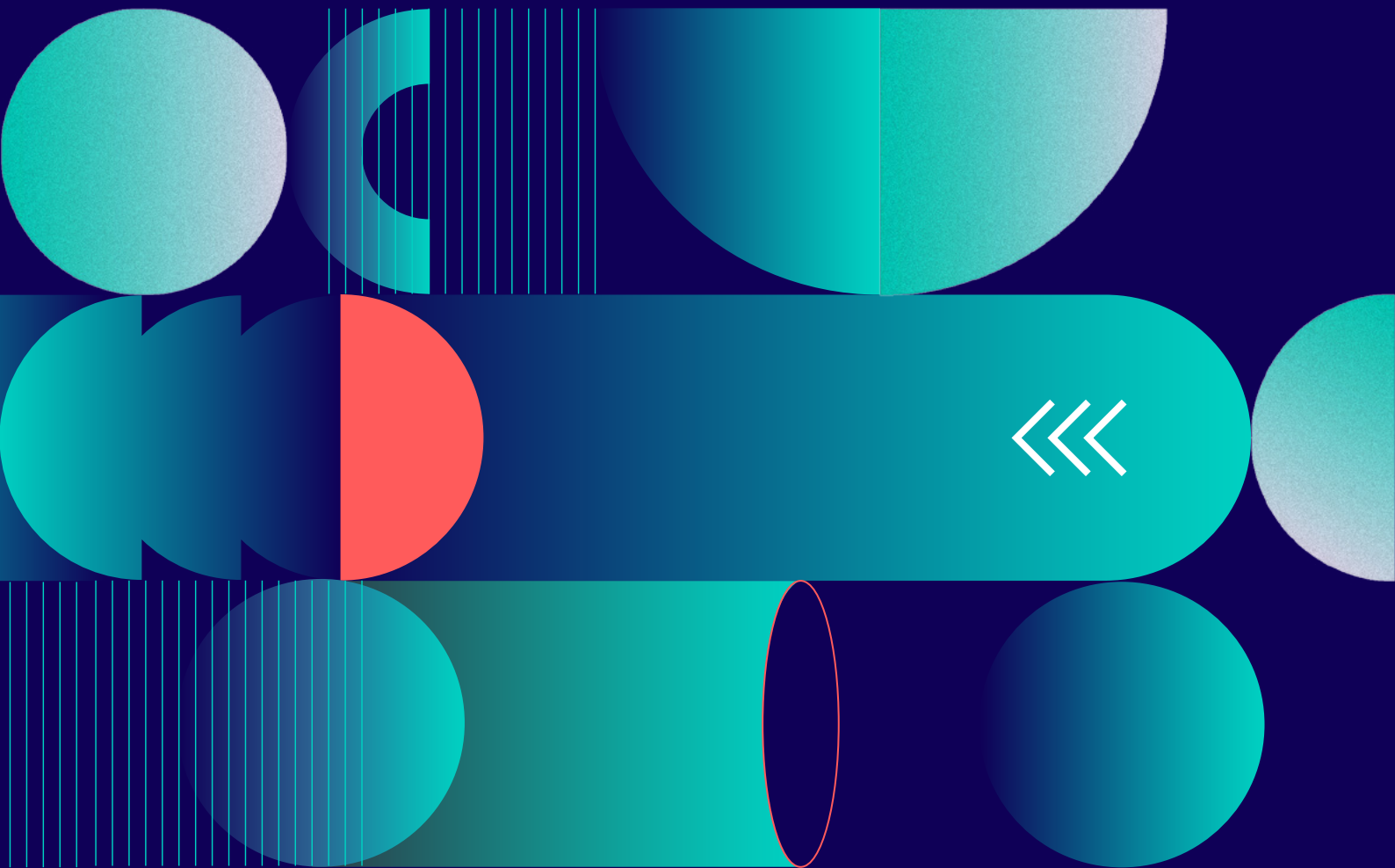
german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

FAIR FORWARD
Artificial Intelligence for all.

RESPONSIBLE AI ASSESSMENTS GUIDE

Identify and assess potential harms and biases in AI systems with a focus on Sub-Saharan Africa and Asia Pacific use cases



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

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

RESPONSIBLE

GIZ - FAIR Forward – Artificial Intelligence for All
Deshni Govender – AI Advisor (deshni.govender@giz.de)
Daniel Brumund – AI Advisor (daniel.brumund@giz.de)

AUTHORS

Eticas: Melissa Robles, Catalina Bernal, Oliver Smith, Gemma Galdon Clavell
FAIR Forward: Deshni Govender, Daniel Brumund, Abeera Dubey

CONTENT REVIEW

FAIR Forward: Nadine Dammaschk, Jonas Gramse, Deshni Govender, Daniel Brumund, Abeera Dubey, Kathleen Ziemann, Balthas Seibold
Community of AI inclusion experts: Meena Lysko, Josia Paska Darmawan, Mercy King'ori.

SPECIAL ACKNOWLEDGMENTS

We are additionally grateful to:

- Nadine Dammaschk and Jonas Gramse for steering the activity in the initial phase, as well as for continuous support and contributions throughout the process.
- The general FAIR Forward team and involved partners for their continuous feedback and contributions.



This document, including the guide and annexes, is available in Open Access under the **Attribution-ShareAlike 4.0 International** license CC BY-SA 4.0 DEED.

ATTRIBUTION

You must give **appropriate credit**, provide a link to the license, and **indicate if changes were made**.

Please cite as follows:
Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) - FAIR Forward (2025). Responsible AI 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** 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, you can notify us briefly by reaching out to **digital-transformation@giz.de**.

DISCLAIMER

The data in this publication has been collected, analysed, and compiled with due care; it 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, April 2026

TABLE OF **CONTENTS**



I. Executive summary 5

II. Acknowledgments 7

01. Introduction 8

02. Case studies 11

03. What can stakeholders achieve with the RAIA guide? 15

04. Risk definitions and approach 17

05. The five phases of RAIA 19

- 5.1. The AI lifecycle 20
- 5.2. Phase 1: Pre-survey 21
- 5.3. Phase 2: Scoping Call 22
- 5.4. Phase 3: Analysis 23
- 5.5. Phase 4: Deep Dive 24
- 5.6. Phase 5: Risk mitigation and post-production/continuous monitoring 26

06. Assessment framework and scoring system 27

- 6.1. the assessment lifecycle 28
- 6.2.Guidance on thresholds and benchmarks 28
- 6.3. Assessment scorecard 29

07. Risk assessment across stages 30

- 7.1. Pre-processing: Data collection and preparation 30

  7.1.1. Bias and fairness 31
  7.1.2. Privacy and confidentiality 33
  7.1.3. Reliability 34
  7.1.4. Governance 35
  7.1.5. Security and misuse 36
  7.1.6. Pre-processing scorecard 37

TABLE OF **CONTENTS**

> 7.2. In-processing: Model training and evaluation 39

7.2.1. Bias & fairness 41
7.2.2. Privacy and confidentiality 43
7.2.3. Reliability 44
7.2.4. Governance 45
7.2.5. Security and misuse 45
7.2.6. In-processing scorecard 46

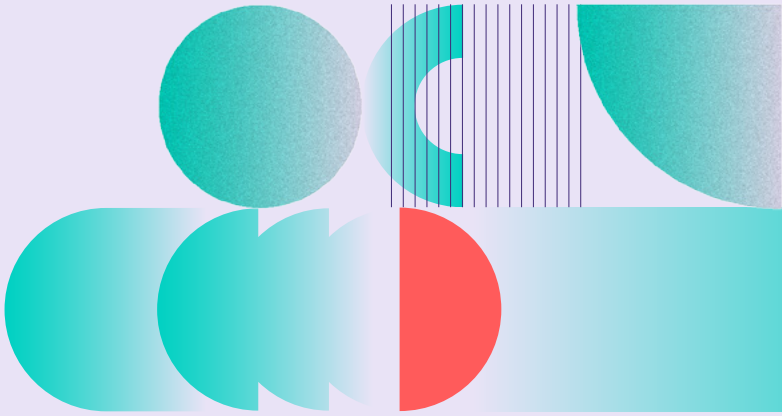
> 7.3. Post-processing: Deployment and monitoring 50

7.3.1. Bias and fairness 52
7.3.2. Privacy and confidentiality 53
7.3.3. Reliability 54
7.3.4. Governance 55
7.3.5. Security and misuse 56
7.3.6. Transparency and explainability 57
7.3.7. Responsibility and redress 58
7.3.8. Post-processing scorecard 59

08. Conclusions 63
09. Biographies: community of AI inclusion experts 64
10. Templates 65
11. References 66

III. Annex A 70

EXECUTIVE SUMMARY



Socio-economic inequalities within and between countries are exacerbated by emerging technologies, producing new forms of discrimination and exclusion. Simultaneously, rapid advances in generative AI increase exposure to harms that undermine the rights of individuals, communities, and institutions. Integrated, holistic, and actionable artificial intelligence (AI) evaluation methods are now more crucial than ever.

The Responsible AI Assessments Guide (RAIA Guide) offers a structured methodology to identify, evaluate, and mitigate risks across an AI system's entire lifecycle from the earliest design and data collection stages through to deployment, monitoring, and continuous improvement. It is a risk and ethics assessment tool designed to guide stakeholders involved in AI systems (evaluators, developers, and deployers of AI, etc.).

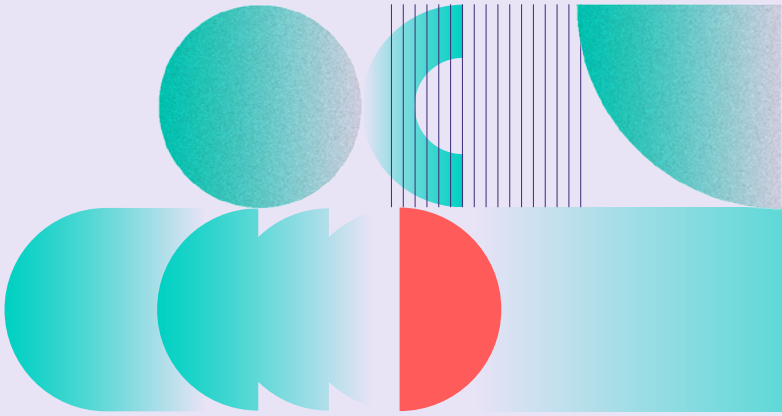
The guide cements human rights and ethics at the core of its assessment methodology, offering guidance to stakeholders dealing with AI systems.

Over a three-year period, the GIZ project FAIR Forward – Artificial Intelligence for All, Eticas, and a diverse community of AI inclusion experts from Sub-Saharan Africa and Asia Pacific collaborated to identify, audit, and simplify methods to inspect AI systems.

The RAIA Guide – a bottom-up exercise – is a tactical approach to this, premised on thirteen AI case study assessments developed for and deployed in Sub-Saharan Africa and Asia Pacific.

1. Energy demand mapping and site identification in Uganda
2. Entrepreneurship policy chatbot in Ghana
3. Micro-climate forecasting and advisory chatbot in Kenya
4. Chatbot for digital government services in Kenya
5. Agricultural Information Exchange Platform in Kenya and India
6. Improving classification accuracy of Indian forests' carbon storage potential
7. Chatbot to access and action public sector audits in Uganda
8. AI-supported coffee yield information in Uganda
9. Comprehensive language corpus for training in local languages in Kenya
10. Landslide detection in Rwanda
11. Crop mapping in the Indian state of Telangana
12. Cashew disease identification in Ghana
13. High carbon stock mapping in Indonesia

EXECUTIVE SUMMARY



The RAIA Guide offers a clear and consistent method to examine AI systems. It supports risk identification and mitigation before, during, and after the deployment of AI systems, enabling teams to uncover risks early, document them systematically, and design interventions effectively.

first version of this guide and framework, covering a subset of use cases, was published in 2024 and is available on the FAIR Forward and Open for Good websites. This revised version (version 2) covers insights from additional use cases, some of which use generative AI capabilities. Version 2 consolidates the original methodology (of separate qualitative and quantitative guides) into a single, integrated, sociotechnical assessment, providing a holistic view of ethical, legal, and technical risk perspectives.

The Version 1 methodology was explicitly structured around the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021). In version 2, these high-level principles are translated into operational risk categories which are a better fit for evaluators, auditors, and developers alike: bias and fairness, privacy and confidentiality, reliability, governance, and security and misuse.

This adaptation expands on the UNESCO principles for practical day-to-day assessment work. The risk categories are consistent with other international frameworks, including the OECD AI Principles, human-rights-based approaches, and emerging standards such as ISO/IEC 42001 and ISO/IEC 23894). They are intended as a "working lens" to implement the underlying ethical commitments in concrete projects.

This guide further introduces system-specific guidance for both Automated Decision-Making (ADM) systems and Large Language Models (LLMs).

Drawing on real-world AI audits and evaluations, the RAIA Guide is intended to be a living framework. It is designed to be adaptable and responsive to AI technological developments, such as the rise of large language models and the risks they pose to individuals, economies, and society.



ACKNOWLEDGMENTS

FAIR Forward and Eticas express their sincere gratitude to the partner organisations and the wider community of AI inclusion experts who contributed to the co-creation of the Responsible AI Assessment Guide.

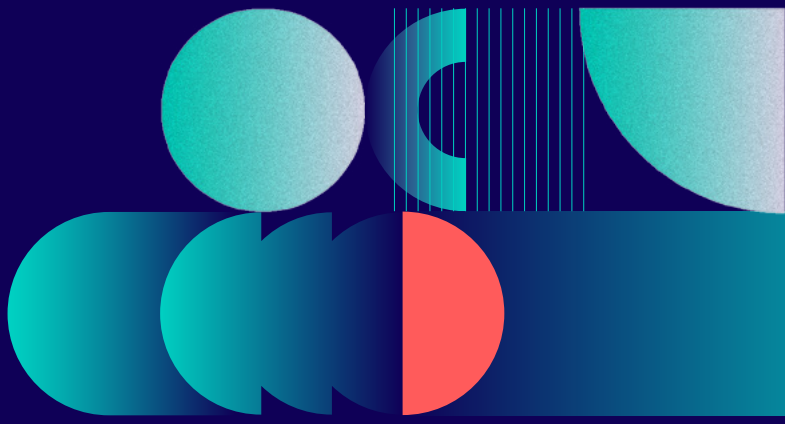
The openness and commitment of FAIR Forward partners made it possible to pilot, test, and iteratively improve early drafts of the assessments within real-world AI initiatives. Through this early engagement, FAIR Forward partners provided essential feedback that informed the design choices and the practical applicability of the Assessments in this Guide.

We further extend special thanks to our group of AI inclusion experts – Josia Paska Darmawan, Mercy King'ori, and Dr. Meena Lysko – for their critical insights and sustained engagement throughout the process, and to the group of experts who helped create version 1.0 of this guide – Favour Borokini, Mohamed Kimbugwe, Raashi Saxena, and Kofi Yeboah. Their perspectives were instrumental in refining and stress-testing the assessment methodology across a range of AI use cases and contexts.

This guide is the result of a collective effort, reflecting the shared expertise, openness, and commitment of the individuals and organisations who contributed their time, knowledge, and experience. This collaboration has made the development of the Responsible AI Assessment Guide possible.

CHAPTER 01

INTRODUCTION



AI systems do not operate in a vacuum. Every AI system is shaped by the social, economic, political and institutional context in which it is designed and used. The same AI model can behave differently depending on its purpose, training data, integration into workflows, and impacted groups. Accordingly, managing AI-related risks means looking beyond algorithms and examining full AI systems in context, including their objectives, data practices, user interfaces (UI), governance arrangements, and real-world use.

The Responsible AI Assessments (RAIA) Guide was created precisely to enable this holistic scrutiny It provides a structured way to identify, evaluate, and mitigate risks across an AI system's entire lifecycle from the earliest design and data collection stages through to deployment, monitoring, and continuous improvement. Ethical principles and technical performance are not mutually exclusive concepts, and this Guide aligns them into a single, integrated process, to bridge human rights considerations with rigorous model testing.

The Guide supports AI stakeholders (including evaluators, developers, and deployers) to critically examine how their systems interact with people and communities, and to help them spot potential problems before they materialise into harm or risk.

The approach is a data-driven, bottom-up and practice-tested, developed through hands-on rigorous assessments of real AI projects. Over a three-year period, the team (spearheaded by GIZ's FAIR Forward initiative, supported by Eticas) worked

The Guide supports AI stakeholders (including evaluators, developers, and deployers) to critically examine how their systems interact with people and communities, and to help them spot potential problems before they materialise into harm or risk.

with local experts in Sub-Saharan Africa and Asia Pacific to inspect and stress-test AI systems in action, as detailed in the case studies (Chapter 2). Insights from these assessments directly shaped the Guide's evaluation techniques and checklists, forming the practical focus of the Guide.

Crucially, the Guide's perspective is grounded in real-world applications in the Global South, especially in domains that are under-studied in mainstream AI ethics research. The pilot assessments tackled high-stakes challenges in sectors such as agriculture, disaster management, energy planning, and public service delivery. These applications are where AI impacts livelihoods and well-being: a farmer's crop advice chatbot, an early warning system for floods, an off-grid energy planning tool, or a digital government service. Such domains have historically received diminished attention compared to entertainment, e-commerce, or productivity tools found in the Global North.

Shifting the focus to development and public interest use cases rather than commercial or corporate contexts, the Guide brings a distinctive perspective to contexts where resources may be limited, data might be scarce or biased, and the social stakes are high. A one-size-fits-all, top-down playbook is outdated. This unique grounding ensures that the Guide's methodology speaks to real problems faced in emerging and low-resource settings, rather than recycling generic best practices. It further highlights issues of inclusion, equity, and local relevance that might be overlooked elsewhere, establishing the Guide's credibility as an assessment framework tested on real-world AI system use cases.

The purpose is to unite ethical foresight with technical rigor in one integrated evaluation process.

By evaluating the key dimensions of a socio-technical system, RAIA Guide enables a 360-degree view of risk: from biases in training data, to interface design pitfalls, to breakdowns in human oversight.

Mitigating actions proportionate to the system's context and impact are preferred over standardised abstract solutions, as responsible AI isn't a one-off checkbox exercise but an ongoing commitment. As an AI system evolves or new risks emerge, the RAIA process supports continuous learning and adaptation.



WHO IS THE RAIA GUIDE INTENDED FOR?

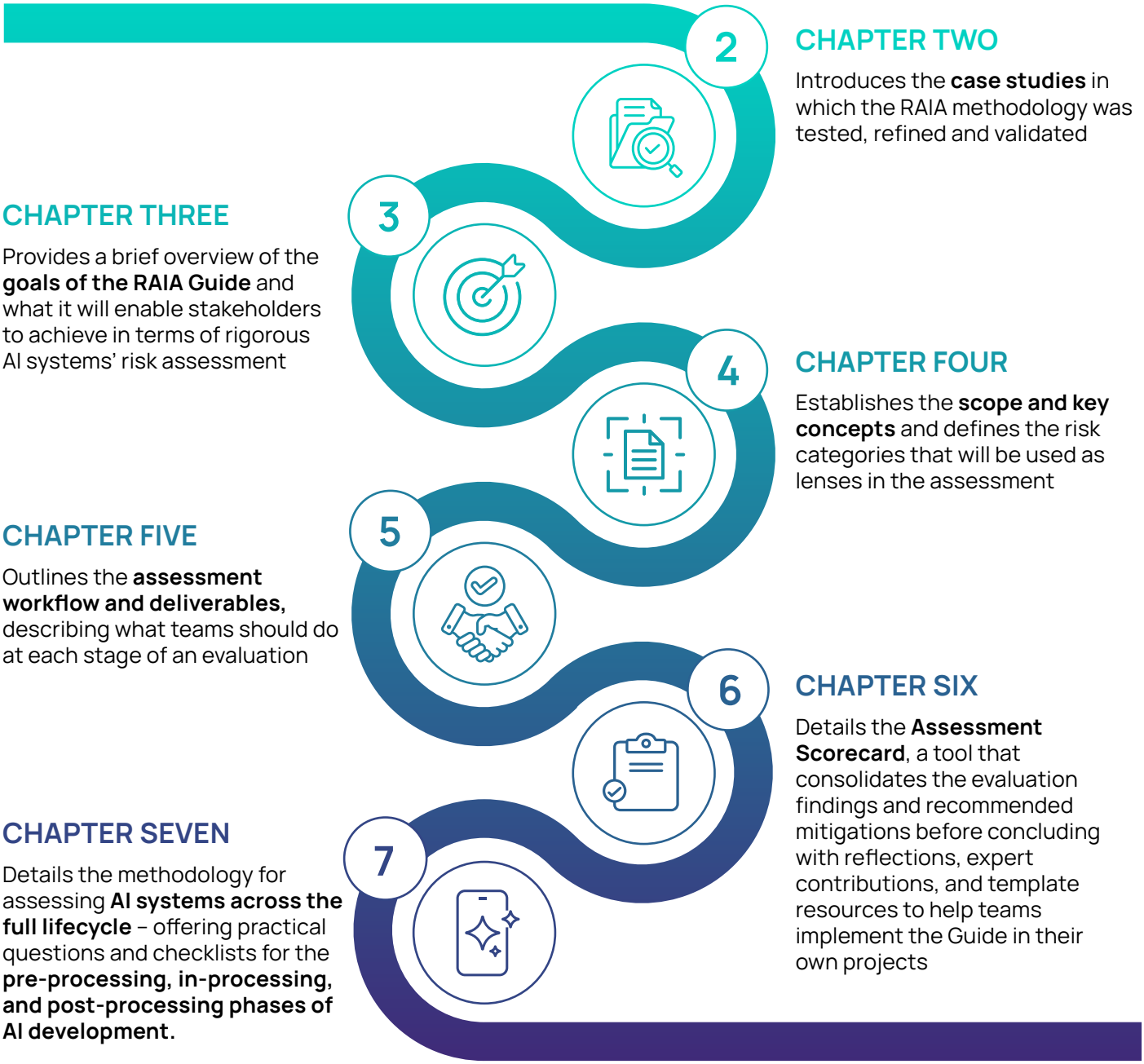
The people and organisations building, deploying and evaluating AI systems – such as AI auditors tasked with inspecting systems, developers designing and training models, project managers integrating AI into services, and policymakers or deployers overseeing AI initiatives.

In these environments (a government agency rolling out an AI-powered public service, a non-profit using machine learning for disaster response, or a startup developing AI for smallholder farmers) the constraints and stakes differ from those in commercial technology (industry) product development.

Projects operate under uncertainty and limited resources, yet their social impact is immense. While the RAIA guide is engineered to support all teams, it is aimed at the under-served, resource-constrained practitioners, providing a clear process and common language to traverse risks and responsibilities. Practitioners in emerging markets or mission-driven projects can systematically work through potential ethical issues and safety hazards, even if they don't have large AI governance departments. RAIA aims to empower AI teams everywhere to build systems that are innovative, safe, fair, and aligned with the communities and purpose they serve.

ROADMAP:
WHAT IS IN THIS GUIDE?

In the chapters that follow, the RAIA methodology is unpacked step by step:



CHAPTER TWO

Introduces the **case studies** in which the RAIA methodology was tested, refined and validated

CHAPTER THREE

Provides a brief overview of the **goals of the RAIA Guide** and what it will enable stakeholders to achieve in terms of rigorous AI systems' risk assessment

CHAPTER FOUR

Establishes the **scope and key concepts** and defines the risk categories that will be used as lenses in the assessment

CHAPTER FIVE

Outlines the **assessment workflow and deliverables,** describing what teams should do at each stage of an evaluation

CHAPTER SIX

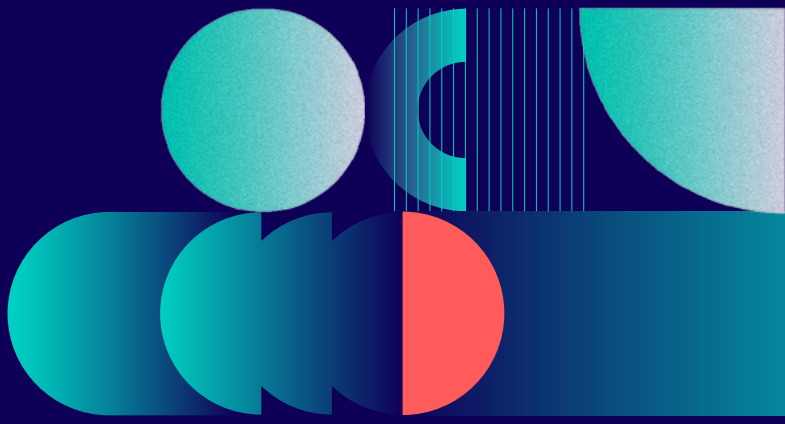
Details the **Assessment Scorecard**, a tool that consolidates the evaluation findings and recommended mitigations before concluding with reflections, expert contributions, and template resources to help teams implement the Guide in their own projects

CHAPTER SEVEN

Details the methodology for assessing **AI systems across the full lifecycle** – offering practical questions and checklists for the **pre-processing, in-processing, and post-processing phases of AI development.**

CHAPTER 02

CASE STUDIES



The RAIA methodology was tested and applied to thirteen AI cases studies across Africa and Asia. Five were assessed under phase 1 (using the previous RAIA methodology) and six assessments were conducted in phase 2 (using the revised methodology).

During the iterative phase of compiling the revised RAIA methodology, two case studies were selected for testing application. Combined, these thirteen AI case studies span different sectors, levels of technical complexity, and deployment contexts, illustrating how AI risks emerge not only from the models themselves but from systems design, integration, and live production.

1. Energy demand mapping and site identification in Uganda

A machine learning–based site-identification tool designed to detect and classify building types in rural areas and estimate their energy demand. The system supports the implementation of Uganda's National Electrification Strategy (NES) by informing decisions about the optimal locations for off-grid energy solutions, such as mini-grids and other decentralised infrastructure. The assessment focused on how design choices, data sources, and assumptions about "typical" buildings could affect the accuracy and fairness of energy planning, particularly for remote or underserved communities. Additionally, it evaluated how uncertainty and model limitations are communicated to decision-makers using the system.

2. Entrepreneurship policy chatbot in Ghana

A natural language processing (NLP)–based chatbot that provides an accessible, conversational interface to entrepreneurship policy documents. The system is intended to help Ghanaian entrepreneurs, small business owners, and other users locate relevant policy information more easily. The assessment examined risks related to reliability and governance, including the risk of outdated or incomplete policy information, uneven performance across languages, and the potential for fabricated or ungrounded responses. Attention was given to how the system signals limitations to users and how responsibility is assigned for content updates and error correction.

3. Micro-climate forecasting and advisory chatbot in Kenya

The system combines data from local Internet of Things (IoT) weather sensors with external climate datasets to generate highly localised weather forecasts. These forecasts are delivered through an SMS-based chatbot that translates predictions into practical, context-aware agricultural advice for smallholder Kenyan farmers. The assessment explored how data quality, sensor coverage, and model uncertainty affect the reliability of the chatbot's advice, especially in high-stakes situations such as planting or harvesting decisions for low-resource individuals. It further examined risks related to unequal user access, given variations in mobile connectivity, literacy, and language, as well as the role of human oversight when automated advice may conflict with local knowledge.

4. Agricultural Information Exchange Platform in Kenya and India

A multi-partner platform that supports the development of minimum viable products aimed at smallholder farmers and agricultural extension officers in Kenya and Bihar (India). It uses a retrieval-augmented generation (RAG) pipeline to deliver contextualised agricultural advice drawn from curated knowledge sources. The assessment addressed risks related to content grounding, source selection, and prompt manipulation, as well as governance challenges arising from multiple organisations contributing data, models, and services. Attention was paid to how responsibilities for accuracy, updates, and harm mitigation are distributed across partners in a shared system.

5. Chatbot for digital government services in Kenya

An AI-powered conversational chatbot designed to help Kenyan citizens discover and navigate digitised government services. The system also serves as a reusable digital layer that can be adapted across different public-service contexts. The assessment focused on governance, transparency, and accountability, such as communication about the system's role, handling errors and misdirection, and the responsibility allocation between technical providers and public institutions. Risks related to exclusion were also examined, particularly for users with limited digital literacy or internet access.

6. Improving classification accuracy of Indian forests' carbon storage potential

An AI system that gathers information about trees to improve the accuracy of forest classification. Developed by HCSA (High Carbon Stock Approach), the system aims to identify forests with high capacity to store carbon that should therefore be prioritised for protection from deforestation. The assessment examined risks related to data quality, representativeness of ground-truth data, and the governance implications of using AI outputs to inform high-stakes conservation and land-use decisions.

7. Chatbot to action and access public sector audits in Uganda

A retrieval-augmented generation (RAG)-based chatbot designed to improve public transparency and service delivery by making government audit reports more accessible and actionable. The system summarises complex audit texts, extracts thematic insights, and enables interactive, user-friendly analysis through a conversational interface. The assessment examined risks related to the accuracy of summarised content, the potential for misleading outputs, and accountability for how audit findings are communicated to the public.

8. A comprehensive language corpus for training in local languages in Kenya

This project aimed at creating a comprehensive local language corpus to serve as a resource for language technology development, linguistic research, and cultural preservation in Kenya. The assessment examined risks related to the representation of minority languages and dialects, consent and ownership of linguistic data contributed by communities, and the governance of a shared resource intended for downstream AI training applications.

9. AI-supported coffee yield information in Uganda

The Croppie AI system is designed to support smallholder coffee farmers in Uganda by estimating crop yields from smartphone images of coffee plants. The system then renders actionable insights, advice, and projections to improve farm management and sustainability. Farmers are incentivised to upload photos through a mobile app, which are analysed using computer vision and convolutional neural network models trained on agricultural imagery to count coffee cherries and infer yield. A significant finding of the assessment was the need to expand and diversify the training data beyond a limited set of regions to reduce geographic and practice-related bias, including biases related to access to smartphones.

10. Landslide detection in Rwanda

The Landslide AI system is designed to support landslide monitoring and early warning by helping authorities and partners better detect, assess, and anticipate landslide risks in vulnerable areas. It collects and processes data from multiple sources (e.g. sensors, high-resolution imagery, weather information, and geospatial data) to identify environmental conditions associated with landslides. The assessment identified potential privacy risks arising from the high-resolution images utilised, recommending privacy-bydesign measures (including anonymisation, pseudonymisation, or reduced image resolution) as recommendations, and robust data governance and access controls. Additional recommendations included the integration of new data sources, like demographic and construction data to reflect evolving risk patterns and reduce errors based on unrepresentative historical data.

11. Crop mapping in India

An AI-based crop mapping system within the Indian state of Telangana aims to support agricultural planning and decision-making by producing an accurate, dynamic, and open-access map of crop types, farm boundaries, and land use. The system works by combining ground-truth data collected by farmers (via mobile apps and surveys), remote sensing and earth observation imagery, and AI-based crop classification models. The assessment addressed several risks related to poor quality data. Improvements included enhancing robustness to low-quality images through model adaptation or higher-resolution satellite data, standardising and incentivising farmer survey inputs, and refining crop area definitions and sampling transparency to improve mapping accuracy.

12. High Carbon Stock Mapping in Indonesia

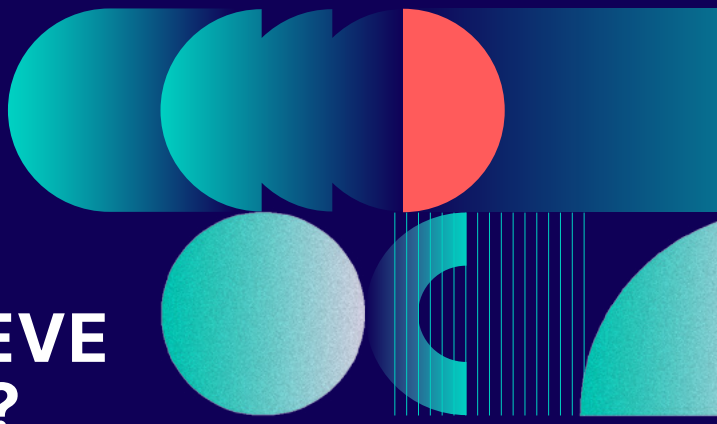
An AI-supported forest mapping and classification tool designed to track carbon storage capacity and provide data-driven decision-making for deforestation prevention. Inspiration was taken from the first version of the RAIA methodology, applying iterative testing approaches to the ground view and incorporating ethical and sociological factors into the AI system. The evaluation considered how social inequalities, historical imbalances and discriminatory patterns could feed into the data that trains the AI system, which impacts smallholder farmers, as well as landscape level and judicial approaches to land use and deforestation. Mitigation measures considered how the early detection of inaccurate or inappropriate data and legal practices could impact downstream applications.

13. Cashew disease identification system in Ghana

This project builds on an earlier machine learning–based cashew disease identification system, developed using an image dataset and a prototype application between 2022 and 2023. The follow-up project focused on piloting the system with farmers and improving existing resources, explicitly incorporating insights from a prior Responsible AI Assessment. The evaluation examined how model performance varied under real-world conditions, how uncertainty and misclassification could affect Ghanaian farmers decisions, and how feedback from end users was integrated into system refinement and governance processes.

CHAPTER 03

WHAT CAN STAKEHOLDERS ACHIEVE WITH THE RAIA GUIDE?



The RAIA Guide is designed to support rigorous, system-level evaluation of AI systems in real-world conditions.

It enables stakeholders to move beyond abstract principles or model-centric checks and to exercise informed judgment about how AI systems operate, whom they affect, and where meaningful risks arise.

Practitioners can:

Create credible, bottom-up evaluations of real AI systems

Equip auditors and evaluators with a methodology that has been developed and refined through the evaluation of deployed AI systems. Support evaluations that reflect how AI systems behave once embedded in organisations, services, and communities. By focusing on systems in operation (including data pipelines, interfaces, organisational processes, and governance arrangements), evaluators can identify risks that usually only manifest in practice, such as misalignments between technical design and institutional use, breakdowns in accountability, or unintended impacts on specific groups.

Translate global AI principles into actionable evaluation criteria

Operationalise high-level ethical and governance frameworks into concrete, testable risk categories and evaluation questions.

One example would be the UNESCO Recommendation on the Ethics of Artificial Intelligence and ISO AI impact assessment standards (e.g. ISO/IEC 42001 – AI Management System; ISO/IEC 42005 – AI Impact Assessment Guidance; ISO/IEC 23894 – AI risk management, and ISO/IEC 37004 – Governance of organisations).

This operationalisation enables auditors and evaluators to:

1. Apply globally recognised principles consistently across diverse projects
2. Produce findings that are intelligible to both technical teams and decision-makers
3. Align evaluation work with emerging regulatory and standards-based requirements without reducing the analysis to a checklist

Rather than treating ethics, compliance, and technical performance as separate exercises, this guide integrates them into a single evaluation process grounded in evidence.

Evaluate AI as a sociotechnical system, not a standalone model

Support the evaluation of full AI systems, recognising that risk often arises from interactions between models, data, people, and institutions. This system-level approach enables evaluators to examine how risks emerge across problem formulation, data practices, deployment choices, and human oversight, in addition to model outputs and accuracy metrics.

This perspective is particularly critical in public-interest and development contexts, where AI systems are often deployed in complex environments characterised by limited resources, high stakes, and diverse user groups.

Support informed judgment, not automated scoring

Strengthen evaluators' professional judgment. Through structured questions, scorecards, and documentation templates, the methodology supports consistent analysis while leaving room for contextual interpretation and expert reasoning.

This approach acknowledges that many AI risks – such as fairness, governance failures, or system misuse – cannot be fully captured through automated tests alone. Instead, they require a combination of technical evidence, contextual understanding, and deliberation.

Produce evaluation outputs that support accountability and follow-up

Provide a clear orientation towards documented outcomes.

The process and templates support the production of clear evaluation records, including identified risks, underlying evidence, mitigation measures, and follow-up actions.

These outputs can be used to:

1. Inform internal decision-making and system redesign
2. Support external scrutiny or oversight
3. Enable longitudinal evaluation as systems evolve over time

By encouraging repeated application across the lifecycle, auditors and evaluators can track whether identified risks are addressed and whether new risks emerge.

Ground AI evaluation in understudied, high-impact contexts

Enables evaluators to work with confidence in contexts that are often under-represented in mainstream AI governance discussions.

Its methodology has been tested on AI systems deployed in areas such as agriculture, climate resilience, energy planning, and digital public services across Sub-Saharan Africa and Asia-Pacific.

This grounding ensures that risk patterns that differ from those typically discussed in Global North settings are captured, where AI is more often associated with consumer services, productivity tools, or leisure applications. Accordingly, assessments contribute to a more globally relevant AI evaluation practice, informed by the realities of AI deployment where social impact is high and margins for error are low.

CHAPTER 04

RISK DEFINITIONS AND APPROACH

Version 1 of the RAIA guide was structured primarily around the principles of the UNESCO Recommendation on the Ethics of Artificial Intelligence and is a central reference for this guide. However, applying high-level principles to real AI systems revealed that assessors struggled to translate them into a concrete risk evaluation. Assessors require an operational risk structure for continued assessments, thus, a repeatable evaluation method that blends ethical, legal, social, human rights, and technical concerns was therefore crafted. High-level ethical principles cut across multiple parts of an AI system and manifest differently depending on data, model behaviour, deployment context, and governance arrangements. This revised guide introduces an integrated taxonomy of risk categories designed specifically to support hands-on assessment work. This taxonomy was developed by drawing on ISO/IEC 42005 (AI Impact Assessment) as a starting point and mapped requirements against UNESCO's principles and other governance frameworks. Overlapping or closely related concepts were deliberately merged to reduce fragmentation and avoid parallel assessments of the same underlying risk through different lenses. The risk categories are grouped using two components: thematic similarity and whether the evaluation methods used to assess them are coherent and comparable. This results in a compact set of risk categories that balances conceptual completeness with practical coherence. Each category combines risks that can be examined using similar types of evidence, questions, tests, and documentation. Assessors can move systematically from qualitative reflection to technical checks and produce findings that are consistent across systems and contexts. This taxonomy captures AI risks across real-world systems and focuses on supporting structured assessments.

The risk categories below are the common vocabulary for this RAIA guide:

Bias and Fairness:

The risk that an AI system produces outcomes that systematically advantage or disadvantage individuals or groups based on protected or sensitive attributes, leading to unequal and unfair treatment, reduced accuracy, or unjust impacts. This includes biases introduced through problem formulation, data collection, model design, or deployment context, and covers both measurable disparities in performance and broader forms of perceived or experienced unfairness.

Privacy and Confidentiality:

The risk that an AI system collects, processes, infers, or exposes personal information in ways that infringe individuals' rights to privacy or confidentiality.

This includes risks related to data leakage, re-identification, unauthorised access or reuse, excessive retention, or insufficient safeguards throughout the AI lifecycle.

Reliability:

The risk that an AI system produces incorrect, fabricated, misleading, or inconsistent outputs, or behaves unpredictably across similar inputs or contexts. This includes phenomena such as confabulation ("hallucination"), misinformation, instability under distribution shifts, and performance degradation over time. Reliability failures undermine trust in AI systems and can lead to harmful or misguided decisions, particularly in high-impact settings.

Governance:

The risk that an AI system lacks clear structures, roles, policies, or processes to ensure accountability across its design, deployment, and use. Weak governance may result in unclear ownership, poor documentation, limited auditability, insufficient oversight, or misalignment with legal, ethical, or organisational standards. Governance risks often shape whether other risks can be detected, addressed, and mitigated in practice.

Security and Misuse:

The risk that an AI system is vulnerable to attacks, manipulation, or unintended use that compromise its integrity, availability, or confidentiality. This includes adversarial inputs, data poisoning, model inversion, prompt injection, unauthorised access, and downstream misuse that can distort outputs, expose sensitive information, or enable harmful behaviour.

Combined, these categories operationalise the core values and principles of the UNESCO Recommendation on the Ethics of Artificial Intelligence.

The RAIA guide focuses on social-impact risks. Organisations may wish to examine additional dimensions: cost, efficiency, or effectiveness (the extent to which a system achieves its intended purpose).

The approach presented can also be extended to cover emerging dimensions like environmental impact, as they become operationally defined.

CHAPTER 05

THE FIVE PHASES OF RAIA

This chapter outlines the step-by-step process for conducting a Responsible AI Assessment. These steps should be combined with Chapter 7, which outlines depth and guidance on the technical aspects of the AI assessment.

1. **Pre-survey:** gathering baseline information on the AI system, its purpose and deployment context
2. **Scoping phase**: defining the focus of the assessment
3. **Analysis**: scrutinising available documentation and data
4. **Deep dive:** examining risks in detail across the risk categories and lifecycle stages (pre-, in-, and post-processing), and identifying potential mitigation measures
5. **Risk and mitigation report**: preparing key results report, and planning for a **f**ollow-up phase to discuss recommendations and next steps (see Figure 1).

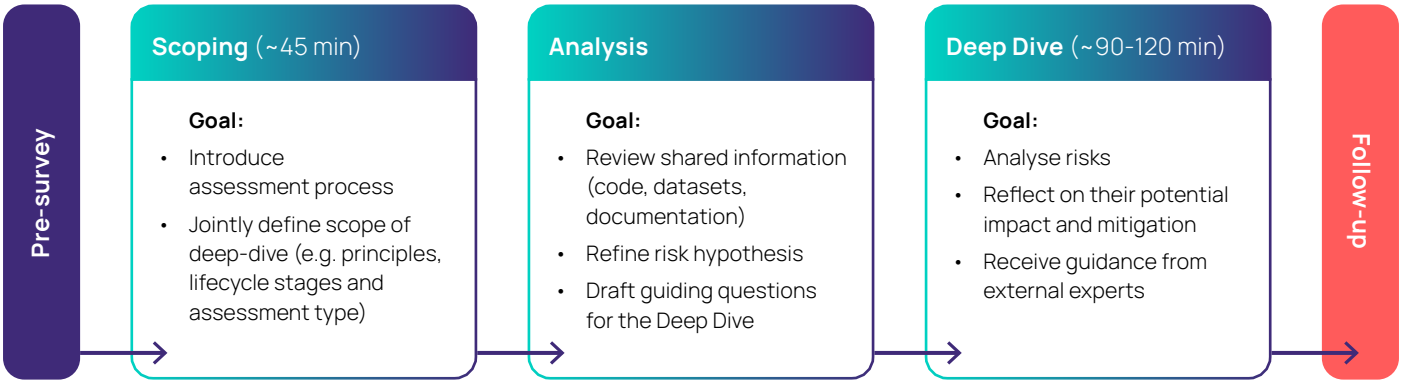


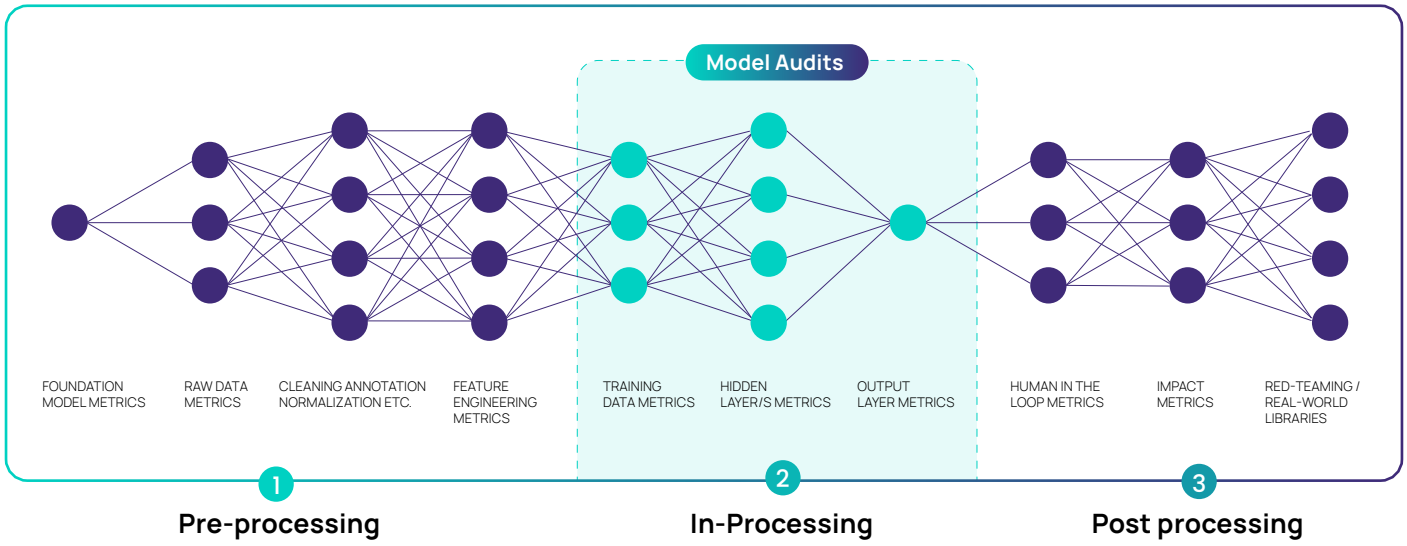
Figure 1. Phases of a Responsible AI Assessment

Combined, these steps form a coherent workflow: beginning with an understanding of the context and scope of the AI system and assessment, moving through structured discussions of risks, and culminating in a final report with concrete recommendations for risk mitigation measures.

Sections 4 to 6 of the RAIA guide build on this workflow, and for each lifecycle stage (pre-processing, in-processing, and post-processing), specify the detailed questions, checks, and tools to be used during the deep-dive phase. For each phase, these sections provide general information on the requirements and tools available in each phase, a step-by-step guide, and a list of best practices.

5.1. The AI lifecycle

Before diving into the five steps, it is important to note that AI systems are always evolving and interacting. The RAIA evaluation process is framed around the AI lifecycle, which consists of three interconnected phases.



Pre-processing

Covering the system's purpose and data, including how it's collected, prepared, and governed.

In-Processing

Covering model development and training.

Post processing

Covering deployment and real-world operation.

Figure 2: Full AI system lifecycle example.

Evaluating AI systems through this lifecycle lens allows auditors and evaluators to assess not only what a system does, but how and why risks arise at different points in time.

It creates a virtuous cycle: by breaking down and framing the lifecycle, it is easier to recommend actions at specific stages where they can have the greatest impact, enabling auditors to see those changes through.

Assessors should conduct evaluations periodically rather than only once.

5.2. Phase 1: Pre-survey

The pre-survey is the first step of a Responsible AI Assessment. It provides an open yet structured way to collect baseline information about the AI system before entering the scoping and deep-dive discussions.

The pre-survey covers key dimensions, including:

1. System development
2. Technical architecture
3. Data sources
4. Human oversight
5. Deployment context
6. Potential risks

By gathering this information in advance, facilitators, internal assessors, and external experts can better tailor the subsequent assessment phases, avoiding delays.

Answers to the pre-survey are also indicative of the development team's awareness of and maturity to tackle AI risks.

The pre-survey also serves as a documentation tool. It creates a consistent record of the systems' purpose, target beneficiaries, intended uses, and safeguards.

This record helps anchor the later risk analysis in concrete facts rather than assumptions.

Optimal Approach:

Assign the right respondent(s):

The pre-survey should be completed by someone with crossfunctional knowledge of the system (e.g., project manager, lead developer, or data governance officer). Where possible, responses should be validated internally by both technical and non technical staff.

Encourage complete answers but accept uncertainty:

Participants may not yet have answers to all questions. It is better to mark "not decided yet" than to leave gaps, as this highlights areas of uncertainty for scoping.

Use the pre-survey to surface early risks:

Some questions are designed to highlight red flags before the deep dive. Teams should pay attention to answers that suggest limited safeguards or missing processes.

Ask for supporting documentation:

When respondents confirm the existence of tests, earlier assessments, or other resources, ask for links to those resources, if possible, to better understand what has already been done and how the assessment can bring value.

5.3. Phase 2: Scoping Call

The scoping call builds directly on the responses to the pre-survey. It serves to review the responses with the development team, the evaluators, and external experts (if relevant). During a 30- 45-minute period, the group clarifies doubts and agrees on the scope of the assessment to prepare for the deep-dive session.

During the call, participants jointly identify:

1. **Final risks to be assessed:** using the risks described above in Section 4 as a reference
2. **Relevant stages of AI lifecycle for the assessment:** focus on pre-, in-, or post-processing stages
3. **Additional documents, data or code** that may be needed for the assessment

The diagram below summarises the main steps around the scoping call: what happens before, during, and after.

It is intended as a practical overview to help structure preparation, facilitation, and follow-up, and to ensure that all necessary inputs for the deep dive are identified and requested in a systematic way.

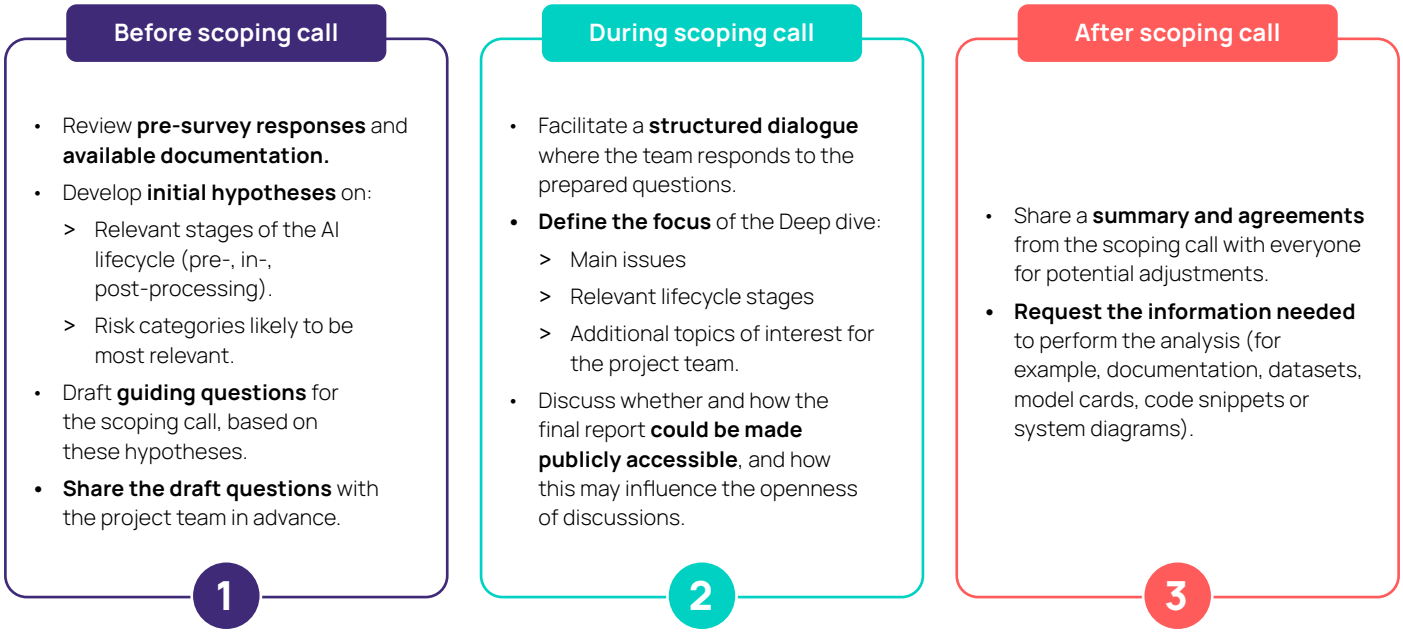


Figure 2: Full AI system lifecycle example.

Optimal Approach:

1. To keep the call manageable, select five to eight questions that allow for open discussions.
2. Keep the scoping call informal so the project team can understand the process of the assessment and feel ready for the deep dive.
3. To avoid delays, hold the call with a smaller group of essential participants.
4. If external experts are invited, they should have advance access to the pre-survey results and be encouraged to bring in additional questions and provide comments based on their experience and expertise.

5.4. Phase 3: Analysis

The analysis phase bridges the scoping call and the deep dive. Its main tasks are to:

1. Systematically review all information gathered (complete pre-survey and additional documents requested after the scoping call) and ensure it is adequate to tackle the risks to be evaluated. If information is missing, incomplete or contradictory, decide if additional content is needed before the scoping call and proceed accordingly.
2. Prepare for the deep dive by planning session questions and sending key draft findings and materials to the project team in advance.
3. Conduct a technical assessment of the AI system using the tools and guidance included in Chapter 7, which includes specific guiding questions, checks, and criteria around data quality, model robustness, evaluation, monitoring, and documentation. The outcomes of the analysis phase are summarised in a scorecard that highlights gaps, and open issues of the AI system.

Optimal Approach:

Triangulate sources of information:

Do not rely only on project team documentation alone; validate insights with code repositories, user feedback, or prior evaluations when possible.

Balance breadth and depth:

While scoping might surface a wide range of issues, the analysis should prioritise three to four key themes to focus the deep dive, aligned with the prioritised risks.

Anticipate external expert contributions:

If external experts are involved, invite them to review draft hypotheses and questions early. This enhances the quality of feedback during the deep dive.



5.5. Phase 4: Deep Dive

The deep dive is the core of a Responsible AI Assessment.

Its goal is to analyse risks and issues identified during the analysis, reflect on the potential impact of those risks, and draft mitigation measures that are suitable and practical for the project team.

A deep dive should last at least 90 minutes, have a clear agenda, and ensure all participants have had access to all relevant materials beforehand. All key parties should participate in the deep dive, as it presents the last opportunity for a team to discuss and refine the assessment scope and preliminary findings. It may be relevant to include external experts as well as end-users and impacted communities, to ensure robustness.

Supporting materials should guide the discussion, with a facilitator and an evaluator responsible for taking notes and connecting concepts. The discussion should always be problem-oriented and leave space to discuss all issues and potential mitigation measures. Before concluding the meeting, next steps should be clear, and these should be distributed among all deep-dive participants, together with the meeting notes.

Optimal Approach:

Content and question design

1. Take the time to **adapt the questions** from chapter 7 of the RAIA guide to the use case at hand as much as possible (see below case study example). This makes the questions tangible and provides for targeted discussions.
   - During the analysis phase, these are independently used by the assessor to conduct technical checks and form initial findings.
   - For the deep dive, these questions are adapted into discussion prompts that reflect findings from the analysis, grounding the conversation in specific evidence from the system.
2. When sharing the initial questions via e-mail, divide them into:
   - Information that is still needed to prepare for the call that the project team can answer quickly
   - Questions for the deep dive that are more open and focused on discussion.
3. Even if a deep dive mainly focuses on a certain stage (e.g. pre-processing), it can still be beneficial to include questions for other stages, (e.g. post-processing) as initial brain teasers.

Table 1: Deep dive questions adapted for Ghana Entrepreneurship Chatbot

Chapter 7 check	Adapted questions for Ghana case study
C2.1.1 Are fairness metrics calculated and reported alongside performance metrics?	Preliminary tests suggested differences in answer quality across languages. How is evaluation and monitoring of performance, and factual consistency done for users interacting in different languages? What steps are taken when a particular language shows significantly lower quality outputs or reliability in testing?
C3.2.1 Do production logs contain personal data, and is it appropriately protected?	The system is publicly accessible without authentication, meaning any user, including small business owners sharing sensitive business details, can interact with it freely. What automated personal information redaction mechanisms are applied to user queries and logs? How are data-retention and user-consent policies defined and communicated to users who may not realise their inputs are being stored?
C3.3.1 Is production performance stable over time or degrading?	The assessment found that outdated or incomplete indexing could lead to users receiving incorrect policy details, wrong eligibility criteria, outdated deadlines, or unsupported references. How often is the policy corpus updated and re-indexed, and how are superseding or conflicting policy documents handled? What fallback messages are shown to users when the system cannot answer reliably or retrieves outdated information?
C3.4.6 Is operational documentation kept up to date?	The assessment found that governance documentation for the system, model cards, defined ownership, update procedures, and incident-management protocols was not provided or made available in an optimal manner. With a publicly deployed system that provides policy guidance to entrepreneurs, what is the status of system documentation? Who is formally responsible for the system's accuracy and maintenance? What process exists for flagging and correcting errors?

Time management and logistics

- A two-week gap between the scoping call and the deep dive has proved to be useful. Since more people are ideally involved in the Deep Dive, schedule the meeting in advance.
- To keep the session manageable, approximately ten to twelve questions for open discussions are recommended.

Expert and end-user involvement

- Experts serve as an additional sounding board that provides otherwise potentially underrepresented perspectives. Their critical input helps to highlight local contextualisation, inclusion, domain expertise or potential effects on end users or beneficiaries. Accordingly, the participation of end users is highly recommended.

5.6. Phase 5: Risk mitigation and post-production/continuous monitoring

This final stage grounds product teams and evaluators for the crucial operational step of how to monitor ongoing compliance and risk mitigation efficacy.

This phase should result in a final assessment report covering:

1. Risks identified and assessment results for each risk
2. Specific mitigation measures tailored for each risk
3. Indicators to assess mitigation effectiveness over time
4. Additional notes: end-user feedback, long-term risks, assessment limitations, etc
5. Follow-up plan: who, what and how often will risks be monitored
6. AI leaflet for external sharing (if required)

Optimal Approach:

To make recommendations more specific and accountable, the following approach is suggested:

1.
   - **What:** Concrete description of necessary mitigation steps
   - **Who:** The person responsible for each step
   - **When:** Due date for each step, based on priority level.
2. Order/prioritise the proposed mitigation measures so that issues requiring more timely intervention are highlighted or mentioned first.
3. Ensure to tailor the recommendations to the use case at hand as much as possible. This makes recommendations actionable for the project teams and facilitates follow-up discussions on concrete mitigation measures outlined in the report.
4. Add process documents, including presentations used, pre-survey results, meetings minutes and other resources as annexes, to ensure robust documentation.

CHAPTER 06

ASSESSMENT FRAMEWORK AND SCORING SYSTEM

This chapter sets out the framework for how risk assessments are conducted and scored across the three lifecycle stages. Assessors should read this chapter before beginning the stage-specific risk assessments in Chapter 7. For each stage of the AI lifecycle (pre-, in-, and post-processing) this chapter unpacks how risks can be identified, with guiding questions to structure a deep dive and suggest example checks or metrics that can be run on ADM and LLM-based systems. This can be done in order, or with focus on specific sections that are relevant to the case or technology. While no assessment should be complete without covering all three stages, systems in very early stages should focus on the pre-processing methodology.

It should be noted that while risk categories overlap, ADM and LLM systems manifest risks differently:

1. ADM systems

ADM systems are AI systems that classify, rank, predict, or score people or cases, directly affecting outcomes such as eligibility or prioritisation. Their main risks concern data representativeness, fairness and non-discrimination, and uneven performance across groups.

2. Generative AI systems

Generative AI systems, especially **LLMs**, create new content such as text, code, or images. They share many of the same risks as ADMs but also introduce issues such as fabricated or ungrounded outputs, prompt manipulation, leakage of personal data, and rapid amplification of misleading content.

Ideally, **an assessment should cover all three stages of the AI lifecycle.** Based on the assessed use cases, while compliance and risk mitigation in pre-processing and in-processing stages are critical, it is not a proxy or a good predictor of actual risk realisation and impact in post-processing.

Assessing throughout the AI lifecycle allows evaluators to have a good understanding of risk sources and mitigation effectiveness. However, **evaluation and harm identification or quantification matter most in protecting users and fundamental rights in the post-processing phase.** Before a system goes live, pre-processing and in-processing assessments serve as stepping stones for the best decisions and to foresee and mitigate harms but should not substitute for post-processing assessment and rigorous post-market monitoring. In some cases, assessments were conducted on AI systems where access was limited to pre-processing and post-processing data only, as in-processing information was controlled by a third-party model provider. Taking this lesson into account, these stages have accordingly been designed as severable, to capture the real-life complexity of existing AI systems; however, this should in no way invite partial assessments.

6.1. the assessment lifecycle

A RAIA assessment is structured around three stages of the AI lifecycle: **pre-processing, in-processing, and post-processing.** Each stage presents distinct risk surfaces, requires different types of evidence, and calls for different analytical approaches.

The depth and focus of each stage will depend on where the system is in its lifecycle.

If the system is...	Primary focus	Secondary focus
In early development (data gathered, model not yet built)	Pre-processing	—
Under active development or fine-tuning	Pre-processing + In-processing	—
Deployed or about to deploy	All three stages	Post-processing for live risk identification
Already in production with issues identified	Post-processing	Pre-processing/In-processing to trace root causes

Important:

Pre-processing and in-processing assessments are valuable for identifying risk sources and guiding development decisions, but they are not substitutes for post-processing assessment. **Real-world impact can only be fully evaluated once a system is operational.**

Pre-processing focuses on how the problem is framed and how data is collected to train, fine-tune, or inform the AI system. Decisions made at this stage (what data to include, how to label it, and whose perspectives are represented) shape what the system can and cannot do and determine which groups may be advantaged or disadvantaged by its outputs.

6.2.Guidance on thresholds and benchmarks

The thresholds used in this guide (i.e. representation ratios below 0.8 or demographic groups representing less than 5% of a dataset) are derived from practitioner experience across the use cases assessed during the development of this guide. These are indicative benchmarks, not universal standards, and should be adapted to the specific system, deployment context, and population being served.

For instance, a 5% threshold may be appropriate for a general-purpose system but inadequate for one deployed in a community where that specific group constitutes a much larger share of the affected population. Similarly, higher-risk systems, such as those making consequential decisions about individuals' access to services, credit, employment, or healthcare, should apply more conservative thresholds and subject borderline cases to additional scrutiny. Where a threshold is applied, assessors should document the rationale for its use or any contextual adjustments made.

6.3. Assessment scorecard

Each stage of the assessment concludes with a scorecard that summarises findings across the five core risk categories. The scorecard uses a 0–5 severity scale. This scale is ordinal (i.e. the distance between 1 and 2 is not the same as the distance between 4 and 5). Scores should be assigned based on the qualitative descriptions below. Scores are assigned to each check and aggregated by risk category, giving assessors and stakeholders a standardised, at-a-glance view of the system's risk profile at that stage of its lifecycle. It is critical to provide comments detailing the rationale behind the assigned rating, including observations, evidence reviewed, and concerns identified (p. 32).

Severity guidance:

Score	Typical findings
0	No issues identified. The risk area is fully addressed. Controls are in place, documented, and functioning.
1-2	Minor concerns Issues identified but limited in scope, well-mitigated, or have a documented remediation plan. Low likelihood or low impact.
3-4	Moderate issues Significant concerns present and requiring attention. Controls are incomplete, inconsistent, or untested. Potential for material impact if unaddressed.
5	Critical gaps or risks. Controls are absent, ineffective, or have failed. Immediate action required.

Of significance to note:

An assessment scorecard is meant to be a living document, just like the RAIA guide itself. Accordingly, it is not a one-off assessment tool or a single review. Rather, the scorecard is envisioned as an assessment overview at a specific point in time. Its results may well change during the AI system's operation and usage. Consequently, scorecards must be updated alongside continuous risk monitoring.

CHAPTER 07

RISK ASSESSMENT ACROSS STAGES

For each lifecycle stage, this chapter provides:

1. Data requests specifying what evidence to collect
2. Assessment questions and checks combining discussion prompts with verifiable indicators. Questions are intended to be a guide to verify signals of risks that were consistently observable, practically verifiable, and predictive of real-world harm across that body of work. These are not exhaustive
3. A stage-specific scorecard summarising findings in a standardised format

7.1. Pre-processing: Data collection and preparation

At this stage, the assessment aims to:

- Identify risks embedded in data collection choices before they are amplified by model training.
- Document data provenance, quality, and limitations.
- Flag gaps in representation, consent, or governance that should be addressed before proceeding.
- Establish a baseline understanding of the data that will inform later stages.

Before conducting the pre-processing risk assessment, request the following from the project team.

Category	Specific items to request
Dataset documentation	Datasheets, data dictionaries, metadata records, README files
Composition data	Breakdown of dataset by relevant demographic or contextual attributes (e.g., age, gender, region, language)
Source information	List of data sources, collection methods, time periods covered
Labelling documentation	Labelling guidelines, annotator instructions, inter-rater agreement scores (if available)
Legal basis/consent records	Privacy notices, consent mechanisms, legal basis documentation, data sharing agreements
Access and security	Data access levels, where it is stored, what controls are in place
Previous assessments	Any prior data quality reviews, bias audits, or compliance checks

Note on data access and verification:

The depth of verification possible depends on the level of access the assessment team has. With direct data access, many checks can be verified through statistical analysis or inspection. With documentation only, verification relies on reviewing evidence that processes exist and were followed. Both approaches are valid; what matters is the clarity regarding what was assessed directly versus what was assessed based on documentation. Where access constraints limit verification, this should be documented as an assessment limitation and noted in the relevant scorecard comments. If the team has direct access to data, additional checks may be possible (e.g., verifying dataset composition directly, inspecting preprocessing pipelines, validating labelling quality through sampling).

Before proceeding, document the access level granted:

Item	Access	Notes (e.g., sample only, documentation only, specific restrictions)
Dataset	☐ No ☐ Yes ☐ Partial	
Preprocessing code/pipelines	☐ No ☐ Yes ☐ Partial	
Labelling data and guidelines	☐ No ☐ Yes ☐ Partial	
Direct contact with data team	☐ No ☐ Yes ☐ Partial	

This record informs how checks can be verified and referenced when documenting assessment limitations. f items are unavailable, document this as a limitation and note the impact on assessment completeness.

7.1.1. Bias and fairness

Data to request:

1. Dataset composition breakdown by protected/ sensitive attributes (where available and appropriate)
2. Target population demographics (who is the system intended to serve, and who might be affected?)
3. Documentation of any known gaps, exclusions, or historical biases in the data
4. Labelling guidelines and any fairness-related instructions given to annotators

Questions and checks:

Check	What to look for	How to verify
C1.1.1 arget variable appropriateness	(Criterion Validity) Is the outcome being predicted the right thing to predict? (Construct Validity) Could the definition itself embed bias?	Review target variable definition; assess whether it captures the intended outcome or a proxy that may disadvantage certain groups (e.g., "success" defined as "promotion within 2 years" may reflect existing promotion biases).
C1.1.2 Selection bias in ground truth	Does the training data only contain [certain] outcomes for a pre-selected population? How was the data sampled?	Identify whether ground truth data excludes populations that were filtered out by prior decisions (e.g., loan repayment data only exists for approved applicants; job performance data only exists for hired candidates). Document what is unknowable due to selection. Calculate ratio of observed outcomes / total affected population
C1.1.3 Population alignment	Does the dataset composition reflect the population the system will serve?	Compare dataset demographics to target population demographics. Calculate representation ratios. Flag if ratio < 0.8
C1.1.4 Under-representation	Are any groups significantly underrepresented or missing entirely? How are relevant groups defined?	Identify groups with < 5% representation or groups present in target population but absent from data. Adjust thresholds based on system risk level. Define groups (e.g., intersectional, age thresholds, etc.)
C1.1.5 Historical bias in sources	Do data sources reflect historical inequities that could be learned by the model? Are there proxies for bias (e.g., postcode)?	Review source descriptions; flag sources known to contain historical discrimination (e.g., past hiring decisions, loan outcomes).
C1.1.6 Labelling consistency	Are labels applied consistently across groups, or could labelling practices introduce bias?	Review labelling guidelines for potential differential treatment; check inter-rater agreement by group (if available. Label flip rate by annotators (if available)
C1.1.7 Fairness criteria defined	Has the team defined what "fair" means for this system?	Ask: What fairness outcome is the system aiming for? For instance, Group vs Individual? Is this documented?
C1.1.8 Exclusions documented	Are deliberate exclusions from the dataset documented with rationale?	Review any exclusion criteria; assess whether exclusions disproportionately affect certain groups. What are the group-wise exclusion rates? Are they equal?

7.1.2. Privacy and confidentiality

Data to request:

1. Data inventory listing personal data fields collected
2. Privacy policy and consent notice
3. Legal basis documentation (e.g., consent, legitimate interest, public interest)
4. Data Privacy Risk or Impact Assessment documentation
5. Deployment location/jurisdiction (impacts applicable privacy regulations)
6. Anonymisation/pseudonymisation procedures and documentation
7. Data access controls and security measures documentation
8. Data retention and deletion policies

Questions and checks:

Check	What to look for	How to verify
C1.2.1 Personal data inventory	Is there a clear record of what personal data is collected?	Review data dictionary; confirm personal data fields are identified and classified
C1.2.2 Sensitive data handling	Is sensitive data (health, ethnicity, political views, etc.) collected? If so, with appropriate safeguards?	Identify sensitive data fields according to applicable data privacy laws; verify enhanced protections are in place
C1.2.3 Data Privacy Risk or Impact Assessment	Do the collected personal data or how they shall be used require a Data Privacy Risk or Impact Assessment?	Review legal requirements of applicable data privacy laws for the necessity of an in-depth risk or impact assessment.
C1.2.4 Legal basis	Is there a documented legal basis for processing? Are privacy notices clear, accurate, and accessible to data subjects?	Review legal basis documentation; confirm it is appropriate for the data types and purposes. Review privacy notices for accuracy, intelligibility, transparency and accessibility.
C1.2.5 Consent quality	If consent is the legal basis, is it informed, specific, and freely given?	Review consent mechanisms and language; assess whether consent covers actual uses
C1.2.6 Anonymisation adequacy	If data is anonymised, is the method adequate to prevent re-identification?	Review anonymisation procedures; assess re-identification risk (consider combination attacks, small groups)
C1.2.7 Access controls	Who can access the data, and are access rights appropriate? Does there exist an auditable trail showing who accessed the data and when?	Review access control documentation; verify principle of least privilege is applied
C1.2.8 Retention limits	Are retention periods defined and enforced?	Review retention policy; confirm deletion procedures exist

7.1.3. Reliability

Data to request:

1. Data quality reports or assessments
2. Labelling process documentation including quality control measures
3. Information on training/deployment alignment (time-period, geography, language, user groups)
4. Sample of data records for spot-checking (if feasible)

Questions and checks:

Check	What to look for	How to verify
C1.3.1 Label accuracy	Are labels accurate, and do they reflect ground truth?	Review labelling process; check for validation against ground truth (gold standard/experts) where available
C1.3.2 Label consistency	Are labels applied consistently across annotators and over time?	Review inter-rater agreement metrics (e.g., Cohen's Kappa scores >0.8, Flip rate etc.), check for labelling drift over time.
C1.3.3 Data quality controls	Are there systematic checks for data quality (completeness, accuracy, consistency, duplicates)? Are there any extreme outliers?	Review data quality procedures; check for automated or manual validation steps. For numeric fields, check the percentage of records with Z-score >3 or outside 1.5× Interquartile Range (target <5%). For categorical fields, check the percentage of rare categories with frequency <1%.
C1.3.4 Training-deployment alignment	Does training data reflect the conditions where the system will be deployed?	Compare training data characteristics (time, geography, language, user types) to expected deployment context. For example, compare training and deployment data distributions using statistical distance metrics (e.g., KL divergence, Wasserstein distance) to assess alignment (target < 0.1).
C1.3.5 Known limitations documented	Are data limitations (gaps, biases, quality issues) documented?	Review dataset documentation for limitations section
C1.3.6 Data provenance	Is the origin and history of the data traceable?	Review provenance documentation; assess whether data lineage can be reconstructed

7.1.4. Governance

Data to request:

1. Data governance policies and procedures documentation
2. Records of key decisions about data collection, selection, and pre-processing
3. Roles and responsibilities documentation
4. Legal and regulatory compliance assessments
5. Audit trail or version history for datasets

Questions and checks:

Check	What to look for	How to verify
C1.4.1 Decision documentation	Are key data decisions (sources, selection criteria, preprocessing) documented with rationale?	Review documentation; assess whether an external reviewer could understand why decisions were made
C1.4.2 Roles and accountability	Is there a named owner accountable for data quality and compliance?	Identify accountable person/team; verify this is documented, not just assumed
C1.4.3 Change management	Is there a process for approving and documenting changes to datasets?	Review change management procedures; check for version control
C1.4.4 Auditability	Could an auditor reconstruct the history of the dataset and key decisions?	Assess completeness of audit trail; identify gaps
C1.4.5 Regulatory alignment	Are relevant legal requirements (data protection, sector regulations) identified and addressed?	Review compliance measures; verify relevant requirements are identified (e.g. human oversight mechanisms and explicit consent in case the system falls under Art. 22 GDPR – Automated decision-making)
C1.4.6 Policy alignment	Are organisational AI/data policies followed?	Compare practices to stated policies; identify deviations

7.1.5. Security and misuse

Data to request:

1. Data security documentation (access controls, encryption, network security)
2. Data pipeline architecture documentation
3. Incident response procedures
4. Risk assessment for data poisoning or tampering
5. Documentation of third-party data sources and their security

Questions and checks:

Check	What to look for	How to verify
C1.5.1 Access control	Are data access rights restricted and reviewed?	Review access control mechanisms; check for principle of least privilege
C1.5.2 Pipeline security	Are data pipelines protected against unauthorised modification?	Review pipeline architecture; identify potential injection points
C1.5.3 Third-party risk	Are third-party data sources vetted for security and integrity?	Review third-party assessment procedures; check for ongoing monitoring
C1.5.4 Data poisoning risk	Could malicious actors inject false or misleading data?	Identify data input channels; assess validation controls
C1.5.5 Encryption and storage	Is data encrypted at rest and in transit where appropriate?	Review encryption practices; verify implementation
C1.5.6 Incident response	Are there procedures to detect and respond to data security incidents?	Review incident response documentation; assess whether data-specific scenarios are covered

7.1.6. Pre-processing scorecard

The scorecard below summarises the findings from the pre-processing risk assessment. For each check, record the severity score (0-5) and brief rationale/justification for assigning that score. This could be based on the potential likelihood of harm and downstream impact. For guidance on how to use the scorecard, refer to Chapter 6.3.

Risk Category	Check	Severity (0-5)	Rationale	Key Findings
Bias and Fairness	**C1.1.1** Target variable appropriateness			
	C1.1.2 Selection bias in ground truth			
	C1.1.3 Population alignment			
	C1.1.4 Underrepresentation			
	C1.1.5 Historical bias in sources			
	C1.1.6 Labelling consistency			
	C1.1.7 Fairness criteria defined			
	C1.1.8 Exclusions documented			
Privacy and Confidentiality	**C1.2.1** Personal data inventory			
	C1.2.2 Sensitive data handling			
	C1.2.3 Data Privacy Risk or Impact Assessment			
	C1.2.4 Legal basis			
	C1.2.5 Consent quality			
	C1.2.6 Anonymisation adequacy			
	C1.2.7 Access controls			
	C1.2.8 Retention limits			

Risk Category	Check	Severity (0-5)	Rationale	Key Findings
Reliability	C1.3.1 Label accuracy			
	C1.3.2 Label consistency			
	C1.3.3 Data quality controls			
	C1.3.4 Training-deployment alignment			
	C1.3.5 Known limitations documented			
	C1.3.6 Data provenance			
Governance	C1.4.1 Decision documentation			
	C1.4.2 Roles and accountability			
	C1.4.3 Change management			
	C1.4.4 Auditability			
	C1.4.5 Regulatory alignment			
	C1.4.6 Policy alignment			
Security and Misuse	C1.5.1 Access control			
	C1.5.2 Pipeline security			
	C1.5.3 Third-party risk			
	C1.5.4 Data poisoning risk			
	C1.5.5 Encryption and storage			
	C1.5.6 Incident response			

Summary Assessment:

Risk Category	Highest Severity	Median Severity	Checks ≥3	Total Checks	Overall Assessment
Bias and Fairness	_/5	_/5	_/7	7	
Privacy and Confidentiality	_/5	_/5	_/7	7	
Reliability	_/5	_/5	_/8	8	
Governance	_/5	_/5	_/6	6	
Security and Misuse	_/5	_/5	_/8	8	

Note on ADM vs. LLM checks:

Some checks are marked as (ADM) or (LLM) to indicate they apply primarily to that system type. Mark as "N/A" if the check does not apply to the system being assessed.

How to interpret the summary:

- **Highest Severity:** The maximum score among all applicable checks in that category.
- **Median Severity:** The median score among all applicable checks in the category.
- **Checks ≥3:** Count of applicable checks scoring 3 or higher.
- **Overall Assessment:** Qualitative summary based on the pattern of findings.

7.2. In-processing: Model training and evaluation

The in-processing stage covers how the model is selected, trained, tuned, and evaluated. Decisions made here such as choice of model architecture, optimization objectives, hyperparameters, evaluation benchmarks, determine how the system behaves in practice and which risks are embedded, amplified, or mitigated.

At this stage, the assessment aims to:

1. Identify risks introduced or amplified during model training and tuning
2. Evaluate whether training processes are documented, reproducible, and auditable
3. Assess model performance and fairness across relevant groups
4. For LLMs: Evaluate safety behaviours, confabulation rates, and robustness to manipulation
5. For ADMs: Evaluate calibration, threshold choices, and disparity across groups

This section covers both Automated Decision-Making (ADM) systems and Large Language Model (LLM) systems. Where checks differ between these system types, the differences are noted explicitly:

Aspect	ADM Systems	LLM Systems
Primary outputs	Scores, classifications, decisions	Text, recommendations, conversations
Key reliability concerns	Calibration, threshold stability, overfitting	Confabulation, hallucination, prompt sensitivity
Key fairness concerns	Disparity in error rates, threshold effects	Stereotyping, differential quality by group
Key security concerns	Adversarial inputs, data poisoning	Prompt injection, jailbreaking, misuse

Before conducting the in-processing risk assessment, request the following from the project team and document availability. This is meant as a minimum subset and can evolve with changing technological landscape.

Category	Specific items to request
Model documentation	Model cards, architecture descriptions, choice rationale
Training configuration	Hyperparameters, optimization settings, loss functions, data splits, pipelines
Evaluation results	Performance metrics on test/validation sets, by group where available
Fairness analysis	Disparity metrics, fairness evaluations, bias testing results
For LLMs: Safety testing	Red-teaming results, safety benchmark scores, refusal rate data
For LLMs: Reliability testing	Confabulation/hallucination rates, factuality assessments
For RAG systems	Retrieval corpus documentation and retrieval accuracy metrics
Training logs	Experiment tracking, version history, checkpoint records
Access documentation	Who has access to training infrastructure, models, and data

If items are unavailable, document this as a limitation. For systems using third-party or proprietary models where training details are not accessible, document what information the vendor provides and note assessment limitations.

Note on data access and verification:

As with pre-processing, the depth of verification depends on access level. For in-house models, direct access to training code, evaluation scripts, and model artifacts allows thorough verification. For third-party or API-based models, assessment may be limited to reviewing vendor documentation and conducting black-box testing. Document what level of access was granted and note limitations in the scorecard.

Before proceeding, document the access level granted:

Item	Access	Notes (e.g., summary metrics only, vendor documentation only, specific restrictions)
Model training code/ notebooks	☐ No ☐ Yes ☐ Partial	
Evaluation results	☐ No ☐ Yes ☐ Partial	
Model artifacts (weights, checkpoints)	☐ No ☐ Yes ☐ Partial	
Experiment tracking logs	☐ No ☐ Yes ☐ Partial	
For LLMs: Red teaming/ safety test data	☐ No ☐ Yes ☐ Partial	
For RAG: Retrieval corpus	☐ No ☐ Yes ☐ Partial	
Direct contact with ML/ data science team	☐ No ☐ Yes	

This record informs how checks can be verified and should be referenced when documenting assessment limitations.

7.2.1. Bias & fairness

Common bias types in in-processing:

The table below explains common types of bias that can arise during model training and evaluation. Understanding these categories helps assessors identify which risks are most relevant for the system under review.

Bias Type	Definition
Algorithmic bias	Arises when the model architecture or optimisation objectives systematically favour certain outcomes or groups, or when objective functions ignore fairness or interpretability constraints (e.g., optimising only overall accuracy so the model sacrifices minority recall).
Aggregation bias	Occurs when a single model is forced to fit a diverse population where different groups have different underlying relationships. (The "one-size-fits-all" error).
Threshold bias	Occurs when a single decision threshold disproportionately benefits, prejudices, or harms specific groups.
Representation bias	Manifests when the model's learned representations (e.g., embeddings or latent vectors) encode social or cultural stereotypes, or reflect imbalanced training data.
Outcome disparity	Differences in predicted outcomes across groups not justified by legitimate factors, persisting even after controlling for legitimate predictive factors.

Bias Type	Definition
Evaluation bias	Arises when benchmark or validation datasets do not represent real-world users or contexts.
Measurement bias	Occurs when input features measure constructs differently across groups (e.g., sensor accuracy or labelling inconsistencies that vary by group).

Data to request:

1. Performance metrics disaggregated by protected/sensitive attributes
2. Fairness metrics (e.g., demographic parity, equalised odds, calibration by group etc.)
3. Threshold selection rationale and impact analysis
4. Evaluation dataset composition and representativeness assessment
5. **For LLMs:** stereotype and bias testing results

Questions and checks:

Check	What to look for	How to verify
C2.1.1 Fairness metrics computed	Are fairness metrics calculated and reported alongside performance metrics? Are metrics disaggregated by intersections of sensitive attributes?	Review evaluation reports; check for group-level breakdowns
C2.1.2 Performance disparity	Are error rates (FPR, FNR) consistent across different demographic groups? Compare group-level metrics	Compare error metrics by group; flag significant disparities (e.g., Equalised Odds or Predictive Rate Parity). Target score depends on individual metrics
C2.1.3 Threshold fairness (ADM)	Does the decision threshold produce equitable outcomes across groups?	Review threshold selection process; assess impact on different groups
C2.1.4 Embedding bias	Do model embeddings or representations encode stereotypes?	Review embedding analysis or bias probing results if available (e.g., WEAT/SEAT) Or feature importance e.g., SHAP values
C2.1.5 Stereotype testing (LLM)	Has the model been tested for stereotyping behaviour across demographic groups?	Review stereotype testing results; check for systematic patterns
C2.1.6 Evaluation set representativeness	Does the evaluation dataset adequately represent all relevant groups?	Compare evaluation set composition to target population; identify gaps
C2.1.7 Fairness-performance trade-offs documented	Are trade-offs between overall performance and group fairness documented and justified?	Review documentation of trade-off decisions; assess justification

7.2.2. Privacy and confidentiality

Data to request:

1. Documentation of privacy-preserving measures in training (e.g., differential privacy, data sanitisation)
2. Access control documentation for training data, checkpoints, and models
3. For LLMs: Personal data leakage testing results, memorisation assessments
4. For RAG systems: retrieval corpus access controls and filtering mechanisms
5. Prompt/query logging and retention policies (if applicable during evaluation)

Questions and checks:

Check	What to look for	How to verify
C2.2.1 Data anonymisation	Has personal data, where possible, been removed or adequately anonymised to avoid re-identification?	Review anonymisation techniques.
C2.2.2 Data minimisation	Is only the absolute necessary personal data used for model training and evaluation, and is it appropriately pseudonymised?	Review categories of personal data used for model training and pseudonymisation techniques.
C2.2.3 Purpose alignment	Can the collected personal data be used for model training?	Review purposes of personal data processing as stated in privacy notice against intended use for model development.
C2.2.4 Privacy preservation in training	Are privacy-preserving techniques used where appropriate?	Review training documentation; check for differential privacy, data sanitisation, or other techniques
C2.2.5 Access controls	Is access to training data, intermediate artifacts and (where applicable) the RAG corpus appropriately restricted?	Review access control mechanisms; verify least-privilege principle
C2.2.6 Personal data leakage testing (LLM)	Has the model been tested for personal data leakage in outputs (e.g. due to memorisation or generalisation)?	Review personal data detection testing; check for systematic leakage
C2.2.7 RAG corpus confidentiality	For RAG systems: Are retrieval corpus access controls appropriate?	Review RAG architecture; check filtering and access rules
C2.2.8 Evaluation data handling	Is personal data in evaluation sets handled appropriately?	Review evaluation data management; check for proper safeguards

7.2.3. Reliability

Data to request:

1. Training vs. validation vs. test performance comparison (overfitting assessment)
2. Calibration metrics and analysis
3. Robustness testing results (adversarial inputs, distribution shift)
4. For LLMs: confabulation/hallucination rate assessments, factuality testing
5. For LLMs: prompt sensitivity analysis
6. Reproducibility documentation (seeds, checkpoints, versioning)

Questions and checks:

Check	What to look for	How to verify
C2.3.1 Overfitting assessment	Is there evidence of overfitting (high training performance, lower test performance)?	Compare training, validation, and test metrics; check for significant gaps
C2.3.2 Calibration (ADM)	Are predicted probabilities well-calibrated (matching actual frequencies)?	Review calibration plots/metrics; check calibration by group
C2.3.3 Output consistency	Does the model produce consistent outputs for similar inputs?	Review consistency testing; check for unexplained variability
C2.3.4 Robustness to perturbations	How does the model perform under input perturbations or distribution shift?	Review robustness testing results; identify vulnerability patterns
C2.3.5 Confabulation rate (LLM)	How often does the model generate false or fabricated information?	Review confabulation/hallucination testing; check rates by topic/domain
C2.3.6 Factuality and grounding (LLM)	Are outputs grounded in reliable sources or verifiable information?	Review factuality assessments; check citation accuracy for RAG systems
C2.3.7 Prompt sensitivity (LLM)	Is the model overly sensitive to prompt phrasing variations?	Review prompt variation testing; check for inconsistent behaviour
C2.3.8 Reproducibility	Can training results be reproduced from documentation?	Review reproducibility documentation; assess completeness

7.2.4. Governance

Data to request:

1. Model development documentation (design decisions, experiments, rationale)
2. Roles and responsibilities for model development and approval
3. Change management procedures for model updates
4. Internal review or approval records
5. Compliance assessments (legal, ethical, organisational policy)

Questions and checks:

Check	What to look for	How to verify
C2.4.1 Model documentation completeness	Are model design decisions, training configuration, and evaluation results documented?	Review model cards, experiment logs, and reports; assess completeness
C2.4.2 Development accountability	Is there clear ownership for model quality, safety, and compliance?	Identify accountable person/ team; verify documentation
C2.4.3 Experiment tracking	Are experiments tracked with sufficient detail for reproducibility and audit?	Review experiment tracking system; assess traceability
C2.4.4 Change management	Is there a process for approving model changes and new versions?	Review change management procedures; check for approval records
C2.4.5 Internal review	Has the model undergone internal review before deployment?	Review internal review records; check scope of review
C2.4.6 Third-party model governance	For models using third-party components: Is vendor governance assessed?	Review vendor assessment; check for ongoing monitoring

7.2.5. Security and misuse

Data to request:

1. Security assessment of training infrastructure
2. Adversarial testing results (for both ADM and LLM systems)
3. For LLMs: prompt injection and jailbreak testing results
4. Model access controls and authentication mechanisms
5. Misuse risk assessment and acceptable use documentation
6. Incident response procedures for security issues

Questions and checks:

Check	What to look for	How to verify
C2.5.1 Training infrastructure security	Is training infrastructure protected against unauthorised access?	Review infrastructure security documentation; check access controls
C2.5.2 Data poisoning resilience	Has the model been assessed for vulnerability to training data poisoning?	Review poisoning risk assessment; check for detection mechanisms
C2.5.3 Adversarial robustness (ADM)	Has the model been tested against adversarial inputs designed to cause misclassification?	Review adversarial testing results; assess coverage
C2.5.4 Prompt injection testing (LLM)	Has the model been tested for prompt injection vulnerabilities?	Review prompt injection testing; check for systematic assessment
C2.5.5 Jailbreak testing (LLM)	Has the model been tested for jailbreak attempts that bypass safety measures?	Review jailbreak/red-team testing; assess coverage and results
C2.5.6 Model extraction risk	Is the model protected against extraction or theft attempts?	Review model access controls; assess extraction risk
C2.5.7 Misuse scenarios identified	Have potential misuse scenarios been identified and assessed?	Review misuse risk assessment; check for documentation
C2.5.8 Incident response readiness	Are there procedures to respond to security incidents during development?	Review incident response procedures; assess completeness

7.2.6. In-processing scorecard

The scorecard below summarises the findings from the in-processing risk assessment. For each check, record the severity score (0-5) and brief rationale/justification for assigning that score. For instance, this could be based on the potential likelihood of harm and downstream impact. For guidance on how to use the scorecard, refer to Section 6.3.

Risk Category	Check	Severity (0-5)	Rationale	Key Findings
Bias and Fairness	**C2.1.1** Fairness metrics computed			
	C2.1.2 Group performance disparity			
	C2.1.3 Threshold fairness (ADM)			
	C2.1.4 Embedding bias (LLM)			
	C2.1.5 Stereotype testing (LLM)			
	C2.1.6 Evaluation set representativeness			
	C2.1.7 Fairness-performance trade-offs documented			
Privacy and Confidentiality	**C2.2.1** Data anonymisation			
	C2.2.2 Data minimisation			
	C2.2.3 Purpose alignment			
	C2.2.4 Privacy preservation in training			
	C2.2.5 Access controls			
	C2.2.6 Personal data leakage testing (LLM)			
	C2.2.7 RAG corpus confidentiality			
	C2.2.8 Evaluation data handling			

Risk Category	Check	Severity (0-5)	Rationale	Key Findings
Reliability	C2.3.1 Overfitting assessment			
	C2.3.2 Calibration (ADM)			
	C2.3.3 Output consistency			
	C2.3.4 Robustness to perturbations			
	C2.3.5 Confabulation rate (LLM)			
	C2.3.6 Factuality and grounding (LLM)			
	C2.3.7 Prompt sensitivity (LLM)			
	C2.3.8 Reproducibility			
Governance	C2.4.1 Model documentation completeness			
	C2.4.2 Development accountability			
	C2.4.3 Experiment tracking			
	C2.4.4 Change management			
	C2.4.5 Internal review			
	C2.4.6 Third-party model governance			

Risk Category	Check	Severity (0-5)	Rationale	Key Findings
Security and Misuse	C2.5.1 Training infrastructure security			
	C2.5.2 Data poisoning resilience			
	C2.5.3 Adversarial robustness (ADM)			
	C2.5.4 Prompt injection testing (LLM)			
	C2.5.5 Jailbreak testing (LLM)			
	C2.5.6 Model extraction risk			
	C2.5.7 Misuse scenarios identified			
	C2.5.8 Incident response readiness			

Summary Assessment:

Risk Category	Highest Severity	Median Severity	Checks ≥3	Total Checks	Overall Assessment
Bias and Fairness	_/5	_/5	_/7	7	
Privacy and Confidentiality	_/5	_/5	_/7	7	
Reliability	_/5	_/5	_/8	8	
Governance	_/5	_/5	_/6	6	
Security and Misuse	_/5	_/5	_/8	8	

Note on ADM vs. LLM checks:

Some checks are marked as (ADM) or (LLM) to indicate they apply primarily to that system type. Mark as "N/A" if the check does not apply to the system being assessed.

How to interpret the summary:

- **Highest Severity:** The maximum score among all applicable checks in that category.
- **Median Severity:** The median score among all applicable checks in the category.
- **Checks ≥3:** Count of applicable checks scoring 3 or higher.
- **Overall Assessment:** Qualitative summary based on the pattern of findings.

7.3. Post-processing: Deployment and monitoring

The post-processing stage covers what occurs after the AI/ML model is trained and deployed into production. Live systems encounter real users, and evolving conditions that may differ from development assumptions and testing environments. Risks that appeared manageable in testing manifest differently in practice, and new risks emerge from user behaviour, feedback loops, and changing contexts.

At this stage, the assessment aims to:

1. Evaluate how the system performs with real users and real data
2. Assess whether monitoring and oversight mechanisms are adequate
3. Identify risks that have emerged or changed since deployment
4. Evaluate human oversight, accountability, and redress mechanisms

For systems already in production: assess incident history and response effectiveness

Post-processing assessment covers both risk categories (the same five as pre- and in-processing) and operational dimensions that become critical once a system is live:

Operational Dimension	What it covers
Human oversight	How humans review, contest, and improve automated decisions
Responsibility and redress	Accountability for harms; channels for objection and appeal
Transparency and explainability	Disclosure of AI use; understandable explanations of outputs
Monitoring and incident response	Ongoing performance tracking; response to issues
Drift and retraining management	Detecting changes in data, concept, or user behaviour that affect model performance/fairness.

These operational dimensions are integrated into the risk category assessments below, with Transparency and Explainability and Responsibility and Redress treated as additional categories given their heightened importance post-deployment.

Before conducting the post-processing risk assessment, request the following from the project team:

Category	Specific items to request
Production performance data	Live performance metrics, disaggregated by user groups where available
Monitoring dashboards/reports	What is tracked, alert thresholds, escalation procedures
Incident logs and Audit procedures	History of issues, complaints, failures since deployment
Human oversight records	Override rates, review frequency, escalation records
User feedback	Complaints, appeals, satisfaction data

Category	Specific items to request
Drift monitoring	Data drift and model drift assessments
For LLMs: Production safety data	Refusal rates, flagged outputs, jailbreak attempts detected
For LLMs: Confabulation monitoring	Production confabulation/hallucination rates if tracked
Redress documentation	Appeal process, response times, resolution outcomes
Transparency materials	User-facing disclosures, explanations provided

If the system is not yet deployed, this assessment focuses on the planned monitoring, oversight, and response mechanisms, scored based on their design adequacy.

Note on data access and verification:

Post-processing assessment ideally involves access to production data, logs, and monitoring systems. The analysis depth depends on what teams can access — from full production logs (allowing detailed statistical analysis) to aggregated metrics or dashboards only. For systems where production data access is restricted (e.g., due to privacy constraints), document any alternative evidence reviewed.

Before proceeding, document the access level granted:

Item	Access	Notes (e.g., aggregated only, sample period, anonymised, specific restrictions)
Monitoring dashboards	☐ No ☐ Yes ☐ Partial	
User feedback/ complaints data	☐ No ☐ Yes ☐ Partial	
Override/appeal records	☐ No ☐ Yes ☐ Partial	
Incident response records	☐ No ☐ Yes ☐ Partial	
Live system access for testing	☐ No ☐ Yes ☐ Partial	
Direct contact with operations team	☐ No ☐ Yes ☐ Partial	

This record informs how checks can be verified and should be referenced when documenting assessment limitations.

7.3.1. Bias and fairness

Data to request:

1. Production performance metrics by group (error rates, outcome rates, quality scores)
2. User feedback and complaints data, analysed for patterns by group
3. Override and appeal data by group
4. Monitoring dashboards showing fairness metrics over time
5. Any post-deployment fairness audits or assessments

Questions and checks:

Check	What to look for	How to verify
C3.1.1 Production outcome equity	Are outcomes equitable across groups in actual use?	Review production metrics by group; compare error rates, approval rates, quality scores
C3.1.2 Complaint patterns	Do complaints or negative feedback concentrate in particular groups?	Analyse complaint data by user characteristics; identify patterns
C3.1.3 Override patterns	Are human overrides more frequent for some groups than others?	Review override data by group; assess whether patterns indicate systematic issues (e.g., higher percentage of overrides for particular group)
C3.1.4 Benchmark vs. production gap	Does real-world performance match benchmark/evaluation results across groups?	Compare production metrics to pre-deployment evaluations; identify divergences
C3.1.5 Feedback loop monitoring	Are feedback loops monitored for fairness effects?	Review monitoring for reinforcement effects; assess whether model updates preserve fairness
C3.1.6 Automation bias indicators	Is there evidence that users over-rely on AI outputs without appropriate scrutiny?	Review human decision patterns; check for rubber-stamping or over-trust
C3.1.7 Drift / real-world changes	Are changes in data or user behaviour affecting model performance or fairness?	Monitor production metrics over time. Flag if FPR/FNR changes >10%, accuracy drops >5%, or fairness metrics exceed thresholds (e.g., parity ratio <0.8)

7.3.2. Privacy and confidentiality

Data to request:

1. Production logging practices (what is captured, retention periods)
2. Handling of personal data in production (detection, redaction, access controls)
3. Data subject request records (access, deletion requests and response)
4. **For LLMs:** Production leakage monitoring of personal data
5. **For RAG**: Production corpus access and query logging
6. Privacy incident history

Questions and checks:

Check	What to look for	How to verify
C3.2.1 Production logging practices	Do production logs contain personal data, and if so, is it appropriately protected?	Review what is logged; check retention, access controls, anonymisation
C3.2.2 Output exposure of personal data (LLM)	Are outputs monitored for leakage of personal data in production?	Review detection mechanisms for personal data; check incident rates
C3.2.3 User data handling	Are user inputs (prompts, queries, data) handled according to privacy policies?	Compare actual practices to stated policies; verify deletion/retention compliance
C3.2.4 Data subject rights	Can users exercise access, correction, and deletion rights effectively?	Review DSR process; check response times and completion rates
C3.2.5 Third-party data sharing	Is production data shared with third parties, and if so, based on appropriate safeguards and legal bases?	Review data sharing practices; verify legal bases, agreements and safeguards
C3.2.6 Privacy incident response	Have privacy incidents occurred, and were they handled appropriately?	Review incident history; assess response adequacy

7.3.3. Reliability

Data to request:

1. Production performance metrics over time (accuracy, error rates, latency)
2. Drift monitoring data (data drift, model drift, concept drift)
3. Incident history (outages, degradations, harmful outputs)
4. For LLMs: Production confabulation/hallucination monitoring
5. For LLMs: Production safety metric trends (refusal rates, flagged outputs)
6. Rollback and incident response records

Questions and checks:

Check	What to look for	How to verify
C3.3.1 Performance stability	Is production performance stable over time, or degrading?	Review performance trends; identify degradation patterns
C3.3.2 Drift detection	Is drift monitored and detected before causing significant impact?	Review drift monitoring; check detection timeliness
C3.3.3 Production confabulation (LLM)	What is the production confabulation/ hallucination rate, and is it monitored?	Review production confabulation metrics; compare to benchmarks
C3.3.4 Harmful output incidents	Have harmful, misleading, or inappropriate outputs occurred in production?	Review incident logs; assess frequency and severity
C3.3.5 Consistency in production	Are outputs consistent for similar inputs in the live system?	Review consistency monitoring; check for unexplained variability
C3.3.6 Rollback capability	Can the system be quickly rolled back if issues are detected?	Review rollback procedures; check if tested and ready
C3.3.7 Incident response effectiveness	When incidents occur, are they detected and resolved promptly?	Review incident history; assess detection time and resolution

7.3.4. Governance

Data to request:

1. Operational governance documentation (roles, responsibilities, decision rights)
2. Change management records for production updates
3. Compliance monitoring records
4. Human oversight procedures and records
5. Audit trail for significant decisions and changes
6. Regulatory reporting and compliance documentation

Questions and checks:

Check	What to look for	How to verify
C3.4.1 Operational accountability	Is there clear ownership for the deployed system's performance, safety, and compliance?	Identify accountable person/team; verify this is documented and active
C3.4.2 Production change control	Are changes to the production system controlled and documented?	Review change management records; check approval workflows
C3.4.3 Human oversight implementation	Are human oversight mechanisms implemented and functioning?	Review oversight records; check review frequency and coverage
C3.4.4 Audit trail completeness	Is there a complete audit trail for system behaviour and decisions?	Review audit logging; assess completeness and tamper-resistance
C3.4.5 Regulatory compliance maintenance	Is ongoing regulatory compliance monitored and maintained?	Review compliance monitoring; check for gaps or violations
C3.4.6 Documentation currency	Is operational documentation kept up to date?	Compare documentation to actual system; identify discrepancies

7.3.5. Security and misuse

Data to request:

1. Security monitoring and alerting configuration
2. Security incident history and response records
3. **For LLMs:** Prompt injection and jailbreak attempt monitoring
4. **For LLMs**: Misuse detection and response records
5. Access control and authentication records
6. Vulnerability assessment and penetration testing results

Questions and checks:

Check	What to look for	How to verify
C3.5.1 Production security monitoring	Is the production system monitored for security threats?	Review security monitoring configuration; check alert coverage
C3.5.2 Adversarial input detection	Are adversarial inputs (attacks, manipulation attempts) detected in production?	Review detection mechanisms; check incident rates
C3.5.3 Prompt injection monitoring (LLM)	Are prompt injection attempts detected and logged?	Review prompt injection monitoring; assess detection effectiveness
C3.5.4 ailbreak attempt tracking (LLM)	Are jailbreak and safety bypass attempts tracked?	Review jailbreak monitoring; check response procedures
C3.5.5 Misuse detection	Is the system monitored for misuse beyond intended purposes?	Review misuse detection mechanisms; check for patterns
C3.5.6 Security incident response	Have security incidents occurred, and were they handled appropriately?	Review incident history; assess response effectiveness
C3.5.7 Access control effectiveness	Are production access controls effective and regularly reviewed?	Review access control logs; check for unauthorised access attempts

7.3.6. Transparency and explainability

Data to request:

1. User-facing disclosures about AI use
2. Explanation mechanisms and their documentation
3. Transparency reports or public disclosures
4. User comprehension or feedback on explanations
5. Documentation available to oversight bodies or auditors

Questions and checks:

Check	What to look for	How to verify
C3.6.1 AI use disclosure	Are users clearly informed that AI is being used?	Review user-facing disclosures and/or consent statements; assess their clarity and prominence if the system performs automated decision-making (see Art. 22 GDPR)
C3.6.2 Purpose and limitations communication	Are the system's purpose and limitations communicated to users?	Review disclosures; check for limitation statements
C3.6.3 Explanation provision	Are explanations for outputs provided where appropriate?	Review explanation mechanisms; assess availability and quality
C3.6.4 Explanation comprehensibility	Are explanations understandable to the intended audience?	Review user feedback; assess language and format appropriateness
C3.6.5 Auditor/ regulator access	Can oversight bodies access necessary information for review?	Review what is available to auditors; identify access gaps

7.3.7. Responsibility and redress

Data to request:

1. Appeal and complaint procedures
2. Appeal/complaint volume and resolution data
3. Remediation procedures and records
4. Accountability documentation (who is responsible for what)
5. Communication procedures for affected individuals

Questions and checks:

Check	What to look for	How to verify
C3.7.1 Appeal mechanism availability	Can affected individuals appeal or contest AI-influenced/ automated decisions?	Review appeal procedures; assess accessibility
C3.7.2 Appeal process effectiveness	Are appeals processed fairly and in reasonable time?	Review appeal data; check resolution times and outcomes
C3.7.3 Harm remediation	Are there procedures to remediate harms when they occur?	Review remediation procedures; check for implementation
C3.7.4 Accountability clarity	Is it clear who is accountable for different types of issues?	Review accountability documentation; assess clarity
C3.7.5 Communication with affected parties	Are affected individuals informed about issues and their resolution?	Review communication procedures; check implementation

7.3.8. Post-processing scorecard

The scorecard summarises the findings from the post-processing risk assessment. For each check, record the severity score (0-5) and brief rationale/justification for assigning that score. For instance, this could be based on the potential likelihood of harm and downstream impact.

For guidance on how to use the scorecard, refer to Section 6.3.

Risk Category	Check	Severity (0-5)	Rationale	Key Findings
Bias and Fairness	**C3.1.1** Production outcome equity			
	C3.1.2 Complaint patterns			
	C3.1.3 Override patterns			
	C3.1.4 Benchmark vs. production gap			
	C3.1.5 Feedback loop monitoring			
	C3.1.6 Automation bias indicators			
Privacy and Confidentiality	**C3.2.1** Production logging practices			
	C3.2.2 Output exposure of personal data (LLM)			
	C3.2.3 User data handling			
	C3.2.4 Data subject rights			
	C3.2.5 Third-party data sharing			
	C3.2.6 Privacy incident response			

Risk Category	Check	Severity (0-5)	Rationale	Key Findings
Reliability	C3.3.1 Performance stability			
	C3.3.2 Drift detection			
	C3.3.3 Production confabulation (LLM)			
	C3.3.4 Harmful output incidents			
	C3.3.5 Consistency in production			
	C3.3.6 Rollback capability			
	C3.3.7 Incident response effectiveness			
Governance	C3.4.1 Operational accountability			
	C3.4.2 Production change control			
	C3.4.3 Human oversight implementation			
	C3.4.4 Audit trail completeness			
	C3.4.5 Regulatory compliance maintenance			
	C3.4.6 Documentation currency			

Risk Category	Check	Severity (0-5)	Rationale	Key Findings
Security and Misuse	C3.5.1 Production security monitoring			
	C3.5.2 Adversarial input detection			
	C3.5.3 Prompt injection monitoring (LLM)			
	C3.5.4 Jailbreak attempt tracking (LLM)			
	C3.5.5 Misuse detection			
	C3.5.6 Security incident response			
	C3.5.7 Access control effectiveness			
Transparency and Explainability	C3.6.1 AI use disclosure			
	C3.6.2 Purpose and limitations communication			
	C3.6.3 Explanation provision			
	C3.6.4 Explanation comprehensibility			
	C3.6.5 Auditor/ regulator access			
Responsibility and Redress	C3.7.1 Appeal mechanism availability			
	C3.7.2 Appeal process effectiveness			
	C3.7.3 Harm remediation			
	C3.7.4 Accountability clarity			
	C3.7.5 Communication with affected parties			

Summary Assessment:

Risk Category	Highest Severity	Median Severity	Checks ≥3	Total Checks	Overall Assessment
Bias and Fairness	_/5	_/5	_/6	6	
Privacy and Confidentiality	_/5	_/5	_/6	6	
Reliability	_/5	_/5	_/7	7	
Governance	_/5	_/5	_/6	6	
Security and Misuse	_/5	_/5	_/7	7	
Transparency and Explainability	_/5	_/5	_/5	5	
Responsibility and Redress	_/5	_/5	_/5	5	

Note:

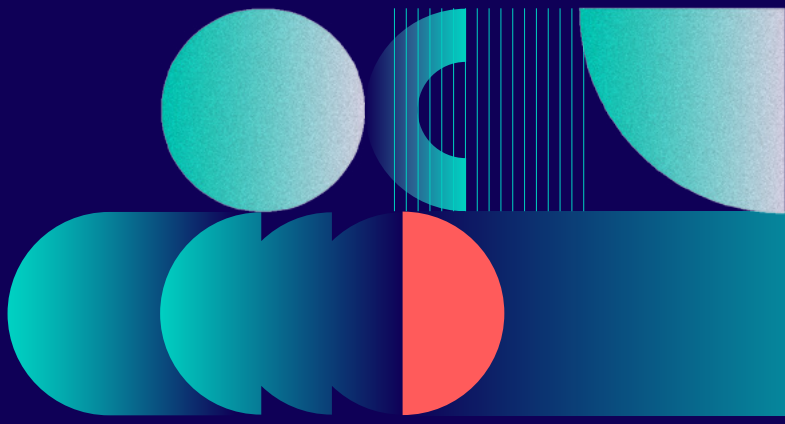
Post-processing includes two additional categories (Transparency and Explainability; Responsibility and Redress) that are particularly critical once a system is deployed and affecting real users.

How to interpret the summary:

- **Highest Severity:**
  Maximum score among all applicable checks in that category.
- **Median Severity:**
  The median score among all applicable checks in the category.
- **Checks ≥3:**
  Count of applicable checks scoring 3 or higher.
- **Overall Assessment:**
  Qualitative summary based on the pattern of findings.

CHAPTER 08

CONCLUSIONS



The RAIA Guide sets out a structured, risk-based approach to assessing AI systems across their lifecycle. It harmonises ethics, human-rights and technical analysis into a single methodology that is applicable to AI systems, including Automated Decision-Making (ADM) systems and LLM-based generative AI systems. Through the translation of high-level principles into operational, implementable risk categories and concrete questions, the RAIA guide empowers practitioners in identifying, documenting and mitigating core risks and other key concerns in AI systems.

The assessment process in the RAIA guide follows a **defined workflow**: a pre-survey and a scoping call, followed by an analysis and deep dive session, and concludes with a risk and mitigation report and a follow-up discussion. The findings emanating from discussions of each stage are **summarised in a scorecard**, rendering a concise overview of the severity of the identified risks and main mitigation measures proposed. The scorecard is a snapshot at a specific juncture of the lifecycle of an AI system, and requires ongoing review and updates, as the system is monitored.

The methodology has been **co-created and tested with diverse partners in Sub-Saharan Africa and Asia Pacific,** demonstrating its relevance across sectors such as energy, agriculture, climate innovation, entrepreneurship support and access to public services. This grounded approach has ensured that the guide reflects real-world constraints and contexts, including varying levels of digital infrastructure, institutional capacity, regulatory maturity and data availability.

The RAIA Guide has clear limitations. It is not a certification scheme, nor a formal audit instrument, and **it does not replace legal, regulatory or domain-specific expert reviews.** The efficacy of the assessment is dependent on the quality of participation, availability of data, accuracy of information and willingness of teams to engage openly with potential risks and trade-offs. The guide should therefore be understood as an aid to structured reflection, documentation and dialogue, across the spectrum of project teams, developers, deployers, regulators, impacted communities and other stakeholders.

The RAIA Guide is intended as a **living, adaptable framework.** As AI practices, regulations and technologies evolve, the guide can be renewed with additional tools, examples, regulatory guidance and sector-specific adaptations. Users are encouraged to provide feedback, share case studies and document how the methodology performs in practice – thus producing a pool of AI commons risk approaches, and commons-based governance. In this way, the RAIA Guide can continue to support more accountable, inclusive, transparent and context-sensitive AI systems that are aligned with human rights and the public interest.

CHAPTER 09

BIOGRAPHIES: COMMUNITY OF AI INCLUSION EXPERTS



Josia Paska Darmawan

Josia Paska is a research activist and social data analyst. Their main interests include diversity, equity, inclusion, digital justice, marginalised community empowerment, and more. Josia has experience in international institutions like the centre for Digital Society, Fairwork Foundation, 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 organisation that advances responsible data practices in support of emerging technologies.

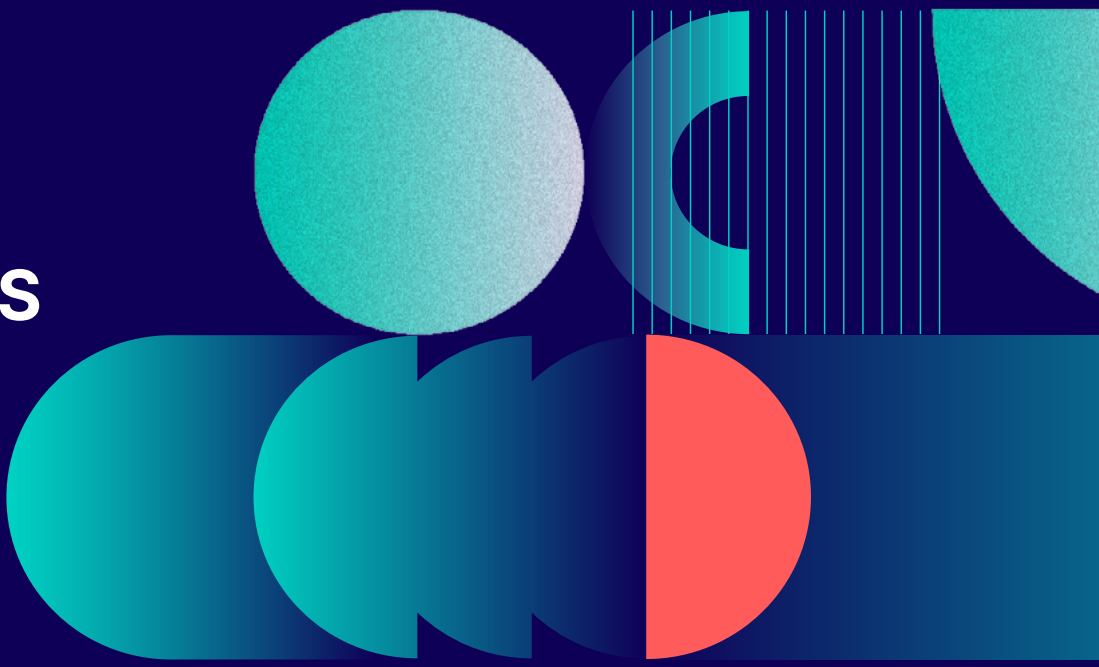


Dr. Meena Lysko

Meena Lysko is an applied physicist working at the intersection of AI, information accessibility, and inclusive innovation. Meena's experience includes 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 a researcher at the Council for Scientific and Industrial Research, and more. Meena is the director of Move Beyond Consulting, a company committed to work that goes beyond conventional consulting, creating lasting, impact-driven changes for communities, industries, and society.

CHAPTER 10

TEMPLATES



The following templates are available to support the practical application of the RAIA Guide. They can be adapted to the context, governance structures, and documentation standards of each organisation.

1. Pre-survey questionnaire template
2. Risk and mitigation measures report

Teams are encouraged to reuse and modify these templates to ensure consistency across assessments while reflecting their specific needs and constraints.

CHAPTER 11

REFERENCES

- Barocas, S., Hardt, M., & Narayanan, A. (2018). Fairness in machine learning limitations and opportunities. The MIT Press. https://api.semanticscholar.org/CorpusID:113402716
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.
- 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
- Castillo, C. (2018). Assessing the impact of machine intelligence on human behaviour: an interdisciplinary endeavor. In Proceedings of HUMAINT Workshop. https://arxiv.org/pdf/1806.03192.pdf.
- Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. ArXiv preprint. 2010.04053 (arxiv.org)
- Criado, N., Ferrer, X., & Such, J. M. (2020). A normative approach to attest digital discrimination. ArXiv.org. https://arxiv.org/abs/2007.07092
- Davidovic, J., Brown, S., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. Big Data & Society, 8(1).
- Eticas, & Castillo, C. (2021). Guide to algorithmic auditing. Eticas. https://eticas.ai/wp-content/uploads/2024/04/Guide-to-Algorithmic-Auditing-EN.pdf
- 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)
- 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
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29 (pp. 3323–3331).

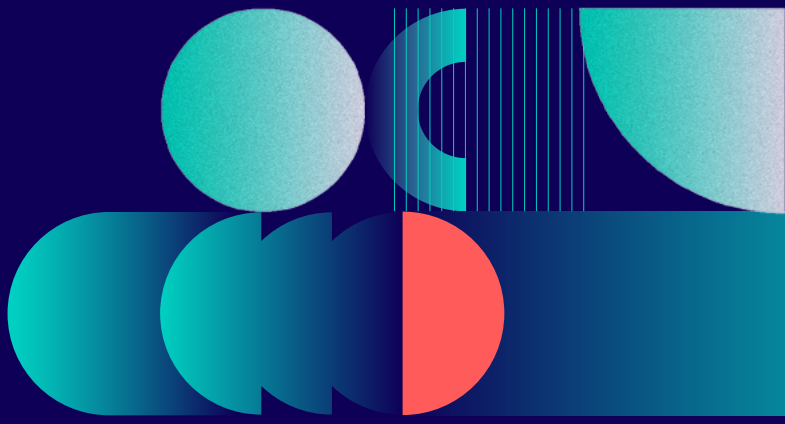
- Hasan, A., Brown, S., Davidovic, J., Lange, B., & Regan, M. (2022). Algorithmic bias and risk assessments: Lessons from practice. Digital Society, 1(1), 14.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. PLoS Biology, 13(3). https://doi.org/10.1371/journal.pbio.1002106
- High-level Advisory Body on Artificial Intelligence (2023). Interim report: governing AI for humanity. United Nations. https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf
- Ho, D. A., & Beyan, O. (2020). Biases in data science lifecycle. arXiv preprint. 2009.09795 (arxiv.org)
- ISO/IEC. (2018). ISO/IEC 20000-1:2018 Information technology — Service management — Part 1: Service management system requirements. ISO/IEC.
- ISO/IEC. (2023). ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management. ISO/IEC.
- ISO/IEC. (2025). ISO/IEC 42005:2025 Information technology — Artificial intelligence — AI system impact assessment. ISO/IEC.
- ISO/IEC. (2022). ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). ISO/IEC.
- ISO/IEC. (2023). ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. International Organization for Standardization / International Electrotechnical Commission.
- ISO. (2023). ISO 37004:2023 Governance of organizations — Governance maturity model — Guidance. ISO/IEC.
- 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, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. ArXiv.org. https://arxiv.org/abs/1910.09700
- Lacoste, Y. (2012). Geography, geopolitics, and geographical reasoning. Herodote, 146-147(3-4), 14-44. https://doi.org/10.3917/her.146.0014
- Laato, S., Tiainen, M., Najmul, A., & Mäntymäki, M. (2022). How to explain AI systems to end users: a systemic literature review and research agenda. Emerald Insights, 32(7), 1-31. https://doi.org/10.1108/intr-08-2021-0600
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020).
- 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

- Lippert-Rasmussen, K. (2013). Born free and equal? A philosophical inquiry into the nature of discrimination. Oxford University Press.
- 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
- Mantelero, A., & Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. Computer Law & Security Review, 41. https://doi.org/10.1016/j.clsr.2021.105561
- 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
- Mitchell, S., Potash, E., Solon, B., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application, 8(1), 141-163. https://doi.org/10.1146/annurev-statistics-042720-125902
- NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.
- OECD.AI Network of Experts. (2024). Catalogue of Tools & Metrics for Trustworthy AI. OECD.AI.
- OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments, OECD/LEGAL/0449.
- 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.
- O'Neill, C., & Gunn, H. (2020). Near-term artificial intelligence and the ethical matrix. In S. Matthew Liao (Ed.), Ethics of artificial intelligence, 237-270. Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0009
- Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. KDD, 560-568. https://doi.org/10.1145/1401890.1401959
- Prabhakaran, V., Mitchell, M., Gebru, T., & Gabriel, I. (2022). A human rights-based approach to responsible AI. ArXiv.org. https://arxiv.org/abs/2210.02667
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Association for Computing Machinery, 33–44. https://doi.org/10.1145/3351095.3372873
- Richardson, R. (2022). Defining and Demystifying Automated Decision Systems. Maryland Law Review, 81(3), 785–819.
- 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
- Scikit-learn developers. (2024). 3.4. Metrics and scoring: quantifying the quality of predictions. In scikit-learn User Guide (version 1.7).
- Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: definition and background. In Mission AI, Research for policy, 15-41. Springer, Cham. https://doi.org/10.1007/978-3-031-21448-6_2

- 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
- Smith, A. L., Greaves, F., & Panch, T. (2023). Hallucination or confabulation? Neuroanatomy as metaphor in large language models. PLOS Digital Health, 2(11), e0000388. https://doi.org/10.1371/journal.pdig.0000388
- 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. (2021a). 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
- UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137
- UNESCO. (2023). Guidance for generative AI in education and research. United Nations Educational, Scientific and Cultural Organization.
- UNESCO. (2025). Red Teaming Artificial Intelligence for Social Good: The Playbook. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000394338.locale=en
- United Nations. (1948). Universal declaration of human rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights
- 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.

ANNEX A

CONSOLIDATED RISK ASSESSMENT SCORECARD



This annex presents a full cross view of all checks, consolidated into one checklist.

For stage-specific scorecards and scoring guidance, refer to Sections: 7.1.6. 7.2.6. 7.3.8.

Pre-processing scorecard

Stage	Risk Category	Severity (0–5)	Rationale	Key Findings
Bias and Fairness	**C1.1.1** Target variable appropriateness			
	C1.1.2 Selection bias in ground truth			
	C1.1.3 Population alignment			
	C1.1.4 Underrepresentation			
	C1.1.5 Historical bias in sources			
	C1.1.6 Labelling consistency			
	C1.1.7 Fairness criteria defined			
	C1.1.8 Exclusions documented			

Stage	Risk Category	Severity (0–5)	Rationale	Key Findings
Privacy and Confidentiality	**C1.2.1** Personal data inventory			
	C1.2.2 Sensitive data handling			
	C1.2.3 Data Privacy Risk or Impact Assessment			
	C1.2.4 Legal basis			
	C1.2.5 Consent quality			
	C1.2.6 Anonymisation adequacy			
	C1.2.7 Access controls			
	C1.2.8 Retention limits			
Reliability	**C1.3.1** Label accuracy			
	C1.3.2 Label consistency			
	C1.3.3 Data quality controls			
	C1.3.4 Training-deployment alignment			
	C1.3.5 Known limitations documented			
	C1.3.6 Data provenance			

Stage	Risk Category	Severity (0–5)	Rationale	Key Findings
Governance	**C1.4.1** Decision documentation			
	C1.4.2 Roles and accountability			
	C1.4.3 Change management			
	C1.4.4 Auditability			
	C1.4.5 Regulatory alignment			
	C1.4.6 Policy alignment			
Security and Misuse	**C1.5.1** Access control			
	C1.5.2 Pipeline security			
	C1.5.3 Third-party risk			
	C1.5.4 Data poisoning risk			
	C1.5.5 Encryption and storage			
	C1.5.6 Incident response			

In-processing scorecard

Stage	Risk Category	Severity (0–5)	Rationale	Key Findings
Bias and Fairness	**C2.1.1** Fairness metrics computed			
	C2.1.2 Group performance disparity			
	C2.1.3 Threshold fairness (ADM)			
	C2.1.4 Embedding bias (LLM)			
	C2.1.5 Stereotype testing (LLM)			
	C2.1.6 Evaluation set representativeness			
	C2.1.7 Fairness-performance trade-offs documented			
Privacy and Confidentiality	**C2.2.1** Data anonymisation			
	C2.2.2 Data minimisation			
	C2.2.3 Purpose alignment			
	C2.2.4 Privacy preservation in training			
	C2.2.5 Access controls			
	C2.2.6 Personal data leakage testing (LLM)			
	C2.2.7 RAG corpus confidentiality			
	C2.2.8 Evaluation data handling			

Stage	Risk Category	Severity (0–5)	Rationale	Key Findings
Reliability	**C2.3.1** Overfitting assessment			
	C2.3.2 Calibration (ADM)			
	C2.3.3 Output consistency			
	C2.3.4 Robustness to perturbations			
	C2.3.5 Confabulation rate (LLM)			
	C2.3.6 Factuality and grounding (LLM)			
	C2.3.7 Prompt sensitivity (LLM)			
	C2.3.8 Reproducibility			
Governance	**C2.4.1** Model documentation completeness			
	C2.4.2 Development accountability			
	C2.4.3 Experiment tracking			
	C2.4.4 Change management			
	C2.4.5 Internal review			
	C2.4.6 Third-party model governance			

Stage	Risk Category	Severity (0–5)	Rationale	Key Findings
Security and Misuse	**C2.5.1** Training infrastructure security			
	C2.5.2 Data poisoning resilience			
	C2.5.3 Adversarial robustness (ADM)			
	C2.5.4 Prompt injection testing (LLM)			
	C2.5.5 Jailbreak testing (LLM)			
	C2.5.6 Model extraction risk			
	C2.5.7 Misuse scenarios identified			
	C2.5.8 Incident response readiness			

Post-processing scorecard

Bias and Fairness	**C3.1.1** Production outcome equity			
	C3.1.2 Complaint patterns			
	C3.1.3 Override patterns			
	C3.1.4 Benchmark vs. production gap			
	C3.1.5 Feedback loop monitoring			
	C3.1.6 Automation bias indicators			

Stage	Risk Category	Severity (0–5)	Rationale	Key Findings
Privacy and Confidentiality	**C3.2.1** Production logging practices			
	C3.2.2 Output exposure of personal data (LLM)			
	C3.2.3 User data handling			
	C3.2.4 Data subject rights			
	C3.2.5 Third-party data sharing			
	C3.2.6 Privacy incident response			
Reliability	**C3.3.1** Performance stability			
	C3.3.2 Drift detection			
	C3.3.3 Production confabulation (LLM)			
	C3.3.4 Harmful output incidents			
	C3.3.5 Consistency in production			
	C3.3.6 Rollback capability			
	C3.3.7 Incident response effectiveness			

Stage	Risk Category	Severity (0–5)	Rationale	Key Findings
Governance	**C3.4.1** Operational accountability			
	C3.4.2 Production change control			
	C3.4.3 Human oversight implementation			
	C3.4.4 Audit trail completeness			
	C3.4.5 Regulatory compliance maintenance			
	C3.4.6 Documentation currency			
Security and Misuse	**C3.5.1** Production security monitoring			
	C3.5.2 Adversarial input detection			
	C3.5.3 Prompt injection monitoring (LLM)			
	C3.5.4 Jailbreak attempt tracking (LLM)			
	C3.5.5 Misuse detection			
	C3.5.6 Security incident response			
	C3.5.7 Access control effectiveness			

Stage	Risk Category	Severity (0–5)	Rationale	Key Findings
Transparency and Explainability	**C3.6.1** AI use disclosure			
	C3.6.2 Purpose and limitations communication			
	C3.6.3 Explanation provision			
	C3.6.4 Explanation comprehensibility			
	C3.6.5 Auditor/ regulator access			
Responsibility and Redress	**C3.7.1** Appeal mechanism availability			
	C3.7.2 Appeal process effectiveness			
	C3.7.3 Harm remediation			
	C3.7.4 Accountability clarity			
	C3.7.5 Communication with affected parties			

Cross-stage Summary

- **Highest Severity:** maximum score in category.
- **Median Severity:** median score.
- **Checks ≥3:** count of checks scoring 3 or higher.
- **Overall Assessment:** qualitative summary based on pattern of findings.

Risk Category	Stage	Highest Severity	Median Severity	Checks ≥3	Total Checks	Overall Assessment
Bias and Fairness	Pre-processing	/5	/5	/8	8	
	In-processing	/5	/5	/7	7	
	Post-processing	/5	/5	/6	6	
Privacy and Confidentiality	Pre-processing	/5	/5	/7	7	
	In-processing	/5	/5	/7	7	
	Post-processing	/5	/5	/6	6	
Reliability	Pre-processing	/5	/5	/6	6	
	In-processing	/5	/5	/8	8	
	Post-processing	/5	/5	/7	7	
Governance	Pre-processing	/5	/5	/6	6	
	In-processing	/5	/5	/6	6	
	Post-processing	/5	/5	/6	6	
Security and Misuse	Pre-processing	/5	/5	/6	6	
	In-processing	/5	/5	/6	8	
	Post-processing	/5	/5	/7	7	
Transparency and Explainability	Post-processing only	/5	/5	/5	5	
Responsibility and Redress	Post-processing only	/5	/5	/6	5	

german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

FAIR FORWARD
Artificial Intelligence for all.

CONTACT US:

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
E fairforward@giz.de

www.giz.de