
  - Facilitators & main assessors
  - Core team (e.g. AI developers, project managers)
  - Additional relevant team members (e.g., operational managers, compliance officers, ethicists)
  - (external) data ethics and inclusion experts
  - (external) domain experts
- Highly recommended:
  - Representatives of end-users or end-beneficiaries

Step-by-step guide

Before Deep Dive

- Based on the scoping call, **refine hypotheses for** biases and human rights **risks**.
- Based on your agreement and hypotheses, **develop initial questions** for the Deep Dive, by using the qualitative and quantitative guides as an inspiration. Adopt those questions to the use case at hand.
- Develop a **draft agenda** for the Deep Dive.
- In advance, **share questions and agenda** with the project team to receive their perspectives on what might be missing and refine them (if necessary).

During Deep Dive

- Facilitators provide the **platform for engaged discussion** on the prepared questions, navigating inputs from the project team and (external) experts,
- Facilitators strive for **problem-oriented** discussion, while also leaving the time for addressing concrete mitigation measures.
- **Experts** provide contextual analysis for the use case at hand. Their active involvement allows to address issues from diverse perspectives.

After Deep Dive

- Prepare Follow-up (see section 4.4)

Best practices

- Take the time to tailor the questions from the Qualitative and Quantitative Guides to the use case at hand as much as possible (see examples below). This allows to make the questions more tangible for the project team and allows for more targeted discussions.
- When sharing the initial questions via e-mail, divide between:
  - Information that you still need for preparing the call that the project team can answer quickly (e.g. what data sources were used for model development)
  - Questions for the scoping call that are more open and focused on discussion.
- To keep the call manageable, approximately 10-12 questions for open discussions are recommended.
- Two weeks of space between the Scoping and Deep Dive proved to be effective. Since more people are ideally involved in the Deep Dive, schedule the meeting accordingly in advance.
- Hypotheses on bias proved to be a tangible tool for the project team to understand potential issues. They were built based on the moments of bias within the Qualitative and Quantitative Guides.
- Even if a Deep Dive mainly focuses on a certain stage, e.g. pre-processing, it can be still beneficial to include questions for other stages, e.g. post-processing, as first brain teasers.
- Experts serve as an additional sounding board that provides additional, otherwise potentially underrepresented perspectives to the discussion. Among others, their critical input helps to highlight local contextualization, inclusion, domain expertise or potential effects on end-users or beneficiaries. In this regard, the participation of experts that also represent the target group of the assessed AI system is highly recommended.

Example of questions for the Deep Dive:

The questions below show how the questions from the Qualitative and Quantitative Guides can be tailored to an AI use case. In this case, they were adapted for an AI use case on coffee yield prediction.

Original questions	Adapted Deep Dive questions
How do you ensure your data to be **representative of the target population**?	How do you ensure that your data **represents a diverse range of coffee crops**, including variations in regions, altitudes, and farming practices?

Have you considered the potential impact of **missing data points** on the system's performance?	Which **data points** are missing to complete the objective of the project and how do you intend to obtain them? Consider the following data points: • At coffee cherry level (type of cherry, color, bud and occlusion) • Historical data (i.e. previous yields) • Socio-demographic (name, age, phone number, province, region, GPS, etc.)
How do you **define a 'fair outcome'** for the users in terms of the AI model's predictions or decisions?	Considering the two main target users: cooperatives and small holder farmers: are there concerns about **unequal access to acquiring loans** based on the crop yield prediction system? • Do cooperatives have an advantage over individual small holder farmers? • Is there also a gendered advantage for male over female small holder farmers? • How could existing benefits for cooperatives affect the outputs of the algorithm? (i.e. regions where cooperatives could have more influence)

2.3 Risk & mitigation measures report

This final stage aims at sharing a short report that documents identified risks and mitigation measures. This report should contain **detailed and actionable recommendations** on the identified issues.

General information:

- Recommended length: report of 3-5 pages
- Required participants:
  - main assessors
  - (external) data ethics and inclusion experts
  - (external) domain experts
- Highly recommended:
  - representatives of end-users or end-beneficiaries

Step-by-step guide

Drafting stage

- Based on Deep Dive, **draft initial recommendations**. It is recommended that the external AI assessors involved drive this process.
- **Share the draft** with (external) experts and – wherever possible – with representatives of end-users or beneficiaries to capture their diverse perspectives in the report. Should privacy reasons or reasons of institutional confidentiality rules prevent you from sharing the full draft, consider sharing parts of the draft.
- **Share the finalized report** with the AI project team and everyone involved.

- Follow-up with the AI project team on earlier discussions **whether and how the final report (or parts of it) can be made publicly accessible**.

Follow-up

- Wherever possible: hold a **final meeting** to acknowledge the joint work, appreciate learnings from the process and provide a chance to clarify remaining questions and comments on the report.
- **Plan for a process to follow-up on the mitigation plan. This might at the minimum involve a follow-up call** to jointly check in on how the issues have been addressed and sketch out ways to create continuous accountability (e.g. by quarterly check-in mechanisms, assignment of AI ethics experts to monitor implementation of mitigation measures, etc.).

Best practices

- Take the time to tailor the recommendations to the use case at hand as much as possible (see examples below). This makes recommendations more actionable for the project teams and may facilitate follow-up discussions on concrete mitigation measures outlined in the report.
- In the report, order/prioritize the proposed mitigation measures so that issues requiring more timely intervention are highlighted or mentioned first. If relevant, summarize important issues and aspects to monitor for the future development stages of the AI system (e.g. deployment stage).
- To make recommendations more specific and accountable, we recommend defining for each recommendation:
  - What: Concrete description of necessary mitigation steps
  - Who: The person responsible for each step
  - When: due date for each step, based on priority level.[5]
- Where possible, agree on a plan for continuous implementation of the report and accountability mechanisms set in place. This might also include a set number of future check-in meetings to assess the implementation of mitigation measures and offer further guidance on the same.

Example of recommendations:

Below you will find exemplary recommendations that have been created for the AI use case on coffee yield prediction.

- The team could integrate a **feedback mechanism** within the application that allows users to provide information on their specific farming practices. This could foster the creation of a dynamic and evolving dataset that adapts to the diversity of agricultural conditions.

[5] The following sources might serve as an initial recommendation: Agbede (2021) or template for SMART Goals like this one.

Associated data points may include:
    - geographical and climate data (different regions where the coffee yield prediction system may be deployed, soil types, temperature variations, etc.),
    - crop varieties and agricultural practices (planting seasons, irrigation methods, pest control strategies), or
    - user demographics and preferences (location, farm size, years of farming experience, etc.).
- If the data is made available publicly, consider implementing **anonymization and aggregation techniques** as these are crucial for protecting individual farmers' identities.
- The team could integrate **bias detection algorithms** or protocols into the system management to identify and rectify any existing biases in the data. For example, techniques like re-sampling, re-weighting, or data augmentation can be employed to balance gender representation in the dataset.

2.4 Responsible AI Assessments: What will happen next?

The method as presented in this Step-by-Step Guide is a pilot approach that proved fruitful during its test phase in 2023. FAIR Forward plans to further refine the method in 2024, with a particular focus on making it more in-depth, iterative and accountable.

To enhance the existing method and structure of the Responsible AI Assessments, we see it as important to:

- Establish a **community for users of the Responsible AI Assessments** to share experiences, best practices, and challenges.
- **Update the Responsible AI Assessments** and its guides regularly, incorporating user feedback and advancements in fairness research.
- **Provide training sessions or tutorials** covering both theoretical concepts behind AI fairness and practical application of the guides.
- Facilitate **integration** of the Responsible AI Assessments **into the AI development workflow**.
- **Cooperate with others** to embed the Responsible AI Assessments within existing AI development environments or frameworks.

3 Conclusion

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 AI resources, emphasizing human rights and ethical considerations throughout the AI lifecycle. They offer a systematic approach to address potentials harms and risks in AI systems, providing both the Qualitative and Quantitative Guides for a holistic assessment.

The Qualitative and Quantitative Guides are intended to foster a reflexive and collaborative design of AI systems, providing actionable insights for developers, researchers, and policymakers. From our perspective, they support and promote alignment with critical core principles for Ethical AI, including inclusivity and collaboration for positive societal impacts.

As a guide for assessing algorithmic fairness in AI activities, the Responsible AI Assessments have proven valuable to FAIR Forward and its partners. Implementing this method has yielded significant outcomes and learnings for FAIR Forward to navigate the complex field of AI fairness by providing actionable recommendations and best practices on how to refine AI activities and mitigate risks.

In 2024, FAIR Forward plans to publish the Responsible AI Assessments open-source and further develop its method and underlying guides. If you want to get involved and contribute, please let us know via fairforward@giz.de.

Nonetheless, it is important to remember that these assessments cannot guarantee absolute harmlessness of an AI solution. Ultimately, their impact will depend on commitment and available resources. Still, the Responsible AI Assessments can play a crucial role in mitigating potential harm.
Implementing these recommendations would facilitate that the Responsible AI Assessments evolve with the dynamic landscape of AI technologies, contributing to a more responsible and inclusive AI future.

Annex

Annex 1: How the Responsible AI Assessments were created

The Responsible AI Assessments were created based on a **co-creative and iterative design**. They represent an algorithmic risk and ethics assessment that prioritizes the involvement of diverse stakeholders. The following subsections briefly detail how the different components and phases of the Responsible AI Assessment were developed and validated.

a. Outlining of AI assessment guides:

The Qualitative and Quantitative Assessment Guides build on an **in-depth documentary review,** delving into the existing discourse, ethical frameworks, and relevant guidelines in the field of AI ethics. Particularly, the Qualitative Guide was structured around the core principles that UNESCO defined for a human-rights based approach to the Ethics of AI (2021).

b. Validation activities:

To test and develop the Responsible AI Assessment and validate its Qualitative and Quantitative Assessment Guides, the following activities were conducted:

- **7 AI activities** from diverse contexts in 6 countries on the African and Asian continent were **selected as test cases** for the Responsible AI Assessments. The project team that worked on these AI use cases shared available documentation (e.g. reports, Github pages, collected training data or models).
- **Scoping calls** were conducted for each AI activity with relevant stakeholders. These calls were based on draft versions of the qualitative and quantitative assessments. The scoping calls aimed to:
    - Identify challenges and risk across the different stages of an AI life cycle (ideation stage, data collection, model development or deployment).
    - Align with participating stakeholders on challenges and issues that should be discussed more in-depth within the so-called “Deep Dives”.
    - Provide an outlook on the Deep Dive session for participating stakeholders.
- **Deep dive sessions** followed that aimed to
    - assess more in-depth challenges that were identified during the scoping call.
    - provide recommendations on how to mitigate issues that were identified.

  These sessions focused on active participation, critical analysis, and solution-oriented discussions between participating stakeholders and facilitators.
- As a follow-up to the deep dive session, Eticas produced **short written reports** that were tailored to each of the examined AI activities and synthesized key recommendations to mitigate identified risks.

For a **detailed step-by-step guide** to these validation activities**, see** section 2**.**

c. Iterative approach:

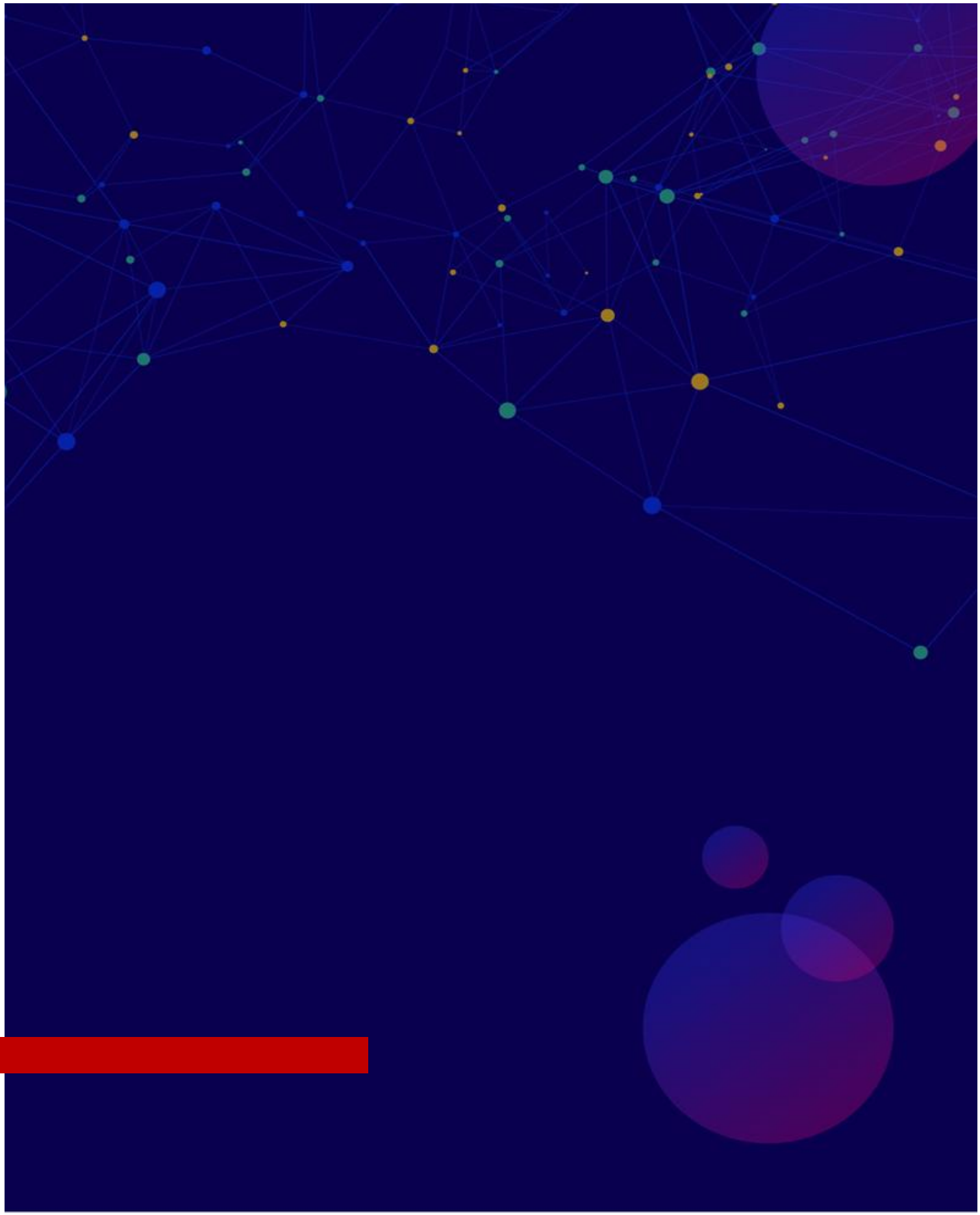
An iterative approach underscored developing the Responsible AI Assessments; feedback from each phase informed subsequent refinements of the Qualitative and Quantitative Assessment Guides as well as the accompanying method to apply the guides.

One pivotal element in developing the Qualitative and Quantitative guide was the involvement of a diverse pool of **AI inclusion experts** from Sub-Saharan Africa and Asia Pacific. They **provided invaluable insights** for developing the Qualitative and Quantitative Guides with regards to digital human rights, inclusion, gender or sector-specific considerations, among others.

As insights were gained from expert input, scoping calls, deep dive sessions, and feedback from participants, the Qualitative and Quantitative Guides evolved, as did the Responsible AI Assessments, ensuring their adaptability and relevance to the contextual landscapes of AI ethics.

Annex 2: Comparative view on two exemplary assessment cases

	Conversational chatbot for public service	Coffee yield prediction
Context	An AI-powered chatbot provides comprehensible information and guidance on adhering to the local Data Protection law, both in English and the local language. It has been co-developed with government partners in one of FAIR Forward's African partner countries.	The project piloted an approach where small-holder coffee farmers leverage computer vision AI-models for generating crop yield estimation. This shall enable the farmers these quantitative insights with coffee cooperatives and other agronomic stakeholders to receive agronomic tips.
Exemplary gaps and high-level mitigation measures	• **Gap 1:** Construction of personas may stigmatize or lack diversity. **Recommendation.:** Expand user categories, ensure dataset diversity, and define fairness goals. • **Gap 2**: Lack of socio-demographic data may lead to bias. **Recommendation:** Increase conversational data, evaluate fairness metrics, and check conversational analytics for bias. • **Gap 3:** Model may not answer all user questions. **Recommendation:** Implement a robust user feedback system, conduct live data monitoring, and encourage user queries.	• **Gap 1:** Potential gender bias in dataset because coffee farming is dominated by male farmers. **Recommendation:** Ensure diverse data collection, incorporate bias detection. • **Gap 2:** Collection and processing of meta data (location, farm size, historic yields) that may indirectly identify farmers. **Recommendation:** Implement robust data protection, pseudonymization/anonymization and encryption measures. • **Gap 3:** Accuracy of the coffee yield model **Recommendation:** Refine CNN architecture, leverage transfer learning, improve outlier detection.
Summary main aspects	The assessment highlighted the importance of diverse data representation, language inclusivity and continuous model improvement that considers literacy levels and socio-demographic factors alike.	The assessment highlights the significance of diverse data sets, robust privacy techniques, and measures to balance gender-related bias.



Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Registered offices
Bonn und Eschborn

Friedrich-Ebert-Allee 32 + 36
53113 Bonn, Deutschland
T +49 228 44 60-0
F +49 228 44 60-17 66

Dag-Hammarskjöld-Weg 1-5
65760 Eschborn, Deutschland
T +49 61 96 79-0
F +49 61 96 79-11 15

E fairforward@giz.de
I www.giz.de