ETHICAL AI LOCALISATION TOOLKIT

A Practical, Research-Based Guide for Ethical, Inclusive, and Context-Sensitive AI in Public and Community Services



Rilla Gusela Sumisra

Contents

Publication Information 3
Preface 4
Who Is This Book For? 5
How to Use This Book 5
PART I Foundations 8
Chapter 1 — When Ethical AI Principles Meet Local Realities 9
Chapter 2 — AI in the Indonesian Context: Ready for Whom? 15
Chapter 3 — From Principles to a Toolkit: How the Framework Was Developed 21
PART II-A Context, readiness, fairness and meaningful participation 27
Running Case — AI Services in Arunika Regency 27
Chapter 4 — Do We Understand the Context? 28
Chapter 5 — Who Is Involved, and Whose Voice Is Missing? 32
PART II-B Data, culture, language and accessibility 37
Running Case — Arunika Regency Takes the Next Step 37
Chapter 6 — Whose Data, Whose Consent, Whose Benefits? 38
Chapter 7 — Does AI Understand Language, Culture, and Its Users? 42
PART II-C Accountability, Remedy and Continuous Oversight 47
Running Case — Arunika After a Limited Pilot 47
Chapter 8 — When AI Gets It Wrong, Who Is Accountable? 48
PART III Using the Toolkit for Real Decision Making 55
Chapter 9 — Using the 15 Domains and 62 Indicators 56
Chapter 10 — From Assessment to Deployment Decisions 60
Methodological Notes for Chapters 9–10 65
Bibliography 66
Appendix A — The 62 Toolkit Indicators 70
Appendix B — Red-Flag Checklist 77
Appendix C — Evidence Checklist 79
Appendix D — Action Plan Template 80
Appendix E — Deployment Decision Record 81
Appendix F — Glossary 82
Appendix G — Participant Methodology and Anonymisation Note 83
Appendix H — Citation and Evidence Conventions 84
Appendix I — Access the Digital Toolkit 85

Publication Information

Title: Ethical AI Localisation Toolkit: A Practical, Research-Based Guide for Ethical, Inclusive, and Context-Sensitive AI in Public and Community Services

Author: Rilla Gusela Sumisra.

First edition, 2026. English-language edition.

© 2026 Rilla Gusela Sumisra. Except where otherwise indicated, this work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0).

Licence and reuse

The author's original material may be copied, redistributed, translated, remixed, transformed, and built upon for non-commercial purposes under the CC BY-NC-SA 4.0 licence.

Reusers must:

- give appropriate credit to Rilla Gusela Sumisra;
- provide a link to the licence;
- indicate whether changes were made;
- distribute adaptations under the same CC BY-NC-SA 4.0 licence; and
- not imply that the author endorses an adaptation, its creator, or its use.

No additional legal or technological restrictions may be applied that would prevent others from exercising the rights granted by the licence.

Licence: https://creativecommons.org/licenses/by-nc-sa/4.0/

Unless expressly indicated otherwise, this licence does not apply to third-party material, quotations owned by others, institutional names or logos, photographs or illustrations credited to other parties, or underlying interview records. Such material remains subject to its own terms and applicable law.

The licence does not waive privacy, publicity, moral, confidentiality, data protection, or other rights that may require separate permission.

Suggested citation

Sumisra, Rilla Gusela. 2026. Ethical AI Localisation Toolkit: A Practical, Research-Based Guide for Ethical, Inclusive, and Context-Sensitive AI in Public and Community Services. First edition.

For an adaptation, retain the source attribution and licence link, identify the adapter, describe the changes made, and release the adaptation under CC BY-NC-SA 4.0. This citation format is recommended; another format may be used provided that it supplies appropriate attribution.

Disclaimer

This book is intended for education, policy learning, and governance support. The toolkit does not replace legal assessment, cybersecurity audits, data protection impact assessments, professional accessibility audits, or applicable regulatory obligations. Users remain responsible for ensuring compliance with relevant laws, policies, contracts, and technical standards.

This is an independently authored educational and governance resource. It is not an official government instrument or certification. The views expressed are the author's own and do not necessarily represent those of any organisation. Acknowledgements do not imply institutional endorsement.

Preface

This book grew out of conversations and friendships across cultures. Through encounters with friends from different cultural backgrounds, including members of Indigenous communities, I learned that technology never operates in a vacuum. It encounters languages, histories, power relations, community practices for safeguarding knowledge, and experiences that global frameworks do not always capture.

These conversations raised the questions that underpin this toolkit: when an AI system is described as fair, safe, or inclusive, does that assessment also hold for people with different languages, geographical circumstances, abilities, and experiences? Who is involved in defining the purpose of the technology? What evidence demonstrates that their contributions actually influence decisions? And who can stop the system when risks remain unresolved?

The Ethical AI Localisation Toolkit was developed to help bridge principles and practice. This book does not propose a new universal definition of ethical AI. It translates literature, policy frameworks, contextual qualitative evidence, and risk-management logic into 15 domains and 62 indicators for examining AI readiness in diverse local contexts.

I thank Globethics for the learning opportunities and ethical reflection that enriched the development of these ideas; the Internet Society for the opportunity to learn and build connections as a Youth Ambassador; and Karine Gentelet, my mentor at Globethics, for her guidance and perspectives on social justice, communities, and technology.

I also thank the interview participants and friends who shared their experiences, questions, and criticisms. I do not treat these conversations as voices representing entire groups or communities, but as reminders that AI governance must remain grounded in human experience and the contexts in which technology is used.

These encounters, support, and learning opportunities enriched the intellectual development of this book, but do not imply institutional endorsement. All methodological choices, interpretations, and errors remain my responsibility as the author. This toolkit should support more accountable decisions, rather than serve as a compliance checklist that produces a single final number; a score must never conceal an unresolved red flag.

The book's working principles: Context before deployment. Evidence before confidence. Participation must influence decisions. A score never overrides an unresolved critical risk.

Terminology

This book uses technical terms such as evidence, red flag, human oversight, contestability, deployment, and reassessment as part of the toolkit's operational vocabulary. Terms are explained in context on their first use in each chapter. Domain names and indicator codes are used consistently throughout the toolkit.

Who Is This Book For?

The book's primary audience is teams in public institutions, local government, and civil society organisations involved in planning, procuring, implementing, overseeing, or evaluating AI systems in public and community-based services.

Reader	How to use the book	Audience category
Public institutions / local government	Pre-deployment review, procurement questions, pilot gate, monitoring, grievance and accountability design.	Primary
Civil society / community organisations	Independent assessment, evidence-based advocacy, participation review, community safeguards.	Primary
Developer / vendor	Translate contextual requirements into design, testing, documentation, accessibility, mitigation and monitoring.	Secondary
Procurement / audit / assurance teams	Request evidence, identify red flags, set conditions and follow-up actions.	Secondary
Academics / students	Study an applied model of contextual ethical-AI operationalisation.	Secondary

How to Use This Book

This book is designed as a professional handbook. Readers do not need to read every section in order. However, the following sequence is recommended for new AI projects and projects entering procurement or pilot testing.

Stage	Practice
1. Define	Define the system's purpose, affected groups, lifecycle stage, geography, type of AI, and the impact of its decisions.
2. Identify triggers	Identify conditions requiring closer scrutiny: high-impact uses, personal or sensitive data, community data, users with disabilities, local languages, limited connectivity in remote areas, children and young people, or external and cloud vendors.
3. Assess	Answer each indicator with Yes, Partial, No, or Not Applicable. A Yes answer still requires evidence.
4. Escalate	Check red flags. A No or Partial answer to a red-flag indicator requires escalation and may prevent deployment.
5. Mitigate	Turn gaps into actions with an owner, deadline, required evidence, and completion status.
6. Decide	Use the assessment to support decisions: Proceed, Proceed with Conditions, Pause & Mitigate, or Do Not Deploy.
7. Monitor	After launch, monitor drift, bias, complaints, accessibility regressions, vendor and library changes, and follow-up mitigation.
8. Reassess	Repeat the assessment when the context, features, data, vendor, or risks change.

Work collaboratively: Ideally, one person should not complete the assessment alone. Form a cross-functional team and involve representatives of the groups actually affected. Evidence matters more than the team's confidence that the system is 'already ethical'.

Practical sequence: READ → ASSESS → MITIGATE → DECIDE → MONITOR → REASSESS

Scope and Limitations

What the toolkit covers

The toolkit focuses on the readiness to implement AI in public services and community-based services, particularly where there is cultural or linguistic diversity, access/infrastructure inequalities, disabled users, community data, marginal groups, or decisions that may impact rights and access to services.

What the toolkit does not claim

- The toolkit is not a national standard, certification, or official compliance instrument.
- Four project interviews are not representative samples of Indonesia and are not used to generate statistical generalisation.
- The toolkit does not claim that all 62 indicators originated in interviews. Their provenance varies: literature/framework-derived, interview-informed, triangulated, or operational design.
- Score thresholds are decision-support thresholds adopted by the toolkit; they are not scientifically validated cut-offs for a population of AI projects.
- Red flags are the gating rules of the design toolkit. Users still need to consider the context, laws, and risk appetite of the organisation.
- The sustainability domain is maintained as part of governance, but it is not one of the main novelty claims of the project at the current evidence stage.

Transferability

Although developed with a focus on the Indonesian context, some concepts can be useful in other multicultural societies. However, users outside of that context should not assume that indicators, red flags, or local examples can be moved without contextual review.

Research and Evidence Notes

The toolkit development uses three streams of evidence: (1) literature and policy frameworks; (2) qualitative interviews to identify implementation gaps and lived experience; and (3) operational risk-management logic to translate both sources into questions, evidence requirements, red flags, mitigation actions, and deployment decisions.

Evidence stream	Function	Limitations
Literature & policy frameworks	Set baseline principles and governance expectations.	Not all frameworks speak at the same level of local context.
Contextual interviews	Identify lived risks, design constraints, implementation practices, and local governance expectations.	Contextual qualitative evidence; not a representation of national statistics.
Operationalisation	Turning principles and evidence into indicators that can be examined and linked to action.	A toolkit design choice that requires piloting and iteration.

Participant Attribution and Anonymity

This book identifies interview participants by role and participant code. Names and contact details are omitted from the participant descriptions. The codes connect contextual findings with their cited sources without presenting an individual as a representative of an entire community.

Code	Public label	Evidence lens
IP-01	Indigenous Participant	Community/Indigenous governance, collective consent, participation, cultural data, source pluralism
EDU-3T-01	Education Expert in 3T Regions (*Tertinggal, Terdepan, dan Terluar* or Underdeveloped, Frontier, and Outermost regions).	education, connectivity, local language, remote/3T readiness
CCE-01	Cultural and Community Expert	grassroots inclusion, women/marginalised groups, AI literacy, social inequality
DP-01	Disability Participant	Lived experience of blindness or low vision, assistive technology, authentication, grievance mechanisms, and human escalation

Interview references use participant codes and, where available, transcript timestamps or question numbers. The underlying interview records are not reproduced in this publication. Their citation does not grant access to, or permission to redistribute, those records.

Statements attributed to a participant describe that person's contribution to this project. They should not be interpreted as the position of all members of the group or community concerned.

Why a Professional Handbook?

The book maintains research traceability without imitating the structure of theses or journal articles. Literature and methods are used to reinforce practical decisions. The main chapters follow a pattern: Problem → Evidence → Local Lens → What This Means → Toolkit Check → Practical Example → Action.

The structure enables professional readers to understand why a safeguard is needed while also knowing what they should do differently in a real project.

Positioning Statement

The Ethical AI Localisation Toolkit is a research-based handbook that helps organisations assess the readiness of AI implementation in the local context of Indonesia. Instead of stopping at the principles of fairness, transparency, accountability, or inclusion, the toolkit translates those principles into practical questions: who should be involved, what evidence should be available, what risks should be red flagged, what mitigation should be done, and when implementation should be delayed or stopped.

PART I

Foundations

Why localisation is needed and how the toolkit was developed

Evidence and the Limits of Interview Use

This section is written as a research-based professional handbook rather than a thesis or research report. The literature discussion therefore addresses implementation questions instead of reviewing the entire history of ethical AI. Claims are distinguished conceptually as literature evidence, contextual interview evidence, author synthesis, or toolkit design rules.

The project's four interviews provide contextual qualitative evidence. They help identify implementation gaps and lived experiences that high-level principles may overlook. They do not establish statistically representative findings about Indonesia. Throughout this book, participants are identified using anonymous labels: Indigenous Participant (IP-01), Education Expert in 3T Regions (EDU-3T-01), Cultural and Community Expert (CCE-01), and Disability Participant (DP-01).

How to interpret claims: When the book says 'the literature indicates', the claim requires support from academic sources or frameworks. When it says 'in this project's interviews', the claim is contextual. When it says 'the toolkit specifies', it refers to an instrument design decision.

Chapter 1 — When Ethical AI Principles Meet Local Realities

1.1 Sound principles do not automatically produce sound practice

The ethics of artificial intelligence are often discussed through a relatively consistent set of principles: fairness, transparency, accountability, privacy, safety, human oversight, inclusion, and respect for human rights. These principles provide a common language for assessing whether AI is being developed and used responsibly. Frameworks such as the UNESCO Recommendation on the Ethics of Artificial Intelligence and the ASEAN Guide on AI Governance and Ethics provide an important foundation for organisations seeking to build a more systematic governance of AI.

However, the hardest work usually begins once the principles are agreed upon. An organisation can declare that its system is "fair", but it still has to answer: fair to whom, in what territory, to what language, and under what conditions of access? An organisation can have a policy of human oversight, but the question of its implementation is whether human officers have the authority, time, information, and institutional courage to reject AI recommendations. Similarly, the commitment to inclusion has not yet answered whether the affected group is merely present in the consultation or actually influences the decision.

Figure 1. Brief implementation concept



The difference between principle and practice is the starting point of this book. The Ethical AI Localisation Toolkit is not trying to replace a global or regional framework. The toolkit works as an operational layer: it helps users test how the principles survive when they encounter local contexts, resource inequalities, language diversity, community structures, accessibility, and real power relations.

The book's central question: Beyond asking whether an AI system follows ethical AI principles, what evidence must be available before we can trust that those principles work in a particular implementation context?

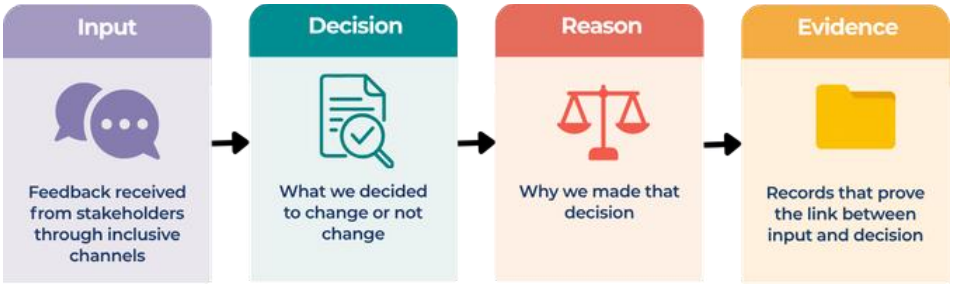
1.2 Localisation is more than translation

In this book, localisation does not mean simply translating the interface into Indonesian or replacing examples from other countries with local examples. Localisation is understood as the process of reassessing design assumptions, data, governance processes, participation mechanisms, complaint pathways, and readiness measures to fit the actual conditions of use.

This approach is in line with the literature that places AI as a socio-technical system, not just a computational model. STELA, for example, demonstrates the value of community-based deliberation to obtain normative perspectives from groups of different social and historical positions. Young et al. also reminded that participation in AI is not just a question of how many people are being asked for opinions, but how decision-making infrastructure allows affected communities to have substantial influence.[1]

With that understanding, localisation can change the questions that organisations ask. Instead of asking if stakeholders have been consulted, we ask if there is any feedback-to-decision trace. Instead of stating that digital services are available nationally, we ask if local conditions are verified and if manual/offline lines are available when connectivity fails. Instead of saying that the interface is accessible because it passes an automated checker, we ask if disabled users can actually accomplish important tasks with assistive technology.

Figure 2. Visual feedback-to-decision trace



1.3 Why high-level principles can lose sight of context

Ethical frameworks provide stable principles that organisations can apply across systems. However, the more general a principle, the greater the scope for interpretation during implementation. Fairness may mean parity in model metrics, fair opportunities, non-discrimination, or equitable access to services. Transparency may mean disclosure of AI involvement, explanations of decisions, model documentation, or source traceability. Participation may involve surveys, consultation, co-design, shared decision-making, or a community's right to refuse a particular use.

This question is no reason to reject the general framework. Instead, it shows why the framework needs to be translated into evidence requirements. ASEAN Guide, for example, has already linked principles to risk assessment, internal governance, human involvement, lifecycle, and monitoring. This toolkit contribution does not imply that such a framework is impractical, but rather details some implementation conditions that require additional granularity: community-level consent, local-language validation, accessibility as lived experience, remote readiness, anti-tokenism, and grievances that can actually come out of the bot to humans.

1 Stevie Bergman et al., "STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment," Scientific Reports 14 (2024): 6616, https://doi.org/10.1038/s41598-024-56648-4; Meg Young et al., "Participation versus Scale: Tensions in the Practical Demands on Participatory AI," First Monday 29, no. 4 (2024), https://doi.org/10.5210/fm.v29i4.13642.

1.4 From confidence to evidence

One of the risks in AI governance is replacing evidence with belief. Project teams can feel confident that the data is representative, the model is transparent, the public is already involved, or the system is accessible. But internal beliefs are not always auditable, and they are not always visible to those who accept the impact of decisions.

Figure 3. From confidence to evidence



Therefore, the toolkit uses a simple principle: the answer "Yes" still requires evidence. Evidence can be in the form of data inventory, bias testing results, local-language benchmarks, accessibility test records, change logs showing stakeholder input changing decisions, collective authorisation records, grievance SLAs, audit clauses, or monitoring dashboards. The goal is not to produce excessive document bureaucracy, but to make sure that governance claims have something to check.

This evidence-based approach also helps organisations distinguish between gaps that can be mitigated and gaps that should stop deployment. A documentation problem may be fixed by adding a process. On the other hand, the absence of human override on high impact decisions, the absence of alternative pathways for groups without digital access, or the failure of authentication for users of assistive technology can be a red flag that requires escalation.

1.5 Human-in-the-loop does not automatically mean human oversight

In many governance discussions, the presence of a human is treated as a safeguard. Research on public decision-making, however, gives reasons for caution. Alon-Barkat and Busuioc examined human–algorithm relationships through three experiments, including a study of Dutch civil servants. They did not find consistent support for automation bias in their experimental designs, but identified selective adherence under certain conditions: advice was more likely to be followed when it aligned with group stereotypes.[2]

The implications are important. Having a human officer at the end of the process does not automatically mean that the officer can correct bias. Human oversight must be designed: who may override a decision, what evidence reviewers can access, how disagreements are documented, when decisions must be escalated, and who has authority to stop the system. This is why the toolkit treats human oversight as an institutional capability rather than a checkbox.

Human oversight also needs to be linked to contestability. Leofante et al. (2024) argue that contestable AI requires a dynamic process that allows users to submit grounds for contestation, obtain explanations, and encourage revisions

when objections prove valid. In the context of government, Landau et al. (2024) places the

2 Saar Alon-Barkat and Madalina Busuioc, "Human–AI Interactions in Public Sector Decision Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice," Journal of Public Administration Research and Theory 33, no. 1 (2023): 153–169, https://doi.org/10.1093/jopart/muac007.

individual's ability to challenge decisions made about himself as a design and procurement need. Therefore, the grievance path in the toolkit does not stop at the complaint channel; it needs to lead to an authorised human reviewer, correction/remedy, and case traceability.

Figure 4. Human oversight



1.6 Source diversity involves more than adding references

Language models learn and operate through data and contexts that are not value-neutral. Feng et al. showed that political bias in pretraining corpora can carry over into language models and affect downstream tasks. Their study matters not because every AI deployment is political, but because it demonstrates a general mechanism: the structure and biases of source material can influence system behaviour.[3]

Stammbach et al. (2024) showed in a Swiss context that alignment with more diverse political perspective data can help models produce more accurate representations of viewpoints than the zero-shot baselines they tested. However, diversity of sources is not a guarantee of neutrality. Tan et al. (2024) even suggested that some LLMs in their experiments may prefer the wrong generated context to the right retrieved context when the two contexts conflict. Therefore, source governance needs to include provenance, conflict handling, uncertainty, and traceability - not just the number of sources.

Practical implication: using many sources is not the same as using sources correctly. The toolkit asks organisations to assess source provenance, conflicts, limitations, and evidence traceability for high-impact uses.

3 Shangbin Feng et al., "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models," in Proceedings of ACL 2023; Dominik Stammbach et al., "Aligning Large Language Models with Diverse Political Viewpoints," in Proceedings of EMNLP 2024, 7257–7267; Hexiang Tan et al., "Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts?," in Proceedings of ACL 2024.

1.7 Participation that does not change decisions is a weak signal

The AI participatory literature increasingly emphasises that participation cannot be judged by presence alone. Young et al. describe the participatory turn in AI and show the tension between local participatory methods and the scale of commercial AI systems. A more meaningful measure is the extent to which decision-making power actually moves closer to affected communities.[4]

Pillai et al., through the Community-Based framework of NLP, used the spectrum of inform, consult, involve, collaborate, and empower to show that engagement can have different levels of depth. STELA also shows how community-centred deliberation can produce normative insights that differ from rules formulated only by developers or experts.[5]

The project's evidence reinforces the issue contextually. IP-01 emphasises the risk of tokenism when communities are involved late or only as legitimacy after important decisions have been made. The toolkit then operationalises the problem through an indicator that asks for evidence that community input is changing the design or policy. In other words, the results of participation must leave a mark on the decision.

Gentelet, Mathieu, and Bahary-Dionne (2026) expand the argument through exploratory studies of digital justice initiatives developed by and for marginal groups. Their findings put social justice, lived experience, local knowledge, community ownership, and horizontal governance before technological novelty. For this toolkit, the implication is that participation evidence is not enough to show that the affected groups are present; evidence must show that they have agency, influence technology choices, and stay close to the decision-making structure.[6]

1.8 What does this book help users do differently?

This question becomes the centre of the book's positioning. The toolkit is more than just a list of principles. It aims to change the behaviour of implementation.

Before using the toolkit	A new question	Changes in practice
"The public has been consulted."	Is there evidence that their input changed the decision?	Create a participation timeline and feedback-to-decision trace.
"The system supports Indonesian."	Is the system tested on the user's actual language/dialect?	Build local benchmarks, error logs, and local validators.
"The website is now accessible."	Can users log in, complete tasks, and complain with assistive technology?	Do manual co-testing and retesting.
There are human officers.	Can officers override, stop, and document reasons?	Set authority, escalation, and stop rules.
" Services are available online."	What happens when connectivity goes out?	Provide low-bandwidth/offline/manual fallback.
"Data has obtained consent."	Does data have a collective or cultural dimension?	Separate individual consent from collective/community authorisation where relevant.

4 Young et al., "Participation versus Scale."

5 M. Pillai et al., "Toward Community-Based Natural Language Processing (CBNLP): Cocreating With Communities," Journal of Medical Internet Research 25 (2023): e48498; Bergman et al., "STELA."

6 Karine Gentelet, Lily-Cannelle Mathieu, and Alexandra Bahary-Dionne, "Social Justice Through Digital Empowerment: Insights from Digital Justice Initiatives by and for Marginalized Groups and Communities," Canadian Journal of Communication 51, no. 2 (2026): 261–287, https://doi.org/10.3138/cjc-2024-0101.

1.9 Limits to the book's claims

This book does not claim that every organisation in Indonesia faces the same problems, or that the project's four interviews represent Indonesian society. It offers a method for making context an explicit object of assessment. Literature provides a baseline; interviews add contextual perspectives; the toolkit translates both into testable questions.

The toolkit's success therefore depends on whether it helps teams identify previously unseen risks, request evidence they had overlooked, involve people previously excluded, and delay deployment while red flags remain unresolved.

Chapter 1 summary: ethical AI localisation translates principles into evidence and context-sensitive decisions. Its practical value lies in specifying who participates, what is tested, what evidence is required, which red flags require escalation, and when deployment must be delayed.

Chapter 2 — AI in the Indonesian Context: Ready for Whom?

2.1 Indonesia is not a single implementation context

Treating the Indonesian context as a single, uniform setting can create new problems. Indonesia encompasses urban and rural areas, economic centres and remote regions, users with different levels of digital literacy, hundreds of local languages, variations in institutional capacity, and diverse accessibility needs. Localisation must therefore avoid creating a new national stereotype.

Figure 5. Geographic distribution of diverse areas in Indonesia



This chapter uses two types of evidence carefully. First, Indonesian sources are used to show that digital access and capacity gaps do have social and geographical dimensions. Second, the project interview evidence is used as a contextual lens to show the types of questions that may arise when AI is applied to 3T, cultural/indigenous communities, grassroots groups, and disabled users. Interviews are not used as a national generalisation.

2.2 The digital divide extends beyond connectivity

One common trend in digital transformation projects is to measure readiness through the availability of devices and connections. Hadi (2018), in a study on rural–urban digital divide in Indonesia, suggests that the digital divide needs to be understood beyond material access. Motivation, skills, usage, education, institutional structures, and governance also play a role in society's ability to leverage technology.

Hadi remains relevant to explaining the structure of the digital divide, but its connectivity figures are not used to describe current conditions. For more recent conditions, BPS's Indonesian Telecommunications Statistics 2024 reports that 72.78% of Indonesia's population accessed the internet in 2024, up from 69.21% in 2023. National indicators, however, do not eliminate regional variation. Komdigi's 2024 Performance Report, using data covering 2023, states that 2,695 villages and urban administrative villages still lacked 4G coverage, including 2,441 in 3T regions; fibre-optic networks reached only 4,714 of 7,230 subdistricts. National connectivity growth and local readiness must therefore be interpreted separately.[7]

The BPS also reported Indonesia's 2024 ICT Development Index at 6.02, up from 5.90 in 2023. This index is useful for reading ICT developments across time and region, but it is not a direct measure of AI readiness. For the toolkit, the data reinforces the principle that readiness must be verified on location, users, service processes, and actual non-digital alternatives.[8]

The implications for AI governance are pretty straightforward. A system can be technically available but practically unusable. Users may have a mobile phone but not have stable

7 Aulia Hadi, "Bridging Indonesia's Digital Divide: Rural-Urban Linkages?," Jurnal Ilmu Sosial dan Ilmu Politik 22, no. 1 (2018): 17–33, https://doi.org/10.22146/jsp.31835; Badan Pusat Statistik, Statistik Telekomunikasi Indonesia 2024 (Jakarta: BPS, 2025).

8 Badan Pusat Statistik, Indeks Pembangunan Teknologi Informasi dan Komunikasi 2024 (Jakarta: BPS, 2025).

connectivity, adequate data plans, the skills to verify AI output, or support lines when the system fails. Therefore, the D10 and D11 in the toolkit are separated: one assesses remote/infrastructure readiness, the other assesses AI literacy and human readiness.

Conceptual synthesis of toolkit: AI readiness is not the same as device ownership. For assessment purposes, readiness should be read at least through technical, infrastructure, human, linguistic, accessibility, institutional, and community readiness.

2.3 Treating 3T (Underdeveloped, Frontier, and Outermost regions) as a readiness question, not a label of deficiency

In the project interviews, EDU-3T-01 describes infrastructure, language, and capacity constraints as implementation factors requiring direct verification. IP-01 also stresses that geography, transport, conflict, outages, and community structures affect system design. This contextual evidence does not establish that all 3T regions share the same conditions.

The toolkit therefore does not use '3T' as a label that automatically means unprepared. Instead, it translates the term into verification questions: has connectivity been tested on site? Is there a low-bandwidth mode? Can the process continue during an outage? Is manual fallback available? Can users access support without relying on the same digital channel? Can operations continue when vendors or cloud services are unavailable?

This approach is important because assuming the wrong readiness can result in exclusion by design. If an important public service can only be accessed through cloud-based AI and does not provide an alternative route, groups with low connectivity can lose access not because they reject technology, but because service design assumes conditions they don't have.

2.4 Bahasa Indonesia does not eliminate local-language challenges

Language is central to localisation because model performance is not always consistent across languages, and languages carry social and cultural context. Cahyawijaya et al. introduced IndoNLG for Indonesian, Javanese, and Sundanese, highlighting the limited NLG benchmarks and resources available for low-resource languages. They demonstrated the benefits of pretraining on related languages for improving performance in languages such as Javanese and Sundanese.[9]

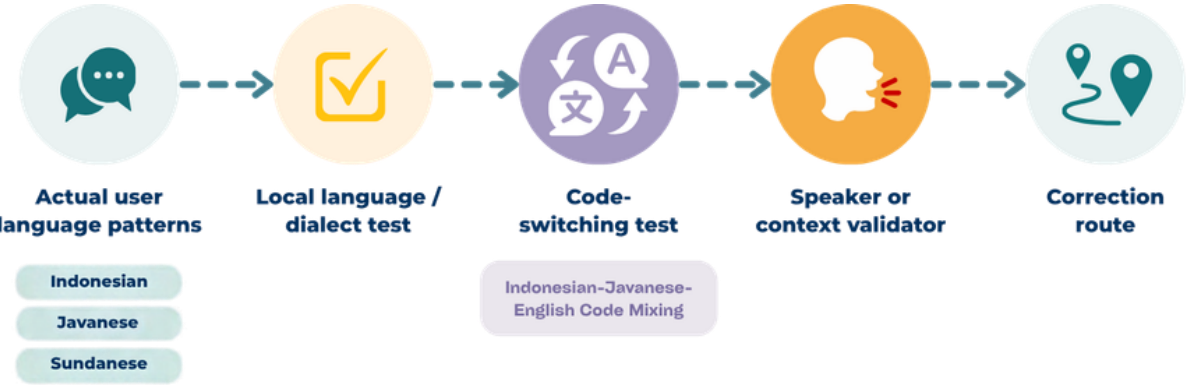
Hidayatullah et al. extended this question to the code-mixed Indonesian–Javanese–English. The study showed that common multilingual or monolingual models do not always handle code-mixing well and that specially designed models can deliver better performance on the tasks they evaluate. For this book, it doesn't mean that every public service should build its own language model. This means that language support must be tested on the user's actual language patterns, including code-switching and local variety where relevant.[10]

EDU-3T-01 and IP-01 also emphasize the need for validation by a speaker or validator who understands the context. The toolkit then distinguishes four questions: whether the dataset/cultural output is validated by the relevant party; whether the system is tested on the local language/dialect; whether language preservation efforts maintain community control; and whether misrepresentation can be corrected quickly by a clear validator.

9 Samuel Cahyawijaya et al., "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," in Proceedings of EMNLP 2021, 8875–8898, https://doi.org/10.18653/v1/2021.emnlp-main.699.

10 Ahmad Fathan Hidayatullah et al., "Pre-Trained Language Model for Code-Mixed Text in Indonesian, Javanese, and English Using Transformer," Social Network Analysis and Mining 15 (2025): 30, https://doi.org/10.1007/s13278-025-01444-9.

Figure 6. Language test path



2.5 Cultural validity involves more than adding cultural symbols

The system can look "local" because it uses local images, terms, or narratives, but remains culturally flawed. Cultural validity means assessing whether representations, categories, sources, and methods of explanation correspond to the context in which they are used. The literature of culturally responsive AI and decolonial AI helps to show that technology can bring certain epistemic assumptions and strengthen power relations if communities are only placed as data objects.

Figure 7. Cultural validity evidence loop

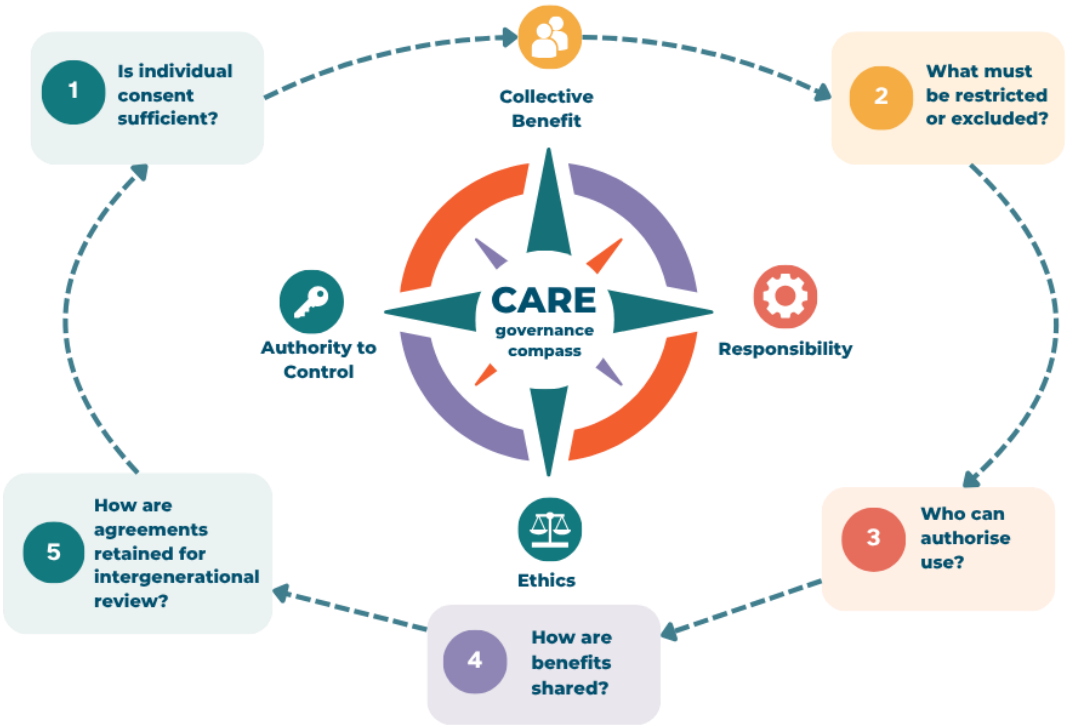


In the toolkit, cultural validation is not treated as decoration. It relates to evidence: who the validator is, what the mandate is, how the error is recorded, whether there is a correction workflow, and who has the authority to decide that a particular representation should not be published or used.

2.6 When data has a collective dimension

Most modern data governance focuses on individuals: who the data subject is, what processing bases are used, how consent is obtained, how long the data is stored, and how individuals can correct or delete the data. Those questions are important, but not always enough for data that has a community, cultural, regional, historical, or traditional knowledge dimension.

Figure 8. CARE governance compass



Some data relates to more than one person. Decisions must consider collective rights, responsibilities, and long-term impacts.

The CARE Principles for Indigenous Data Governance offers a framework that places Collective Benefit, Authority to Control, Responsibility, and Ethics as the principles of governance. Engstrom et al., through a scoping review of 85 health studies using routinely collected Indigenous data, showed that reporting on Indigenous guiding principles, data sovereignty, and consent still varies. These sources come from an international context and should not be treated as direct representations of indigenous Indonesian people.[11]

These sources provide a conceptual basis for questions also raised by IP-01 and EDU-3T-01: when is individual consent insufficient for collectively held knowledge? Which data must be restricted or excluded? Who can authorise its use? How should benefits be shared, and how can agreements be retained for intergenerational review?

D6 operationalises these questions. It complements personal data protection law by addressing data and knowledge with a collective dimension.

2.7 Accessibility means being able to complete a service journey

Digital accessibility is often reduced to compliance with technical lists. Technical standards remain important, but user experience suggests that obstacles can arise on a combination of seemingly small features: keyboard focus, screen reader announcement, CAPTCHA, OTP, 2FA, form validation, error messages, or buttons that can only be accessed via a pointer.

11 Stephanie Russo Carroll et al., "The CARE Principles for Indigenous Data Governance," Data Science Journal 19 (2020): 43, https://doi.org/10.5334/dsj-2020-043; Teyl Engstrom et al., "Indigenous Data Governance Approaches Applied in Research Using Routinely Collected Health Data: A Scoping Review," npj Digital Medicine 7 (2024): 68, https://doi.org/10.1038/s41746-024-01070-3.

Disability-inclusive AI literature reinforces the importance of user involvement. El Morr et al. (2024) identify limitations in user involvement and emphasise user-centred design and inclusive participation. Umucu (2025) positions persons with disabilities as co-designers. Maun et al. (2024) show that participatory design requires an understanding of user needs and relationships that support equal partnership. Van Toorn (2024), in a study of an Australian government digital assistant, shows that co-design can fail without sufficient institutional support and shared power.

The Indonesian context also shows why the accessibility of government digital services cannot be assumed. Lestari, Hafiar, and Priyatna evaluated 21 ministry websites using aXe DevTools against WCAG 2.1 and reported 1,098 accessibility violations across the websites analysed. This finding supports treating accessibility as a substantive public-service requirement. Because the evaluation relied on automated tools, however, this book does not treat it as evidence that automated checking is sufficient. It combines this evidence with participatory design literature and DP-01's lived experience to support manual testing and direct co-testing with users.[12]

DP-01 provides a more operational level of detail for the project: a combination of mouse + screen reader, CAPTCHA/OTP/2FA, accessible grievance, human escalation, direct co-testing, and regression monitoring. These details are not claimed to be universal research findings; they are used as lived-experience evidence that helps the toolkit test the service journey that users actually have to go through.[13]

The practical principle of accessibility is not only "can a page be opened", but "can users complete critical tasks safely, independently, and without workarounds that undermine their security or dignity?"

2.8 Grassroots inclusion and human readiness

CCE-01 emphasises AI literacy, understandable consent, data minimisation, community engagement, and human readiness. This contextual evidence helps avoid assuming that adoption automatically means readiness.

Human readiness needs to be tested through real capabilities: whether users understand that AI output can be wrong; whether they know when to verify information; whether officers can explain decisions; whether there are local trainers; whether material fits the context; and whether non-AI paths remain available to people who haven't or can't use the system.

It also changes the way we view literacy. AI literacy is not a one-time task before launch. It becomes part of the lifecycle, because models, features, data, interfaces, and risks can change. When the system is updated, the capacity of users and staff needs to be updated.

2.9 From the Indonesian context to implementation triggers

Rather than applying an identical checklist to every project, the toolkit uses prescreening triggers. These do not remove other domains; they identify areas requiring closer attention. High-impact uses call for scrutiny of rights, fairness, transparency, oversight, and grievances. Community or Indigenous data requires attention to collective governance. Users with disabilities bring accessibility, grievances, and monitoring into focus. Multiple or local languages heighten the need for cultural and linguistic validation

12 Dini Tri Lestari, Hanny Hafiar, and Centurion Chandratama Priyatna, "Analisis Aksesibilitas Website Kementerian di Indonesia Sebagai Implementasi dari E-Government," Journal of Digital Communication Science 2, no. 1 (2024): 16–26, https://doi.org/10.56956/jdcs.v2i1.274.

13 DP-01 (Disability Participant), interview result, August 21, 2026, Ethical AI Localisation Toolkit research project. Question references: Q6 supports accessible grievance, response-time expectations, and remedy, while Q8 supports post-launch accessibility regression monitoring.

and literacy. Unstable connectivity calls for attention to remote readiness, human readiness, and remedy. External or cloud vendors require closer scrutiny of privacy and vendor governance.

The trigger helps the localisation stay proportional. Low-risk internal systems do not need to be treated the same as AI that affects access to public services. In contrast, high-impact systems should not use the absence of an incident as an excuse to bypass evidence that should be available prior to deployment.

2.10 What we cannot yet claim

There are still areas that need more up-to-date Indonesian evidence. Hadi (2018) provides a conceptual foundation for the digital divide, but not enough to describe the connectivity conditions of 2026. The local-language literature of NLP has already provided strong technical evidence, but it does not automatically prove cultural harms in all its uses. Literature Indigenous Data Governance provides conceptual and comparative support, but should not be projected directly to all indigenous communities in Indonesia.

This chapter uses Indonesian evidence for claims about Indonesia, interview results for contextual illustration, as well as global literature and Global South for concepts and comparisons. The power of any claim remains limited by the scope and quality of available evidence, including on 3T, digital readiness, AI/digital literacy, and disability digital access.

Chapter 2 summary: Localisation does not mean looking for a single "Indonesian version" of ethical AI. It means making contextual variation an explicit part of the assessment: infrastructure, language, culture, accessibility, community data, human readiness, and institutional capacity.

Chapter 3 — From Principles to a Toolkit: How the Framework Was Developed

3.1 Why the development method needs to be explained

A governance toolkit gains credibility from the ability to trace why its indicators exist. Without provenance, a checklist may appear arbitrary: who decided that a question mattered? Does a requirement come from literature, interviews, author experience, or a design decision? Do all indicators have equally strong evidence?

The toolkit therefore uses an evidence model that distinguish between literature/frameworks, contextual qualitative interviews, and operational risk-management logic. The three streams do not have the same function. Literature sets the baseline; interviews help find contextual gaps and lived experience; operationalisation transforms both into questions, evidence requirements, red flags, and mitigation actions.

Figure 9. The toolkit combines three evidence streams



3.2 Stage one: literature and policy frameworks as a baseline

The development of the toolkit starts with the literature and the ethical/responsible AI framework that is already available. The UNESCO Recommendation provides a foundation for human rights, dignity, cultural diversity, inclusion, education, the environment, and governance throughout the lifecycle. The ASEAN Guide adds a regional perspective through the principles of governance, risk assessment, human involvement, internal structures, deployment, and monitoring.

In addition to framework governance, the development corpus also includes AI decolonial literature, cultural responsiveness, Indigenous/community data governance, accessibility, and intercultural ethics.

The goal is not to find the most accurate framework, but to identify recurring principles and areas that need contextual refinement.

For procurement and vendor governance, the book's evidence base is also strengthened with more operational public-sector guidance. The OECD (2025) identifies data/vendor lock-in as the challenge of AI implementation in public procurement and emphasises contract terms for IP/data rights, ongoing testing, monitoring, vendor performance, as well as lock-in protection. UK Government guidance and World Bank procurement guidance add open standards, data portability, auditability, knowledge transfer, exit management, and government control as mechanisms to reduce dependency on a single vendor.

The supporting literature serves two purposes: informing the development of indicators and strengthening the interpretation of findings. Sources used to corroborate or qualify an argument are distinguished from those that originally informed an indicator. This distinction preserves the provenance of the framework without treating every later reference as an original design source.

3.3 Stage two: four contextual interview perspectives

The project interviews were not designed as a representative survey of Indonesia. They provide contextual qualitative evidence from four deliberately different perspectives: education in remote/3T regions, cultural and community practice, Indigenous and community governance, and lived experience of disability and low vision. DP-01's experience of blindness or low vision does not represent Deaf people or people with motor, cognitive, learning, speech, or multiple disabilities. D8's broader coverage draws on literature, accessibility standards, and direct co-testing with users relevant to each use case.

Participant code	Contextual lenses	Example of contribution to toolkit
EDU-3T-01	Education Expert in 3T Regions	remote readiness, connectivity, language validation, non-digital access, human readiness
CCE-01	Cultural and Community Expert	Grassroots inclusion, AI literacy, data minimisation, clear consent, meaningful participation
IP-01	Indigenous Participant	Anti-tokenism, collective consent, restricted data, source pluralism, benefit-sharing, grievance mechanisms, and service continuity during conflict
DP-01	Disability Participant	screen reader + mouse/keyboard, CAPTCHA/2FA, accessible grievance, human escalation, co-testing, regression monitoring

The choice of these four lenses does not mean that they cover all of Indonesia's diversity. They serve as a stress test against framework assumptions. If a principle seems sufficiently clear at an abstract level but generates new questions when read through lived experience, those questions can be considered for operationalisation.

For example, "participation" becomes more concrete when IP-01 emphasises tokenism: when communities are involved, whether they can influence decisions, and what evidence that their input is not just collected. "Accessibility" becomes more concrete when DP-01 moves the focus from automated testing to user journeys: pointer + screen reader, CAPTCHA, authentication, grievance, and human escalation. "Readiness" becomes more concrete when EDU-3T-01 and IP-01 emphasize field conditions, not institutional assumptions.

3.4 Stage three: operationalisation

Literature and interviews don't automatically generate checklists. It takes a design decision to turn a principle into an item that can be evaluated. This process is called operationalisation.

Operationalisation in the toolkit follows a pattern: principle or concern → question assessment → evidence required → red flag when critical → mitigation when unmet → owner/action plan → deployment decision. With this structure, the toolkit tries to avoid two extremes: too abstract principles and too mechanical checklists.

Example operationalisation Participation principle → concern about tokenism → "Is there evidence that community input changes design/policy?" → change log linked to stakeholder input → red flag if participation is only formal → mitigation: feedback-to-decision trace.

3.5 Indicators do not all have the same provenance

One important methodological decision was not to claim that all 62 indicators came from four interviews. Most core principles have a literature/framework baseline. Some indicators are reinforced by interviews. Some are born of triangulation. Some mechanisms, especially scoring, thresholds, red-flag gates, and action-tracking logic, are operational design choices.

Indicator provenance is distinguished as literature- or framework-derived, interview-informed, operational design, or a combination of these. These categories make clear whether a requirement draws on established principles, contextual experience, or a design choice. They do not imply that every indicator has the same strength of evidence.

3.6 Turning a principle into a testable requirement

Example A — Accessibility

Baseline: digital services should be accessible. Disability-inclusive AI literature emphasises user engagement and participatory design. Van Toorn points out that inclusion in digital government is an institutional process, not just a technical outcome.[14]

Contextual lens: DP-01 identifies a combination of very operational barriers: the use of pointers with screen readers, authentication barriers, bot-only complaints, and regressions after updates.

Operationalisation: D8 takes a cross-disability and cross-modality approach. Critical user journeys should not depend solely on vision, hearing, speech, memory, complex movement, or one input device. Evidence may include keyboard and screen-reader tests, alternative-input tests, caption and transcript checks, accessible authentication, cognitive walkthroughs, and direct co-testing with relevant users. D14 adds regression monitoring after changes to features, vendors, libraries, or service modalities.

Example B — Participation

Baseline: affected stakeholders need to be involved. Young et al. show that meaningful participation relates to the transfer of decision-making power, whereas CBNLP differentiates the level of involvement from inform to empower. STELA demonstrates the benefits of deliberative community-centred input.[15]

---

14 Christo El Morr et al., "Exploring the Intersection of AI and Inclusive Design for People with Disabilities," Studies in Health Technology and Informatics (2024); Rachael Maun, Marc Fabri, and Pip Trevorrow, "Participatory Methods to Engage Autistic People in the Design of Digital Technology: A Systematic Literature Review," Journal of Autism and Developmental Disorders 54 (2024): 2960–2971; Georgia van Toorn, "Inclusion Interrupted," Government Information Quarterly 41, no. 1 (2024): 101900.

15 Young et al., "Participation versus Scale"; Pillai et al., "Toward Community-Based Natural Language Processing"; Bergman et al., "STELA."

Contextual lens: IP-01 emphasises anti-tokenism and engagement from an early stage. CCE-01 adds attention to grassroots inclusion.

Operationalisation: D9 not only requests participation records, but change logs that show input changing design/policy. Thus, attendance is not the primary evidence; influence evidence is sought.

Example C — Community data

Baseline: CARE Principles shift attention from data that is only understood as a technical object to collective benefit and authority to control. Engstrom et al. show that Indigenous Data Governance practices in research are still highly variable.[16]

Contextual lens: IP-01 and EDU-3T-01 raise issues of individual consent that are not always sufficient for collective knowledge, restricted cultural information, and long-term governance needs.

Operationalisation: D6 separates individual consent and collective/community authorisation, adding restricted-data registers, non-coercive/local-language consent, benefit-sharing, and intergenerational review.

Example D — Source diversity and traceability

Baseline: transparency and explainability are the general principles of AI governance. Feng et al. (2023) show that data pretraining bias can propagate to models and downstream tasks. Stammbach et al. (2024) suggests that the model can be directed to represent a more diverse political viewpoint in a Swiss context. Tan et al. (2024) show that models may prefer the wrong generated context to the right retrieved context.

Contextual lens: IP-01 raises concerns about source dominance on human rights issues, conflicts, or sensitive topics.

Operationalisation: D3 requests disclosure, local-language explanations, source/limitation traceability for high-impact use cases, as well as source diversity and conflict handling. The toolkit does not claim that diversity automatically eliminates bias; it makes source governance something to be checked.

3.7 From 15 domains to 62 indicators

The toolkit contains 15 domains and 62 indicators. They cover context and human rights; fairness; transparency; accountability; privacy; collective consent and community data; cultural and linguistic validation; accessibility; meaningful participation; local readiness; AI literacy and human readiness; grievances; vendors, procurement and data sovereignty; monitoring; and sustainability.

D15 Sustainability and Environmental Impact draws on literature addressing the full lifecycle of AI. UNEP highlights energy, water, mineral resources, emissions, and electronic waste across software and hardware lifecycles. Jiang et al. map the lifecycle of LLM-powered chatbots from research and hardware manufacture to operation and disposal. Bolón-Canedo et al. examine more efficient models, algorithms, hardware, and infrastructure. In this toolkit, D15 requires evidence of resource use, efficiency choices, lifecycle and vendor considerations, mitigation, and monitoring. It is not a carbon-accounting standard.[17]

The 15-domain structure is maintained because it helps organisations see that the ethical risks of AI are not at one point. A project can be strong on security but weak on accessibility; strong on model fairness

16 Carroll et al., “CARE Principles”; Engstrom et al., “Indigenous Data Governance Approaches.”

17 United Nations Environment Programme, Artificial Intelligence (AI) End-to-End: The Environmental Impact of the Full AI Life Cycle Needs to Be Comprehensively Assessed (Nairobi: UNEP, 2024); Peng Jiang et al., “Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots,” Engineering 40 (2024): 202–210; Verónica Bolón-Canedo et al., “A Review of Green Artificial Intelligence: Towards a More Sustainable Future,” Neurocomputing 599 (2024): 128096, https://doi.org/10.1016/j.neucom.2024.128096.

but weak on distributional access; strong on documentation but weak on grievance. Domain-level view prevents a single score from hiding the pattern.

3.8 Red flags as gating criteria

The toolkit treats certain conditions as red flags because their severity cannot be offset by high scores elsewhere. Examples include missing human override for high-impact decisions, missing collective authorisation for relevant community data, inaccessible critical functions, or grievance channels without human escalation.

It's important to clarify that red-flag logic is a design rule toolkit. It's not a clinical trial or a scientific cut-off. The reason is governance: aggregate scores can produce false reassurance when one critical safeguard fails.

So the principle of the toolkit is: a score never overrides an unresolved red flag. Scores help read readiness profiles; red flags help determine when escalation is needed.

3.9 Scoring supports decisions; it does not provide certification

The toolkit uses Yes = 2, Partial = 1, No = 0, and Not Applicable is removed from the denominator. Threshold then directs the initial decision: Proceed, Proceed with Conditions, Pause & Mitigate, or Do Not Deploy.

These thresholds should be read as decision-support thresholds adopted by the toolkit, not scientifically validated cut-offs to the AI project population. Therefore, the results of the assessment should not be marketed as certification that the system is “ethical”. The function of scoring is to help organisations find patterns of readiness, document conditions, and link gaps to action plans.

3.10 Evidence requirements as a governance discipline

One of the most fundamental design decisions is that “Yes” should have supporting evidence. It forces teams to distinguish between policy intent and operational reality. For example, claims that the public is involved should have a timeline and decision trace. Claim that the accessible interface must have a manual test record. Claims that the vendor can be audited must have a contract clause. Claims that the grievance works should have case tracking and SLA.

Evidence requirement also increases transferability. When a project changes people, vendors, or lifecycle phases, the next reviewer doesn't rely on the team's previous memory. It can see documents, test results, and unresolved actions.

3.11 Methodological limitations

The toolkit has several limitations. First, the small, purposive interview sample does not provide statistical representation. Second, DP-01 primarily contributes lived experience of blindness or low vision. Coverage of Deaf and hard-of-hearing users, motor and speech disabilities, cognitive and learning disabilities, neurodivergence, and multiple disabilities draws on literature and standards; relevant users must still participate in testing before accessibility claims are made. Third, evidence strength varies across domains. Participation, community data, local-language AI, and disability-inclusive AI have relatively strong literature triangulation. Sustainability draws on UNEP and Green AI and lifecycle literature but is not the toolkit’s main original contribution. Fourth, some operational rules have not undergone large-scale piloting.

Fifth, regulation, technical standards, and AI technologies change rapidly, so the toolkit requires regular updates. Sixth, organisations may interpret requirements differently; reviewer training and calibration are necessary when assessments inform formal decisions.

3.12 Further development and validation

The next stage is more than just adding an indicator. The toolkit needs to be tested in real use case or controlled simulation to see if the questions are understood consistently, evidence requirements are realistic, red flags are too sensitive or actually too weak, and action plans really help decisions.

Chapter 3 summary: the toolkit was developed through literature/framework triangulation, contextual qualitative interviews, and operational risk-management logic. Not all indicators have the same provenance. Transparency about origins, evidence strength, and design choices are all part of the credibility toolkit.

Closing Part I — From Foundations to Five Implementation Questions

The first three chapters build one major argument: ethical AI isn't finished on principle, and localisation isn't finished on translation. Responsible implementation requires evidence that the system suits the context, that the affected groups have influence, that data and resources have governance, that accessibility can be used in practice, and that organisations are able to respond when the system fails.

Part II will turn that foundation into five practical questions: Do we understand the context? Who is involved and who is not heard? Whose data, whose consent, whose benefits? Does AI understand language, culture, and users? When the AI is wrong, who's responsible?

Bridge to Part II: Principle → Context → Evidence → Assessment → Mitigation → Decision → Monitoring.

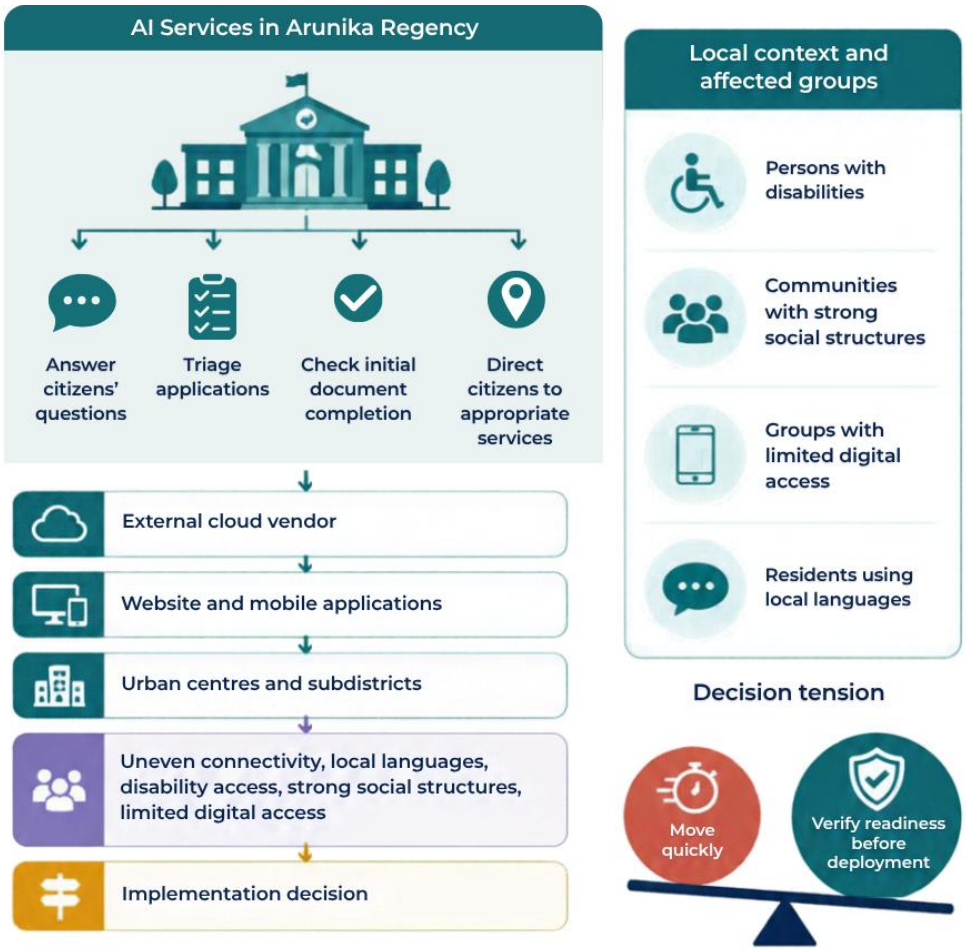
PART II-A
Context, readiness, fairness and meaningful participation

Running Case — AI Services in Arunika Regency

Throughout Part II, the book uses one consistent simulation case. The Arunika Regency Government plans to launch an AI-assisted public service to answer citizens' questions, triage applications, check the initial completion of documents, and direct citizens to appropriate services. The system comes from an external cloud vendor and will be used on websites as well as mobile applications.

The deployment area includes urban centres and subdistricts with uneven connectivity. Some residents use local languages in everyday conversation. Users include persons with disabilities, communities with strong social structures, and groups with limited digital access. The local government wants to move quickly because the project is part of its public-service transformation programme.

Figure 10. Fictional Case AI Services in Arunika Regency



Case note: Arunika Regency is fictional. It does not represent a specific locality. The scenario illustrates how the toolkit changes implementation decisions.

Chapter 4 — Do We Understand the Context?

4.1 Why context must be established through evidence

AI projects often begin with a technical question: what models are used, what data is available, and how much does it cost to implement. However, before that question is answered, organisations need to know the context of use: goals, who are affected, risk levels, infrastructure conditions, human capabilities, language, accessibility, and alternative pathways when systems fail.

Figure 11. Example of the context of use



The UNESCO Recommendation and the ASEAN Guide both place human rights, risk, accountability, and lifecycle governance as the baseline. This toolkit adds a more operational question: whether the context is actually documented, whether affected groups are mapped, whether readiness is verified in the field, and whether there is a fallback when digital assumptions are not met.

Principle Chapter 4: Context is not background information. Context is part of the control environment.

4.2 Context mapping before risk scoring

In the toolkit, D1 starts the assessment with system objectives, affected groups, human-rights implications, and risk classification. This is important because one technology can have very different risk profiles depending on the use case. An internal chatbot for document search is not the same as a system that affects citizens' access to social assistance, education, or administrative services.

Figure 12. A minimum context map

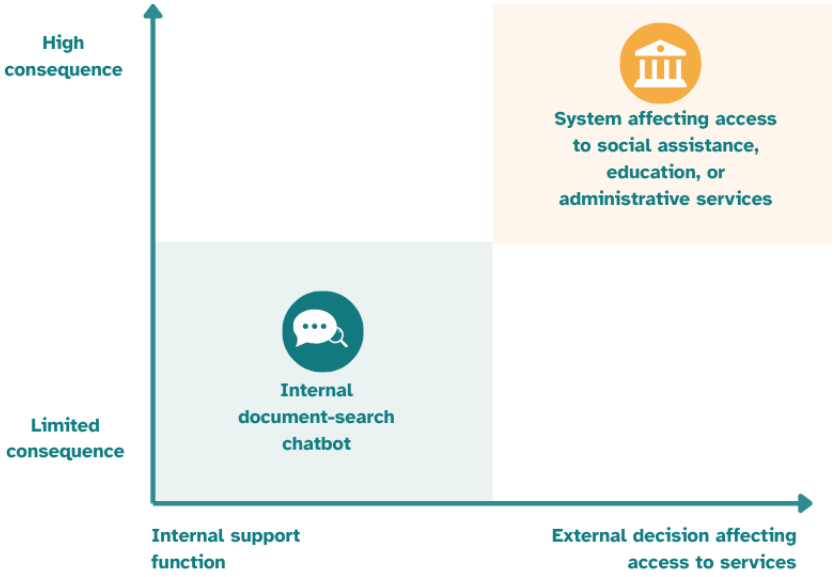


4.3 High impact is not just about technology

Risk classification should focus on the consequences of the decision, not just the sophistication of the model. Simple systems can be high-impact when determining access to essential services. Conversely, advanced models can be less risky if used only for internal functions without external decisions.

High-impact uses therefore trigger closer scrutiny of D1, D2, D3, D4, and D12. Beyond asking whether a system uses generative AI, teams must ask what happens if it is wrong, biased, inexplicable, or impossible to challenge.

Figure 13. Matrix of consequences



Impact is determined by the consequences of outcomes for people and communities, and by how easy it is to challenge or correct errors.

4.4 Readiness must be verified on site

BPS reports increasing internet access nationally, but national indicators do not establish local readiness. Komdigi's data also identify coverage gaps in villages and urban administrative villages, particularly in 3T regions. D10 therefore requires verification of actual local conditions rather than relying on regional labels.

EDU-3T-01 and IP-01 reinforce these needs in project interviews: readiness must take into account connectivity, transport/logistics, geographical conditions, outage risk, local operator capabilities, and the existence of non-digital lines. This evidence is contextual in nature and is not intended as a generalisation of the entire 3T region.

4.5 Offline and manual fallback as ethical safeguards

Fallback is often treated as a continuity or user experience issue. In public service, it is also a matter of fairness and rights. If the only way to get essential services requires cloud-based AI, then connectivity failure can turn into exclusion.

Therefore, D10 R2 asks organisations to assess whether there is low-bandwidth, offline, or manual fallback. A good fallback isn't just an unanswered phone number. It needs to be tested: whether officers are available, whether manual processes have SLAs, whether case status can be transferred from digital channels to humans, and whether citizens are not punished for using alternative channels.

No critical public service should become inaccessible solely because the AI channel is unavailable.

4.6 Human readiness: the system may be ready before its users are

D11 separates technical readiness from human readiness. These include AI literacy, training, output verification capabilities, local support, as well as the ability to run manual processes when systems are unavailable.

CCE-01 highlights the relationship between AI literacy, social conditions, and grassroots inclusion. A single awareness session is insufficient. Training should reflect each role: users need to understand AI's function and complaint routes; staff need to verify outputs, recognise failures, and override recommendations; managers need visibility of risks and authority to stop the system.

A single awareness session is insufficient.
AI literacy is an ongoing capability, not a one-time launch activity.

4.7 Toolkit in Action — D1, D10, D11

Domain	Core question	Examples of evidence	Red flag	Early mitigation
D1 Context / Rights / Risk	Are goals, affected groups, high-impact status, and rights implications documented?	Project charter; impact map; risk classification; affected-group register	Use case has high impact but risk classification does not exist	Pause scoping; perform an impact/risk review
D10 remote readiness	Are local conditions verified and fallback available?	Field-readiness report; connectivity tests; fallback SOP; outage drill	Essential services are only available via digital/AI channels	Set up manual/offline path and test before launch
D11 Human readiness	Do users, operators, and reviewers understand the role and limitations of AI?	Training records; role-based guidance; local support plan; manual-runbook	Officers don't know how to override failure.	Role training + escalation drills + refresher

4.8 Evidence to Look For

- a project profile that explains the objectives and limits of use;
- affected-group maps and stakeholder maps;
- risk classification linking use cases to potential harm;
- field-readiness verification, not just desk assessment;
- connectivity log or pilot test at target location;
- manual/offline fallback SOP and test results;
- role-based AI literacy/training records;
- human override and escalation matrix;
- review schedule when context or vendor changes.

4.9 Red Flags

- the system's role changes from assisting decisions to determining them without risk reassessment;
- affected groups have not been mapped;
- deployment in regions with unstable connectivity without fallback;
- officers are asked to follow AI output but do not have the authority to deviate;
- manual service was discontinued before the AI channel proved stable;
- citizens are not informed that AI affects service travel.

4.10 Running Case — Initial Findings in Arunika Regency

Arunika's team originally referred to the project as an 'information chatbot'. When context mapping was done, it was found that the system would also triage requests and mark documents considered incomplete. The output will affect the order of citizen service. Thus, the function of the system is more impactful than the original label.

Field-readiness checks also showed two districts experiencing unstable connectivity at a given time. Manual tracks are available, but undocumented and have no SLA. Front office officers haven't been trained on when AI recommendations should be ignored.

Interim decision: resolve the gaps in risk classification, fallback procedures and service levels, and role-based human-override training before expanding the pilot. Where these gaps leave a red-flag indicator at No or Partial, the toolkit requires Do Not Deploy until the critical safeguard is verified. Pause and mitigation describe the corrective work, not permission to bypass that gate.

4.11 A short reader checklist

- Do we document what AI actually does, not just the name of the project?
- Are affected groups and potential harms mapped?
- Is readiness verified at the target location?
- Are there non-AI pathways that actually work?
- Can an officer override and know when to do so?
- Is reassessment triggered when use cases, data, vendors, or context change?

Chapter 4 summary: context is not a decoration. It determines the level of risk, the evidence needed, who should be involved, what fallback is required, and whether organisations are prepared to implement AI.

Chapter 5 — Who Is Involved, and Whose Voice Is Missing?

5.1 Fairness goes beyond model metrics

Fairness often starts with a technical question: do the error rates differ between groups? This question is important, but it's too narrow when AI systems are used in public service. Injustices can arise before a model works, for example when a certain group cannot access a digital channel, or after output is generated, when some users have no way to understand, challenge, or correct a decision.

Therefore, D2 treats fairness more broadly: data representation, performance across groups, distributional access, and exclusion risks. Literature like The Forgotten Margins of AI Ethics and participatory AI reminds us that whoever is not present in a dataset or decision process can be just as important as metric fairness itself.

5.2 From representation to distributional fairness

A model can have good average performance but still produce an uneven impact. The fairness assessment needs to ask who benefits, who carries the error, who has to take additional steps, and who loses access if the system fails.

In the context of Indonesia, the digital divide shows why access fairness is relevant. If one group has to travel physically because it cannot use digital channels, while another group enjoys instant automation, service inequalities may arise even though models do not show differences in accuracy based on demographic attributes.

5.3 Source bias and fairness

Source governance also affects fairness. Feng et al. show how political bias in pretraining data can propagate to language models and downstream tasks. Stammbach et al. show that alignment with more diverse political perspectives can improve viewpoint representation in the Swiss context studied. These findings do not establish that adding sources automatically makes a model neutral; they support examining provenance and diversity of viewpoints.[18]

5.4 Meaningful participation: from consultation to influence

Participation means more than just attending a workshop. Young et al. emphasised the importance of decision-making power in participatory AI. Pillai et al. showed a spectrum of involvement from inform to empower. STELA suggests that community-centred deliberation can produce normative insights that differ from rule-sets created only by developers or experts.[19]

The IP-01 interview reinforces concerns about anti-tokenism: engagement that comes after major decisions are made only gives little room for the community to influence outcomes. CCE-01 adds attention to grassroots inclusion and the importance of engaging processes that can be understood by groups with different literacy levels.

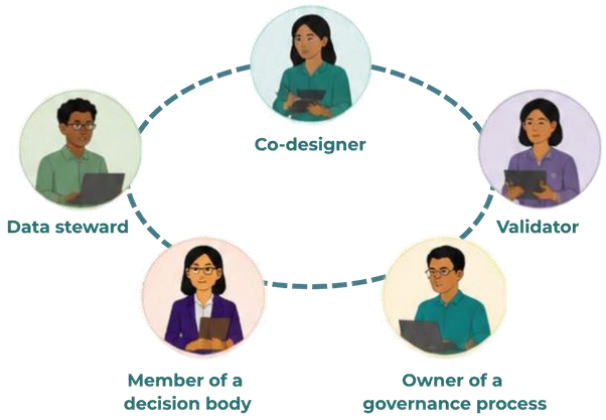
The by-and-for model provides a stronger benchmark. The affected group can act as a co-designer, data steward, validator, member of a decision body, or owner of a governance process. Gentelet et al. (2026) suggests that legitimacy and trust are not primarily born of technology, but rather of social proximity, representation,

18 Feng et al., “From Pretraining Data”; Stammbach et al., “Aligning Large Language Models with Diverse Political Viewpoints.”

19 Young et al., “Participation versus Scale”; Pillai et al., “Toward Community-Based Natural Language Processing”; Bergman et al., “STELA.”

and community ability to maintain control over goals, data, and deployment.

Figure 14. The by-and-for model: connected roles



Evidence standard Documentation that a person is 'invited' is weak participation evidence. The stronger evidence is that there are changes in decision making that can be traced back to their input.



5.5 Tracing feedback to decisions

One of the most important mechanisms in D9 is feedback-to-decision tracing. An organisation should be able to show not only what input is received, but how that input is processed.

Stakeholder input	Decisions	Reason	Evidence
Low vision user: CAPTCHA cannot be solved	Replace authentication with an accessible alternative option	Critical access barrier	Issue log + design change + retest
Local community: terms of service not understood	Revision of terminology and addition of local-language explanation	Risk of misunderstanding	Workshop note + change log
Subdistrict officer: the network frequently disconnects	Add offline/manual fallback	Continuity risk	Field test + SOP
The user group requested full auto-approval	Not accepted.	High-impact decisions still require human review	Decision record + rationale

5.6 Who should be involved?

Not all stakeholders need to be at every meeting. What is needed is coverage of people who have relevant knowledge, interests, and lived experience.

- the person who will use the service;
- persons affected by the decision but not using the interface directly;
- frontline staff dealing with failure cases;
- representative communities or organisations where the data/impact has a collective dimension;
- disabled users and assistive-technology users;
- local-language/cultural validators;
- procurement, privacy, security, legal/compliance, and service owners;
- vendor/developer - but not as the only requirement determinant.

5.7 Intersectionality: one category is not enough

Users don't live in one category. A citizen can simultaneously live in a region with low connectivity, use the local language, have a disability, and rely on certain public services. Intersectional review helps to see failure modes that are not visible when assessment separates categories rigidly.

The toolkit does not require statistical models for every combination of identities. It requires testing and participation plans that do not overlook people facing intersecting risks.

5.8 Accessible participation: testing with users

Indonesian evidence from Lestari, Hafiar, and Priyatna (2024) suggests that accessibility of government websites is still a real issue. But automated tools only catch some of the problems. The disability-inclusive literature of AI and DP-01 reinforces the reason for direct co-testing.

Disability-inclusive participation means users are engaged before the design is locked, test tasks reflect the actual service journey, issues have an owner, and retesting is done after the change. It connects D8 accessibility to D9 participation: accessibility is not just a technical domain, but also a question of who has the right to determine whether a service is actually usable.

5.9 Toolkit in Action — D2 and D9

Domain	Core question	Evidence	Red flag	Mitigation
D2 Fairness	Are performance, access, and adverse impacts tested in relevant groups/contexts?	Disaggregated test results; access analysis; error logs; impact review	One group carries a much higher error/access burden without mitigation	Retest; redesign; alternative channel; policy constraint
D9 Participation	Were affected communities involved from the start and did their input change decisions?	Participation timeline; stakeholder map; change log; decision rationale	Consultation is carried out after procurement/design is locked	Re-open requirements; structured co-design; document influence

5.10 Evidence to Look For

- stakeholder maps listing affected - not only powerful - groups;
- the participation plan before the requirement is locked;
- disaggregated model/performance testing where relevant;
- access and distribution impact analysis;
- the feedback-to-decision change log;
- records of accepted and rejected input, with reasons;
- direct co-testing records for accessibility/local-language issues
- benefit/burden analysis for different groups;
- reassessment after a change in model, policy, or channel.

5.11 Red Flags

- new stakeholder consultations are carried out after the selected vendor and the main design is locked;
- a community representative is present but does not have sufficient information or time to provide input;
- fairness is only measured as overall accuracy;

- group-level testing shows gaps but no mitigation owner;
- critical feedback recorded but no decision response;
- disabled users are not involved in critical function testing;
- the most vulnerable groups have no way to complain or challenge the decision.

5.12 Mitigation patterns

Problem	Mitigation patterns	Evidence closure
Tokenistic participation	Repeat engagement at the requirement stage with clear decision authority	Updated requirements + signed decision log
Intergroup performance gaps	Data review, threshold adjustment, workflow constraint, or human review	Retest report + residual-risk sign-off
Digital access burden	Low-bandwidth/manual channel; equivalent SLA	Channel test + service-level evidence
Language/cultural misunderstanding	Native/local validators; correction workflow	Validation report + change log
Accessibility barrier	Direct user testing + remediation + regression testing	Issue closure + retest evidence

5.13 Running Case — Participation in Arunika Regency

Arunika's team initially had a stakeholder list that mainly consisted of service providers, vendors, and consultants. After D9 was used, the list expanded to include frontline staff, residents of districts with low connectivity, screen reader/low vision users, local community representatives, and complaint officers.

The first workshop revealed two issues. Chatbot terms for application status were not understood consistently across subdistricts. The complaint channel also sent residents back to the same bot when they wanted to report its errors.

Feedback-to-decision trace results in two changes: terminology is changed after local-language validation, and grievance flow is redesigned so users can directly choose human escalation. The vendor initially declined because of the increased costs of integration. The decision was made to include the change as a condition before the pilot.

Participation case studies become meaningful when it changes a requirement, contract, workflow, or deployment decision. If there is no traceable change, the organisation needs to explain why.

5.14 From fairness to legitimacy

Fairness and participation are connected. A technically fairer model may still lack legitimacy if affected groups have no voice, cannot understand the system, or have no remedy. Meaningful participation also cannot substitute for technical testing. The toolkit treats both as complementary safeguards.

5.15 A short reader checklist

- Is fairness tested on performance and access?
- Was the affected group involved before the decision was sealed?
- Is there any evidence that stakeholder input changed anything?
- Is the decision not to follow the feedback documented for a reason?
- Does testing include relevant intersectional risks?

- Are direct users with accessibility needs involved?

- Do the groups at greatest risk also have a grievance/contestability path?

Chapter 5 summary: Fairness is not just numbers, and participation is not just attendance. Stronger governance connects representation, access, lived experience, decision power, and traceable influence.

Closing Part II-A — From Context to Influence

Chapters 4 and 5 show two steps that must be taken before an organisation is comfortable with deployment. First, the context has to be verified: goals, risks, infrastructure, human readiness, and fallback. Second, fairness and participation should be tested not just on formal documentation, but on who gains access, who bears the error, and who can actually influence decisions.

The next section will move on to more specific questions: who's using data, who's approving it, how benefit-sharing is understood, and how cultural-linguistic validity and accessibility are tested before AI is considered ready.

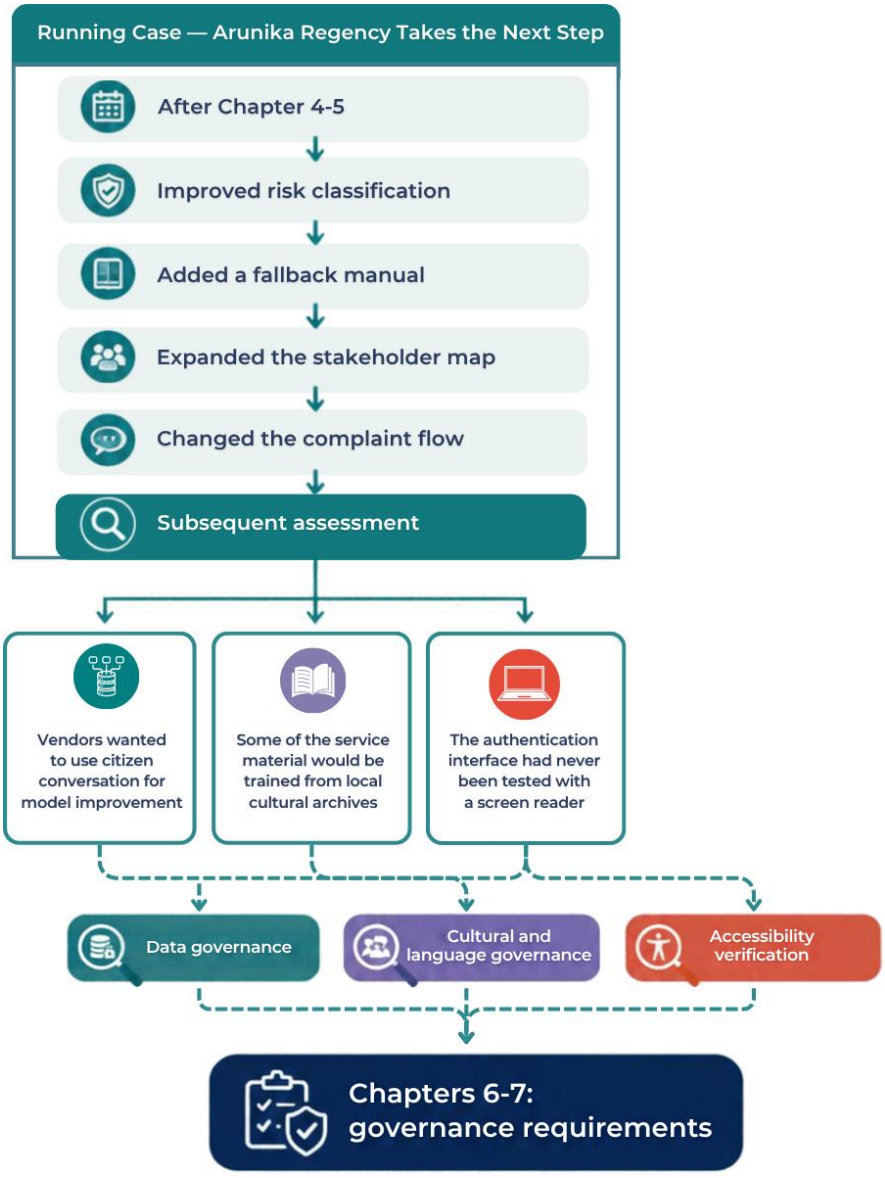
Bridge to Chapters 6–7: Context → Participation → Data Governance → Cultural and Linguistic Validation → Accessibility.

PART II-B
Data, culture, language and accessibility

Running Case — Arunika Regency Takes the Next Step

After Chapter 4-5, the Arunika Regency team has improved the risk classification, added a fallback manual, expanded the stakeholder map, and changed the complaint flow. However, subsequent assessment found that vendors wanted to use citizen conversation for model improvement, some of the service material would be trained from local cultural archives, and the authentication interface had never been tested with a screen reader. Chapters 6-7 show how the toolkit transforms those findings into governance requirements.

Figure 15. Arunika Regency Takes the Next Step



Chapter 6 — Whose Data, Whose Consent, Whose Benefits?

6.1 Privacy alone does not resolve community-data questions

Data governance usually starts with the individual: what data is collected, for what purposes, who can access it, how long it is stored, and how the data is corrected or deleted. D5 maintains the baseline through minimisation, legal basis, retention, security, access control, and user controls.

Interview evidence CCE-01 reinforces the importance of informed consent, purpose specification, access governance, encryption, and continuity of data custody. But interviews and literature Indigenous data governance also show that some data doesn't just have individual meanings. Knowledge, location, practice, story, language, or cultural artefacts can have a collective dimension.

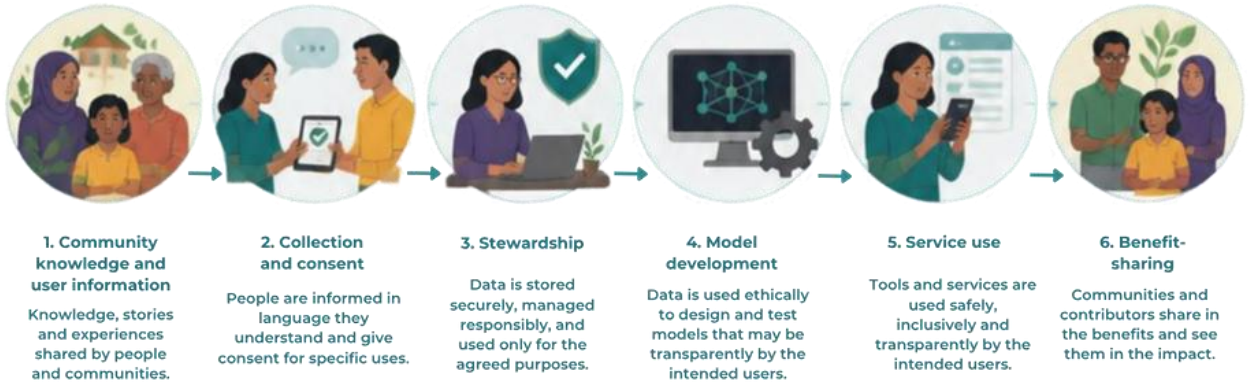
6.2 Individual consent and collective authorisation are distinct questions

The D6 was built to prevent one simple mistake: assuming that an individual's consent automatically entitles them to use knowledge that the community perceives as their own or shared responsibility. The CARE Principles provide conceptual anchor through Collective Benefit, Authority to Control, Responsibility, and Ethics. Engstrom et al. pointed out that Indigenous Data Governance and consent practices vary across contexts; therefore, this model should not be automatically transferred to entire communities in Indonesia.[20]

In project evidence, IP-01 emphasises layered collective consent, restricted cultural data, and intergenerational records. EDU-3T-01 also raises the limitations of individual consent when data contains collective knowledge. The toolkit then separates individual consent from collective/community authorisation when the context does require it.[21]

Key question: who can authorise this use of data—individuals, communities, customary authorities, institutions, or a combination of these?

Figure 16. Data Journey: From Community Knowledge to Shared Benefits



20 Carroll et al., "CARE Principles"; Engstrom et al., "Indigenous Data Governance Approaches."

21 IP-01 (Indigenous Participant), interview transcript, August 2026, Ethical AI Localisation Toolkit research project, esp. 32:11 and 33:51.

6.3 Restricted data: not everything that can be digitised belongs in AI

D6 G2 requests prohibited/restricted-data registers for categories that may cause harm when included in a dataset or published by AI. Interview IP-01 identifies examples such as sensitive locations, cemeteries, sacred practices, conflict areas, trauma information, and data that allows doxxing. The categories should be defined contextually with the parties having legitimacy, not imposed from the outside.

Restricted-data registries are different from just access controls. The question isn't just who can see the data, but whether certain data should be collected, transferred, or used for modelling at all.

Data justice also asks organisations to assess data omissions. When official data does not represent the experiences of a particular group, such absences can be a form of bias and invisibility. However, closing data gaps shouldn't automatically mean collecting more data. Organisations need to assess together with the affected groups whether new collection is needed, who controls it, what benefits result, and whether community-owned or low-tech alternatives are more appropriate (Gentelet et al., 2026).

6.4 Consent must be understandable and capable of being refused

CCE-01 stresses that the public needs to know what data is being collected, for what, and who will access it. Evidence interviews also show that culturally appropriate consent does not always have to be in the form of digital recordings or applications; low-tech documentation may be more appropriate in some contexts. The toolkit therefore requires familiar, understandable, non-coercive consent, and protects the right to refuse.[22]

Local-language materials, in-person deliberation, and non-retaliation statements are more powerful examples of evidence than formally complete but unintelligible consent forms.

6.5 Benefit-sharing involves more than payment

D6 G4 assesses whether the use of community data results in long-term benefits and improves community control/capacity. Interview IP-01 cites infrastructure, capacity building, access, long-term funding, and local stewardship as more sustainable benefits than one-off payments.

This doesn't mean that one benefit-sharing model is universal. What's important is transparency: who gets value from the data, who carries the risk, and how the benefits are negotiated and reviewed.

6.6 Intergenerational governance

The G5 requested documented collective agreements for cross-generational review. This idea stems from the concern that decisions about the use of cultural knowledge may have longer consequences than the project cycle or the term of office of the manager.

Intergenerational record doesn't mean consent never changes. The record needs to contain what was agreed, for what purpose, who agreed, when the review was done, and how the change or withdrawal could be processed.

6.7 Vendors and data sovereignty

Community data governance doesn't stop when data moves to vendors. D13 assesses secondary use, data residency/cross-border processing, vendor lock-in, audit rights, exit plans, knowledge transfer, and fair licensing.

---

22 CCE-01 (Cultural and Community Expert), interview transcript, August 2026, Ethical AI Localisation Toolkit research project, 26:41; see also 27:39 on encryption/access custody and 42:00 on low-tech consent documentation.

The public-sector procurement guidelines that we've included in the evidence base - the OECD, the UK Government, and the World Bank - support the need for contract terms that clarify data/IP rights, auditability, portability, ongoing testing, vendor performance, and exit. In the context of community/language data, the toolkit adds the question of whether communities still have control or fair licensing.

Figure 17. Checkpoints at Every Stage



6.8 Toolkit in Action - D5, D6, D13

Domain	Core question	Evidence	Red flag	Mitigation
D5 Privacy and security	Is the data minimised, legal, secure, and user controllable?	Data inventory; retention; access matrix; threat model; privacy workflow	Sensitive data is collected without obvious need or excessive access	Minimise; restrict; encrypt; define retention and user controls
D6 Community data	Are collective authorisation and restricted-data governance available when relevant?	Collective consent; governance map; restricted-data register; benefit plan	Community data is collected only with individual consent	Stop collection; establish collective process
D13 Vendor/Data sovereignty	Does the contract protect data rights, auditability, portability, exit, and community interests?	DPA; subprocessor list; data-flow map; audit clause; exit plan; licensing terms	Vendors can use secondary data without authorisation	Revise the contract before transfer

6.9 Evidence to Look For

- data inventory and purpose specification;
- retention schedule and access matrix;
- collective/community governance maps where relevant;
- restricted/prohibited-data registers;
- local-language consent materials and record of refusal rights;
- benefit-sharing plans and community stewardship arrangements;
- vendor/subprocessor list and data-flow map;
- audit rights, portability, exit and knowledge-transfer clauses;
- licensing terms for community/language data.

6.10 Red Flags

- community or cultural data enters a model without relevant collective authorisation;
- sacred/sensitive information has no restricted-data rule;
- consent is only formal and not comprehensible;

- the vendor may use the data for secondary purposes without basis/authorisation;
- the location/identity data of marginalised groups is too easily accessible;
- no exit plan or data portability;
- benefit-sharing is one-off and does not give community capacity/control.

6.11 Running Case — Arunika and Local Data

Vendor Arunika proposed using conversation logs for 'continuous improvement'. Data-flow reviews suggest that logs will be sent to cloud environment vendors, with subprocessors not listed in the contract. In addition, the content team wants to incorporate local story archives into the knowledge base so that the chatbot feels more contextual.

The toolkit separates two issues. Conversation logs entered D5/D13: minimisation, purpose, access, secondary use, processor/subprocessor, retention, and user controls. Local story archives enter D6/D7: who has authority, what data is restricted, whether there is collective authorisation, and how attribution/benefit-sharing is regulated.

Case decision: the pilot remains on hold until contract clauses are corrected, data flows are documented, and the community-data governance process is agreed.

6.12 A short checklist

- Is all the data really necessary?
- Is individual consent sufficient for this type of data?
- Is there any knowledge that should be restricted or prohibited?
- Can consent be understood and rejected without unreasonable consequences?
- Who benefits economically, socially, or institutionally from the data?
- What happens to the data when the vendor's contract expires?
- Does the community still have control or fair licensing over cultural/language data?

Chapter 6 summary: Data governance needs to distinguish between individual privacy, collective/community authority, benefit-sharing, and vendor control. Legally processed data is not necessarily socially or culturally legitimate.

Chapter 7 — Does AI Understand Language, Culture, and Its Users?

7.1 Cultural validity is not decoration

AI can look local because it uses Indonesian, visual motifs, or certain terms, but still misunderstands the context. D7 therefore assesses cultural representation, language/dialect performance, community control, and correction authority.

The EDU-3T-01 interview provides an example of cultural hallucinations in AI-generated representations that do not reflect local cultural characteristics. IP-01 adds the need for local-dialect testing and authority to correct misrepresentation. This evidence is used as a contextual lens, not as evidence that all AI produces cultural harm.

Figure 18. local and cultural validity



7.2 Supporting a language does not necessarily mean understanding it

IndoNLG and the Indonesian-Javanese-English code-mixed literature indicate that model performance should not be assumed uniformly in low-resource language or code-switching patterns. For the toolkit, the implications are simple: language support needs to be tested on the language/dialect and task the user actually uses.[23]

D7 L2 requires local benchmarks, error logs, and dialect testing. Benchmarks don't always have to be big. For public service, a small but representative test set of critical service intents is often more useful than a demo that only shows a general conversation.

23 Cahyawijaya et al., “IndoNLG”; Hidayatullah et al., “Pre-Trained Language Model for Code-Mixed Text.”

7.3 Who has the authority to validate cultural appropriateness?

Validation should not depend on one person who is considered to represent the entire community. Evidence interviews show that language offices, elders, customary figures, educators, and community members may hold different authority. D7 L1 therefore requests validators for relevant and documented roles.

The validator doesn't just give the final approval. It needs to have correction authority, escalation path, and SLA so misrepresentation isn't allowed to stay online once it's discovered.

7.4 Preserving language without taking control

D7 L3 assesses whether language preservation is done without transferring ownership/control from the community. Efforts to digitise languages can have great benefits, but corpuses, recordings, annotations, and derived models can also create value for institutions or vendors.

Therefore, language governance needs to be reconnected to D6 and D13: who authorises, who holds the repository, how licensing is organised, who gets access, and whether the community can restrict certain uses.

7.5 Accessibility: compliance does not guarantee a usable experience

D8 treats accessibility as a lived experience across disabilities. WCAG 2.2 addresses visual, auditory, physical, speech, cognitive, language, learning, neurological, and combined needs, while recognising that conformance does not meet every user need. The toolkit therefore calls for testing across modalities, accessible authentication, caption and transcript quality checks, cognitive accessibility, alternative inputs, and direct co-testing with relevant users.[24]

Coverage includes people who are blind or have low vision; Deaf or hard-of-hearing people; people with motor, speech, cognitive, learning, or multiple disabilities; neurodivergent users; and people experiencing age-related changes in ability. Teams should document a reasoned coverage plan based on service modalities, critical journeys, risks, and the user population.

DP-01 provides contextual evidence for specific operational details. The participant emphasises that automated checks are insufficient, complaint channels must themselves be accessible, and high-impact services require human escalation when AI fails.

Lestari, Hafiar, and Priyatna add Indonesian evidence through their evaluation of ministry websites using aXe DevTools against WCAG 2.1. Their automated assessment shows why accessibility cannot be assumed; it does not establish that automated testing alone is sufficient.[25]

7.6 Testing critical user journeys

Accessibility testing should follow the whole service journey. Critical tasks may include logging in, authenticating, completing forms, uploading documents, understanding and resolving errors, reading results, requesting corrections, and reaching human support.

- ☐ cross-disability critical-journey matrix

---

24 World Wide Web Consortium, Web Content Accessibility Guidelines (WCAG) 2.2 (2024), https://www.w3.org/TR/WCAG22/; Sara Gartland et al., "The State of Web Accessibility for People with Cognitive Disabilities," Behavioral Sciences 12, no. 2 (2022): 26, https://doi.org/10.3390/bs12020026; Emre Umucu, "Artificial Intelligence and Health Equity for People with Disabilities," INQUIRY 62 (2025): 1–6, https://doi.org/10.1177/00469580251365472.

25 Lestari, Hafiar, and Priyatna, "Analisis Aksesibilitas Website Kementerian di Indonesia."

- ☐ Keyboard, screen-reader, and alternative-input tests
- ☐ Caption and transcript QA report
- ☐ visual/non-audio notification test
- ☐ Voicebot/IVR alternative-channel test
- ☐ Cognitive walkthrough and plain-language review
- ☐ Accessible-authentication test
- ☐ disability coverage rationale and participant matrix
- ☐ Feedback-to-decision trace and retest evidence

Journey	Test	Evidence	Failure consequence
Login to the site	Keyboard + screen reader + pointer; CAPTCHA/OTP/2FA	Test record + issue log	User cannot enter service
Application	Labels, focus order, validation, document uploads	Manual QA + co-test	Submission failure
Results	Readable status, explanation, visual to text alternative	Screen-reader output + content QA	User cannot understand decision
Complaint	Accessible form/channel + human escalation	Complaint test + SLA	No meaningful remedy

7.7 Accessibility regression is a lifecycle risk

A system that is accessible during a pilot may become inaccessible after library updates, redesigns, authentication changes, or vendor integration. Therefore, DP-01 informs accessibility regression monitoring and RACI lifecycle. D14 then links management updates to retesting.

7.8 Toolkit in Action — D7 and D8

Domain	Core question	Evidence	Red flag	Mitigation
D7 Cultural/Linguistic	Is cultural/language output validated by competent parties and correctable?	Validator record; local benchmark; error log; licensing; correction SLA	Cultural output published without local validation	Pause publication; validate and correct
D8 Accessibility	Is critical journey usable across modalities and has it been tested with relevant users?	Cross-disability matrix; multimodal test; direct co-testing; retest evidence	Critical journey inaccessible for a relevant disability profile	Fix before deployment; provide an equivalent accessible alternative

7.9 Evidence to Look For

- native/local-language validator roles;
- local-language/dialect benchmarks and error logs;
- cultural content approval and correction workflow;
- community-controlled language repository/licensing where relevant;
- manual keyboard + screen-reader test record;
- mouse/pointer + screen-reader co-test;

- the CAPTCHA/OTP/2FA accessibility test;
- direct disabled-user testing record;
- issue ownership, remediation, and retest evidence;
- The accessibility regression release gate.

7.10 Red Flags

- cultural content published without relevant validators;
- models are only tested in formal Indonesian while users use dialect/code-mixing;
- the vendor takes ownership/control over language data without governance agreement;
- critical function failure with screen reader or keyboard;
- CAPTCHA/OTP/2FA requires an unsafe workaround;
- an automated accessibility scanner is treated as the only test;
- disabled users are not involved in critical QA functions.

7.11 Running Case — Two Failures Hidden by Arunika's Demonstration

Arunika's chatbot performs well in the vendor demonstration. Local-language testing, however, finds literal and confusing translations of application-status terms. In accessibility co-testing, the screen reader reads the home page but does not correctly announce CAPTCHA information or errors following document upload.

If the assessment looked only at visual demos, both problems would not appear. D7 produces local-language correction workflow; D8 produces remediation for authentication and error messaging, followed by retest with users.

Case decision: resume the pilot only after critical accessibility issues are resolved and local-language validation is completed for the main service intents.

7.12 A short checklist

- Is the output tested on the user's actual language/dialect?
- Who has the authority to declare cultural representation wrong?
- Are the correction workflow and SLA available?
- Will the language data remain in the agreed governance?
- Is critical journey tested across modalities and with users with a range of relevant needs, not just through automated scans?
- Is authentication accessible without a workaround that lowers security?
- Are disabled users directly involved in testing and retesting?

Chapter 7 summary: real localisation requires language/cultural validation and accessibility that can be proven through task, user, evidence, and correction workflow - not visual resemblance or compliance claims alone.

Closing Part II-B — From Data to User Experience

Chapters 6-7 show that data governance, cultural validity, language performance, and accessibility are not completely separate domains. Community data can be the material of language models; language models can produce cultural misrepresentation; interfaces that are linguistically accurate can still exclude users when authentication or grievance is not accessible.

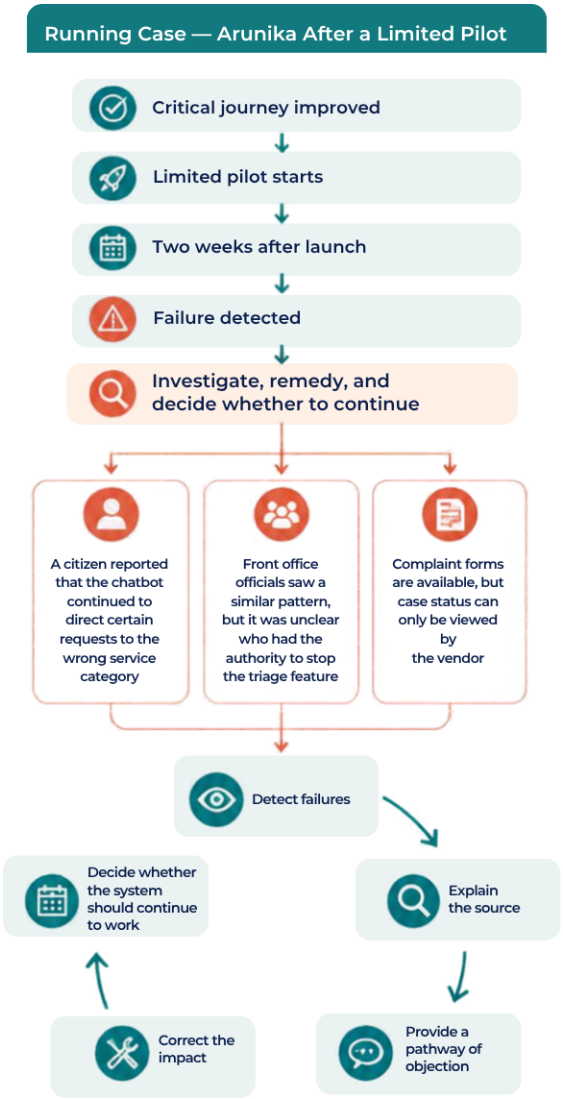
The toolkit links these safeguards through a sequence: who authorises data use, which data may be used, how models are tested, who validates outputs, who can correct errors, and who can access the resulting service.

Bridge to Chapter 8: Data & Community Authority to Cultural/Linguistic Validity to Accessibility to Accountability, Remedy & Human Escalation.

PART II-C
Accountability, Remedy and Continuous Oversight

Running Case — Arunika After a Limited Pilot

Figure 19. Arunika After a Limited Pilot



Arunika Regency has improved fallback, participation, community-data governance, local-language validation, and accessibility on the critical journey. The limited pilot then starts. Two weeks after launch, a citizen reported that the chatbot continued to direct certain requests to the wrong service category. Front office officials saw a similar pattern, but it was unclear who had the authority to stop the triage feature. Complaint forms are available, but case status can only be viewed by the vendor.

The situation illustrates a problem that is often late to be seen: governance is not complete when the system is successfully launched. Organisations should be able to detect failures, explain the source of the problem, provide a pathway of objection, correct the impact, and decide whether the system should continue to work.

Chapter 8 — When AI Gets It Wrong, Who Is Accountable?

8.1 Accountability begins before an incident

Accountability is often seen as the process after something goes wrong: who has to answer, who investigates, and who corrects. In AI governance, accountability should be designed before deployment. D4 assesses who owns the risk, who has the authority, how human oversight works, and how the system is managed throughout the lifecycle.

ASEAN Guide places accountability, human involvement, internal governance, deployment controls, and monitoring as part of responsible AI. UNESCO also places human oversight and responsibility as important elements throughout the lifecycle. The toolkit operates these principles into auditable evidence: RACI, override authority, escalation path, incident log, audit trail, and stop criteria.

Chapter 8 principle: accountability requires knowing who must act before an incident escalates, as well as who must answer for it afterwards.

8.2 Useful transparency is transparency people can act on

D3 treats transparency as purpose-specific. Users need notice of AI involvement, understandable explanations, and routes to question or challenge results. Internal reviewers need provenance, limitations, model and vendor information, and an evidence trail. Auditors and regulators may require more technical documentation. The aim is useful disclosure, not disclosure of every technical detail to everyone.

Therefore, useful transparency must answer three questions: what happens, on what basis, and what users can do if the results go wrong.

Figure 20. Useful transparency lets people act



8.3 Source traceability when the model and its sources disagree

Source traceability is becoming increasingly important for generative and retrieval-augmented systems. Tan et al. showed that some LLMs in their experiments can choose the wrong generated context over the right retrieved context when both are conflicting. Feng et al. also showed how biases in data pretraining can creep into models and downstream tasks.[26]

The implication of governance is not that every output should display the entire chain-of-thought or data training. What is needed is provenance relevant to the use case: what resources are used, when they are updated, how much trust or limitation is known, and how inter-source conflicts are handled.

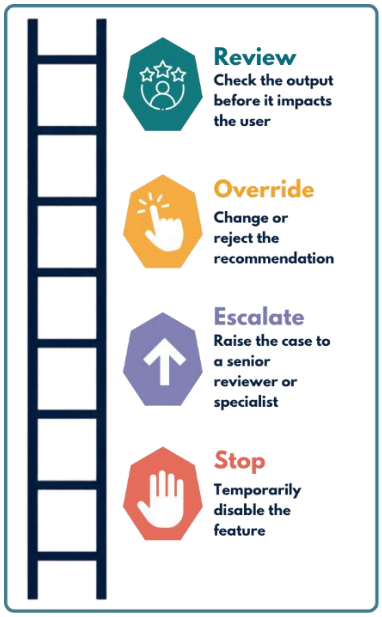
Safe rule For high-impact use cases, organisations need to be able to trace evidence that underlies output or decision - not just trust the model's answers because they sound convincing.

8.4 Human-in-the-loop is not the same as human authority

Alon-Barkat and Busuioc show that human presence in public sector decisions does not automatically eliminate risk. They found patterns of selective adherence to certain conditions, so that the human reviewer himself could be influenced by how algorithmic advice is presented.[27]

Therefore, human oversight must be judged through authority, competence, information, and process. Officers need to know when AI output should be ignored, how disagreements are recorded, when cases should be escalated, and who has stop authority. If the human reviewer simply presses 'approve' without sufficient time or evidence, the oversight is formal, not substantive.

Figure 21. Intervention ladder



8.5 Four levels of human intervention

Level	Human role	Example	Risk when weak
Review	Checking the output before impacting the user	Officer checks the application classification	Rubber-stamping
Override	Changing/rejecting the AI recommendation	Officers correcting service categories	Authority is unclear.
Escalate	Raising the case to a senior reviewer/specialist	High impact case transferred to supervisor	Cases stall at the frontline
Stop	Temporary disabling of system features	Automated triage is suspended after a pattern of errors emerges	No emergency authority

26 Tan et al., “Blinded by Generated Contexts”; Feng et al., “From Pretraining Data.”

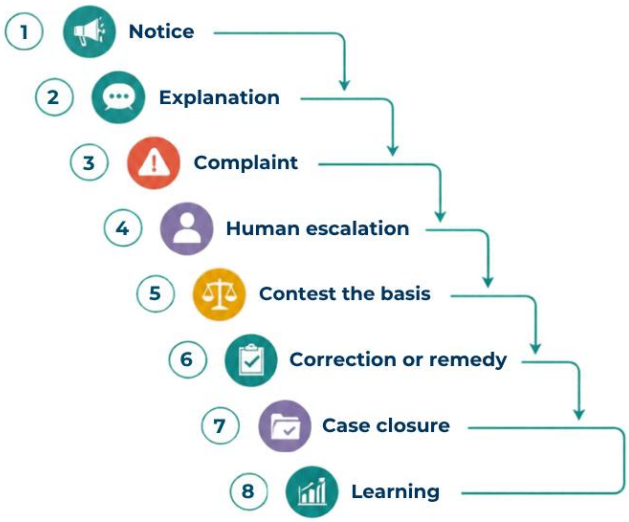
27 Alon-Barkat and Busuioc, “Human–AI Interactions in Public Sector Decision Making.”

8.6 Contestability: users need to challenge decisions, not just complain

Complaint and contestability are not identical. Complaint channels can only record dissatisfaction without changing the decision. Contestability means that the user has a way to question the basis of the decision, submit evidence or grounds for challenge, reach an authorised reviewer, and obtain correction or remedy if the objection is valid.

Leofante et al. framed contestable AI as a dynamic process that should be able to accept grounds for contestation and revise decisions/processes. Landau et al. put contestability as a design requirement on government AI systems. Alfrink et al. suggested that a public AI that is contestable should be responsive to dispute and connected to institutional capacity.[28]

Figure 22. Operational synthesis Notice to Explanation to Complaint to Human Escalation to Contestation to Correction/ Remedy to Case Closure to Learning



8.7 Grievance pathways must extend beyond AI

D12 Q1-Q4 assesses whether the complaint channel is available online and offline, accessible, has human escalation, SLA, correction/takedown/remedy, and traceability to resolution. A very simple but important principle is: don't use AI as the only way to complain about AI failure.[29]

DP-01 reinforces the requirement that the grievance itself should be accessible to assistive-technology users. IP-01 and EDU-3T-01 also add to the need for offline paths, especially when connectivity or conflict makes digital channels unreliable.

8.8 Remedy requires more than an apology

Remedy must be adjusted to harm. In some cases, data or status corrections are sufficient. In other cases, remedy may require re-processing, human review, reinstatement of access, takedown, notification to the affected party, or change of policy/workflow so that the error does not recur.

The toolkit doesn't set a single universal remedy. What is assessed is whether the organisation has a mechanism to identify harm, determine corrective action, respond to users, and document closure.

---

28 Francesco Leofante et al., "Contestable AI Needs Computational Argumentation," in Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning (2024), 888–896, https://doi.org/10.24963/kr.2024/83; Susan Landau et al., "Challenging the Machine: Contestability in Government AI Systems," arXiv:2406.10430 (2024); K. Alfrink et al., "Contestable Camera Cars," arXiv:2302.04603 (2023).

29 DP-01, interview result, Q6; IP-01 and EDU-3T-01, interview evidence used contextually for offline access and escalation scenarios.

8.9 Vendor accountability: organisations cannot outsource responsibility

When a model, cloud, monitoring tool, or complaint backend is managed by a vendor, the organisation still needs visibility. D13 and D14 therefore need to be read together: contracts should provide auditability, data access, incident reporting, change notification, performance expectations, and exit rights.

The OECD emphasises lifecycle procurement controls, vendor performance, ongoing monitoring, lock-in protection, and access to data/derived products. The UK Government and World Bank guidance also support open standards, portability, audit rights, knowledge transfer, and exit planning. In this book, the evidence is used as a governance baseline for public-sector procurement.[30]

8.10 Monitoring: what to track after launch

D14 shifts attention from whether a system passed pre-launch checks to whether the organisation can detect subsequent changes. Monitoring should cover technical drift and changes in social or service outcomes.

Monitoring area	Example of a signal	Trigger	Action
Performance	Error rates increase for specific service intents	Threshold is exceeded	Investigate + retrain/restrict
Fairness	The gap between groups grows	Disaggregated metric alert	Bias review + mitigation
Accessibility	Regression after library/auth update	Critical journey test failed	Block release / hotfix
Grievance	Repeated complaints on the same category	Complaint cluster	Root-cause review
Vendor/model	Model/version changed	Change notification	Reassessment
Source and content	Knowledge base outdated or conflicting	Stale source flag	Update/validate sources
Operational	Human override is rising sharply	Override pattern	Review of workflow

8.11 Monitoring without action only produces a dashboard

Monitoring becomes governance only when the metric has an owner, threshold, response, and escalation path. A dashboard that shows a red indicator but no one is required to act only gives a false sense of control.

Therefore, D14 needs to be connected to the action plan. Each critical metric should have: owner, trigger, maximum response time, evidence of action, and condition for reassessment or pause.

8.12 Incident management and learning from root causes

AI incidents can come from data, models, interfaces, vendors, sources, human processes, or a combination of several factors. Incident review shouldn't stop at the wrong model. A more useful question is why the error reaches the user, which safeguards fail, why monitoring doesn't catch it early, and what system changes are needed.

30 OECD, Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions (Paris: OECD Publishing, 2025); UK Government, Guidelines for AI Procurement; World Bank, Artificial Intelligence in the Public Sector: Maximizing Opportunities, Managing Risks.

- What's the actual impact and who's affected?
- What immediate containment is required?
- Should the output/feature be temporarily stopped?
- Are there any data/source/model/version changes?
- Is human oversight working or just formal?
- Did the complaint channel find the pattern before?
- What is corrective action and prevention action?
- When is a full reassessment needed?

8.13 Toolkit in Action - D3, D4, D12, D13, D14

Domain	Core question	Evidence	Red flag	Mitigation
D3 Transparency	Can the user/reviewer understand the relevant AI involvement, source, and limitation?	Disclosure; source log; explanation; provenance	High-impact output not traceable	Restrict use; add traceability and source validation
D4 Accountability	Does the human reviewer have authority, competence, override/escalation/stop power?	RACI; override log; escalation matrix	No human override or stop authority	Pause high-impact use until governance exists
D12 grievance	Can users complain, reach out to humans, and get correction/remedy?	Complaint channels; SLA; case tracking; resolution evidence	Complaint only bot-only or inaccessible	Add accessible human path + remedy workflow
D13 Vendor	Does the organisation have audit/data/change/exit visibility?	Contract; incident clause; audit rights; version notice	Vendor refuses visibility on incident-critical data	Renegotiate / restrict deployment
D14 Monitoring	Do drift, complaints, accessibility, vendor changes trigger action?	Dashboard; alert thresholds; issue log; reassessment records	Critical signals have no owner/action	Define thresholds, owners and pause rules

8.14 Evidence to Look For

- AI disclosure and explanation templates for users;
- source/provenance logs for high-impact outputs;
- RACI and named risk owner;
- the human override, escalation, and stop-authority matrix;
- override/disagreement logs
- accessible online and offline complaint channels;
- case-tracking, SLA, correction/remedy record;
- vendor incident-notification and audit clauses;
- the model/version/change log;

- monitoring thresholds, owners, and response runbooks;
- incident review and reassessment records.

8.15 Red Flags

- high-impact decision has no human override;
- users are unaware that AI is involved in decisions affecting them;
- source/provenance cannot be traced in a critical use case;
- the complaint channel directs users back to the bot without human option;
- critical accessibility grievance cannot be filed in an accessible manner;
- the vendor holds incident data but the organisation does not have access/audit rights;
- the model changes without reassessment;
- monitoring finds an error pattern but no owner or stop rule.

8.16 Running Case — Arunika Chooses a Targeted Pause

In the Arunika pilot, the complaint cluster showed that the AI triage misdirected some kind of request. Vendors can see the full log, but the management team only accepts the aggregate dashboard. Frontline staff have the ability to correct one case, but they don't have the authority to stop the triage feature.

The toolkit produces three actions. First, the triage feature is temporarily discontinued but the general information function is still running. Second, vendors are required to provide case-level evidence and source/version information for root-cause review. Third, human stop authority and incident escalation are included in the SOP as well as the operational contract.

Root-cause analysis found that changes to the vendor's knowledge-base mapping did not trigger the reassessment. Once the mapping is fixed, the affected case is reviewed by the officer, the user receives a correction notice, and the new pilot is reopened after targeted retest.

A good governance case study doesn't always mean shutting down the whole system. It means being able to limit function, repair harm, and reopen only when evidence shows the safeguards are working.

8.17 A short reader checklist

- Does the user know when the AI is involved?
- Can high-impact output be traced to relevant sources/evidence?
- Do human reviewers have authority to override and stop?
- Can the complaint channel reach humans?
- Is remedy more than just acknowledgment?
- Does the organisation have audit/change visibility to the vendor?
- Do the monitoring metrics have an owner and response threshold?
- Is the model/vendor/source change triggering the reassessment?
- Did the incident result in learning and prevention action?

Chapter 8 summary: Accountability is not a single person or a single document. It is a set of capabilities: usable transparency, human authority, contestability, remedy, vendor visibility, monitoring, incident response, and reassessment.

Closing Part II — From Principles to Testable Safeguards

Chapter 4-8 transforms ethical AI from a set of values into an implementation question. Context and readiness determine the risk. Fairness and participation determine who benefits or is left behind. Data governance determines who has authority. Cultural-linguistic validation and accessibility determine whether the system actually works for the user. Accountability and remedies determine what happens when the system fails.

Part III will turn the whole question into an evaluation process that can be run: how to evaluate 15 domains and 62 indicators, how to evaluate evidence, how to read red flags and scores, how to formulate mitigation/action plans, and how to turn assessment into deployment decisions.

Bridge to Section III: Assess to Verify Evidence to Escalate Red Flags to Mitigate to Decide to Monitor to Reassess.

PART III
Using the Toolkit for Real Decision Making

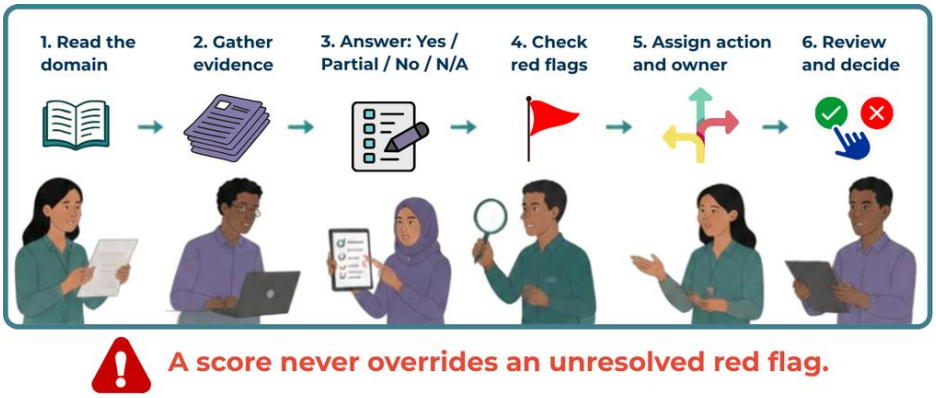
Chapter 9 — Using the 15 Domains and 62 Indicators

Methodological note: the domain structure, indicator set, red-flag rules, scoring formula, and decision bands are design choices in this toolkit. They have not been validated as population-level scientific cut-offs.

9.1 The toolkit is not a multiple-choice test

The four answer options - Yes, Partial, No, and Not Applicable - seem simple. But good assessment doesn't stop at choosing answers. Every answer must be read with evidence. Thus, the real question is not just 'have we done this?' but 'What evidence can other reviewers examine?'

Figure 23. Using the toolkit



The rule of thumb is yes, without evidence is not a strong yes. Evidence is part of the answer, not an attachment after the assessment is complete.

9.2 Who should complete the assessment?

A toolkit isn't ideal when it's filled by one person. Many indicators require knowledge spread across a range of functions. The minimum team should include service owner, technical/vendor representative, privacy/security, procurement, frontline operations, and parties who understand affected users.

For use cases involving local languages, community data, disabilities, or marginalised groups, the reviewer also needs to involve validators or users with relevant lived experience. The goal is not to amplify the meeting, but to make sure that the people with the evidence and the people affected are present.

9.3 Four response options

Response	Meaning	Minimum evidence	Common Mistakes
Yes	Requirements are met and evidence is available.	Document/test/log/decision record that can be verified.	Answer yes based on team beliefs.
Partial	The requirement is partly met or supporting evidence is incomplete.	Evidence of what is in place, the remaining gap, and the action owner.	Considering Partial as a safe condition without mitigation.
No	Requirements are not met.	Note the gap + risk/impact + mitigation decision.	Empty the answer so the score doesn't fall.
N/A	Requirement is irrelevant after a documented reason.	A documented rationale explaining why the requirement does not apply.	Using N/A to avoid difficult questions.

9.4 When can N/A be used?

N/A is removed from the score denominator in the toolkit. Therefore, its use must be conservative. N/A means that the requirement does not apply to use cases, not that the team has not yet had data or has not had time to check.

- N/A makes sense when the system never processes personal data for a particular privacy indicator.

- N/A may be relevant to the collective-consent indicator when there is no community/cultural data or collective knowledge.

- N/A is not appropriate if the organisation doesn't know if community data is being used.

- N/A should not be used just because the vendor hasn't provided evidence.

9.5 The structure of the 15 domains

Domain	Theme	Indicators	Red flags	Questions of governance
D1	Context, Purpose, Human Rights and Risk	4	2	Are use cases and risks properly understood?
D2	Fairness and Non-Discrimination	4	1	Are benefits, errors, and access burdens distributed fairly?
D3	Transparency, Source & Explainability	4	1	Is AI involvement, source, and limitation traceable?
D4	Accountability, Human Oversight & Lifecycle	4	3	Who has the authority to review, override, escalate, and stop?
D5	Privacy, security and personal data	4	2	Is personal data processed proportionately and securely?
D6	Collective consent, community data & benefit-sharing	5	2	Does community data have proper collective governance?
D7	Cultural and linguistic validation	4	1	Are languages and cultural representations validated?
D8	Accessibility and Assistive Technology	5	4	Is the critical journey really accessible?
D9	Meaningful Participation & Intersectional Inclusion	5	2	Does participation affect decisions?
D10	3T, Local Readiness & Continuity	4	2	Have the local conditions and the fallback been verified?
D11	AI literacy, education and human readiness	4	0	Are humans ready to use, monitor, and replace AI if necessary?
D12	Complaints, Redress and Human Escalation	4	2	Can users complain, challenge, and get remedies?

Domain	Theme	Indicators	Red flags	Questions of governance
D13	Vendors, Procurement and Data Sovereignty	4	1	Is the vendor/data rights/audit/exit strong enough?
D14	Monitoring, Audit and Continuous Improvement	4	0	Does monitoring trigger action and reassessment?
D15	Sustainability and Environmental Impact	3	0	Are environmental/resource impacts considered throughout the lifecycle?

9.6 Red flags are not ordinary indicators

Of the toolkit's 62 indicators, 23 are designated as red flags. A No or Partial response to one of these indicators means that a critical safeguard remains unresolved. Under the toolkit's rules, the resulting decision is Do Not Deploy, regardless of the aggregate score.

This is a design rule toolkit, not a scientific cut-off that has been validated on AI project populations. The rule is deliberately conservative so that the critical safeguard is not overshadowed by many yes answers to less critical indicators.

For example, if a system achieves a score of 88% but a high-impact decision has no human override, aggregate score should not be used to justify deployment.

9.7 How to interpret evidence

Strong evidence usually has four characteristics:

- Specific - directly answer the requirement, not the general document.
- Current - according to model, vendor, version, region, and process being evaluated.
- Traceable - have an owner, date, version, or case reference.
- Testable - other reviewers may check or repeat the test.

9.8 A practical evidence hierarchy

Level	Example	Strength	Notes
A - Direct operational evidence	Test result, signed decision, contract clause, incident log, field verification	High	Showing real practice.
B - Process evidence	SOP, policy, training record, governance map	Medium–high	It has to be related to implementation.
C - Stated assurance	Vendor statement, self-declaration, slide deck	Low–medium	It needs corroboration.
D - Assumption	"We are confident that the system is safe."	Low	Not enough for Yes.

9.9 An evidence gap does not always mean system failure

A requirement may be met in practice while its evidence has not been retained. This does not justify a full Yes. The distinction informs mitigation: a system gap may require redesign, while an evidence gap may require retesting, documentation, or restoration of the audit trail.

9.10 How to work through each domain

- Read all the domain indicators before answering one by one.
- Identify the evidence owner for each indicator.
- Separate policy evidence from implementation evidence.
- Mark the red flags first.
- Record uncertainty and conflicting evidence.
- Make an action for every Partial/No.
- Do a second domain review by someone who doesn't produce key evidence when a decision has a big impact.

9.11 A short walkthrough of three indicators

Indicators	Response	Evidence	Interpretation	Action
D10 R2 - low-bandwidth/offline/manual fallback	Partial	The manual SOP is available, not tested.	Process exists but readiness unverified	Run the outage drill and document the SLA
D9 M2 - community input changes decisions	Yes	Change log + revised requirements	Meaningful influence can be traced	Monitor the next phase
D8 X3 - CAPTCHA/OTP/2FA is accessible	No	Co-test failed on the CAPTCHA	Critical access barrier	Fix before pilot; retest with users

9.12 Quality checks before calculating results

- Is supporting evidence available for every red-flag indicator?
- Does every N/A response have a documented rationale?
- Does every Partial or No response have an action owner?
- Does the evidence match the system version being assessed?
- Does the review include affected users' evidence as well as vendor evidence?
- Is there a disagreement between the reviewers? If so, is it documented?
- Will the decision be reviewed by the accountable owner?

Chapter 9 summary: assessment is not a race to answer 62 questions. It builds an evidence trail sufficient to support decisions, explain their reasons, and repeat the review when circumstances change.

Chapter 10 — From Assessment to Deployment Decisions

10.1 A score is a summary, not a verdict

The toolkit assigns Yes = 2, Partial = 1, and No = 0, excluding N/A responses from the denominator. The percentage score is the total points earned divided by the maximum applicable points, multiplied by 100. Four decision bands guide interpretation.

Assessment condition	Decision labels	Practical Meaning
Unresolved red flag or score < 50%	Do not deploy	Critical safeguards have not been met or readiness is very low.
50-69% without an unresolved red flag	Pause and Mitigate	Significant gaps; make repairs before deployment.
70-84% without an unresolved red flag	Proceed with Conditions	You can proceed on a limited basis with terms, owners, and deadlines.
≥85% without unresolved red flag	Proceed	Readiness is relatively strong, remaining subject to monitoring/reassessment.

Methodological caution: these decision bands are toolkit design thresholds, not scientifically validated safety limits or proof that a system is ethical. Organisations may apply stricter requirements based on law, risk appetite, or sector needs. Under the toolkit, an unresolved red flag remains a deployment barrier regardless of the aggregate score.

10.2 Why the domain profile matters more than a single number

Two projects can both get 80% but they have very different profiles. One may be weak in sustainability and documentation; the other weak in accessibility and grievance. The consequences of governance are not the same.

Therefore, the decision should read three layers: red flags, domain profile, and aggregate score - in that order.

Read order of results: 1) Red flags to 2) Critical domain gaps to 3) Aggregate score to 4) Mitigation feasibility to 5) Deployment decision.

10.3 Full Simulation — Arunika Regency's Initial Assessment

To demonstrate how the toolkit works, the Arunika case was simulated after the discovery but before all mitigation was completed. The figures below are not the results of field research; they were made to explain the assessment mechanism.

Domain	Early assessment	Red flag?	Major mitigation	Reassessment
D1 - Context, purpose, human rights and risks	75%	—	Closing evidence gaps and conditions	100%
D2 - Fairness and Non-Discrimination	75%	—	Closing evidence gaps and conditions	88%
D3 - Transparency, Source and Explainability	75%	—	Closing evidence gaps and conditions	88%

Domain	Early assessment	Red flag?	Major mitigation	Reassessment
D4 - Accountability, Human Oversight & Lifecycle	63%	Human stop authority is unclear.	Set the override/escalation/stop authority	88%
D5 - Privacy, security and personal data	75%	—	Closing evidence gaps and conditions	88%
D6 - Collective consent, community data and benefit-sharing	60%	Collective authorisation has not been completed	Complete community governance + restricted-data register	80%
D7 - Cultural and linguistic validation	63%	—	Closing evidence gaps and conditions	88%
D8 - Accessibility and Assistive Technology	60%	CAPTCHA failed the co-test	Repair the auth flow + direct retest	90%
D9 - Meaningful participation and intersectional inclusion	70%	—	Closing evidence gaps and conditions	90%
D10 - 3T, Local Readiness & Continuity	63%	—	Closing evidence gaps and conditions	88%
D11 - AI literacy, education and human readiness	75%	—	Closing evidence gaps and conditions	88%
D12 - Complaints, Redress and Human Escalation	63%	Bot-only escalation on one flow	Added accessible human escalation + SLA	88%
D13 - Vendors, Procurement and Data Sovereignty	63%	Vendor incident-data access clause is weak	Change the contract clause to incident/audit visibility	88%
D14 - Monitoring, Audit and Continuous Improvement	75%	—	Closing evidence gaps and conditions	88%
D15 - Sustainability and environmental impact	67%	—	Closing evidence gaps and conditions	83%

Illustratively, domain averages in the early stages are about 68%. But five unresolved red flags mean the toolkit's decision remains Do Not Deploy. This shows why scores shouldn't be read separately from the gating criteria.

10.4 Turning the assessment into an action plan

Any relevant Partial/No needs to be converted into a closed action. A good action plan isn't about 'improving accessibility', but about the owner, the deliverable, the deadline, the closure evidence, and who accepts the residual risk.

Gap	Action	Owner	Deadline	Closure evidence	Gate
D8 X3 authentication inaccessible	Redesign CAPTCHA/OTP path and co-test again	Digital service + vendor	Two weeks.	Retest report, 0 critical barrier	Before the pilot
D12 human escalation is weak	Add direct human route + SLA	Service operations	Two weeks.	Complaint test + SLA log	Before the pilot

Gap	Action	Owner	Deadline	Closure evidence	Gate
D13 incident visibility	Amendment to the vendor contract	Procurement/legal	Three weeks	Signed clause + data access test	Before the expansion
D4 stop authority	Approve the incident-response and stop-authority RACI	Service owner	One week.	Signed RACI + drill	Before the pilot
D6 community authorisation	Complete governance process	Programme lead + community liaison	Four weeks	Authorisation record + restricted register	Before cultural-data use

10.5 A targeted pause is more useful than binary thinking

Deployment decisions may be made for clearly separable functions. In Arunika, general information can remain available where safeguards are verified, while triage or cultural-data functions are held back pending mitigation.

This approach is acceptable only when separation can be enforced, users cannot be routed into suspended functions, and monitoring verifies those boundaries.

10.6 Reassessment after mitigation

After the five red flags were resolved, Arunika's simulated reassessment produced a domain average of about 88%, placing it in the Proceed band. Staged deployment is nevertheless appropriate: a high score does not eliminate uncertainty in a new use case, and the accountable organisation may impose stricter conditions.

If an organisation wants to be more conservative, management decisions can remain 'Proceed with Conditions' even though the band toolkit shows Proceed. The toolkit provides decision support; the accountable owner remains responsible for final decisions and stricter conditions.

10.7 Four types of deployment decision

Decision	When to use	Minimum conditions	Example
Proceed	Evidence is strong and there's no unresolved red flag	Monitoring, ownership, reassessment triggers	Information service with stable safeguards
Proceed with Conditions	The non-critical gap still exists.	Named conditions + deadline + owner	Pilot limited by strict monitoring
Pause and Mitigate	The gap is significant but repairable	Action plan prior to deployment	Fallback hasn't been tested; training hasn't been completed.
Do not deploy	Critical safeguards unresolved or readiness very low	No deployment until the gate is closed.	No human override; inaccessible critical function

10.8 Conditions must be auditable

Proceed with Conditions can easily be an empty decision if the conditions are not connected to the deadline or the evidence. Each condition should mention what needs to be completed, who the owner is, when due, evidence closure, and what the consequences of being late are.

- Condition without owner = hope.
- Condition without deadline is a permanent suspension.

- Condition without closure evidence = difficult to audit.
- Condition without consequence = not really a gate.

10.9 Reassessment triggers

The assessment should be reopened when there are changes that could alter the risk profile:

- the model or major version changes;
- vendor/subprocessor/cloud location changes;
- expanded use cases or AI starting to influence new decisions;
- the data set or source base changes materially;
- critical incident or complaint cluster appears;
- accessibility regression was found;
- policy/regulation changes;
- expanded deployment to new regions, languages or user groups;
- the monitoring shows the performance/fairness drift.

10.10 Assessment frequency

Not all indicators should be checked at the same frequency. Procurement clauses may be reviewed at renewal/change events; accessibility needs regression testing at release; complaints may be monitored continuously; community agreements may have their own review cycle.

Trigger/cadence	Focus	Examples of evidence
Pre-design / procurement	Context, participation, data and vendor governance	Impact map, requirements, contract
Pre-pilot	Readiness, language, accessibility, oversight	Field test, benchmark, co-test, RACI
Pre-deployment gate	All red flags + unresolved actions	Closure evidence + decision record
Continuous / monthly	Performance, complaints, drift	Dashboard, complaint log
Per release/change	Accessibility, source/model/vendor changes	Regression test + change log
Periodic strategic review	Community agreements, sustainability, long-term benefits	Review record + updated conditions

10.11 How to maintain a decision record

The deployment decision should leave a short but complete record:

- the date and version of the system assessed;
- scope/use case and deployment area;
- score + domain profile;
- red flags and closure status;
- unresolved non-critical gaps;
- conditions, owners, deadlines;
- residual risk statement;

- accountable decision-maker;
- reassessment triggers and next review date.

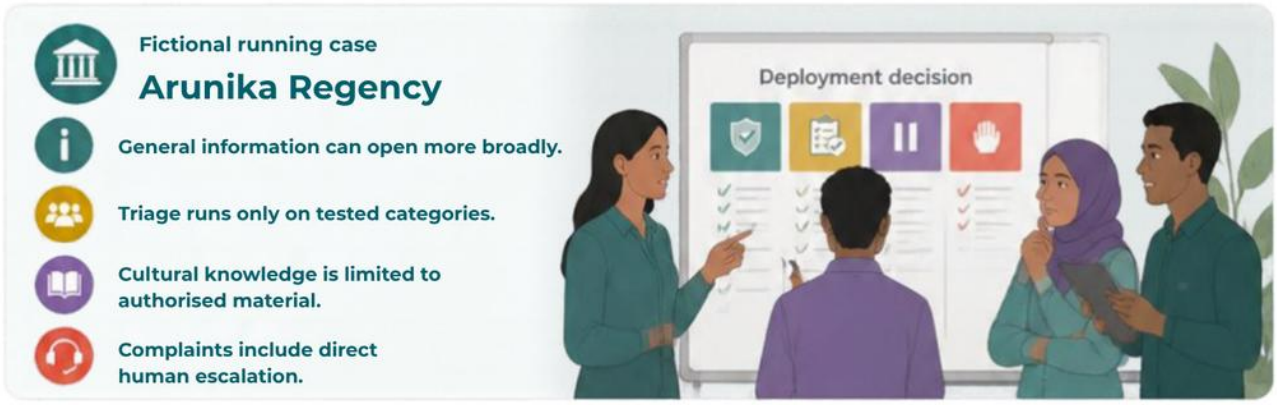
Decision rule: the score summarises evidence. The team exercises judgement. The accountable owner decides.

10.12 Concluding the Arunika Running Case

After the mitigation cycle, Arunika decided on a staged deployment. The general information function is opened more broadly; triage only runs on tested categories; cultural knowledge feature is limited to authorised material; and complaint flow provides direct human escalation.

Monitoring showed the first two months were stable, but one vendor authentication update triggered regression accessibility. The release was stopped before full rollout because D14 monitoring is connected to the D8 regression gate. This case captures one key message of the book: responsible AI is not a final state of affairs. It's the organisation's ability to detect change, respond to evidence, and repeat decisions.

Figure 24. Concluding the Arunika Running Case



10.13 A reader's decision checklist

- Are all the red flags closed?
- Is the score read with the domain profile?
- Does every Partial/No have owner and closure evidence?
- Do conditions have a deadline and a consequence?
- Did the accountable owner sign the decision record?
- Can deployments be limited per function when necessary?
- Have monitoring and reassessment triggers been set?
- Does the user still have a human grievance/remedy path after launch?

Chapter 10 summary: The toolkit does not answer 'Is AI ethical?' with one number. It helps organisations make understandable decisions: what is known, what is not finished, what is not safeguarded, what action to take, and when the system can be continued, restricted, delayed, or discontinued.

Closing Part III — The Toolkit as a Decision-Making System

Chapters 9-10 show that the value of the toolkit is not in the 62 checkboxes, but in the inter-component relationships: requirement, evidence, red flag, mitigation, ownership, decision, monitoring, and reassessment.

With that approach, ethical AI localisation becomes a repeatable process of governance. Organisations don't have to wait for a perfect system; they need to know the gaps that are still open, limit the risk, and have the discipline not to proceed when critical evidence is not yet available.

Final model: Principle → Context → Evidence → Assessment → Red Flag → Mitigation → Decision → Monitoring → Reassessment.

Methodological Notes for Chapters 9–10

The scoring system, decision bands, red-flag logic, and Arunika simulation in Chapters 9–10 are decision-support designs. They have not been statistically validated against a population of AI projects. This qualification applies to all uses of the toolkit.

Bibliography

Alfrink, K., I. Keller, N. Doorn, and G. Kortuem. 2023. “Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute.” arXiv:2302.04603.

Alon-Barkat, Saar, and Madalina Busuioc. 2023. “Human–AI Interactions in Public Sector Decision Making: ‘Automation Bias’ and ‘Selective Adherence’ to Algorithmic Advice.” Journal of Public Administration Research and Theory 33 (1): 153–169. https://doi.org/10.1093/jopart/muac007.

ASEAN. 2024. ASEAN Guide on AI Governance and Ethics. Association of Southeast Asian Nations.

Badan Pusat Statistik. 2025. Indeks Pembangunan Teknologi Informasi dan Komunikasi 2024. Jakarta: BPS.

Badan Pusat Statistik. 2025. Statistik Telekomunikasi Indonesia 2024. Jakarta: BPS.

Bergman, Stevie, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, William Isaac, et al. 2024. “STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment.” Scientific Reports 14: 6616. https://doi.org/10.1038/s41598-024-56648-4.

Bolón-Canedo, Verónica, Leticia Morán-Fernández, Beatriz Cancela, and Amparo Alonso-Betanzos. 2024. “A Review of Green Artificial Intelligence: Towards a More Sustainable Future.” Neurocomputing 599: 128096. https://doi.org/10.1016/j.neucom.2024.128096.

Cahyawijaya, Samuel, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, et al. 2021. “IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation.” In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 8875–8898. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.699.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. “The CARE Principles for Indigenous Data Governance.” Data Science Journal 19: 43. https://doi.org/10.5334/dsj-2020-043.

El Morr, Christo, et al. 2024. “Exploring the Intersection of AI and Inclusive Design for People with Disabilities.” Studies in Health Technology and Informatics. https://doi.org/10.3233/SHTI240475.

Engstrom, Teyl, Elton H. Lobo, Kristie Watego, Carmel Nelson, Jinxiang Wang, Howard Wong, Sungkyung Linda Kim, et al. 2024. “Indigenous Data Governance Approaches Applied in Research Using Routinely Collected Health Data: A Scoping Review.” npj Digital Medicine 7: 68. https://doi.org/10.1038/s41746-024-01070-3.

Feng, Shangbin, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. “From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models.” In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.

Gartland, Sara, et al. 2022. “The State of Web Accessibility for People with Cognitive Disabilities: A Rapid Evidence Assessment.” Behavioral Sciences 12 (2): 26. https://doi.org/10.3390/bs12020026.

Gentelet, Karine, Lily-Cannelle Mathieu, and Alexandra Bahary-Dionne. 2026. “Social Justice Through Digital Empowerment: Insights from Digital Justice Initiatives by and for Marginalized Groups and Communities.” Canadian Journal of Communication 51 (2): 261–287. https://doi.org/10.3138/cjc-2024-0101.

Hadi, Aulia. 2018. “Bridging Indonesia’s Digital Divide: Rural-Urban Linkages?” Jurnal Ilmu Sosial dan Ilmu Politik 22 (1): 17–33. https://doi.org/10.22146/jsp.31835.

Hidayatullah, Ahmad Fathan, et al. 2025. “Pre-Trained Language Model for Code-Mixed Text in Indonesian, Javanese, and English Using Transformer.” Social Network Analysis and Mining 15: 30. https://doi.org/10.1007/s13278-025-01444-9.

Hughes, Sarah E., et al. 2026. “Assessing the Role of Medical Caption Technology to Support Physician-Patient Communication for Patients With Hearing Loss: Mixed Methods Pilot Study.” JMIR Rehabilitation and Assistive Technologies 13: e79073. https://doi.org/10.2196/79073.

James, Tyler G., et al. 2022. “‘They’re Not Willing to Accommodate Deaf Patients’: Communication Experiences of Deaf American Sign Language Users in the Emergency Department.” Qualitative Health Research 32 (1): 48–63. https://doi.org/10.1177/10497323211046238.

Jiang, Peng, Christian Sonne, Wangliang Li, Fengqi You, and Siming You. 2024. “Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots.” Engineering 40: 202–210. https://doi.org/10.1016/j.eng.2024.04.002.

Landau, Susan, James X. Dempsey, Ece Kamar, Steven M. Bellovin, and Robert Pool. 2024. “Challenging the Machine: Contestability in Government AI Systems.” arXiv:2406.10430.

Leofante, Francesco, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, et al. 2024. “Contestable AI Needs Computational Argumentation.” In Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, 888–896. https://doi.org/10.24963/kr.2024/83.

Lestari, Dini Tri, Hanny Hafiar, and Centurion Chandratama Priyatna. 2024. “Analisis Aksesibilitas Website Kementerian di Indonesia Sebagai Implementasi dari E-Government.” Journal of Digital Communication Science 2 (1): 16–26. https://doi.org/10.56956/jdcs.v2i1.274.

Maun, Rachael, Marc Fabri, and Pip Trevorrow. 2024. “Participatory Methods to Engage Autistic People in the Design of Digital Technology: A Systematic Literature Review.” Journal of Autism and Developmental Disorders 54: 2960–2971. https://doi.org/10.1007/s10803-023-06015-5.

Mohamed, Shakir, Marie-Therese Png, and William Isaac. 2020. “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence.” Philosophy & Technology 33 (4): 659–684.

OECD. 2025. Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions. Paris: OECD Publishing.

Pillai, M., A. C. Griffin, C. A. Kronk, and T. McCall. 2023. “Toward Community-Based Natural Language Processing (CBNLP): Cocreating With Communities.” Journal of Medical Internet Research 25: e48498.

Pudjiastuti, Tri Nuke, et al. 2022. Naskah Kebijakan Pengembangan Riset Teknologi Alat Bantu bagi Penyandang Disabilitas. Jakarta: Badan Riset dan Inovasi Nasional. https://doi.org/10.55981/brin.679.

Stammbach, Dominik, Philine Widmer, Eunjung Cho, Caglar Gulcehre, and Elliott Ash. 2024. “Aligning Large Language Models with Diverse Political Viewpoints.” In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 7257–7267. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.412.

Tan, Hexiang, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. “Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts?” In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6207–6227. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.337.

UK Government and World Economic Forum. 2020. Guidelines for AI Procurement. London: Office for Artificial Intelligence and World Economic Forum. https://www.gov.uk/government/publications/guidelines-for-ai-procurement.

Umucu, Emre. 2025. “Artificial Intelligence and Health Equity for People with Disabilities: An Integrated Framework for Disability-Inclusive AI Design.” INQUIRY 62: 1–6. https://doi.org/10.1177/00469580251365472.

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO.

United Nations Environment Programme. 2024. Artificial Intelligence (AI) End-to-End: The Environmental Impact of the Full AI Life Cycle Needs to Be Comprehensively Assessed. Nairobi: UNEP.

van Toorn, Georgia. 2024. “Inclusion Interrupted: Lessons from the Making of a Digital Assistant by and for People with Disability.” Government Information Quarterly 41 (1): 101900. https://doi.org/10.1016/j.giq.2023.101900.

World Bank. 2020. Artificial Intelligence in the Public Sector: Maximizing Opportunities, Managing Risks. Washington, DC: World Bank. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/809611616042736565.

World Wide Web Consortium. 2021. Making Content Usable for People with Cognitive and Learning Disabilities. W3C Working Group Note. https://www.w3.org/TR/coga-usable/.

World Wide Web Consortium. 2024. Captions and Subtitles. Web Accessibility Initiative. https://www.w3.org/WAI/media/av/captions/.

World Wide Web Consortium. 2024. Web Content Accessibility Guidelines WCAG 2.2. W3C Recommendation. https://www.w3.org/TR/WCAG22/.

Young, Meg, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. 2024. “Participation versus Scale: Tensions in the Practical Demands on Participatory AI.” First Monday 29 (4). https://doi.org/10.5210/fm.v29i4.13642.

Interview Sources and Qualitative Evidence

IP-01 (Indigenous Participant). Interview transcript, August 2026, Ethical AI Localisation Toolkit research project.

EDU-3T-01 (Education Expert in 3T Regions). Interview transcript, August 2026, Ethical AI Localisation Toolkit research project.

CCE-01 (Cultural and Community Expert). Interview transcript, August 2026, Ethical AI Localisation Toolkit research project.

DP-01 (Disability Participant). Interview responses, 21 August 2026, Ethical AI Localisation Toolkit research project.

APPENDICES & TOOLKIT REFERENCE

Reference tables, checklists, templates, glossary, and access to the digital toolkit

Appendix A — The 62 Toolkit Indicators

This appendix presents the 62 indicators in this edition of the toolkit, together with evidence requirements, red-flag designations, mitigation guidance, and their stated basis.

D1. Context, Purpose, Human Rights and Risk

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
C1	Yes	Are the goals of AI systems clear, legitimate, proportionate, and not in conflict with human rights?	Purpose statement, affected groups, use boundaries, risk registers.	Delay until the purpose and limits of use are agreed upon.	Literature
C2	—	Were the contexts of the field - geography, conflict, social norms, basic services, and community structures - assessed before design?	Context assessment, site visit/field notes, stakeholder map.	Do a contextual/field assessment before the final design.	Indigenous Participant (IP-01) + literature
C3	Yes	Are clear red lines defined for uses that could undermine freedom of expression, safety, or fundamental rights?	Prohibited use policy, rights impact analysis, escalation rule.	Limit or discontinue high-risk use cases until adequate safeguards are in place.	Indigenous Participant (IP-01) + literature
C4	—	Was the toolkit/risk review used from pre-design to post-deployment evaluation?	Lifecycle governance plan, gate approvals.	Set the review gate on design, build, test, launch, and post-launch.	Indigenous Participant (IP-01) + Disability Participant (DP-01)

D2. Fairness and Non-Discrimination

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
F1	Yes	Is the system being tested for biases against region, ethnicity, language, gender, disability, age, and economic status?	Disaggregated bias tests, sample distribution, fairness metrics.	Don't use it for important decisions before bias is tested and reduced.	Literature + interviews
F2	—	For political, human-rights, or conflict-related topics, is the system tested for dominant-narrative or viewpoint bias?	Prompt tests, source diversity report, independent review.	Add adversarial testing with an independent/local perspective.	Indigenous Participant (IP-01)
F3	—	Are the benefits and burdens of the system analysed intersectionally (e.g. indigenous women with disabilities)?	Intersectional risk matrix, user research.	Separate the analysis by vulnerability combination, not just a single category.	Indigenous Participant (IP-01) + Cultural and Community Expert (CCE-01)
F4	—	Are alternative services available so that groups without devices/internet/assistive technology do not lose access?	Alternate service path, offline/manual SOP.	Provide a non-AI/manual route with equivalent service quality.	Education Expert 3T Regions (EDU-3T-01) + Disability Participant (DP-01)

D3. Transparency, Sources and Explainability

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
T1	—	Are users notified when they interact with AI or synthetic content?	Disclosure labels, user notices.	Add disclosure to the interaction point.	Literature

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
T2	—	Is the explanation of the decision available in a language/dialect that the user understands?	Plain-language/local-language versions, understanding test.	Translate and test understanding with local users.	Education Expert 3T Regions (EDU-3T-01) + Indigenous Participant (IP-01)
T3	Yes	For high-impact uses, can the main sources, limitations, and levels of uncertainty be traced?	The model/data card, the source log, the limitation notice.	Do not use output as primary reference without adequate traceability.	Literature + Indigenous Participant (IP-01)
T4	—	Are independent and local sources represented alongside government and dominant vendor sources?	Source diversity audit, coverage benchmark.	Add a plural source benchmark and disclosure if coverage is uneven.	Indigenous Participant (IP-01)

D4. Accountability, Human Oversight and Lifecycle

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
A1	Yes	Is there a clear owner of accountability if AI is wrong or harmful?	RACI, named owner, escalation SOP.	Set the owner before deployment.	Literature
A2	Yes	Can high-impact decisions be reviewed or overridden by humans?	Human-review SOP, override authority, audit trail.	Add human-in-the-loop with real authority.	Literature + Disability Participant (DP-01)
A3	—	Are responsibilities for design, frontend, backend, content, quality assurance, project management, legal review, and ethics/accessibility auditing defined throughout the lifecycle?	Lifecycle RACI, definition of done.	Set ownership per phase and acceptance criteria.	Disability Participant (DP-01)
A4	Yes	Is there any formal authorisation to pause/stop deployment when the red flag is not fixed?	Stop criteria, approval gate, incident process.	Set a stop authority and document the decision.	Logic toolkit + interviews

D5. Privacy, Security and Personal Data

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
P1	Yes	Is data collection minimised, supported by a lawful basis for processing, and subject to limited retention?	Data inventory, lawful basis, retention schedule.	Reduce the data and set the retention.	Literature
P2	Yes	Are sensitive data, locations, AI conversations, and marginalised group identities protected from over-access/doxxing?	The access matrix, the encryption, the privacy threat model.	Restrict access, encrypt data, and prohibit sensitive profiling.	Indigenous Participant (IP-01) + Disability Participant (DP-01)
P3	—	Does security testing include hacking, data leakage, prompt/data injection, espionage, and misuse?	Pen-test, red-team report, incident plan.	Run a security/red-team test before launch.	Indigenous Participant (IP-01) + literature
P4	—	Are users given practical controls to delete, correct, or restrict the use of their data?	User controls, privacy workflow, response SLA.	Add the privacy controls that are accessible.	Literature + Disability Participant (DP-01)

D6. Collective Consent, Community Data and Benefit-Sharing

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
G1	Yes	For cultural or communal data, is collective consent obtained in addition to individual consent?	Collective consent record, individual consent, governance map.	Stop collecting until collective and individual consent is met.	Education Expert 3T Regions (EDU-3T-01) + Indigenous Participant (IP-01)
G2	Yes	Does the prohibited or restricted-data register address sacred knowledge, sensitive locations, burial sites, rituals, conflict areas, and data that could enable doxxing?	Restricted-data register, access protocol.	Create a restricted data registry with the community.	Education Expert 3T Regions (EDU-3T-01) + Indigenous Participant (IP-01)
G3	—	Is consent sought through familiar methods, in an understandable language, without pressure, and with protection for the right to refuse?	Consent process, local-language materials, non-retaliation statements.	Use in-person/local-language deliberation and the non-retaliation rule.	Indigenous Participant (IP-01) + Cultural and Community Expert (CCE-01)
G4	—	Does benefit-sharing provide long-term benefits and strengthen community control and capacity?	Benefit plan, capacity building, community stewardship.	Prioritise infrastructure, capacity building, access, and community custody.	Indigenous Participant (IP-01) + CARE
G5	—	Are collective agreements documented for review and understanding across generations?	Intergenerational record, review and renewal schedule.	Add renewal/review points and community-managed archives.	Indigenous Participant (IP-01)

D7. Cultural and Linguistic Validation

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
L1	Yes	Are the cultural data sets/outputs validated by the relevant native speakers/cultural validators?	Validation record, named validator roles.	Do not publish/deploy until local validation is complete.	Education Expert 3T Regions (EDU-3T-01) + Indigenous Participant (IP-01)
L2	—	Are models tested on local languages/dialects and cultural representation errors?	Local benchmarks, error logs, and dialect testing.	Build a local benchmark with the community.	Education Expert 3T Regions (EDU-3T-01) + Indigenous Participant (IP-01)
L3	—	Does the system support language preservation without taking over community ownership or authority?	Language governance, licensing, community custody.	Use a community-controlled repository/licensing.	Indigenous Participant (IP-01) + Education Expert 3T Regions (EDU-3T-01)
L4	—	Can cultural misrepresentation be corrected promptly by a clearly designated validator with the necessary authority?	Correction workflow, validator authority, SLA.	Set the validator authority and correction SLA.	Education Expert 3T Regions (EDU-3T-01) + Indigenous Participant (IP-01)

D8. Accessibility and Assistive Technology

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
X1	Yes	Can the entire critical journey be completed without relying solely on vision, hearing, speech, complex movement, or one type of input device?	Cross-disability journey matrix; keyboard, screen-reader, alternative-input, and non-audio notification tests.	Provide equivalent alternative inputs and outputs; improve semantic structure, focus order, target size, and status messages.	WCAG 2.2 + disability-inclusive design literature
X2	Yes	Can text, controls, focus, status, and visual information be understood through magnification, contrast, reflow, and screen readers, including by users with combined needs such as deafblindnessness?	Zoom/reflow, contrast/focus, screen-reader, caption customisation, and multiple-disability scenario tests.	Improve reflow, contrast, focus visibility, scalable text/captions, and compatibility with assistive technology.	WCAG 2.2 + W3C Diverse Abilities
X3	Yes	Do login, CAPTCHA, OTP, recovery, and 2FA provide equivalent methods without relying on any particular sensory modality, memory, or motor ability?	Accessible-authentication test; password-manager and copy-and-paste compatibility; multimodal recovery test.	Provide several accessible authentication methods that are equivalent; do not make audio or visual alternatives the only option.	WCAG 2.2 + Disability Participant (DP-01)
X4	—	Are information, media, notifications, voice services, and AI outputs available in an accurate and understandable alternative modality?	Caption/transcript QA; visual-alert test; text alternative for voice channels; plain-language review; sign-language plan when relevant.	Provide human-verified captions and transcripts, visual alerts, text alternatives, plain-language content, and sign language access where relevant.	WCAG 2.2 + W3C Captions + cognitive-accessibility evidence
X5	Yes	Are critical user journeys tested with users whose needs are relevant to the service, and do the findings influence design or deployment decisions?	Disability coverage rationale; participant matrix; issue log; feedback-to-decision trace; retest evidence.	Do cross-disability co-testing based on use cases, close critical issues, and document decision changes.	Disability-inclusive literature + W3C COGA + DP-01

D9. Meaningful Participation and Intersectional Inclusion

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
M1	Yes	Have affected communities been involved from the design stage, rather than only invited to awareness sessions after decisions were made?	Participation timeline, meeting records, decision trace.	Do not proceed if consultation begins only after the main decisions have been made.	Indigenous Participant (IP-01) + Cultural and Community Expert (CCE-01)
M2	Yes	Is there any evidence that community input changes design/policy, so participation is not tokenism?	Change log linked to stakeholder input.	Make a feedback-to-decision trace and explain input received/refused.	Indigenous Participant (IP-01)
M3	—	Does representation include directly affected citizens, not just urban/formal representatives?	Stakeholder map by geography/impact.	Extend the invitation to affected areas and local communities.	Indigenous Participant (IP-01)

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
M4	—	Have participation costs and logistics been budgeted for so they do not become grounds for exclusion?	Participation budget, travel/access support.	Use public facilities and proportional logistical support.	Indigenous Participant (IP-01)
M5	—	Do women, young people, persons with disabilities, low-income groups, and people with intersecting identities have a meaningful voice?	Disaggregated participation records, influence evidence.	Audit intersectional representation and its impact on decision making.	Cultural and Community Expert (CCE-01) + Indigenous Participant (IP-01) + Disability Participant (DP-01)

D10. 3T Regions, Local Readiness and Continuity

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
R1	Yes	Is the readiness of the region verified directly, not just on the basis of institutional claims?	Readiness assessment, field verification.	Verification of infrastructure, human resources, language, and basic services on site.	Education Expert 3T Regions (EDU-3T-01) + Indigenous Participant (IP-01)
R2	Yes	Is low-bandwidth/offline or manual mode available if the network is unstable?	Offline testing, manual SOP, continuity plan.	Build offline/manual fallback before deployment.	Education Expert 3T Regions (EDU-3T-01) + Indigenous Participant (IP-01)
R3	—	Have conflict scenarios, fibre cuts, transport disruption, and outages been tested?	BCP/DR scenario test, the service continuity plan.	Add the scenario-based continuity testing.	Indigenous Participant (IP-01)
R4	—	Is local/edge hosting or design that reduces external connectivity dependency considered proportionally?	The architecture options analysis, the resilience rationale.	Evaluate edge/local options without compromising security/privacy.	Indigenous Participant (IP-01) + Education Expert 3T Regions (EDU-3T-01)

D11. AI Literacy, Education and Human Readiness

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
E1	—	Do users get AI literacy on hallucinations, deepfake, privacy, and verification of information?	Training materials, attendance, understanding tests.	Do literacy onboarding before adoption.	Cultural and Community Expert (CCE-01) + Education Expert 3T Regions (EDU-3T-01)
E2	—	In education, are there rules to prevent overreliance and maintain independent thinking?	Classroom guidance, alternative assessments.	Set rules for AI use and use assessments without AI where necessary.	Education Expert 3T Regions (EDU-3T-01)
E3	—	Is the local train-the-trainer/peer learning model used to strengthen capacity?	Local trainer plans, community materials.	Train local facilitators and develop simple, context-appropriate materials.	Disability Participant (DP-01) + Cultural and Community Expert (CCE-01)

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
E4	—	Is human readiness measured by the actual ability to use a system, not just the ownership of a device?	Usability/readiness test, support plan.	Do a task-based readiness test.	Indigenous Participant (IP-01)

D12. Grievances, Redress and Human Escalation

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
Q1	Yes	Are complaint channels available online and offline and accessible with assistive technology?	Accessible complaint test, offline channel SOP.	Provide dual channels and testing with disabled users.	Indigenous Participant (IP-01) + Education Expert 3T Regions (EDU-3T-01) + Disability Participant (DP-01)
Q2	Yes	Can users escalate to a human when bots or AI fail?	Human escalation path, staffing, SLA.	Don't use bot-only grievances for serious cases.	Indigenous Participant (IP-01) + Disability Participant (DP-01)
Q3	—	Are response times, case status, corrections/takedown, and form of redress clearly explained?	SLA, case tracker, remedy policy.	Set the SLA and real outcome, not an automatic apology.	Indigenous Participant (IP-01) + Disability Participant (DP-01)
Q4	—	Is each complaint handled on a case-by-case basis and traceable to resolution?	Case log, owner, evidence of closure.	Use case-level tracking and root-cause review.	Indigenous Participant (IP-01)

D13. Vendors, Procurement and Data Sovereignty

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
V1	Yes	Does the vendor contract prohibit secondary use of data without an appropriate basis or authorisation?	DPA/contract clauses, subprocessor list.	Revise the contract before transferring data.	Literature + interviews
V2	—	Are the data residency, cross-border transfers, and processing locations known?	Data flow map, residency/transfer assessment.	Map the data flow and transfer controls.	Literature + Indigenous Participant (IP-01)
V3	—	Are vendor lock-in safeguards, audit rights, an exit plan, and knowledge or technology transfer arrangements in place?	The exit plan, the portability, the audit clause.	Add portability, audit rights, and exit criteria.	Decolonial literature + interviews
V4	—	For community/language data, does the community still have fair control or license?	Licensing terms, governance rights, benefit plans.	Use community-controlled licensing and attribution.	Education Expert 3T Regions (EDU-3T-01) + Indigenous Participant (IP-01)

D14. Monitoring, Audit and Continuous Improvement

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
N1	—	Is the system monitored for model drift, emerging bias, source imbalances, and sociocultural impacts?	Monitoring the dashboard, the audit schedule.	Schedule periodic audits and community reviews.	Literature + interviews
N2	—	Is every feature, library, or vendor update retested for accessibility regressions?	Regression test, release checklist.	Add the accessibility regression test to the release gate.	Disability Participant (DP-01)
N3	—	Does the action plan record the owner, deadline, evidence, and mitigation status?	The action log, the evidence links.	Use the action tracker before approval.	Toolkit design
N4	—	Did the community/users get updates and feedback opportunities after launch?	Post-launch engagement logs.	Maintain a feedback loop throughout the operation.	Indigenous Participant (IP-01)

D15. Sustainability and Environmental Impact

ID	Red flag	Question	Evidence Required	If no/partial	Toolkit basis
S1	—	Are the system's computing and energy requirements assessed in proportion to its benefits?	Compute/energy estimate, architecture comparison.	Choose a more efficient architecture/model if the benefits are equal.	Literature
S2	—	Do infrastructure locations and resource consumption consider the impact on local communities/ecosystems?	Site/environment/community impact review.	Do an environmental/community impact review.	Literature
S3	—	Does the procurement take into account hardware lifecycle, e-waste, and data centre/cloud efficiency?	The procurement criteria, the lifecycle plan.	Add sustainability criteria to procurement.	Literature

Appendix B — Red-Flag Checklist

The toolkit designates 23 indicators as red flags. A No or Partial response means a critical safeguard remains unresolved and requires escalation. Deployment cannot proceed until the critical requirement is met and supported by evidence.

ID	Domain	The red-flag question	Evidence that must be there	Closure status
C1	D1 - Context, purpose, human rights and risks	Are the goals of AI systems clear, legitimate, proportionate, and not in conflict with human rights?	Purpose statement, affected groups, use boundaries, risk registers.	☐ Open ☐ Mitigating ☐ Closed
C3	D1 - Context, purpose, human rights and risks	Are clear red lines defined for uses that could undermine freedom of expression, safety, or fundamental rights?	Prohibited use policy, rights impact analysis, escalation rule.	☐ Open ☐ Mitigating ☐ Closed
F1	D2 - Fairness and Non-Discrimination	Is the system being tested for biases against region, ethnicity, language, gender, disability, age, and economic status?	Disaggregated bias tests, sample distribution, fairness metrics.	☐ Open ☐ Mitigating ☐ Closed
T3	D3 - Transparency, Source and Explainability	For high-impact uses, can the main sources, limitations, and levels of uncertainty be traced?	The model/data card, the source log, the limitation notice.	☐ Open ☐ Mitigating ☐ Closed
A1	D4 - Accountability, Human Oversight & Lifecycle	Is there a clear owner of accountability if AI is wrong or harmful?	RACI, named owner, escalation SOP.	☐ Open ☐ Mitigating ☐ Closed
A2	D4 - Accountability, Human Oversight & Lifecycle	Can high-impact decisions be reviewed or overridden by humans?	Human-review SOP, override authority, audit trail.	☐ Open ☐ Mitigating ☐ Closed
A4	D4 - Accountability, Human Oversight & Lifecycle	Is there any formal authorisation to pause/stop deployment when the red flag is not fixed?	Stop criteria, approval gate, incident process.	☐ Open ☐ Mitigating ☐ Closed
P1	D5 - Privacy, security and personal data	Is data collection minimised, supported by a lawful basis for processing, and subject to limited retention?	Data inventory, lawful basis, retention schedule.	☐ Open ☐ Mitigating ☐ Closed
P2	D5 - Privacy, security and personal data	Are sensitive data, locations, AI conversations, and marginalised group identities protected from over-access/doxxing?	The access matrix, the encryption, the privacy threat model.	☐ Open ☐ Mitigating ☐ Closed
G1	D6 - Collective consent, community data and benefit-sharing	For cultural or communal data, is collective consent obtained in addition to individual consent?	Collective consent record, individual consent, governance map.	☐ Open ☐ Mitigating ☐ Closed
G2	D6 - Collective consent, community data and benefit-sharing	Does the prohibited or restricted-data register address sacred knowledge, sensitive locations, burial sites, rituals, conflict areas, and data that could enable doxxing?	Restricted-data register, access protocol.	☐ Open ☐ Mitigating ☐ Closed
L1	D7 - Cultural and linguistic validation	Are the cultural data sets/outputs validated by the relevant native speakers/cultural validators?	Validation record, named validator roles.	☐ Open ☐ Mitigating ☐ Closed

ID	Domain	The red-flag question	Evidence that must be there	Closure status
X1	D8 - Accessibility and Assistive Technology	Can the entire critical journey be completed without relying solely on vision, hearing, speech, complex movement, or one type of input device?	Cross-disability journey matrix; keyboard, screen-reader, alternative-input, and non-audio notification tests.	☐ Open ☐ Mitigating ☐ Closed
X2	D8 - Accessibility and Assistive Technology	Can text, controls, focus, status, and visual information be understood through magnification, contrast, reflow, and screen readers, including by users with combined needs such as deafblindnessness?	Zoom/reflow, contrast/focus, screen-reader, caption customisation, and multiple-disability scenario tests.	☐ Open ☐ Mitigating ☐ Closed
X3	D8 - Accessibility and Assistive Technology	Do login, CAPTCHA, OTP, recovery, and 2FA provide equivalent methods without relying on any particular sensory modality, memory, or motor ability?	Accessible-authentication test; password-manager and copy-and-paste compatibility; multimodal recovery test.	☐ Open ☐ Mitigating ☐ Closed
X5	D8 - Accessibility and Assistive Technology	Are critical user journeys tested with users whose needs are relevant to the service, and do the findings influence design or deployment decisions?	Disability coverage rationale; participant matrix; issue log; feedback-to-decision trace; retest evidence.	☐ Open ☐ Mitigating ☐ Closed
M1	D9 - Meaningful participation and intersectional inclusion	Have affected communities been involved from the design stage, rather than only invited to awareness sessions after decisions were made?	Participation timeline, meeting records, decision trace.	☐ Open ☐ Mitigating ☐ Closed
M2	D9 - Meaningful participation and intersectional inclusion	Is there any evidence that community input changes design/policy, so participation is not tokenism?	Change log linked to stakeholder input.	☐ Open ☐ Mitigating ☐ Closed
R1	D10 - 3T, Local Readiness & Continuity	Is the readiness of the region verified directly, not just on the basis of institutional claims?	Readiness assessment, field verification.	☐ Open ☐ Mitigating ☐ Closed
R2	D10 - 3T, Local Readiness & Continuity	Is low-bandwidth/offline or manual mode available if the network is unstable?	Offline testing, manual SOP, continuity plan.	☐ Open ☐ Mitigating ☐ Closed
Q1	D12 - Complaints, Redress and Human Escalation	Are complaint channels available online and offline and accessible with assistive technology?	Accessible complaint test, offline channel SOP.	☐ Open ☐ Mitigating ☐ Closed
Q2	D12 - Complaints, Redress and Human Escalation	Can users escalate to a human when bots or AI fail?	Human escalation path, staffing, SLA.	☐ Open ☐ Mitigating ☐ Closed
V1	D13 - Vendors, Procurement and Data Sovereignty	Does the vendor contract prohibit secondary use of data without an appropriate basis or authorisation?	DPA/contract clauses, subprocessor list.	☐ Open ☐ Mitigating ☐ Closed

Red-flag rule Aggregate scores should not be used to cover unresolved red flags.

Appendix C — Evidence Checklist

This checklist helps the team gather evidence before a formal assessment.

Context and Risk

- ☐ Project charter / system purpose
- ☐ affected-group map
- ☐ risk classification
- ☐ human rights impact review
- ☐ Lifecycle stage

Data and privacy

- ☐ data inventory
- ☐ Purpose specification
- ☐ Retention schedule
- ☐ Access matrix
- ☐ security/threat model
- ☐ User controls

Community and Participation

- ☐ Stakeholder map
- ☐ Participation timeline
- ☐ Feedback to decision trace
- ☐ Collective authorisation record
- ☐ Restricted-data register
- ☐ benefit-sharing plan

Language & Culture

- ☐ Local-language benchmark
- ☐ Native/local validator record
- ☐ cultural validation log
- ☐ Correction SLA
- ☐ language-data licensing/governance

Accessibility

- ☐ Keyboard test
- ☐ Screen-reader test
- ☐ Mouse/pointer + screen-reader co-test
- ☐ the CAPTCHA/OTP/2FA test
- ☐ Direct disabled-user test
- ☐ Retest evidence

Human Oversight and Remedy

- ☐ RACI
- ☐ override log
- ☐ Escalation matrix
- ☐ Stop authority
- ☐ complaint channels
- ☐ SLA
- ☐ case-tracking
- ☐ Correction/remedy record

Vendor & Procurement

- ☐ DPA/contract
- ☐ Subprocessor list
- ☐ data-flow map
- ☐ Audit clause
- ☐ Incident-notification clause
- ☐ portability/exit plan
- ☐ transfer of knowledge

Monitoring and Sustainability

- ☐ Monitoring thresholds
- ☐ Change log
- ☐ incident log
- ☐ Regression test
- ☐ evidence of resource use
- ☐ Efficiency choices
- ☐ Environmental mitigation plan

Appendix D — Action Plan Template

Use one line for each gap that needs mitigation.

Gaps	Risk	Action	Owner	Due date	Closure Evidence	Status	Decision Gate
						☐ In progress ☐ Closed	
						☐ In progress ☐ Closed	
						☐ In progress ☐ Closed	
						☐ In progress ☐ Closed	
						☐ In progress ☐ Closed	
						☐ In progress ☐ Closed	
						☐ In progress ☐ Closed	
						☐ In progress ☐ Closed	

Completed example

Gaps	Risk	Action	Owner	Due date	Closure Evidence	Status	Decision Gate
D8 X3	Critical access barrier	Redesign CAPTCHA/OTP and retest	Digital service + vendor	14 days	Direct co-test + issue closure	Closed	Before the pilot
D12 Q2	No human escalation	Add direct human route + SLA	Service operations	14 days	Complaint test + SLA log	Closed	Before the pilot

Appendix E — Deployment Decision Record

Field	Record
System / Use case	
Version / Vendor	
Date of the assessment	
Deployment Scope / Geography	
Aggregate Score	
Red Flags Open / Closed	
Critical Domain Gaps	
Decision	☐ Proceed ☐ Proceed with Conditions ☐ Pause & Mitigate ☐ Do Not Deploy
Conditions	
Residual risk statements	
Accountable Decision-Maker	
Next Review Date	
Reassessment triggers	

Appendix F — Glossary

Term	Working definition
AI readiness	The readiness of organisations, users, infrastructure, processes, language, accessibility, and governance to use AI responsibly.
Affected group	The group that receives the direct or indirect impact of using AI.
Collective authorisation	Authorisation at the community/group level when data or knowledge has a collective dimension.
Contestability	The ability of a user to challenge the basis or outcome of a decision and obtain meaningful review/correction.
Critical journey	A sequence of important tasks that a user must complete, such as logging in, filling out a form, receiving results, and complaining.
Deployment gate	The conditions that must be met before the system or feature can be carried on to the next stage.
Evidence requirement	Evidence must be available to support the assessment answers.
Feedback-to-decision trace	The trail that shows how stakeholder input influences or does not influence the decision, and why.
Human escalation	The ability of users or officers to transfer cases from AI/automation to humans in authority.
Human oversight	The human capacity to review, override, escalate, or stop AI systems/decisions.
Intersectionality	The ways in which overlapping identities and circumstances shape people's experiences of risk and impact.
Localisation	Adjustment of governance, evidence, testing, participation, language, accessibility, and controls to the context of actual use.
Mitigation	Action to reduce or eliminate the gaps/risks found in the assessment.
N/A	Not Applicable; only used when the requirement is completely irrelevant and the reason is documented.
Red flag	Critical conditions that trigger gating and should not be covered by aggregate scores.
Reassessment	Repetition of assessment after material change, incident, or mitigation.
Remedy	Correction or recovery provided when the user is affected by error/harm.
Residual risk	Risk that remains after mitigation.
Source provenance	Information on the origin, version, and trace of the source used by the system.
Vendor lock-in	Dependencies that make it difficult for organisations to move vendors, access data, audit, or maintain continuity.

Appendix G — Participant Methodology and Anonymisation Note

Project interviews are used as contextual qualitative evidence to identify implementation gaps, lived experience, and requirements that may be less visible in high-level frameworks. Interviews are not designed to represent the population of Indonesia statistically.

Code	Public attribution	Primary contextual lens
IP-01	Indigenous Participant	community/Indigenous governance, collective consent, cultural data, participation, continuity, grievance
EDU-3T-01	Education Expert in 3T Regions	remote readiness, connectivity, local language, human readiness, offline access
CCE-01	Cultural and Community Expert	Grassroots inclusion, AI literacy, consent, data minimisation, and community engagement
DP-01	Disability Participant	Lived experience of blindness or low vision, assistive technology, authentication, grievance mechanisms, and human escalation

Participant names and contact details are omitted from this public edition. Role-based attribution and participant codes are used consistently. Interview evidence is contextual: one person's account cannot establish a claim about an entire group. The publication licence does not release the underlying interview records or waive participants' privacy rights.

Appendix H — Citation and Evidence Conventions

This book follows Chicago Notes and Bibliography conventions. Concise tables may use consistent author–year references, with complete entries in the bibliography.

Evidence type	Example of attribution	Interpretation
Peer-reviewed literature	Author, year	Supports academic claims within the scope of the cited study; transferability depends on context.
Official framework/report	Institution, year	Provides a governance baseline, policy position, or official statistic within the source's scope.
Interview	Participant code + timestamp/question	Provides contextual qualitative evidence rather than statistically representative findings.
Toolkit rule	Ethical AI Localisation Toolkit, first edition	Identifies an operational design choice rather than an empirically validated threshold.
Author synthesis	'This book proposes...'	Distinguishes the author's interpretation from findings reported in literature or interviews.

Evidence integrity: additional sources can corroborate, challenge, or qualify an argument. They are not retrospectively treated as original sources of an indicator unless the development record supports that attribution.

Closing Note on the Appendices

This appendix is designed so that the reader can use the toolkit without having to go back to the whole chapter for each step. However, tables and checklists should be read together with explanations in Chapters 4-10 so that the assessment does not turn into a mechanical compliance exercise.

Core principle: evidence before confidence. Context before deployment. A score never overrides an unresolved critical risk.

Appendix I — Access the Digital Toolkit

The book can be used independently through the 62 indicators, checklists, and templates available in the appendix. For a more interactive experience, readers can use the companion website as a tool for assessment and follow-up planning.

Key functions

- Fill out the assessment based on 15 domains.
- Identify evidence gaps and red flags.
- Review results without treating scores as the sole basis for decisions.
- Draft an action plan for follow-up.

Scan the QR code or visit aiethicskit.org

The use of the website is optional. Avoid entering personal data, confidential information, or sensitive documents. Learn the purposes of data processing, storage, and use through the Privacy Notice on the website.