D3.4 Mechanism for Evaluating Ethics Readiness Levels and Algorithmic Impact Assessment

AIOLIA DELIVERABLE D3.4

Horizon Europe Grant Agreement N° 101187937
AIOLIA

Project Name	AIOLIA
Deliverable Title/Number	D3.4
Description	Mechanism for evaluating ethics readiness levels and algorithmic impact assessment, developed under T3.4.
Lead beneficiary	RISE
Lead Authors	Laurynas Adomaitis
Contractual delivery date:	M16 / 31 May 2026
Actual delivery date:	01 June 2026
Sensitivity	PU - Public

Document History

Name	Organisation	Role	Action	Date
Laurynas Adomaitis	RISE	Lead author	First version for internal review	19/05/2026
Petra Bayerl	CENTRIC	Contributor	Reviewed and improved the AIA Subtool	21/05/2026
Daniele Miorandi	Afliant	Reviewer	Provided feedback and deliverable QA	22/05/2026
Artur Bogucki	CEPS	Reviewer	Provided feedback and deliverable QA	26/05/2026
Laurynas Adomaitis	RISE	Lead author	Final version	31/05/2026
Alexei Grinbaum	CEA	Coordinator	Final check	01/06/2026

Configuration Management

Nature of Deliverable	
R	Report

Dissemination level	
PU	Public, fully open

Acronym/abbreviations	
AI	Artificial Intelligence
AIA	Algorithmic Impact Assessment
AI Act	Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
DPO	Data Protection Officer
ERL	Ethics Readiness Level
GDPR	General Data Protection Regulation
HUDERIA	Human Rights, Democracy and Rule of Law Impact Assessment methodology
IRL	Integration Readiness Level

LEA	Law Enforcement Agency
LED	Law Enforcement Directive
LPERL	Legal, Privacy and Ethics Readiness Level
RRI	Responsible Research and Innovation
TRL	Technology Readiness Level
WP	Work Package

How to cite

Adomaitis, L. (2026). Mechanism for Evaluating Ethics Readiness Levels and Algorithmic Impact Assessment. AIOLIA Deliverable 3.4.

Acknowledgements

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein. Reproduction is authorised provided the source is acknowledged.

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Use of AI

During preparation of this report the author has used Claude Code and Perplexity LLM-based coding agents for coding assistance, and Claude and ChatGPT chatbots for editorial and bibliographic research purposes.

TABLE OF CONTENTS

ERL System Card 7

Executive Summary 8

1. Introduction 9

1.1. Purpose of the Deliverable 9

1.2. Scope of the Deliverable 9

2. Background: From MultiRATE to AIOLIA 10

2.1. Origin of the ERL Approach 10

2.2. Limits of Flat Evaluation 10

2.3. AIOLIA's Added Value 11

3. AIOLIA Activities Under T3.4 12

3.1. Initial Preparation 12

3.2. First Expert Panel 12

3.3. Development Phase 12

3.4. AIOLIA Plenary Demonstration 13

3.5. Healthcare Validation Activities 13

3.6. Public-Administration AIA Review 14

4. From Evaluation to Ethics-by-Design 15

4.1. Conceptual Shift 15

4.2. Design-Time Function of ERLs 15

4.3. Dialogue as a Methodological Requirement 16

5. Healthcare AI Module 17

5.1. Rationale for Selecting Healthcare 17

5.2. Main Ethics Issues Addressed 17

6. Public-Administration AIA Sub-Tool 19

6.1. Rationale for Selecting Public Administration 19

6.2. Main AIA Dimensions 19

7. Tool Definition and Methodology 21

7.1. Ethics Readiness Levels 21

7.2. Dynamic Questionnaire Design 21

7.3. Indicator Logic 22

7.4. Scoring Logic 23

7.5. Tracking Over Time 24

8. Code Implementation and Replication _____ 25

8.1. Implementation Availability _____ 25

8.2. High-Level Architecture _____ 25

8.3. Reference Implementation Journey _____ 26

8.4. Technical Guidance for Replicating and Hosting the Implementation _____ 33

8.5. Why Public Hosting Was Not Chosen _____ 34

9. Limitations and Next Steps _____ 35

10. Conclusion _____ 36

11. Appendices _____ 37

11.1. Appendix A: Overview of the ERL Methodology _____ 37

11.2. Appendix B: AI for Healthcare Indicator Block _____ 37

11.3. Appendix C: AI Impact Assessment Indicator Block _____ 39

11.4. Appendix D: AIOLIA Technical Measures with linked indicators _____ 42

12. References _____ 47

LIST OF TABLES

Table 1. ERL score-to-level mapping
Table 2. Healthcare AI indicator block
Table 3. Public-administration AI impact assessment indicator block
Table 4. AIOLIA technical measures pairings

LIST OF FIGURES

Figure 1. Healthcare AI module topic map
Figure 2. Public-administration AIA sub-tool topic map
Figure 3. Dynamic questionnaire and scoring logic
Figure 4. Reference implementation architecture
Figure 5. Reference implementation welcome and artefact attribution screen
Figure 6. Participant setup screen
Figure 7. Question card with linked AIOLIA technical measure
Figure 8. Completion screen and expert recommendation field
Figure 9. Top areas needing attention view
Figure 10. Advanced ethics analytics dashboard
Figure 11. Per-indicator contribution waterfall
Figure 12. Exit-report score trajectory
Figure A1. Overview of the ERL methodology and repeated assessment loop

ERL System Card

```
---
tool_name: Ethics Readiness Levels Tool
version: 0.2-dev reference implementation
repository: https://github.com/LA-NS/ethics-readiness-levels/tree/v0.2-dev
maintainer_context: RISE / AIOLIA Task 3.4
task: expert-led ethics readiness and algorithmic impact assessment session
tags: ethics-readiness, ethics-by-design, algorithmic-impact-assessment, AI-governance, dialogue-evaluation
production_status: development reference
optional_llm_endpoint: http://127.0.0.1:1234/v1/chat/completions
dependencies: requirements.txt
local_url: http://127.0.0.1:8080
database_seed: schema.sql
database: SQLite local file lperl_local.sqlite
entrypoint: app.py
runtime: Python Flask local web application
---
```

Tool summary. The ERL tool artefact is a reference implementation of the Ethics Readiness Levels mechanism. It is a dynamic questionnaire, scoring engine, session flow, analytics view, and report-generation pathway for expert-led ethics readiness sessions.

Intended use. The tool is intended for facilitated sessions in which an ethics expert and a technical or domain expert jointly assess an AI system. It is suitable for design reviews, maturity tracking, use-case validation, training exercises, and structured follow-up between assessment sessions.

Inputs. The artefact mandates a dialogue between at least two experts, one representing ethics, another knowing technical details of the implementation. Together, they discuss and answer binary questions of relevance, mitigation, and validation. The tool prioritizes custom feedback and recommendations.

Outputs. The artefact produces a score trajectory, final ERL/LPERL score, readiness-level interpretation, negative and positive event counts, top unresolved contributors, block attribution, per-indicator contribution views, expert recommendations, and an exportable assessment report.

Out-of-scope use. The artefact is not a certification system, legal compliance checker, public self-assessment portal, or product-ranking benchmark. It should not be used to compare unrelated AI systems, and it should not be treated as evidence of regulatory compliance without separate legal and technical analysis.

Technical shape. The artefact is a single-machine Flask application. The user runs `python app.py`, the application creates or reads a local SQLite database, and the assessment is opened in a browser at `http://127.0.0.1:8080`. The main runtime files are `app.py`, `schema.sql`, `requirements.txt`, `templates/`, `static/`, and the generated `lperl_local.sqlite` database.

Optional services. The core ERL workflow does not require a cloud service. Optional local LLM help in `app.py` expects a separate local API at `http://127.0.0.1:1234/v1/chat/completions`. If that service is absent, questionnaire navigation, scoring, analytics, and report export remain usable.

License. This project is licensed under the GNU General Public License v3.0.

Executive Summary

This deliverable reports the work carried out under AIOLIA Task 3.4 to develop a mechanism for evaluating Ethics Readiness Levels (ERLs) and integrating algorithmic impact assessment (AIA). The developed ERL tool is a web-implemented support tool to conduct semi-structured dialogues between ethics experts and technical experts with the goal of improving the design of AI systems.

The mechanism builds on MultiRATE readiness-level work (Unzueta et al., 2026) and adapts it into a dialogue-led ethics-by-design instrument. It uses contextual onboarding, activated indicator blocks, hierarchical parent and follow-up indicators, yes/no scoring, evidence recording, and session-level score trajectories. The mechanism generates maturity recommendations by tracing readiness gaps to activated ethics concerns, negative scoring events, missing mitigations, and safeguards that have not yet been validated.

The mechanism is intended for expert-led sessions involving ethics, technical, and domain expertise through dialogue. Dialogue is a key part of the mechanism because it elicits ethical reflection. The healthcare AI module was validated with AIOLIA Healthcare Use Case 1 and AIOLIA Healthcare Use Case 2. The public-administration AIA sub-tool was reviewed by CENTRIC, revised into version 2, and prepared for further validation in T4.3 citizen engagement activities.

Although the DoA originally anticipated a modular Excel-based tool to reduce user burden, the implementation evolved into a reference mechanism encoded in structured indicator files and questionnaire logic (however, Excel-style indicator tables are also provided in the Appendices for transparency and reuse). The chosen solution preserves the modularity, scoring transparency, and replicability envisaged in the DoA, while enabling branching, session tracking, and repeated assessment more robustly than a static spreadsheet.

1. Introduction

1.1. PURPOSE OF THE DELIVERABLE

This deliverable presents the mechanism developed under AIOLIA Task 3.4 for evaluating Ethics Readiness Levels (ERLs) and for integrating algorithmic impact assessment (AIA) into AIOLIA's operational ethics work. The deliverable is intended for the AIOLIA consortium, internal reviewers, and European Commission project officers.

The present deliverable has two purposes. First, it records the AIOLIA activities undertaken to adapt and validate the mechanism. Second, it defines the tool and methodology through the readiness scale, indicator blocks, branching logic, scoring model, healthcare module, public-administration AIA sub-tool, and reference implementation.

The deliverable also clarifies the methodological position of the mechanism. It is a structured instrument for dialogue-led ethics-by-design. Its primary function is to help teams see where ethical reflection is missing, where mitigations have been implemented, where validation is still required, and how maturity changes across repeated assessment sessions.

The deliverable should therefore be read as a practical assessment and replication mechanism. It supports ethical maturity planning, but it does not certify systems, provide legal advice, or replace stakeholder engagement.

1.2. SCOPE OF THE DELIVERABLE

The scope of the deliverable covers four connected outputs.

The first output is a general ERL evaluation mechanism. It uses a 0-4 readiness scale to describe the maturity of ethics integration in an AI system or AI-related research area. The scale is inspired by readiness-level approaches such as Technology Readiness Levels (TRLs) and Integration Readiness Levels (IRLs), but it is adapted to ethics.

The second output is a healthcare AI module developed in the AIOLIA context. This module addresses healthcare-specific concerns. It reflects both established healthcare AI ethics concerns and concrete feedback from AIOLIA healthcare partners.

The third output is a public-administration AIA sub-tool. It applies the ERL/AIA logic to AI systems used by or for public administrations, where algorithmic systems may affect citizens' legal rights, entitlements, obligations, access to services, or democratic relationship with the state. It draws on the logic of algorithmic impact assessment and fundamental rights impact assessment, especially as reflected in the EU AI Act and existing public-sector AIA approaches.

The fourth output is a replication logic, which explains how the mechanism can be implemented on another platform by expert teams. The implementation is represented in code and indicator structures, but the mechanism is intentionally not reduced to a public self-service web portal. Replication requires preserving the dialogue-led structure of the method.

Although the DoA originally anticipated a modular Excel-based tool to reduce user burden, the implementation evolved into a reference mechanism encoded in structured indicator files and questionnaire logic. The chosen solution preserves the modularity, scoring transparency, and replicability envisaged in the DoA, while enabling branching, session tracking, and repeated assessment more robustly than a static spreadsheet. The tables provided in the Appendices retain an Excel-compatible representation of the indicator blocks and technical-measure mappings. These tables allow expert teams to inspect, export, or reimplement the mechanism even where they do not use the reference codebase.

The report begins by showing how the goals of T3.4 have been achieved. It then explains the background of the ERL approach, including its relationship to previous MultiRATE readiness-level work and why AIOLIA reframed it as an ethics-by-design mechanism. The following section documents AIOLIA activities and validation work. The report then defines the conceptual shift from evaluation to ethics-by-design, followed by dedicated chapters on the healthcare module and public-administration AIA sub-tool.

2. Background: From MultiRATE to AIOLIA

2.1. ORIGIN OF THE ERL APPROACH

The ERL mechanism developed in AIOLIA expands previous readiness-level work by project partners, especially CEA and RISE. The direct predecessor was developed in the context of MultiRATE, a project concerned with multi-dimensional readiness assessment. (Unzueta et al., 2026). In addition to TRLs, MultiRATE considered related readiness scales such as integration, commercialisation, societal, security, privacy, legal, and ethics readiness. This broader readiness logic is important because AI systems often fail not only due to technical immaturity but also because legal, ethics, social, organisational, or security conditions are not ready for deployment.

The ERL (or LPERL – Legal, privacy, ethics readiness level in the MultiRATE nomenclature) component of this earlier work provided a way to evaluate legal, privacy, and ethics readiness (Adomaitis et al., 2024). However, in its earlier form it was primarily an evaluation tool. It was not embedded in development practice. This is where AIOLIA T3.4 picked up.

The conceptual foundation of ERLs is readiness-level thinking. Technology Readiness Levels were originally developed to support large technical programmes by making technological maturity visible and trackable (Mankins, 1995; Olechowski et al., 2015). The ERL method adapts this insight to ethics. Ethics maturity also develops over time, can be underestimated by teams, and should be assessed at points when design can still change. ERLs are closely aligned with an integration logic. Ethics readiness concerns the integration of a system with users, institutions, social environments, legal frameworks, values, and accountability structures. AI and ML-specific readiness work makes a similar move for systems whose maturity cannot be captured by generic software readiness alone (Eljasik-Swoboda et al., 2019; Lavin et al., 2022; de Jong, 2025).

For that reason, ERLs follow a progression from absence of ethics consideration to identification, characterisation, harmonisation through design, and control through accountability. This structure is reflected both in the readiness levels and in the indicator trees.

TRLs ask whether a technology works as a technology. ERLs ask whether the AI system is integrated into the human, institutional, regulatory, and social setting in which it will operate. For example, the integration and use of a facial recognition or clinical decision-support system cannot be read from technical performance alone. It depends on how the system is embedded in workflows, oversight arrangements, user practices, rights protections, and accountability structures. This integration emphasis follows systems-readiness and integration-readiness thinking (Sauser et al., 2006).

Insights to Experts (IEEE ZINC 2026)

"Ethics readiness is about integration too. An AI system's profile is not determined by its internal technical properties alone but by how it interfaces with users, social contexts, regulatory frameworks, and institutional processes. The ethics status of a facial recognition system depends on how it integrates with policing practices, judicial oversight, and civil liberties protections, not on its technical performance alone." (p. 2)

2.2. LIMITS OF FLAT EVALUATION

Even with that transformation from technology readiness to ethics readiness, there is another missing piece. That is the difference between a flat evaluation and an ethics-by-design process. The critique of flat evaluation is grounded in the wider AI ethics literature showing that principles and policy frameworks such as the HLEG guidelines and ALTAI (High-Level Expert Group on AI, 2019; European Commission, 2020), regulatory and legal-readiness instruments such as the adopted EU AI Act (European Parliament and Council of the European Union, 2024), and soft law alone (Terpan, 2015) do not guarantee responsible AI practices. This is consistent with critiques that principles and guideline inventories do not, by themselves, ensure implementation (Mittelstadt, 2019; Hagendorff, 2020).

In this deliverable, flat evaluation means an assessment approach where a system is checked against a list of concerns or readiness indicators without sufficient contextual branching, without explicit support for repeated assessment, and without converting the assessment into design action. Flat evaluation can be useful for producing a snapshot, but it is limited as an ethics-by-design mechanism.

Ethics issues in AI development do not appear once and remain static. A system's ethics profile changes as data sources change, models are updated, interfaces are redesigned, the intended users shift, new deployment contexts emerge, or regulation develops. This point requires anticipatory responsibility and anticipatory governance (Jonas, 1985; Guston, 2014). A one-off assessment may therefore provide false reassurance. It can also miss whether a system is improving or deteriorating over time. The ERL mechanism addresses this by logging sessions and comparing score trajectories across repeated assessments. It also reflects work on moving anticipatory ethics from speculative analysis into governance routines (Umbrello et al., 2023).

Static checklists are vulnerable to a box-ticking mentality. Teams may answer questions in a way that confirms existing assumptions rather than prompts deeper reflection. Generic checklists also tend to be either too broad to be meaningful or too long to be usable. AIOLIA's own findings on ethics principles (Bayerl et al., 2026) show that principles and components vary across use cases. This supports the use of more specific, context-sensitive principles (Kundu et al., 2023). For example, transparency may be an independent principle in one context and a component of non-bias or autonomy in another. This supports the need for a modular and contextual tool rather than a universal flat list.

Insights to Experts (IEEE ZINC 2026)

"An onboarding process collects contextual information before any indicators are presented. This information drives all subsequent contextual adaptation." (p. 3)

2.3. AIOLIA'S ADDED VALUE

AIOLIA's contribution is the adaptation of ERLs from readiness evaluation to ethics-by-design support. This adaptation has several dimensions.

First, AIOLIA enriches ERLs with project-specific work on ethics principles and use cases. D3.1 informed the indicators through principles and components such as privacy, autonomy or agency, non-discrimination and non-bias, transparency and explainability, oversight, safety and non-maleficence, accountability, and welfare or human well-being. D3.1 also informed the technical measures mapped to ERL indicators in Appendix D.

Second, D3.4 uses D3.1 not only as a source of technical measures, but also as a source of domain issues and validation needs. The mechanism asks whether a concern is relevant, whether a mitigation has been implemented, and whether that mitigation is evidenced or validated.

Third, D3.3 complements D3.4. As D3.3 generalises ELSE and organisational measures across research areas, D3.4 provides an assessment mechanism that can help evaluate whether such measures are identified, implemented, validated, and revisited over time. This was especially important for the public-administration AIA sub-tool.

Fourth, AIOLIA adds new modules for priority application areas. The healthcare module reflects healthcare use-case validation, while the public-administration AIA sub-tool reflects the need to assess systems that may affect citizens' rights, public services, administrative justice, and democratic accountability.

Finally, D3.4 can support WP4 as a training and learning resource. ERL sessions work well with innovators and technically minded participants because they make ethical reflection concrete. The first training-oriented use is planned during the AIOLIA trainings in Novi Sad on 2-3 June 2026, after which the mechanism can support other technically oriented learning activities.

3. AIOLIA Activities Under T3.4

3.1. INITIAL PREPARATION

The work under T3.4 began by reviewing previous ERL/LPERL work from MultiRATE and identifying how it could be adapted to the AIOLIA context. The starting point was that an evaluation mechanism already existed, but AIOLIA required it to become more operationally useful for ethics-by-design. The task therefore focused on adapting the tool, expanding its content, and connecting it to AIOLIA's use-case and training agenda.

The ERL mechanism was reviewed as an inherited methodology that could capture legal, privacy, and ethics readiness. Its strengths were the readiness-level structure, indicator-based assessment, and scoring model. Its limitations were the need for stronger domain adaptation, stronger ethics-by-design framing, and clearer pathways from evaluation results to design recommendations.

AIOLIA's previous work on operational ethics, especially the results reflected in D3.1 was used to enrich the ERL mechanism. This was important because AIOLIA studies how principles such as human oversight, autonomy, privacy, transparency, non-bias, safety, non-maleficence, accountability, and human well-being are interpreted in practical use cases.

3.2. FIRST EXPERT PANEL

The first expert panel took place online on 8 January 2026 and lasted 90 minutes. It brought together ethics, legal, policy, and responsible innovation expertise from Sheffield Hallam University, McGill University, CEA, and CEPS.

The agenda began with an introduction to previous MultiRATE work and the ERL methodology. The second part was an open forum on AIOLIA contributions and on how the inherited ERL approach should be adapted for the project.

The panel decided that a major AIOLIA contribution would be to derive technical measures from D3.1 and match them to concrete ERL indicators. The panel also confirmed the need for a healthcare module, a public-administration AIA module, a web implementation, and a Berlin plenary demonstration that would serve as the second expert consultation required by T3.4.

The outputs were integrated by translating the discussion into indicator changes, the AIOLIA technical-measures mapping, new domain modules, and the implementation agenda for the web-accessible reference mechanism.

3.3. DEVELOPMENT PHASE

After the expert panel, the development phase translated the decisions into tool content and structure. This phase included module drafting, indicator adaptation, scoring review, and preparation of demonstration material.

The healthcare AI block was drafted as a standalone AIOLIA module. In the indicator schema reproduced in Appendix B, it includes parent indicators numbered 24-34 and covers patient agency and informed decision-making, consent for AI training or modelling, clinician judgement, automation bias, bias in healthcare outcomes, patient health data, clinical validation, commercial or secondary data use, clinically useful justification, ethics oversight, liability, clinical validity, performance in subpopulations, workflow integration, documentation accessibility, and measurable operational or clinical value.

The public-administration AIA block was drafted as a standalone AIA sub-tool. In the version 2 indicator schema reproduced in Appendix C, it includes indicators numbered 101-114 and covers rights-affecting decisions, legal basis, proportionality, non-discrimination, citizen notice, explainability and explanation limits, human review, meaningful human authority, data rights, accuracy and reliability, democratic scrutiny, independent audit, vendor governance, digital inclusion, accountability and redress, and public-servant impact.

The software implementation itself was created during AIOLIA as part of the T3.4 development work. Implementation used manual coding together with Perplexity and Claude Code coding agents for coding support,

debugging, and implementation assistance. The methodological design, indicator content, scoring logic, and final review remained under human control.

The tool retained the ERL scoring logic but adapted it to the new modules. Indicators are stored with a number, text, yes score, no score, and block. The hierarchical numbering allows the questionnaire to branch dynamically. The score begins at 4 and is adjusted as the evaluation proceeds. Negative scores identify risks or missing capabilities; positive scores recognise mitigation; deeper negative scores can penalise missing validation.

3.4. AIOLIA PLENARY DEMONSTRATION

The mechanism was demonstrated on 7 March 2026 during the in-person AIOLIA consortium meeting in Berlin. Thirty minutes were dedicated to ERLs. All consortium members were represented, together with Scientific Advisory Board members and European Commission representatives.

The demonstration used a mock evaluation case with a reduced number of indicators to show a complete web-accessible implementation of the ERL methodology. The demo flow covered onboarding, evaluation, and analysis stages, including how selected blocks activate, how answers affect score progression, and how results are translated into reflection prompts.

The plenary served as the second expert consultation required by T3.4 and focused on implementation. The demonstration was received positively as a non-checklist method that encourages reflection. In the discussion, NIT noted similarities between ERL analytics and safety-evaluation analytics and suggested adding metrics such as the largest contributors to unresolved issues. This suggestion was later implemented in the software analytics.

The demonstration also supported the development of the methodology paper accepted for presentation at IEEE ZINC 2026 in Novi Sad on 3-4 June 2026 (Adomaitis, Israel-Jost, and Grinbaum, 2026).

3.5. HEALTHCARE VALIDATION ACTIVITIES

Two healthcare validation meetings were conducted in April 2026 with AIOLIA Healthcare Use Case 1 and AIOLIA Healthcare Use Case 2. The validation was immersive: the actual use cases were evaluated in their current state, and both partners responded to the first version of the AIOLIA healthcare module by working through the indicators.

The validation materials consisted of the draft healthcare indicator block, the current state of each healthcare use case, and the scoring logic. Each indicator was reviewed during the process, so the meetings tested both the content of the module and the usability of the dialogue-led method.

The first validation, on 27 April 2026 with AIOLIA Healthcare Use Case 1, produced several changes. It clarified that automation can itself be a quality improvement in some healthcare settings, but that adoption may depend more on bottlenecks, throughput, reimbursement, and operational value than on pure quality arguments. This insight informed the indicator on measurable operational or clinical value.

The same validation clarified that training alone is insufficient to address over-reliance. Operational safeguards such as mandatory verification, second-reader checks, override documentation, and monitoring may be required. It also distinguished general technical explainability from clinically useful justification and led to more operational wording of liability and clinical justification indicators.

The second validation, on 29 April 2026 with AIOLIA Healthcare Use Case 2, provided further wording and conceptual feedback. Participants noted that "patient autonomy" can be interpreted as physical autonomy, while "agency" or "decision-making" better captures the intended concern. The meeting also distinguished a tool designed to replace clinician judgement from a tool that becomes a replacement in practice through over-reliance.

The feedback also distinguished self-explainability from internal explainability and suggested that diagnostic indicators should not assume that every healthcare AI system is diagnostic. Indicator 31 was reformulated toward justification in general use, including intended use, demonstrated outcome impact, and statistically valid evidence. The phrase "patient values" was removed because it was undefined and variable.

Engagement governance was proportionate to the activity. No separate ethics approval was required because the validation was conducted with internal consortium partners for tool-development purposes and did not

collect personal or sensitive data from research subjects. Separate consent procedures were therefore not applicable. Each validation session is accompanied by freeform text notes stored locally and securely on RISE equipment.

3.6. PUBLIC-ADMINISTRATION AIA REVIEW

The public-administration AIA sub-tool was reviewed by CENTRIC, with results returned to RISE on May 21, 2026. Following this review, the tool was revised into version 2, and the resulting changes have been implemented in the indicator table reproduced in Appendix C. The review focused on whether the AIA block was sufficiently adapted to public-administration contexts, where AI systems may affect citizens through eligibility decisions, prioritisation, service allocation, case processing, risk scoring, or administrative recommendations.

The earlier citizen-information indicator, formerly treated as a separate 101.3 item, was moved into the later transparency and notification indicator 103.1 so that citizen-facing information duties are grouped together. This makes the structure more coherent. The separate 101.3 row was removed, and the scoring of 101.1 and 101.2 was rebalanced to +0.15 each. In 103.1, the earlier wording “processing their case” was replaced because it was too narrow and implied a specific administrative workflow. The revised formulation, “processing their data or contributing to a decision that affects them,” better covers the range of public-administration AI uses, including systems that influence decisions indirectly.

Several changes were also made to the explainability indicators. The indicator on plain-language explanations, 104.1, was clarified so that explanations must be appropriate to the affected citizens’ expected knowledge, needs, and context. This avoids assuming a single generic public audience and recognises that explanation needs differ depending on the service. A new indicator, 104.3, was added to address the limits of explanation. This is important because public authorities may face different constraints on what can be explained. The revised tool does not assume that full explanation is always possible, but asks whether such limits are justified. The scores for 104.1, 104.2, and 104.3 were adjusted to -0.1 each to distribute the explainability burden across accessibility, meaningfulness, and transparency about limits.

The human oversight and accountability indicators were slightly revised to avoid specific institutional assumptions. The procurement and vendor-governance indicators were substantially reframed. Indicator 111 was changed so that external procurement is not treated as inherently negative. The revised formulation instead focuses on the governance risks that can arise when public authorities depend on external providers. This makes the indicator more balanced and better aligned with real public-sector procurement practice. Indicator 111.2 was likewise generalised. Rather than focusing narrowly on government ownership or the ability to switch vendor, it now asks whether the public authority retains sufficient control. This better captures the underlying requirement that citizens should not be exposed to unaccountable or unreviewable systems simply because critical knowledge or operational control sits outside the public institution.

After these changes were incorporated, a clean v2 version was made, is now live on the Github repository, and is reproduced here as Appendix C. The public-administration AIA sub-tool is ready for further validation in T4.3 citizen engagement workshops organised by CENTRIC. This next validation route is appropriate because the module concerns public-sector AI systems that may directly affect citizens. Expert review supports methodological quality and public-administration fit, while citizen engagement is needed to test whether the tool adequately captures notice, accessibility, appeal, redress, democratic legitimacy, and institutional trust from the perspective of those affected.

4. From Evaluation to Ethics-by-Design

4.1. CONCEPTUAL SHIFT

The main conceptual shift in T3.4 is from evaluation to ethics-by-design. In a flat evaluation model, a tool produces a judgement about a system's readiness or compliance. The output is often a score, checklist result, or traffic-light classification. Such outputs can be useful, but they may arrive too late or remain too detached from design. In an ethics-by-design model, the assessment is embedded in the development process. This follows ethics and values-in-design approaches, where values are operationalised through design decisions rather than appended after deployment (Van den Hoven et al., 2015; Friedman and Hendry, 2019). It helps identify issues while the system is still malleable, translates ethics concepts into operational design questions, and tracks whether mitigation and validation are improving over time. It also follows the view that embedded values need to be addressed during design (Moor, 1985; Brey, 2010).

AIOLIA's ERL/AIA mechanism therefore treats evaluation as a **structured intervention** in design. The score is not a verdict but a way to organise the conversation. The mechanism can be used at different TRL stages. At early TRLs, a team may not yet have implemented mitigations, but the tool can still reveal whether ethics issues have been identified and characterised. At later TRLs, the tool can test whether mitigations have been implemented, whether they are validated, and whether control mechanisms exist.

This allows ethics to influence design before deployment. For example, if a healthcare AI system has not addressed automation bias, the tool points toward the design need for operational safeguards, human verification, override documentation, monitoring, or second-reader workflows.

The mechanism uses indicators that are organised into blocks and trees. The onboarding process selects relevant blocks. The answers to parent indicators determine whether follow-up indicators appear. This reduces unnecessary burden and prevents irrelevant questions from dominating the session.

Most importantly, the indicators are prompts for dialogue. A yes/no answer is required for scoring and branching, but the value of the session lies in the discussion that leads to the answer. Participants are expected to clarify what the system does, what evidence exists, what assumptions are being made, and what design changes would improve readiness. AI ethics principles such as transparency, fairness, accountability, privacy, autonomy, and human oversight are important but often too abstract to guide design directly (Floridi et al., 2018). AIOLIA's use-case work shows that these principles become meaningful only when specified in context.

4.2. DESIGN-TIME FUNCTION OF ERLS

Parent indicators identify whether a category of ethics concern applies. For example, a healthcare indicator asks whether the AI system automates clinical recommendations or risks replacing clinician judgement. A public-administration indicator asks whether the AI system makes or substantially influences decisions that directly affect citizens' legal rights, entitlements, or obligations.

Once an issue is identified, the mechanism helps characterise its specific form. This is analogous to methods that identify and specify concrete ethical issues and impacts in emerging technology assessment (Adomaitis et al., 2022). For example, public-administration discrimination risk is connected to testing training data and model outputs across demographic groups, addressing identified biases, and monitoring outcomes after deployment. Healthcare over-reliance is connected to the relation between clinician judgement, tool design, override possibilities, and operational safeguards.

Mitigation indicators ask whether concrete measures are in place. These measures may be technical, organisational, procedural, or communicative. Examples include bias audits, cybersecurity protocols, clinician override mechanisms, citizen appeal mechanisms, procurement clauses, or independent audit access. Because mitigation indicators contribute positive score recovery, the mechanism makes design progress visible. This is where Responsible Research and Innovation becomes operational in governance and practice (von Schomberg, 2013; Stahl et al., 2017). Teams can see which actions improve readiness. Responsibility is therefore treated as a feature of how innovation processes are organised (Grinbaum and Groves, 2013). Technical and organisational measures also matter because architecture itself can regulate possible behaviour (Lessig, 2000).

Validation indicators ask whether the mitigation works. A claimed safeguard that has not been tested may create false confidence. The mechanism therefore penalises unvalidated claims in some subtrees. This helps prevent ethics-washing, where ethics language substitutes for substantive organisational or technical change (Bietti, 2021).

Control mechanisms include oversight, audit, accountability, redress, reporting processes, training, and periodic reassessment. They are essential for higher ERL levels because ethics readiness is not only about initial design but about maintaining responsible operation over time. This is important for AI because responsibility can be distributed across developers, deployers, institutions, users, and automated systems (Floridi, 2016; Grinbaum, 2019).

4.3. DIALOGUE AS A METHODOLOGICAL REQUIREMENT

The ERL method requires dialogue. This is a hard requirement for the validity of the mechanism. An ethics expert brings knowledge of ethics concepts, regulatory context, and risk patterns. A technical or domain expert brings knowledge of how the system works, what data it uses, how users interact with it, what constraints exist, and what mitigations are realistic. The tool is designed to structure the interaction between them. The dialogue requirement also recognises that ethical labour is distributed across technical and ethics expertise (Politi and Grinbaum, 2020). Its philosophical basis is close to dialogical ethics, where ethical understanding arises through encounter with perspectives other than one's own (Buber, 1958; Levinas, 1991).

Self-assessment is discouraged because it is vulnerable to confirmation bias, cognitive dissonance (Festinger, 1957), social desirability, Dunning-Kruger effects, and lack of awareness. A developer or product owner may answer "yes" based on intention rather than evidence, or may interpret ambiguous indicators in a favourable way. Conversely, an ethics expert alone may misunderstand technical constraints. Dialogue reduces these risks by bringing assumptions into the open and introducing an external perspective that cannot simply be assumed away.

Insights to Experts (IEEE ZINC 2026)

"Each participant's understanding is enriched and challenged, producing a shared evaluation more accurate than either could achieve alone. […] Another motivation for requiring dialogue is bias correction.
Self-evaluation is distorted by cognitive dissonance, confirmation bias (interpreting ambiguous indicators), social desirability effects (answering as one thinks one should rather than as one has), and Dunning-Kruger effects (overestimating ethics awareness in areas of limited expertise). The dialogue format mitigates all of these biases by introducing an external perspective that cannot be assumed away." (p 3-4).

The score is useful because it tracks progress, highlights gaps, and motivates improvement. However, the primary value is the structured reflection produced by the session. This is grounded in the philosophical traditions of Martin Buber and Emmanuel Levinas, both of whom argue that ethics understanding arises through encounter with the Other (the person whose perspective and concerns are different from one's own). Each participant's understanding is enriched and challenged, producing a shared evaluation more accurate than either could achieve alone.

5. Healthcare AI Module

5.1. RATIONALE FOR SELECTING HEALTHCARE

Healthcare was selected as a priority module because AI in healthcare is a high-impact domain where ethics, legal, technical, clinical, and organisational concerns are tightly connected. A healthcare AI system may affect diagnosis, treatment planning, patient safety, clinician responsibility, hospital workflow, resource allocation, and trust in medical institutions.

Healthcare also fits AIOLIA's broader focus on operational ethics. AIOLIA aims to co-create practical ethics guidelines for use cases and AI research areas. Healthcare provides a test case for whether abstract principles can be translated into domain-specific indicators. Concepts such as autonomy, explainability, fairness, safety, and accountability take distinct forms in clinical settings.

The module was also selected because AIOLIA Healthcare Use Case 1 and AIOLIA Healthcare Use Case 2 could provide relevant feedback. The validation meetings directly influenced the healthcare module. It was refined against practical concerns such as clinical workflow, reimbursement, adoption incentives, blind review practices, over-reliance, and the difference between technical explainability and useful clinical justification.

5.2. MAIN ETHICS ISSUES ADDRESSED

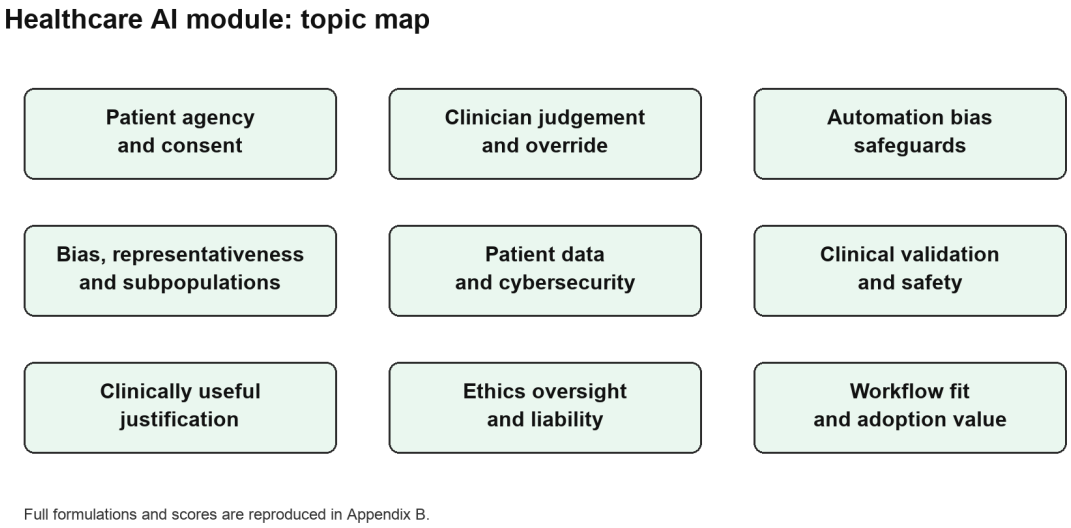

Figure 1. Healthcare AI module topic map.

The healthcare module contains indicators 24-34 in the indicator schema reproduced in Appendix B. It is designed as a standalone healthcare AI block that translates healthcare AI concerns into operational yes/no indicators with scores.

The module begins with patient agency and informed decision-making. The feedback from partners suggested that "autonomy" may be ambiguous in healthcare because it can sound like physical autonomy. The intended concern is patient agency and informed decision-making. The module therefore asks whether the AI system affects patient agency or decision-making and whether patients are provided with meaningful information about the AI tool to support informed decisions.

It also asks whether informed consent is obtained before using patient data for AI training or modelling and whether electronic or dynamic consent mechanisms allow patients to opt out. This connects patient agency to data use, not only to clinical decisions.

A central healthcare concern is whether AI complements or replaces clinician judgement. The module asks whether the AI system automates clinical recommendations or replaces clinician judgement, whether it is

designed as a complement, and whether clinicians can override recommendations without professional penalty (Kelly et al., 2019; Rajkomar et al., 2019).

The module also addresses automation bias and the risk that clinicians become over-reliant on the AI. The over-reliance concern has a long history in human-computer interaction, visible already in Weizenbaum's ELIZA case, where users attributed more understanding to a simple program than it possessed (Weizenbaum, 1966). Validation feedback strongly indicated that training alone is insufficient. The Appendix B indicator therefore asks whether operational safeguards exist beyond training, such as mandatory human verification, override documentation, second-reader checks, or monitoring.

Healthcare AI can produce biased or discriminatory outcomes if training data are not representative or if model performance differs across patient groups. The module asks whether outputs can result in biased or discriminatory healthcare outcomes, whether representative datasets were used across demographics, whether multi-institutional data collaborations were used to reduce underrepresentation, and whether regular audits detect and correct racial or gender bias after deployment (Obermeyer et al., 2019).

It also asks whether performance differs across clinically relevant subpopulations or atypical anatomies, whether the tool was assessed for disparities in patients with rare comorbidities or atypical anatomies, and whether identified performance disparities have been addressed through retraining, recalibration, or additional data collection.

The module asks whether the system processes or stores patient health data and whether cybersecurity protocols and privacy-preserving measures are implemented. It also asks whether there is a risk that data could be used for unauthorised commercial or secondary purposes and whether institutional safeguards prevent such use.

This goes beyond general GDPR logic by recognising the particular sensitivity of health data and the trust relationship between patients, clinicians, hospitals, and AI developers.

The module distinguishes between technical explainability and clinically useful justification. Validation feedback made clear that an explanation can be technically sophisticated but practically useless for clinical decision-making. The module therefore asks whether model complexity or design choices prevent clinically useful justification of outputs at the point of care (Amann et al., 2020), whether clinicians can access clear case-level rationale they can use to defend and communicate AI-informed decisions, and whether developers have published documentation on assumptions, limitations, and data provenance.

The module also asks whether the AI system can justify its clinical recommendations, whether outputs are supported by reasons aligned with professional medical standards, and whether individual recommendations can be justified by intended use, demonstrated outcome impact, and statistically valid evidence.

The module asks whether the AI system was developed without formal ethics oversight and whether bioethicists or ethics experts are part of the development team. It also asks whether an algorithmic or human rights impact assessment was conducted before clinical deployment, whether standardised healthcare AI ethics checklists and guidelines are used, and whether there is a framework for assigning liability and compensation for AI-induced medical errors.

Clinical validity and workflow fit are also addressed. The module asks whether the system is clinically valid and technically accurate, whether it reflects the patient's clinical state, whether it has been tested for specific tasks, whether outputs and documentation are accessible within the clinical workflow, and whether risk scores, heatmaps, or image overlays fit into physicians' daily practice.

Finally, the module includes an adoption-value indicator. This indicator reflects partner feedback that hospitals often invest to remove bottlenecks and improve throughput, not to add abstract quality.

6. Public-Administration AIA Sub-Tool

6.1. RATIONALE FOR SELECTING PUBLIC ADMINISTRATION

Public administration was selected because AI systems in this domain can directly affect citizens' rights, entitlements, obligations, access to services, and ability to contest decisions. Public-sector AI raises issues that go beyond ordinary consumer AI. It involves state power, administrative justice, democratic legitimacy, public accountability, and equal access to services.

The EU AI Act strengthens the relevance of this domain. It introduces obligations for high-risk AI systems and, under Article 27, requires fundamental rights impact assessments for a narrower set of deployers, including bodies governed by public law and private actors providing public services, when they deploy certain high-risk systems listed in Annex III. Recital 96 frames this as a tool for identifying effects on fundamental rights before and during deployment. The public-administration AIA sub-tool is therefore aligned with the adopted AI Act's concern with rights-facing uses, deployer obligations, oversight, and redress (European Parliament and Council of the European Union, 2024).

Existing AIA approaches support this direction, notably Canada's Algorithmic Impact Assessment and the Council of Europe's HUDERIA methodology, developed under the CAI in support of the Framework Convention on AI. The AIOLIA public-administration AIA sub-tool adapts these ideas to the ERL mechanism by using indicators, branching, scoring, and dialogue.

6.2. MAIN AIA DIMENSIONS

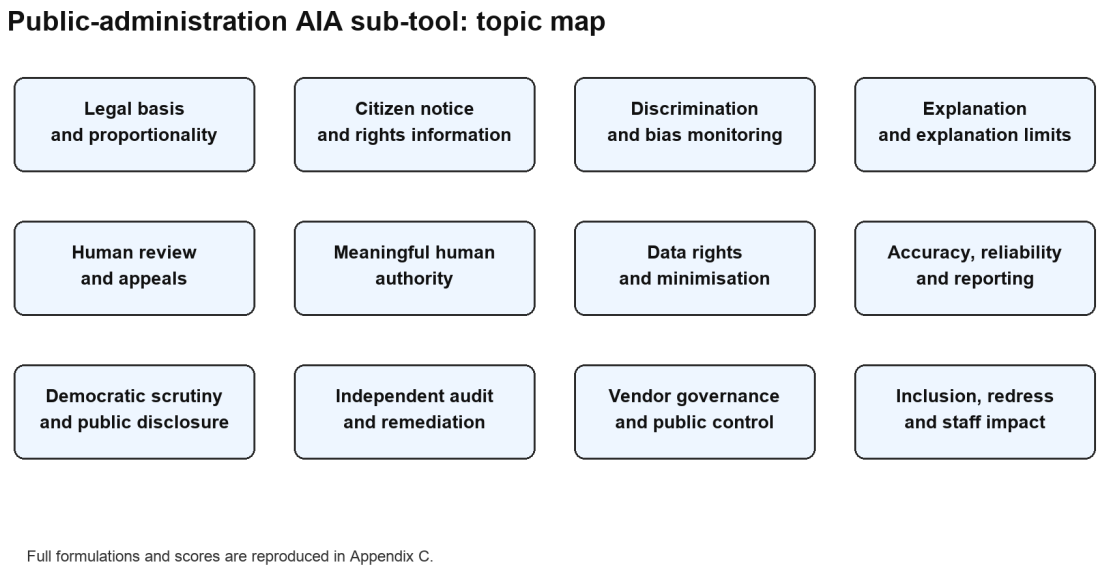

Figure 2. Public-administration AIA sub-tool topic map.

The public-administration AIA block or sub-tool contains indicators 101-114 in the version 2 indicator schema reproduced in Appendix C. It is a standalone AIA-oriented block that can be used for public-administration systems and further validated through T4.3 citizen engagement workshops.

The block begins with decisions that directly affect citizens' legal rights, entitlements, or obligations. If such decisions are in scope, the tool asks whether there is an explicit legal basis and whether a proportionality assessment has justified the use of AI given the stakes for citizens (Citron and Pasquale, 2014; Kroll et al., 2017).

Citizen information duties are grouped together. The block asks whether citizens are given sufficient information about the role of AI, whether they are explicitly informed when AI processes their data or contributes to a

decision that affects them, whether known limitations and error possibilities are communicated, and whether citizens are informed of their rights.

Explainability is treated as a practical condition for contestability (Kroll et al., 2017). The block asks whether the AI system's contribution to a specific decision can be explained, whether explanations are appropriate to affected citizens' expected knowledge, needs, and context, whether explanations are available upon request without legal action, and whether legal, technical, or organisational limits on explanation are identified, justified, and communicated where relevant.

The block gives strong weight to non-discrimination. It asks whether the system can produce discriminatory outcomes, whether training data and outputs have been tested across demographic groups, whether identified biases have been addressed, and whether discriminatory patterns are monitored after deployment (Selbst et al., 2019).

Human review and meaningful human authority are assessed separately. Citizens should be able to request review by a human official, and officials or staff responsible for the decision should be trained to critically evaluate AI recommendations rather than accept them uncritically. The block also asks whether there is documented evidence that oversight is exercised in practice.

The remaining dimensions address personal-data rights, accuracy and reliability in the deployment context, reporting to designated oversight or governance bodies where available, democratic or parliamentary scrutiny, consultation with affected communities or civil society, public disclosure, independent audit, remediation timelines, vendor-governance risks, digital inclusion, accountability, redress, and impacts on public servants' work (Busuioc, 2021; Raji et al., 2020; Wieringa, 2020).

7. Tool Definition and Methodology

7.1. ETHICS READINESS LEVELS

The ERL scale describes the maturity of ethics integration in an AI system or AI-related research activity. It has five levels: ERL 0 through ERL 4.

ERL 0: ethics considerations lacking

ERL 0 indicates that ethics, legal, and privacy considerations are lacking.

ERL 1: ethics issues identified

ERL 1 indicates that the team has identified key ethics and privacy issues.

ERL 2: ethics interactions characterised

ERL 2 indicates that ethics interactions and tensions have been characterised.

ERL 3: ethics-by-design compatibility

ERL 3 indicates that ethics and privacy considerations have been identified, characterised, and integrated into a coherent design.

ERL 4: control over ethics issues

ERL 4 indicates that the system has sufficient control mechanisms to manage ethics, legal, and privacy issues and ensure accountability.

7.2. DYNAMIC QUESTIONNAIRE DESIGN

The questionnaire is dynamic because it adapts to the evaluated system. The assessment begins with contextual questions. These identify whether the system is a healthcare AI system, law enforcement system, personal-data-processing system, AI system, or public-administration system. The answers determine which indicator blocks are activated.

Indicator blocks group related indicators by application area, technology type, or regulatory/ethics domain. The core blocks include:

- general ethics; (AIOLIA-improved)
- GDPR/privacy; (AIOLIA-improved)
- law enforcement; (AIOLIA-improved)
- AI Ethics; (AIOLIA-improved)
- healthcare AI; (AIOLIA-owned)
- public-administration AIA. (AIOLIA-owned).

This block structure reduces user burden by preventing irrelevant indicators from being shown.

Within each block, indicators are organised by hierarchical numbers. For example, indicator 2 may have children 2.1, 2.2, and 2.2.1. A "yes" answer may lead deeper into a subtree when a risk or mitigation is present. A "no" answer may skip follow-up questions or trigger penalties when a required capability is missing. The tree structure has two benefits. It saves time by skipping irrelevant questions, and it creates logical progression from broad issue identification to specific mitigation and validation.

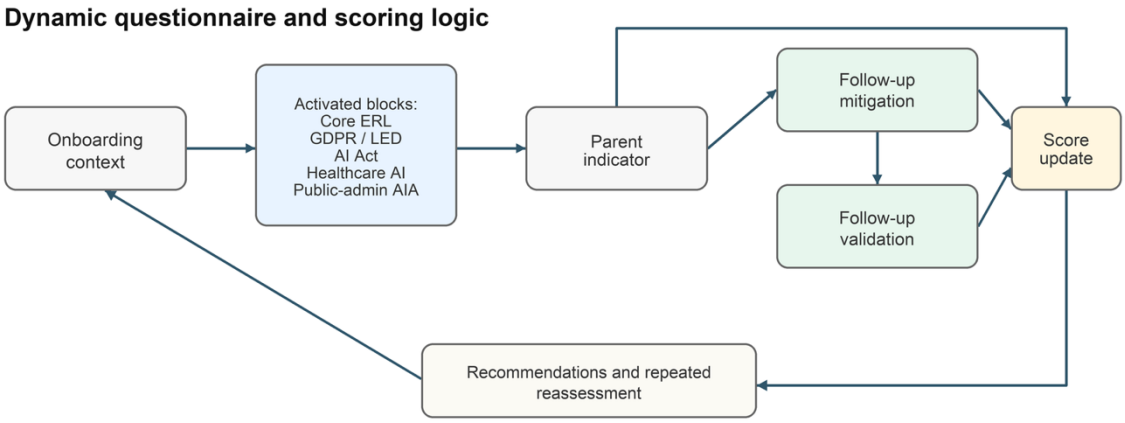

Figure 3. Dynamic questionnaire and scoring logic.

7.3. INDICATOR LOGIC

The indicator logic has three layers: relevance, mitigation, and validation/control. The concepts operationalised by the tool are "thick concepts" (Williams, 2006). Terms such as fairness, transparency, privacy, and accountability describe a state of affairs while also carrying a normative judgement. This matters for indicator design. Indicators must be concrete enough for a technical team to answer, but open enough to trigger discussion about how the concept applies in the specific system under review. Their translation into indicators follows the idea of specifying norms for concrete ethical problems (Richardson, 1990). The same term, for example transparency, can therefore require different operational meanings in AI-assisted criminal identification and blockchain systems (Mahesh Kumar et al., 2025; Tang et al., 2020). For this reason, the method avoids loading each indicator with long definitions. When a participant asks "how does this apply to our product?", the correct answer is often not a pre-written definition but a structured dialogue.

Relevance indicators ask whether an issue applies. For example:

- Can the product influence the user's decision-making?
- Does the AI system impact patient autonomy?
- Can the AI system produce discriminatory public-administration outcomes?

If the answer indicates that a risk is present, the score may drop and follow-up indicators become relevant.

Mitigation indicators ask whether measures exist to address the issue. For example:

- Have safeguards been implemented to prevent unintended effects on autonomy?
- Is the healthcare AI tool designed as a complement to clinician judgement?
- Were training data and model outputs tested for bias?

Positive answers to mitigation indicators can recover points lost by risk identification.

Validation indicators ask whether mitigations are effective, evidenced, or under control. For example:

- Have penetration tests or red-team exercises been conducted?
- Have identified performance disparities been addressed?
- Is there documented evidence that human oversight is exercised in practice?

This layer prevents teams from claiming ethics maturity based only on intentions or untested safeguards.

For each positive answer, assessors should record the evidence basis, such as a policy document, technical artefact, audit report, validation study, training record, user-facing documentation, deployment log, or expert testimony. Where evidence is missing, the answer should be treated as planned or unvalidated rather than fully implemented.

Insights to Experts (IEEE ZINC 2026)

"The methodology operationalizes thick ethical concepts into indicators. It uses a risk-mitigation-validation scoring mechanism that rewards validated mitigation. It requires dialogue between an ethics expert and a technical expert to encourage ethical reflection and correct for self-evaluation bias." (p. 1)

7.4. SCORING LOGIC

The scoring system converts the qualitative dialogue into a quantitative trajectory. Each assessment starts at a score of 4. The system is initially assumed to be fully ready, and the assessment tests that assumption. Risks and missing capabilities reduce the score; mitigation and validation can recover it.

The maximum-start scoring choice does not remove the evidential burden from the assessed team. It is used because ERL branching first asks whether a risk is applicable to the specific system. A skipped subtree means that the risk was not activated in that session, not that the team earned maturity credit. Where a risk is identified, points are deducted, and recovery requires evidenced mitigation or validation. If the assessor lacks evidence for a positive answer, the conservative interpretation is that the measure is planned or unvalidated, not fully implemented.

When a risk is identified, the score decreases. The weight depends on the severity of the risk. For example, in the general ethics block, technical design causing significant damage has a high negative score. In the healthcare block, biased or discriminatory healthcare outcomes carry a strong negative score. In the public-administration block, discrimination also has a strong negative score.

When mitigation measures exist, the score increases. This recovery is often partial because mitigation does not erase all residual risk. For example, cybersecurity standards, additional security measures, and misuse reporting can recover part of a damage-risk deduction, but not necessarily all of it. However, all mitigation measures collectively can always regain all the negative of the parent question.

If a team claims mitigation but cannot validate it, the score may decrease again. This is intended to avoid ethics-washing, where ethics language substitutes for substantive organisational or technical change (Bietti, 2021). A claimed safeguard without evidence may be less mature than an acknowledged gap because it creates false confidence.

The mechanism generates maturity recommendations by tracing readiness gaps back to activated parent indicators, negative scoring events, missing mitigation indicators, and safeguards that have not yet been validated. It then prioritises recommendations by identifying which ethics domains contribute most to reduced readiness at block level. In repeated sessions, earlier priorities can be checked to determine whether they have been addressed or remain unresolved.

Final ERL classification

Table 1. ERL score-to-level mapping (s = score achieved in the assessment)

Score range	ERL level	Meaning
<= 0	ERL 0	Ethics, legal, and privacy considerations lacking
0 < s <= 1	ERL 1	Ethics and privacy issues identified
1 < s <= 2	ERL 2	Ethics interactions characterised
2 < s <= 3	ERL 3	Ethics-by-design compatibility
s > 3	ERL 4	Control over ethics, legal, and privacy issues

7.5. TRACKING OVER TIME

The mechanism is designed for repeated use. Each assessment session records the selected blocks, answers, score changes, and final score. This allows the same system to be reassessed later. The score progression within a session shows how the score changes question by question. This reveals which indicators produce the largest losses, which mitigations recover readiness, and where validation gaps remain.

The most meaningful output is the trajectory rather than a single score. A system may regress after an architectural change, staff turnover, data change, or new deployment context. Conversely, a lower-scoring early-stage system may show genuine ethics maturation if repeated sessions show that it is identifying issues, implementing mitigations, and validating controls. The score is therefore useful primarily as a time-sensitive management signal.

Insights to Experts (IEEE ZINC 2026)

"A single ERL score is informative but insufficient. The trajectory between successive evaluations carries the most important information. The tool is designed for repeated evaluation at regular intervals (every six months is recommended) and stores the complete score progression of each session." (p. 4)

8. Code Implementation and Replication

8.1. IMPLEMENTATION AVAILABILITY

A code implementation/reference implementation exists in the GitHub repository:

`https://github.com/LA-NS/ethics-readiness-levels/tree/v0.2-dev`

The public repository `LA-NS/ethics-readiness-levels` describes Ethics Readiness Levels as an iterative method to track how ethical reflection is implemented in AI system design. It presents the methodology as a dynamic, tree-like questionnaire built from context-specific indicators and emphasises that the tool facilitates structured dialogue. The repository also describes modular indicator blocks, progressive scoring, and the importance of tracking progress over time. The open-source implementation and documentation is a public basis for reuse. The `v0.2-dev` branch is the relevant development reference for the AIOLIA-adapted work described in this deliverable.

For the purpose of this deliverable, the implementation should be understood as a **reference mechanism**. It shows how the method can be encoded in indicator files, database schema, questionnaire navigation, scoring logic, and result tracking. It is not presented as a publicly hosted service for individual self-assessment.

8.2. HIGH-LEVEL ARCHITECTURE

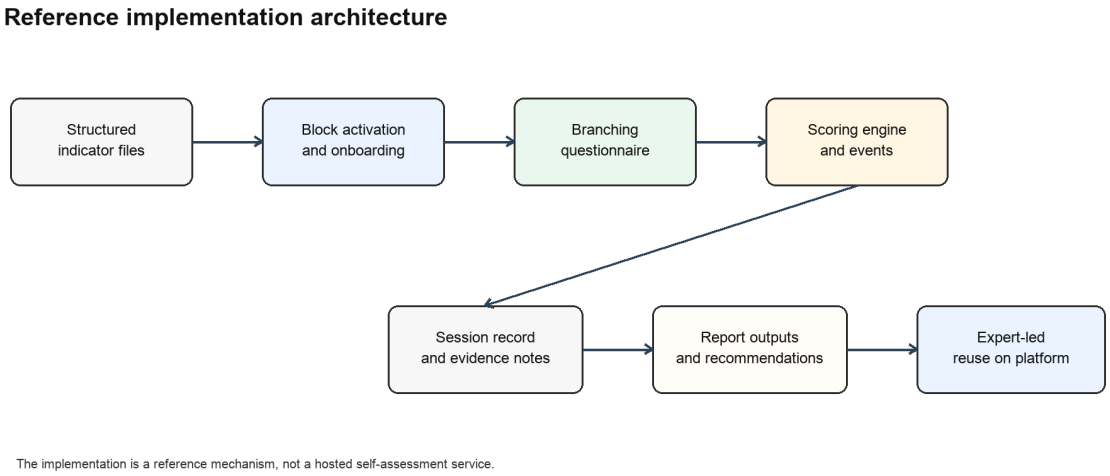

Figure 4. Reference implementation architecture.

The implementation can be described at a high level through five components.

Indicators are stored with:

- a hierarchical number;
- question or indicator text;
- yes score;
- no score;
- block identifier.

In the local schema, these fields are represented in a `questions` table. The hierarchical number creates the tree structure. The block identifier determines which module the indicator belongs to.

The tool selects active blocks based on onboarding information. A general ethics block is always included. Additional blocks are activated depending on whether the system involves personal data, law enforcement, AI, healthcare AI, or public administration.

The healthcare route can activate a specialised healthcare block instead of the generic AI route. The public-administration AIA route activates the public-sector AIA block.

Each assessment is treated as a session. A session records start and end time, answers, score progression, and final score. Answers are linked to the session and to the question. At the end of the session, the users can download an exit report, and the data is not further saved to the cloud.

The branching engine uses hierarchical indicator numbers. A "yes" answer can lead to a child indicator, such as from 2 to 2.1. If no child exists, the system moves to the next sibling or climbs back up the tree to find the next available indicator. If no further indicator exists in the block, the assessment moves to the next active block.

Each answer updates the score by applying the corresponding yes or no score. The score progression is stored and visualised at the end. Results can include final ERL, score trajectory, negative and positive score events, top concerns, block-level contributions, and improvement priorities.

8.3. REFERENCE IMPLEMENTATION JOURNEY

The screenshots below document the user journey of the v0.2-dev reference implementation. They show how the artefact moves from methodological framing, to participant setup, to indicator cards with AIOLIA technical measures, to final scoring, expert recommendations, and post-assessment analytics.

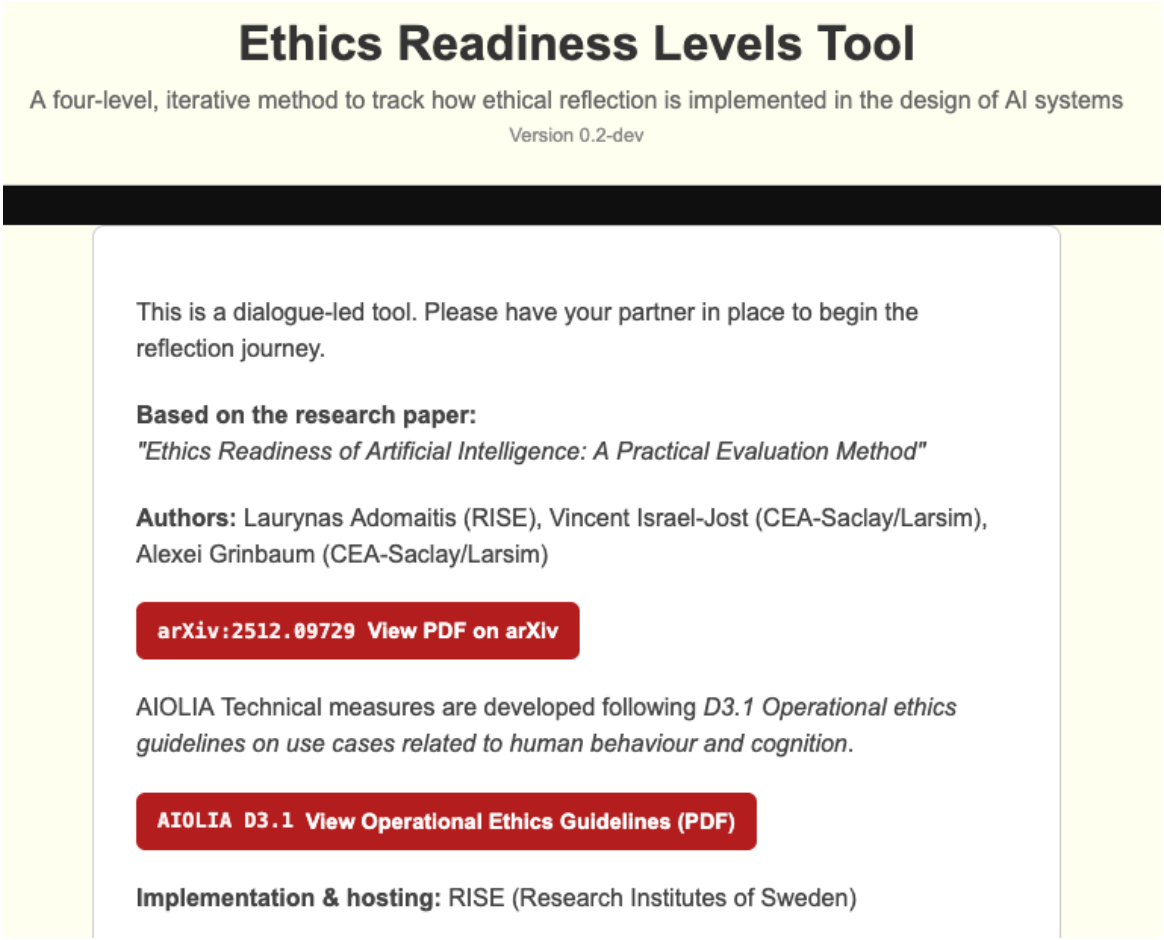

Figure 5. Reference implementation welcome and artefact attribution screen.

Welcome and attribution. The opening screen frames the implementation as the Ethics Readiness Levels Tool, identifies the 0.2-dev implementation version, and anchors the artefact in the methodology paper and AIOLIA D3.1. The design choice is deliberately documentary. Before a user answers anything, the tool states that the assessment is dialogue-led, points to the research basis, and identifies implementation and hosting responsibility.

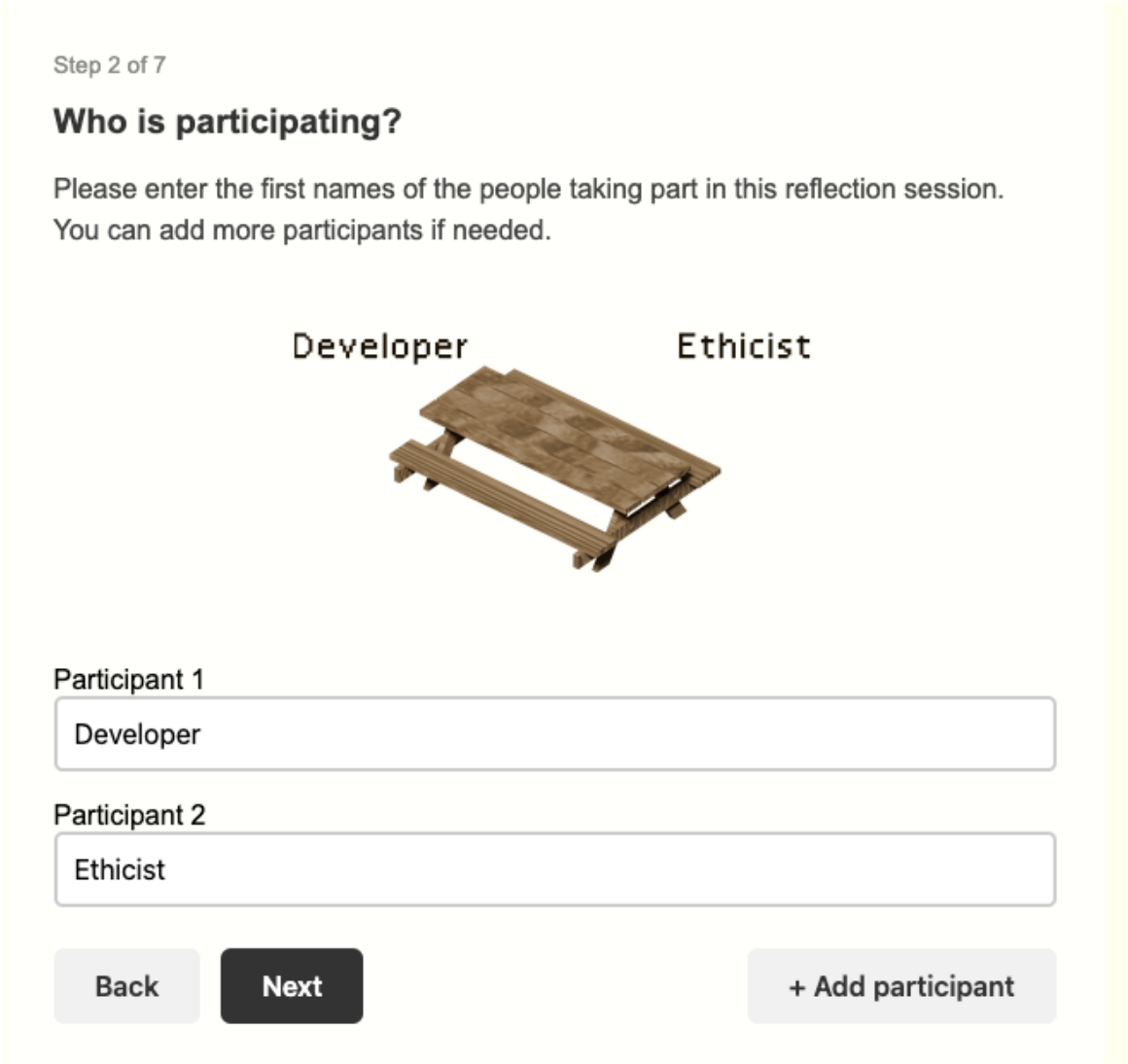

Figure 6. Participant setup screen.

Participant setup. The second step asks for the people taking part in the reflection session. The default Developer and Ethicist labels express the methodological requirement that ERL assessment is not a solitary checklist. The small shared-table visual reinforces the idea that the session is a conversation between complementary forms of expertise, while the add-participant control keeps the design open to domain experts, legal experts, or other relevant participants.

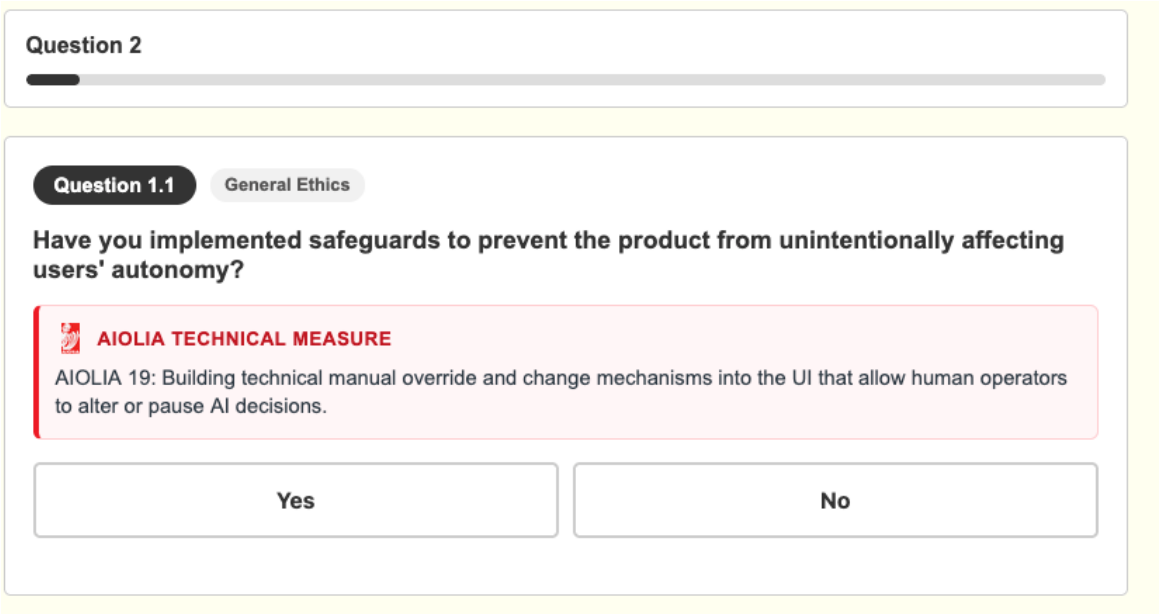

Figure 7. Question card with linked AIOLIA technical measure.

Question card and technical measure. The question view shows the current step, progress bar, indicator number, indicator block, question text, and binary answer controls. The important design choice is the red AIOLIA Technical Measure panel. It connects an abstract ethics indicator to a concrete measure distilled from AIOLIA D3.1, for example technical manual override and change mechanisms that allow human operators to alter or pause AI decisions. This is how the reference implementation makes ethics-by-design operational rather than leaving the user with a principle alone.

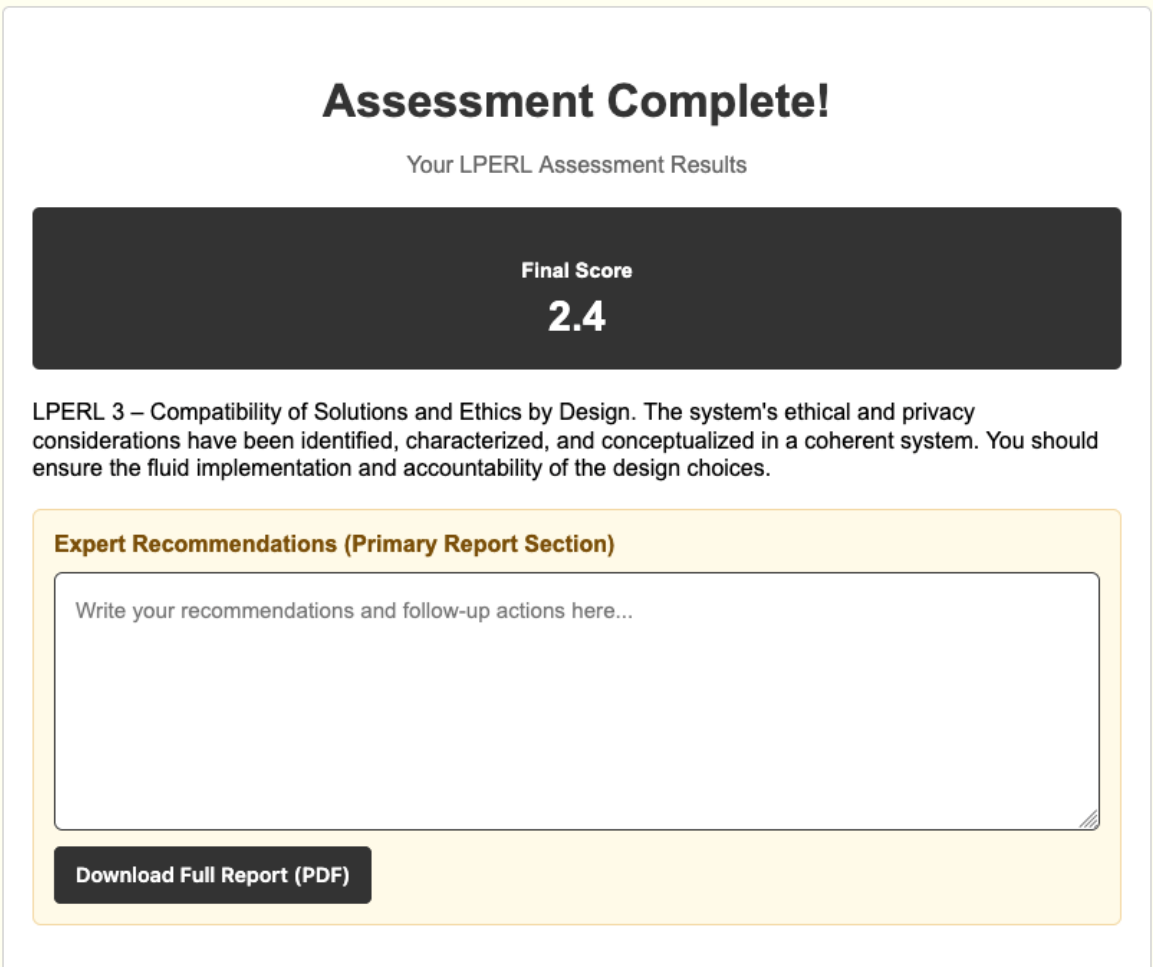

Figure 8. Completion screen and expert recommendation field.

Completion and recommendations. The completion screen reports the final score and corresponding readiness level, but the primary report section is the expert recommendation field. This reflects the methodological position of the deliverable. The session's value lies in the expert-led interpretation of gaps, follow-up actions, and design changes. The PDF export makes the session output portable without turning the score into a certification mark.

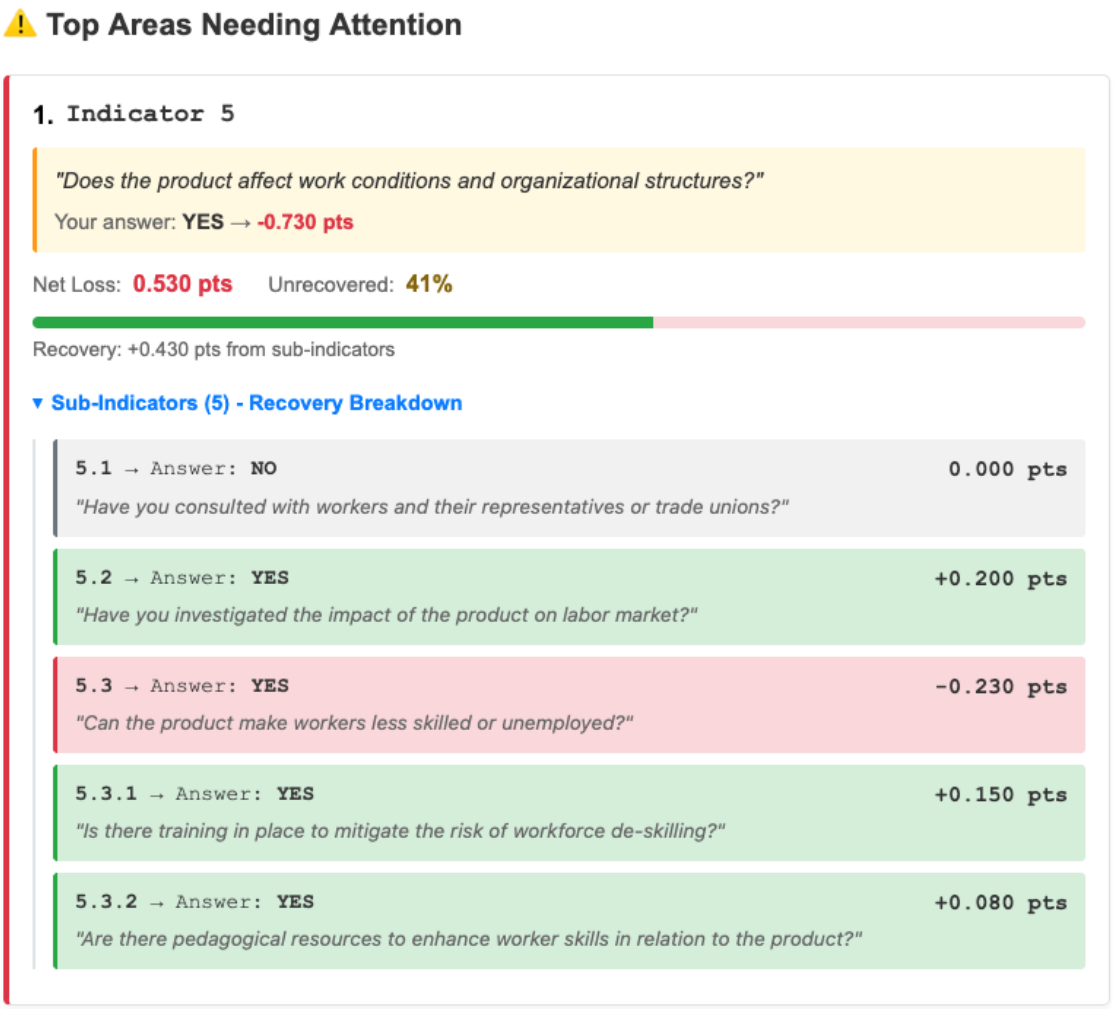

Figure 9. Top areas needing attention view.

Top areas needing attention. This view translates scoring events into improvement priorities. It identifies the parent indicator with the largest unresolved loss, shows the answer that triggered the loss, reports net loss and unrecovered share, and expands sub-indicators to show where mitigation recovered points and where risk remained. The green and red contribution rows make the recommendation logic inspectable rather than opaque.

Figure 10. Advanced ethics analytics dashboard.

Advanced analytics. The dashboard provides post-assessment telemetry, step count, negative and positive events, neutral events, total loss, total recovery, volatility, worst drop, cumulative score dynamics, risk hotspots, and block attribution. The design supports expert review after the session by making it possible to see not only the final score, but also the sequence, concentration, and source of readiness losses.

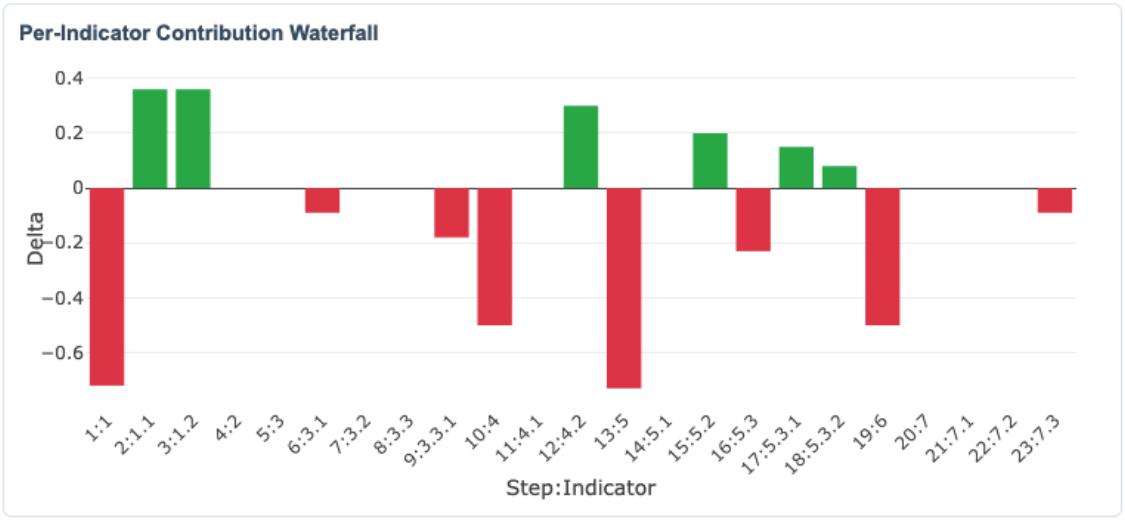

Figure 11. Per-indicator contribution waterfall.

Contribution waterfall. The waterfall view shows every indicator-level score delta in sequence. Positive events are shown in green, negative events in red, and the zero line separates recovery from loss. This design supports auditability because the final score can be traced back to each indicator-level contribution rather than treated as a black-box result.

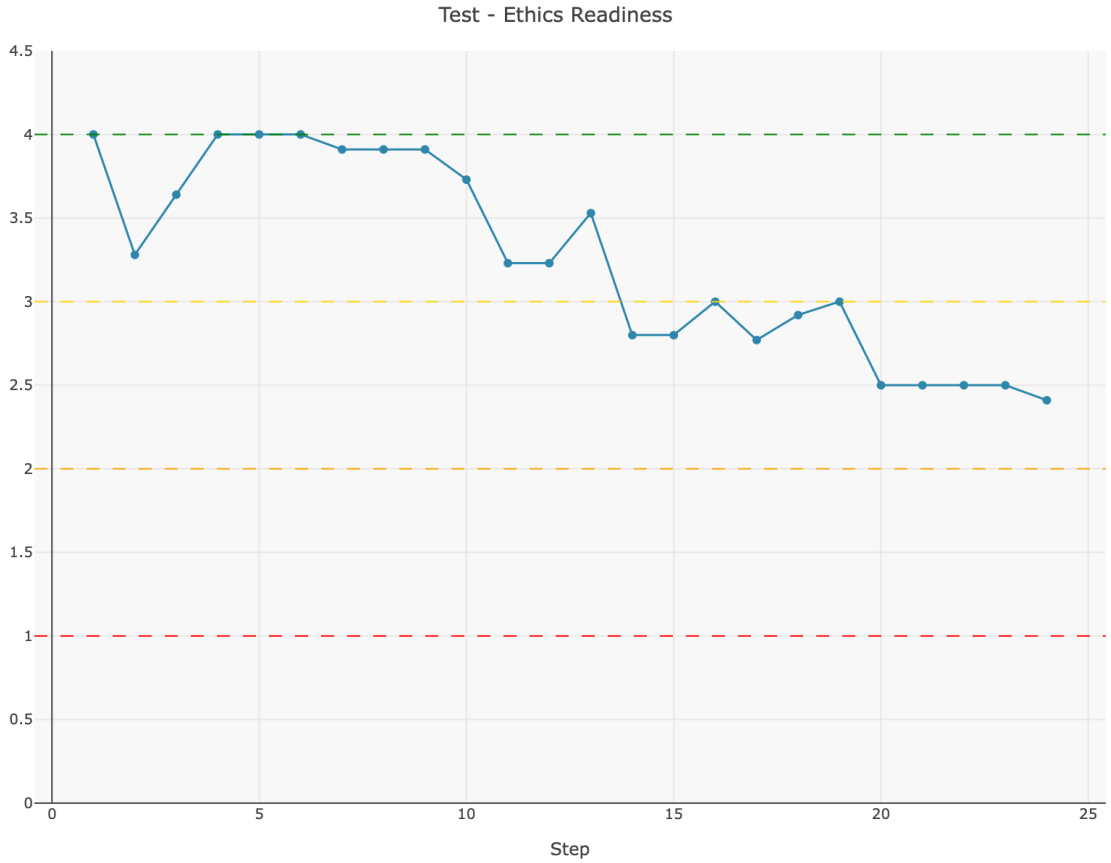

Figure 12. Exit-report score trajectory.

Exit-report trajectory. The generated report plots the readiness score over assessment steps against readiness-level reference lines.

8.4. TECHNICAL GUIDANCE FOR REPLICATING AND HOSTING THE IMPLEMENTATION

This section gives a practical, non-specialist route for running the reference implementation locally. The implementation is a Python Flask application. It uses SQLite as a local database, reads its questions from `schema.sql` on first run, stores local session data in `lperl_local.sqlite`, and serves the tool in a web browser at `http://127.0.0.1:8080`. No separate cloud database or Node.js frontend is required for the reference implementation described here.

Step 1 - Install the basic tools. A user needs Git, Python 3, and a web browser. Git is used to download the repository. Python runs the application. The browser opens the local web interface. On macOS, Git can be installed through the Xcode command-line tools or from git-scm.com; Python should be installed from python.org or through an institutional software manager. On Windows, install Git for Windows and Python from python.org, making sure that Python is added to PATH.

Step 2 - Choose a folder for the tool. Open Terminal on macOS/Linux or PowerShell on Windows. Move to a place where project files can be stored, such as Documents. The folder should be on the user's machine, not in a shared public web directory.

Step 3 - Clone the GitHub repository. Cloning means downloading a working copy of the code. The AIOLIA reference branch is `v0.2-dev`.

```
git clone -b v0.2-dev https://github.com/LA-NS/ethics-readiness-levels.git
cd ethics-readiness-levels
```

Step 4 - Create a Python virtual environment. A virtual environment is a small isolated Python workspace for this tool. It prevents the tool's packages from being mixed with other Python projects on the same computer.

```
python3 -m venv venv
```

On Windows, the equivalent command is usually:

```
py -3 -m venv venv
```

Step 5 - Activate the virtual environment. Activation tells the terminal to use the Python environment created for this tool.

```
source venv/bin/activate
```

On Windows PowerShell, the equivalent command is usually:

```
venv\\Scripts\\Activate.ps1
```

After activation, many terminals show `(venv)` at the beginning of the line. This is a useful visual confirmation.

Step 6 - Install the required Python packages. The repository includes `requirements.txt`; for the v0.2-dev implementation this covers the Flask web framework and supporting analysis/visualisation packages such as Plotly and pandas. If the local environment reports a missing `requests` package when using optional local LLM support, install it in the same environment.

```
python -m pip install --upgrade pip
python -m pip install -r requirements.txt
python -m pip install requests
```

Step 7 - Start the application. The repository may include a `start.sh` launcher for macOS/Linux. If present, the simplest route is to run it from the repository folder. It creates or activates the virtual environment, installs dependencies, and starts the local server.

```
./start.sh
```

The manual route is:

```
python app.py
```

The terminal should print that the Ethics Readiness Levels Tool is starting and that the browser should open `http://localhost:8080`. Currently, the Flask application is explicitly configured to run on `127.0.0.1` at port `8080`.

Step 8 - Open the tool in a browser. Open Chrome, Firefox, Safari, Edge, or another browser and go to:

```
http://127.0.0.1:8080
```

The address `127.0.0.1` means the user's own computer. Other people on the internet cannot access this local session unless the user deliberately changes the hosting configuration.

Step 9 - Run an assessment. The tool starts a local assessment session, asks onboarding questions, activates the relevant indicator block, shows one question card at a time, displays linked AIOLIA technical measures where available, updates the score after each answer, and finally presents the result screen, recommendations field, analytics, and report export.

Step 10 - Stop and restart. To stop the server, return to the terminal where `python app.py` is running and press `Ctrl+C`. To use the tool again later, open the repository folder, activate the virtual environment, and run `python app.py` again.

```
source venv/bin/activate
python app.py
```

Step 11 - Update the implementation. To fetch later changes from the repository, stop the app and run:

```
git pull
```

After updates, rerun package installation if the dependency file has changed.

```
python -m pip install -r requirements.txt
```

Step 12 - Optional local LLM support. The supplied `app.py` contains optional endpoints for local AI assistance and recommendations. These call a local service at `http://127.0.0.1:1234/v1/chat/completions`. This is not required for the ERL assessment itself. If no local LLM server is running, the core questionnaire, scoring, analytics, and report workflow can still be used. Only those optional help features will fail or remain unavailable. The LLM “guide” is usually a gimmick to lighten up long silences or illustrate a certain point.

8.5. WHY PUBLIC HOSTING WAS NOT CHOSEN

Public hosting was not chosen because it could encourage self-evaluations and score self-attribution. This would undermine the ERL methodology. The tool requires dialogue. The most effective assessment involves an ethics expert and a technical or domain expert. The ethics expert helps interpret ethics principles, regulatory concerns, and social risks. The technical or domain expert explains how the system actually works, what evidence exists, what constraints are present, and what mitigations are feasible. The assessment can be supported by accompanying training materials, but it becomes meaningful through the interaction between these perspectives.

This decision is compatible with open-science reuse because openness is provided through the repository, indicator schema, appendix tables, and replication guidance rather than through an ungated public scoring portal. A hosted expert-gated or token-protected service could still be misunderstood as a public self-assessment or certification route and would require ongoing access governance, data-protection arrangements, versioning, and maintenance beyond this deliverable. Replication by another expert team requires use of the repository under its licence, export or local adoption of the structured indicator schema, configuration of the scoring logic, governance for session records and evidence notes, and clear communication that ERL results are not certification outputs.

A public self-service website could make the tool appear easier to use, but it could also encourage a single person to click through the indicators alone. That would risk superficial answers, overconfident scoring, and box-ticking. It could also create the false impression that an ERL score is a certification or compliance mark. For that reason, the implementation is better treated as a reference mechanism for expert-led sessions.

This decision is consistent with the method's core principle that the score is useful only when it is the trace of structured ethical reflection. Without dialogue, the score loses much of its meaning.

Recurring maturity gaps across ERL sessions can inform learner profiles, training priorities, and scenario-based exercises, especially where teams repeatedly struggle with evidence, validation, oversight, or accountability. The first training-oriented use is planned during the AIOLIA trainings in Novi Sad on 2-3 June 2026.

9. Limitations and Next Steps

The ERL/AIA mechanism is not a legal certification tool and does not provide legal advice. It includes indicators inspired by legislation such as the GDPR, LED, and AI Act, but answering these indicators does not prove legal compliance. Legal compliance requires qualified legal analysis.

The scoring weights are expert-informed and practical, but they remain open to refinement. Ethics severity cannot be derived from a purely objective formula. The paper presents weights as beginning from expert consensus and being refined through validation against real use cases. Weights should therefore be reviewed as the tool is used across more cases.

Because different systems activate different blocks and encounter different indicators, scores across unrelated use cases are not directly comparable. A score of 3.0 in one domain may represent a different assessment path than a score of 2.5 in another. The most meaningful comparison is within the same system over time. This incomparability is not a defect but a consequence of context-sensitive assessment. Public administration, healthcare, law enforcement, robotics, and consumer AI do not raise identical ethics profiles. Forcing cross-product ranking would weaken the contextual sensitivity the mechanism was designed to enhance.

The quality of the result depends on the quality of the dialogue and the quality of evidence. If participants lack relevant knowledge, withhold information, or answer aspirationally rather than evidentially, the score will be less reliable. Facilitation, documentation, and evidence recording are therefore part of the method rather than administrative extras.

The healthcare module has received partner feedback and has been adjusted. The public-administration AIA sub-tool has been reviewed by CENTRIC, revised into version 2, and prepared for further T4.3 citizen engagement validation. Continued validation will improve clarity, weight adequacy, and domain relevance.

10. Conclusion

D3.4 delivers a mechanism for evaluating Ethics Readiness Levels and algorithmic impact assessment in the AIOLIA project. The mechanism builds on prior ERL/LPERL readiness-level work but adapts it for AIOLIA's ethics-by-design objectives. It combines a 0-4 ERL scale, modular indicator blocks, dynamic tree-like questionnaire navigation, score progression, session tracking, evidence guidance, maturity recommendations, and domain-specific modules.

The work produced three major AIOLIA-specific extensions: the first Python-based and web-enabled implementation of the tool, a healthcare AI module, and a public-administration AIA sub-tool. The healthcare module reflects validation with AIOLIA Healthcare Use Case 1 and AIOLIA Healthcare Use Case 2. The public-administration AIA sub-tool reflects CENTRIC review and version 2 revisions, and is prepared for further validation in T4.3 citizen engagement workshops.

The mechanism is supported by an implementation reference in the LA-NS/ethics-readiness-levels repository. It should be reused as an expert-led assessment and training support mechanism, not as a certification tool or public self-assessment portal.

11. Appendices

11.1. APPENDIX A: OVERVIEW OF THE ERL METHODOLOGY

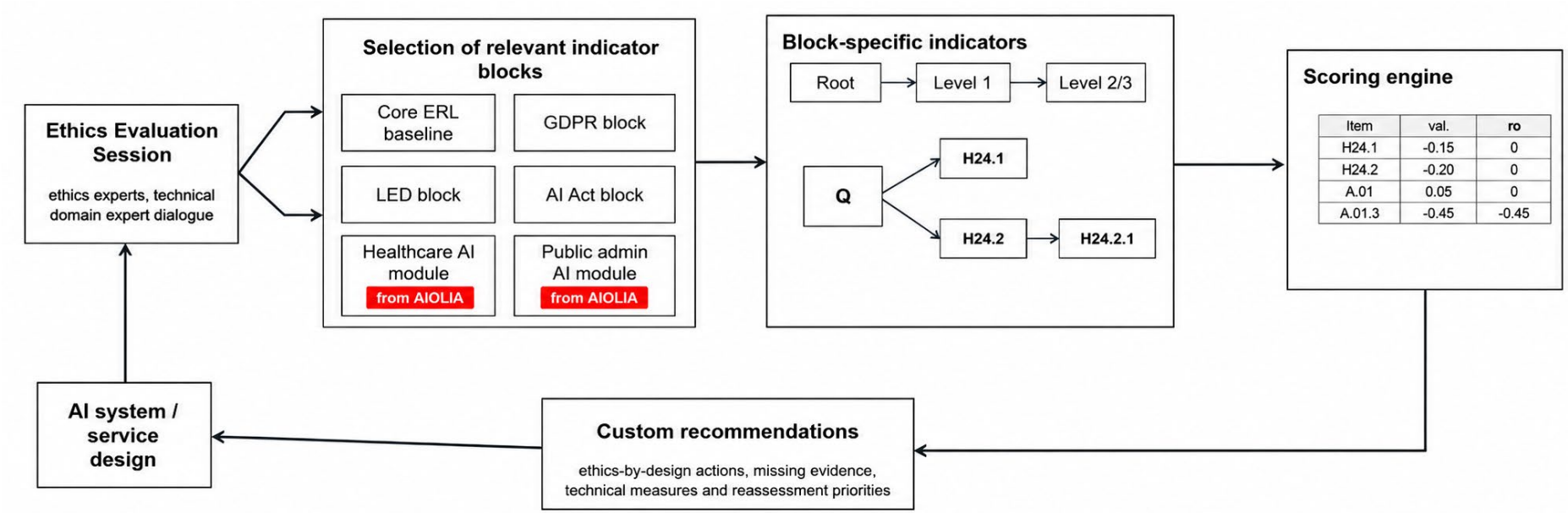

Figure A1. Overview of the ERL methodology and repeated assessment loop.

11.2. APPENDIX B: AI FOR HEALTHCARE INDICATOR BLOCK

Table 2. Healthcare AI indicator block

No.	Complete formulation	Yes score	No score
24	Does the AI system affect patient agency or decision-making?	-0.25	0
24.1	Have patients been provided with meaningful information about the AI tool to support informed decision-making?	0.25	0
24.2	Is informed consent obtained before using patient data for AI training or modelling?	0	-0.15
24.2.1	Are there electronic or dynamic consent mechanisms in place allowing patients to opt-out?	0	-0.05
25	Does the AI system automate clinical recommendations or replace clinician judgment?	-0.3	0
25.1	Is the AI tool designed as a complement to clinician judgment rather than a replacement?	0.15	0

No.	Complete formulation	Yes score	No score
25.2	Can clinicians override the AI system's recommendations without professional penalty?	0.15	0
25.3	Is there a risk of automation bias, where clinicians become over-reliant on the AI?	-0.2	0
25.3.1	Beyond training, are there operational safeguards (e.g., mandatory human verification, override documentation, second-reader checks, monitoring) to mitigate automation bias?	0.2	0
26	Can the AI system's outputs result in biased or discriminatory healthcare outcomes?	-0.65	0
26.1	Were representative datasets used to train the algorithm across different demographics?	0.25	0
26.1.1	As an additional step, have you engaged in multi-institutional data collaborations to further reduce underrepresentation of minority groups?	0.2	0
26.2	Are there regular audits to detect and correct racial or gender bias after deployment?	0.2	0
27	Does the system process or store patient health data?	-0.3	0
27.1	Are cybersecurity protocols and privacy-preserving measures (such as anonymization) implemented?	0.15	0
27.2	Has the AI system undergone prospective clinical trials or independent validation to ensure patient safety?	0.15	0
27.3	Is there a risk the data could be used for unauthorized commercial or secondary purposes?	-0.2	0
27.3.1	Are there institutional safeguards preventing the unauthorized commercialization of patient data?	0.2	0
28	Do model complexity or design choices prevent clinically useful justification of outputs at the point of care?	-0.4	0
28.1	Can clinicians access clear, case-level rationale they can use to defend and communicate AI-informed decisions?	0.2	0
28.2	Have developers published documentation on the algorithm's assumptions, limitations, and data provenance?	0.2	0
29	Was the AI system developed without formal ethical oversight?	-0.4	0
29.1	Are bioethicists or ethics experts part of the AI development team?	0.1	0
29.2	Has an algorithmic or human rights impact assessment been conducted before clinical deployment?	0.1	0
29.3	Do developers use standardized AI ethics checklists and guidelines specific to healthcare?	0.1	0
29.4	Is there a framework for assigning liability and compensation for AI-induced medical errors?	0.1	0
30	Is the AI system clinically valid and technically accurate?	0	-0.5
30.1	Does the tool accurately reflect the patient's clinical state (e.g., identifying findings in chest x-rays or CT scans)?	0	-0.3
30.2	Has the system been tested for accuracy in specific tasks like vessel measurements or treatment planning?	0	-0.2
31	Can the AI system justify its clinical recommendations?	0	-0.2
31.1	Are the AI outputs supported by reasons that align with professional medical standards?	0	-0.12
31.2	For individual recommendations, can the system provide justification grounded in intended use, demonstrated outcome impact, and statistically valid evidence?	0	-0.08
32	Is there evidence that performance differs across clinically relevant sub-populations or atypical anatomies?	-0.3	0

No.	Complete formulation	Yes score	No score
32.1	Has the tool been assessed for performance disparities in patients with rare comorbidities or atypical anatomies?	0.3	0
32.1.1	Have the identified performance disparities been addressed? For example, through retraining, recalibration, or additional data collection?	0	-0.15
33	Are the AI outputs and documentation accessible within the clinical workflow?	0	-0.15
33.1	Are risk scores, heatmaps, or image overlays presented in a format that fits into a physician's daily practice?	0	-0.09
33.2	Is the documentation on data provenance and validation accessible to hospital managers and medical regulators?	0	-0.06
34	Is there evidence that the AI delivers measurable operational or clinical value that can justify adoption in settings with limited or no direct reimbursement?	0	-0.3

11.3. APPENDIX C: AI IMPACT ASSESSMENT INDICATOR BLOCK

Table 3. Public-administration AI impact assessment indicator block

No.	Complete formulation	Yes score	No score
101	Does the AI system make or substantially influence decisions that directly affect citizens' legal rights, entitlements, or obligations?	-0.3	0
101.1	Is there an explicit legal basis authorising the use of AI for this type of decision?	0.15	0
101.2	Has a proportionality assessment been conducted to justify the use of AI given the stakes involved for citizens?	0.15	0
102	Can the AI system produce outcomes that are discriminatory or systematically disadvantage certain demographic groups?	-0.6	0
102.1	Were the training data and model outputs tested for bias across demographic groups such as age, gender, ethnicity, or socioeconomic status?	0.3	0
102.1.1	Have identified biases been addressed? For example, through retraining, reweighting, or process redesign?	0	-0.15
102.2	Are there ongoing mechanisms to monitor for discriminatory patterns in outcomes after deployment?	0.3	0
103	Are citizens given sufficient information about the role of AI in decisions that affect them?	0	-0.3
103.1	Are citizens explicitly informed when an AI system is involved in processing their data or contributing to a decision that affects them?	0	-0.1
103.2	Are citizens informed of the AI system's known limitations and the possibility of error?	0	-0.1
103.3	Are citizens informed of the rights available to them in relation to AI-assisted decisions?	0	-0.1
104	Can the AI system's contribution to a specific decision be explained to the citizen it affects?	0	-0.3

No.	Complete formulation	Yes score	No score
104.1	Are explanations provided in plain language appropriate to the affected citizens' expected knowledge, needs, and context?	0	-0.1
104.2	Is an explanation available to the citizen upon request without requiring them to take legal action?	0	-0.1
104.3	Are any legal, technical, or organisational limits on the explanation identified, justified, and communicated to the citizen where relevant?	0	-0.1
105	Can citizens request a review of an AI-assisted decision by a human official?	0	-0.45
105.1	Is the right to human review communicated to the citizen at the time of the decision?	0	-0.2
105.2	Is there a formal appeal mechanism through which citizens can contest an AI-assisted decision?	0	-0.15
105.2.1	Is the appeal process accessible to citizens regardless of their level of digital literacy or internet access?	0	-0.1
106	Do human officials retain meaningful decision-making authority, rather than routinely deferring to AI outputs?	0	-0.3
106.1	Are the officials or staff responsible for the decision trained to critically evaluate AI recommendations rather than accept them uncritically?	0	-0.15
106.2	Is there documented evidence that human oversight is exercised in practice, not only stated in policy?	0	-0.15
107	Does the system respect citizens' rights in relation to the personal data it processes?	0	-0.3
107.1	Is there a clear legal basis for processing each category of personal data used by the system?	0	-0.1
107.2	Is personal data collection limited to what is strictly necessary for the stated purpose?	0	-0.1
107.3	Do citizens have enforceable rights to access, correct, and request erasure of their data?	0	-0.1
107.3.1	Is there a procedure for citizens to exercise these rights without undue administrative burden?	0	-0.05
108	Has the AI system been validated for accuracy and reliability in the specific public administration context where it is deployed?	0	-0.3
108.1	Has the system been tested on data representative of the actual population it will affect, including vulnerable groups where relevant?	0	-0.15
108.2	Is performance continuously monitored and reported to designated oversight or governance bodies where such bodies are available?	0	-0.1
108.2.1	Are performance reports made publicly accessible in an understandable format?	0	-0.05
109	Was the deployment of this AI system subject to democratic or parliamentary scrutiny?	0	-0.25

No.	Complete formulation	Yes score	No score
109.1	Were affected communities or civil society organisations consulted before the system was deployed?	0	-0.15
109.2	Is the existence, purpose, and scope of this AI system publicly disclosed in an accessible format?	0	-0.1
110	Can the AI system be independently audited for fairness, accuracy, and legal compliance?	0	-0.2
110.1	Is access to algorithmic audits granted to independent oversight or regulatory bodies?	0	-0.1
110.2	Is there a defined process for acting on audit findings, including remediation timelines?	0	-0.1
111	Does the use of an external vendor or development partner create governance risks, such as reduced transparency, auditability, public-sector control, or liability clarity?	-0.15	0
111.1	Does the procurement or collaboration agreement include clauses requiring transparency, auditability, and liability for harm to citizens?	0.08	0
111.2	Does the public authority retain sufficient control over data, documentation, and system operation to change, replace, or discontinue the vendor arrangement without unacceptable service disruption?	0.07	0
112	Does the AI system risk creating or deepening digital exclusion among citizens?	-0.2	0
112.1	Are alternative non-digital channels available for citizens who cannot or choose not to interact with the AI system?	0.1	0
112.2	Has the system been assessed for accessibility by citizens with disabilities or limited digital literacy?	0.1	0
113	Is there a designated accountability framework specifying who is responsible when the AI system causes harm to a citizen?	0	-0.25
113.1	Is there a formal redress or compensation mechanism for citizens harmed by errors in AI-assisted decisions?	0	-0.15
113.2	Are the accountability and redress arrangements publicly communicated in plain language?	0	-0.1
114	Does deployment of this AI system significantly affect the roles, workload, or required skills of public servants?	-0.1	0
114.1	Have public servants been consulted on the system's impact on their work?	0.05	0
114.2	Is training and upskilling support provided to public servants whose roles are affected by the system?	0.05	0

11.4. APPENDIX D: AIOLIA TECHNICAL MEASURES WITH LINKED INDICATORS

Table 4. AIOLIA technical measures pairings

Indicator	Original indicator	Technical measure
1.1	Have you implemented safeguards to prevent the product from unintentionally affecting users' autonomy?	Building technical manual override and change mechanisms into the UI that allow human operators to alter or pause AI decisions.
1.2	Is there a risk of users becoming overly reliant on the product?	Forcing users via the digital interface to input documented text justification before the system allows them to accept specific AI recommendations.
1.2.1	Are there measures to discourage users from over-relying on the product?	Triggering automated explanation prompts in the UI to force human reflection before confirming an AI-generated decision.
2	Is it possible that technical design decisions will result in significant damage?	Performing comprehensive, technical hazard and failure-mode analysis on the system's underlying architecture.
2.1	Does the product comply with recognized cybersecurity standards?	Designing and deploying resilient hardware and software architectures technically capable of withstanding errors, data corruption, or attacks.
2.2	Are there additional security measures against potential attacks throughout the product's lifecycle?	Building explicit technical safeguards directly into the code designed to detect and block attempted system misuse or manipulation.
2.2.1	Have you conducted penetration tests or red-team exercises on the product?	Conducting active, aggressive security penetration tests to computationally identify and patch technical vulnerabilities.
2.4	Can the product be misused for malicious or illegal purposes?	Building explicit technical safeguards directly into the code designed to detect and block attempted system misuse or manipulation.
2.4.1.1	Have you implemented preventive measures to mitigate the risks of misuse?	Integrating automated control mechanisms that instantly pause, downgrade, rollback, or constrain system outputs when predefined risk thresholds are met.
3	Is the product designed to adapt to diverse user/operator preferences and abilities?	Using diverse and representative datasets to prevent bias during AI model design and training.
3.1	Have you evaluated accessibility by individuals with special needs or those at risk of exclusion?	Performing computational subgroup performance analysis to detect and mitigate bias across different demographic or user groups.
3.3	Have you studied and evaluated possible discrimination against affected persons?	Performing computational subgroup performance analysis to detect and mitigate bias across different demographic or user groups.
3.3.1	Have you implemented measures to minimize unfair or discriminatory effects?	Utilizing automated fairness drift detection to continuously monitor for emergent biases as the system operates in the real world.
5	Does the product affect work conditions and organizational structures?	Running automated, continuous fairness reviews of AI-generated risk scores in workplace and HR systems.
5.3	Can the product make workers less skilled or unemployed?	Displaying technical comparison views on-screen that explicitly contrast the AI's assessment with a human's assessment.
5.3.1	Is there training in place to mitigate the risk of workforce de-skilling?	Providing simplified, dynamically generated technical explanations of AI tools tailored specifically for employees using workplace systems.
5.3.2	Are there pedagogical resources to enhance worker skills in relation to the product?	Providing simplified, dynamically generated technical explanations of AI tools tailored specifically for employees using workplace systems.

Indicator	Original indicator	Technical measure
6	Have you established mechanisms to facilitate audits of the product?	Maintaining automated audit logs to ensure total technical traceability of AI outputs and decisions.
6.1	Can the product be audited by independent third parties for assigning responsibility?	Utilizing immutable, append-only logs that are cryptographically secured so they technically cannot be altered or deleted once written.
7	Are there oversight processes for ethical concerns and assigning responsibility?	Hardcoding Human-in-the-loop workflows directly into the software's operational pipeline so processes cannot advance without human input.
7.1	Is there ongoing oversight by a third party beyond the product's development phase?	Automatically logging all human validation actions to maintain a transparent, verifiable record of human-in-the-loop oversight.
8	Does the product gather specialized categories of personal data, such as biometric or health-related information?	Applying robust data encryption protocols to technically protect sensitive information and user privacy across the system.
8.1	Are there security protocols in place for safeguarding specialized categories of personal data?	Designing backend infrastructure with restricted, isolated data access environments for processing highly sensitive personal information.
8.1.2	Have you implemented measures to ensure only the essential sensitive personal data is collected?	Hardcoding data minimization rules directly into the system's data collection APIs and storage architecture.
8.3	Are the objectives for collecting personal data clearly defined, explicit, and legitimate?	Building large-scale technical consent and data lifecycle management architectures, particularly for virtual assistants and HR platforms.
8.3.1	Is the acquired personal data strictly relevant and limited to what is essential for the intended purposes?	Hardcoding data minimization rules directly into the system's data collection APIs and storage architecture.
8.3.2	Is the personal data you gather accurate and regularly updated?	Enforcing automated data quality, integrity, and validation protocols within the machine learning training pipeline.
8.3.3	Is personal data stored in a way that allows for identification of subjects only for the duration needed for its intended use?	Building large-scale technical consent and data lifecycle management architectures, particularly for virtual assistants and HR platforms.
10	Are there methods to authenticate the identity of a data subject requesting access?	Embedding cryptographic artefacts like secure hashes and keys throughout the infrastructure to mathematically protect data integrity.
10.1	Are the identity verification methods you employ both secure and reliable?	Applying robust data encryption protocols to technically protect sensitive information and user privacy across the system.
10.1.1	Do you employ pseudonymization techniques as part of the identity verification process?	Executing automated data anonymization and pseudonymization processes on datasets wherever technically feasible.
10.1.2	Is personal data only retained for the purpose of responding to potential future requests?	Building large-scale technical consent and data lifecycle management architectures, particularly for virtual assistants and HR platforms.
11	Does the product support the "right to be forgotten," allowing for the erasure of personal data?	Building dynamic technical consent portals that allow users to actively toggle and adjust their data tracking preferences over time.
11.1	Can a data subject request erasure under some specified conditions?	Building large-scale technical consent and data lifecycle management architectures, particularly for virtual assistants and HR platforms.
12	Is all data processing information transparent and easily accessible to data subjects?	Programming the UI to generate user-facing explanations that clarify the AI's functioning and limitations directly on the screen.
12.1	Is data processing information also available in electronic formats?	Automating the programmatic generation and delivery of explanatory messages sent to users exactly when their content is restricted.

Indicator	Original indicator	Technical measure
12.1.1	Is data processing information articulated in clear language that is easily understandable?	Creating automated, patient-facing plain-language summaries for AI-generated outputs in clinical settings.
12.1.2	Is the consent mechanism presented separately from other terms or conditions?	Designing user interfaces with direct, built-in consent mechanisms to protect user autonomy.
12.1.3	Is it possible for data subjects to revoke their consent at any time?	Building dynamic technical consent portals that allow users to actively toggle and adjust their data tracking preferences over time.
13	Does the product consider a minor's consent (below the age of 13-16)?	Implementing digital age verification API mechanisms to protect minors and ensure human safety at the point of access.
13.1	Is the processing of data for subjects under 13-16 years of age based on parental consent?	Implementing digital age verification API mechanisms to protect minors and ensure human safety at the point of access.
13.1.1	Are reasonable technological methods employed to confirm that parental consent is genuine?	Implementing digital age verification API mechanisms to protect minors and ensure human safety at the point of access.
13.1.2	Does the product provide clear and age-appropriate information to subjects under 13-16 years about the data being collected?	Creating automated, patient-facing plain-language summaries for AI-generated outputs in clinical settings.
13.1.3	Is an option available for any minor to remove personal data?	Building dynamic technical consent portals that allow users to actively toggle and adjust their data tracking preferences over time.
14	Does the product process personal data that is not needed for identification?	Executing automated data anonymization and pseudonymization processes on datasets wherever technically feasible.
14.1	If the processing objectives do not require identification, do you refrain from collecting additional identifying data?	Hardcoding data minimization rules directly into the system's data collection APIs and storage architecture.
14.1.1	Is the restriction of such non-identifiable data clearly marked within the product?	Applying clear, hardcoded UI labelling to explicitly distinguish AI-generated outputs from human-generated content.
15	Does the product handle personal data?	Hardcoding data minimization rules directly into the system's data collection APIs and storage architecture.
15.1	Does the product handle personal data only for a specific, explicit, and legitimate purpose?	Building large-scale technical consent and data lifecycle management architectures, particularly for virtual assistants and HR platforms.
15.1.1	Are there measures in place to periodically review and remove unnecessary data?	Enforcing automated data quality, integrity, and validation protocols within the machine learning training pipeline.
15.1.2	Can your product share personal data if the law requires it?	Utilizing role-based digital permissions to securely manage access and intervention capabilities within the system.
15.2	Does your product process data related to a person's racial or ethnic background?	Using diverse and representative datasets to prevent bias during AI model design and training.
15.2.1	Is there extra security for such sensitive data?	Applying robust data encryption protocols to technically protect sensitive information and user privacy across the system.
16	Can you correct personal data in your product?	Building large-scale technical consent and data lifecycle management architectures, particularly for virtual assistants and HR platforms.
16.1	Can you delete data in your product if it violates the Law Enforcement Directive (LED) regulations?	Building large-scale technical consent and data lifecycle management architectures, particularly for virtual assistants and HR platforms.

Indicator	Original indicator	Technical measure
16.1.1	Can you limit data usage while someone challenges its accuracy?	Utilizing role-based digital permissions to securely manage access and intervention capabilities within the system.
16.1.2	Does your product keep data for legal evidence?	Generating and digitally storing case-level evidence chains and appeal records for AI decisions in healthcare settings.
17	Can users ask to verify the lawfulness of the data processing?	Maintaining automated audit logs to ensure total technical traceability of AI outputs and decisions.
17.2.1	Is there a mention of LED in your product documentation or official release?	Implementing automated backend generation of documentation regarding model behaviour to support technical explainability.
17.3	Do you have a procedure for the cases in which a user's request for data access or correction is denied?	Building technical manual override and change mechanisms into the UI that allow human operators to alter or pause AI decisions.
17.3.1	If data access or correction is denied, are users informed they can seek a regulatory review?	Triggering automated UI explanations precisely at the moment user content is actively restricted, flagged, or taken down.
18	Is the AI system considered high-risk as per relevant regulation in the European Union (AI Act)?	Executing formal technical model auditing against established baseline computational metrics.
18.1	Does it comply with relevant requirements for high-risk systems in the AI Act or includes a compliance plan for the upcoming regulation?	Implementing automated backend generation of documentation regarding model behaviour to support technical explainability.
19	Is performance of the AI system evaluated and documented?	Running automated validation cycles that computationally test the model's weights and outputs against standardized benchmark datasets.
19.1	Is the AI system performance continuously evaluated during normal operation?	Running continuous technical performance monitoring pipelines to mathematically ensure the system operates within defined safe parameters.
19.1.1	Is a description of the risk management system included in the technical documentation?	Performing comprehensive, technical hazard and failure-mode analysis on the system's underlying architecture.
19.1.2	Does technical documentation include changes made to the system through its lifecycle?	Applying strict technical version control mechanisms for both the AI models and their underlying datasets.
19.3	Does technical documentation address the monitoring, functioning, and control of the AI system?	Implementing automated backend generation of documentation regarding model behaviour to support technical explainability.
20	Do users get sufficient information about the methods and capabilities of the AI system?	Programming the UI to generate user-facing explanations that clarify the AI's functioning and limitations directly on the screen.
20.1	Does the AI system make users aware that they are communicating or interacting with the AI system?	Applying clear, hardcoded UI labelling to explicitly distinguish AI-generated outputs from human-generated content.
20.2	Does the AI system inform users of its limitations?	Displaying visual safety boundaries directly within the user interface for real-time applications like automotive AI.
20.3	Are affected persons informed about their rights?	Automating the programmatic generation and delivery of explanatory messages sent to users exactly when their content is restricted.
21	Do decisions or outputs of the AI system influence or affect humans directly?	Enforcing bounded automation via hardcoded technical limits that restrict exactly what the AI can execute without a human override.
21.1	Did you establish detection and response mechanisms for undesirable effects, for example false negatives or false positives?	Deploying algorithmic anomaly detection to automatically flag abnormal data patterns or potential system misuse.

Indicator	Original indicator	Technical measure
21.1.1	Is there a 'stop button' (if relevant) or a procedure for humans to safely abort operation of the AI system?	Implementing technical safe-fail protocols that automatically shut down or default the system to a safe state during critical errors.
21.2	Can the AI system be controlled or overseen by humans during normal operation?	Hardcoding Human-in-the-loop workflows directly into the software's operational pipeline so processes cannot advance without human input.
21.2.1	Was the AI system designed so as to be controlled or overseen by a human during operation?	Building technical manual override and change mechanisms into the UI that allow human operators to alter or pause AI decisions.
21.2.1.1	Have you evaluated the efficacy of oversight measures during normal operation?	Creating specialized, interactive UI dashboards engineered explicitly to make the human review of algorithmic data drift highly efficient.
21.2.1.1.1	Were humans offered training on how to exercise control?	Providing simplified, dynamically generated technical explanations of AI tools tailored specifically for employees using workplace systems.
21.3	Have you implemented technical measures to facilitate explicability of the outputs?	Rendering mathematical confidence and uncertainty indicators on the screen alongside AI-generated outputs.
21.3.1	Do these technical measures meet benchmarks and industry standards?	Running automated validation cycles that computationally test the model's weights and outputs against standardized benchmark datasets.
22	Can the AI system be attacked resulting in unintended or unexpected harm?	Conducting computational robustness and resilience testing against adverse conditions, corrupted data, or malicious attacks.
22.1	Is the training and application data stored securely with standard and robust authorisation and encryption requirements?	Applying robust data encryption protocols to technically protect sensitive information and user privacy across the system.
22.2	Is the AI system robust against AI-specific adversarial attacks?	Implementing explicit algorithmic jailbreak detection mechanisms to prevent users from bypassing the AI's safety filters.
22.3	Is the AI system resilient against model extraction and model replication attacks?	Embedding cryptographic artefacts like secure hashes and keys throughout the infrastructure to mathematically protect data integrity.
22.4	Does the AI system have a functionality to report incidents or breaches to relevant authorities?	Automating the ongoing monitoring of algorithmic risk scores and generating instant digital alerts when thresholds are exceeded.
23	Was there training to ensure sufficient AI literacy of the users?	Providing simplified, dynamically generated technical explanations of AI tools tailored specifically for employees using workplace systems.
23.1	Was the training specifically adapted and contextualized for the use case?	Providing simplified, dynamically generated technical explanations of AI tools tailored specifically for employees using workplace systems.
23.4	Was the AI literacy of affected persons taken into account during the system's deployment?	Creating automated, patient-facing plain-language summaries for AI-generated outputs in clinical settings.

12. References

Adomaitis, L., Israel-Jost, V., and Grinbaum, A. (2026). Ethics Readiness Levels for AI Systems: A Practical Evaluation Framework. Accepted for presentation at IEEE ZINC 2026; passed IEEE Xplore publication review.

AIOLIA. (2026). Operational context-sensitive guidelines in Canada, China, Japan, and South Korea. AIOLIA Deliverable D3.2.

AIOLIA. (2026). Context-enriched Operational Non-technical Guidelines for AI Research Areas. AIOLIA Deliverable D3.3.

Adomaitis, L., Grinbaum, A., and Lenzi, D. (2022). TechEthos D2.2: Identification and specification of potential ethical issues and impacts and analysis of ethical issues of digital extended reality, neurotechnologies, and climate engineering. PhD Thesis, CEA Paris Saclay.

Adomaitis, L., Hoog, B., and Grinbaum, A. (2024). Security and ethics readiness levels: Two new scales. 2024 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), 1-8. https://doi.org/10.1109/ICTMOD63116.2024.10878193

Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20, 310. https://doi.org/10.1186/s12911-020-01332-6

Bayerl, P. S., Lawlor, E., and Akhgar, B. (2026). Operational ethics guidelines on use cases related to human behaviour and cognition. CENTRIC, AIOLIA D3.1. https://aiolia.eu/wp-content/uploads/2026/02/AIOLIA-D3.1-certified.pdf

Bietti, E. (2021). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. SSRN.

Brey, P. (2010). Values in technology and disclosive computer ethics. In The Cambridge handbook of information and computer ethics, 41-58.

Buber, M. (1958). I and Thou (2nd ed.). Scribner.

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. Public Administration Review, 81(5), 825-836. https://doi.org/10.1111/puar.13293

Citron, D. K., and Pasquale, F. (2014). The scored society: Due process for automated predictions. Washington Law Review, 89(1), 1-33.

de Hert, P., and Papakonstantinou, V. (2016). The new Police and Criminal Justice Data Protection Directive: A first analysis. New Journal of European Criminal Law, 7(1), 7-19. https://doi.org/10.1177/203228441600700102

de Jong, E. (2025). Ethics readiness of technology: The case for aligning ethical approaches with technological maturity. https://doi.org/10.48550/arXiv.2504.03336

Eljasik-Swoboda, T., Rathgeber, C., and Hasenauer, R. (2019). Assessing technology readiness for artificial intelligence and machine learning based innovations. In DATA, 281-288.

European Commission. (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission.

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations and Directives. Official Journal of the European Union.

Festinger, L. (1957). A Theory of Cognitive Dissonance. Stanford University Press.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. Philosophical Transactions of the Royal Society A, 374(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L., et al. (2018). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707.

Friedman, B., and Hendry, D. (2019). Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press.

Grinbaum, A. (2019). Les robots et le mal. Desclee de Brouwer.

Grinbaum, A., and Groves, C. (2013). What is 'responsible' about responsible innovation? Understanding the ethical issues. In Responsible innovation: Managing the responsible emergence of science and innovation in society, 119-142.

Guston, D. H. (2014). Understanding 'anticipatory governance'. Social Studies of Science, 44(2), 218-242. https://doi.org/10.1177/0306312713508669

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120.

High-Level Expert Group on Artificial Intelligence. (2019). Ethics Guidelines for Trustworthy AI. European Commission.

Jonas, H. (1985). The Imperative of Responsibility: In Search of an Ethics for the Technological Age. University of Chicago Press.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195. https://doi.org/10.1186/s12916-019-1426-2

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165, 633-705.

Kundu, S., et al. (2023). Specific versus general principles for constitutional AI. https://doi.org/10.48550/arXiv.2310.13798

Lavin, A., et al. (2022). Technology readiness levels for machine learning systems. Nature Communications, 13(1), 6039. https://doi.org/10.1038/s41467-022-33128-9

Lessig, L. (2000). Code is law. Harvard Magazine, 1.

Levinas, E. (1991). Totality and Infinity. Springer Netherlands. https://doi.org/10.1007/978-94-009-9342-6

Mahesh Kumar, S., et al. (2025). AI-powered face sketching for criminal identification. In Intelligent Computing and Communication, 415-425. Springer Nature. https://doi.org/10.1007/978-981-96-1267-3_35

Mankins, J. C. (1995). Technology readiness levels: A white paper. NASA Office of Space Access and Technology.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501-507.

Moor, J. H. (1985). What is computer ethics? Metaphilosophy, 16(4), 266-275. https://doi.org/10.1111/j.1467-9973.1985.tb00173.x

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

Olechowski, A., Eppinger, S. D., and Joglekar, N. (2015). Technology readiness levels at 40: A study of state-of-the-art use, challenges, and opportunities. 2015 Portland International Conference on Management of Engineering and Technology (PICMET), 2084-2094. https://doi.org/10.1109/PICMET.2015.7273196

Politi, V., and Grinbaum, A. (2020). The distribution of ethical labor in the scientific community. Journal of Responsible Innovation, 7(3), 263-279.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380, 1347-1358. https://doi.org/10.1056/NEJMra1814259

Richardson, H. S. (1990). Specifying norms as a way to resolve concrete ethical problems. Philosophy & Public Affairs, 19(4), 279-310.

Sauser, B., Verma, R., Ramirez-Marquez, J., and Gove, R. (2006). From TRL to SRL: The concept of systems readiness levels. Proceedings of the Conference on Systems Engineering Research, Los Angeles, CA.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68. https://doi.org/10.1145/3287560.3287598

Stahl, B. C., Timmermans, J., and Flick, C. (2017). Ethics of emerging information and communication technologies: On the implementation of responsible research and innovation. Science and Public Policy, 44(3), 369-381.

Tang, Y., Xiong, J., Becerril-Arreola, R., and Iyer, L. (2020). Ethics of blockchain. Information Technology & People, 33(2), 602-632. https://doi.org/10.1108/ITP-10-2018-0491

Terpan, F. (2015). Soft law in the European Union - The changing nature of EU law. European Law Journal, 21(1), 68-96. https://doi.org/10.1111/eulj.12090

Umbrello, S., et al. (2023). From speculation to reality: Enhancing anticipatory ethics for emerging technologies (ATE) in practice. Technology in Society, 74, 102325.

Unzueta, L., et al. (2026). The MultiRATE Holistic Readiness Level framework for civil security technologies. Open Research Europe, 6, 42. https://doi.org/10.12688/openreseurope.22711.1

Van den Hoven, J., Vermaas, P. E., and Van de Poel, I. (2015). Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains. Springer.

von Schomberg, R. (2013). A vision of responsible research and innovation. In R. Owen, J. Bessant, and M. Heintz (Eds.), Responsible Innovation, 51-74. Wiley. https://doi.org/10.1002/9781118551424.ch3

Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language communication between man and machine. Communications of the Association for Computing Machinery, 9, 36-45.

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1-18. https://doi.org/10.1145/3351095.3372833

Williams, B. (2006). Ethics and the Limits of Philosophy. Routledge.