Accelerating the digital transformation of the health sector in the Americas

AI-GUARD Tool

Artificial Intelligence Governance & Use Assessment for Responsible Deployment

PAHO
Pan American Health Organization
World Health Organization
Americas Region
IDB
Inter-American Development Bank

AI-GUARD Tool

Artificial Intelligence
Governance & Use Assessment for
Responsible Deployment

Washington, D.C., 2026

PAHO Pan American Health Organization World Health Organization Americas Region IDB

Table of contents

Acknowledgements

This publication was developed with the support of the following individuals and partners:

Pan American Health Organization
Marcelo D'Agostino, Myrna Marti, Sebastian Garcia-Saiso, Juan Carlos Diaz, Jaime Pedrosa Comino, Joao Paulo Souza, Marcos Luis Mori, Maria Alejandra Farias, Francisco Barbosa Junior.

Inter-American Development Bank
Jennifer Nelson, Pablo Oreficce.

Italian Hospital, Buenos Aires, Argentina, PAHO/WHO Collaborating Center for Information Systems and Digital Health
Daniel Luna, Fernando Plazzotta.

Foreword

Artificial intelligence is rapidly transforming health systems across the Americas. From strengthening disease surveillance and early outbreak detection to improving service delivery, optimizing resource allocation, enhancing diagnostic support, strengthening research, and streamlining administrative processes, AI technologies offer significant opportunities to advance public health outcomes. When aligned with clearly defined health priorities, AI can contribute to reducing waiting times, improving access to specialized services, strengthening primary health care performance, enhancing health workforce planning, and supporting data-driven policy decisions. However, the integration of AI into health systems is not without complexity. AI deployment may introduce governance, equity, transparency, and accountability challenges if not implemented with appropriate safeguards. Risks may arise from insufficient oversight, unclear accountability structures, biased training data, lack of subgroup performance evaluation, limited transparency regarding system limitations, or inadequate monitoring once deployed. High-impact AI systems, particularly those influencing diagnosis, eligibility, or allocation of public resources, require structured governance mechanisms to prevent unintended consequences and protect vulnerable populations.

Recognizing both the transformative potential and the governance challenges of AI in health, Artificial Intelligence Governance & Use Assessment for Responsible Deployment (AI-GUARD) has been developed as a PAHO-IDB technical guidance tool to support Member States and health institutions in making structured, transparent, and responsible decisions regarding AI adoption, procurement, development, and scale-up. The instrument provides a tiered, proportional framework that aligns oversight intensity with the level of impact and risk associated with each initiative.

AI-GUARD forms part of a continuous effort by PAHO and IDB to strengthen Member States' capacity to responsibly embrace artificial intelligence in health. Through technical cooperation, policy dialogue, capacity-building initiatives, AI readiness assessments, and the development of practical implementation tools, PAHO, in collaboration with IDB, has supported countries in advancing digital transformation while safeguarding equity, transparency, and public trust. AI-GUARD complements these efforts by offering a structured governance instrument that translates strategic principles into operational decision-making practice.

AI-GUARD is also a core operational component of the PAHO-IDB AI Readiness Assessment Toolkit, contributing to a broader, integrated approach that supports countries in evaluating their preparedness for artificial intelligence adoption across governance, data, workforce, and technological domains. Within this framework, AI-GUARD provides a practical decision-support instrument focused specifically on the assessment of individual AI initiatives, ensuring that readiness considerations are translated into concrete, risk-informed implementation decisions.

AI-GUARD is designed to complement and strengthen the governance frameworks established by PAHO Resolution CD60.R9 on Digital Transformation of Health Systems, and to align with national regulatory developments across the region. The instrument promotes governance that is proportionate, evidence-informed, and aligned with national public health priorities. It encourages institutions to assess strategic value alongside readiness, to embed meaningful human oversight in AI-supported decisions, to evaluate potential bias and differential impacts across populations, and to establish monitoring mechanisms that ensure sustained performance over time.

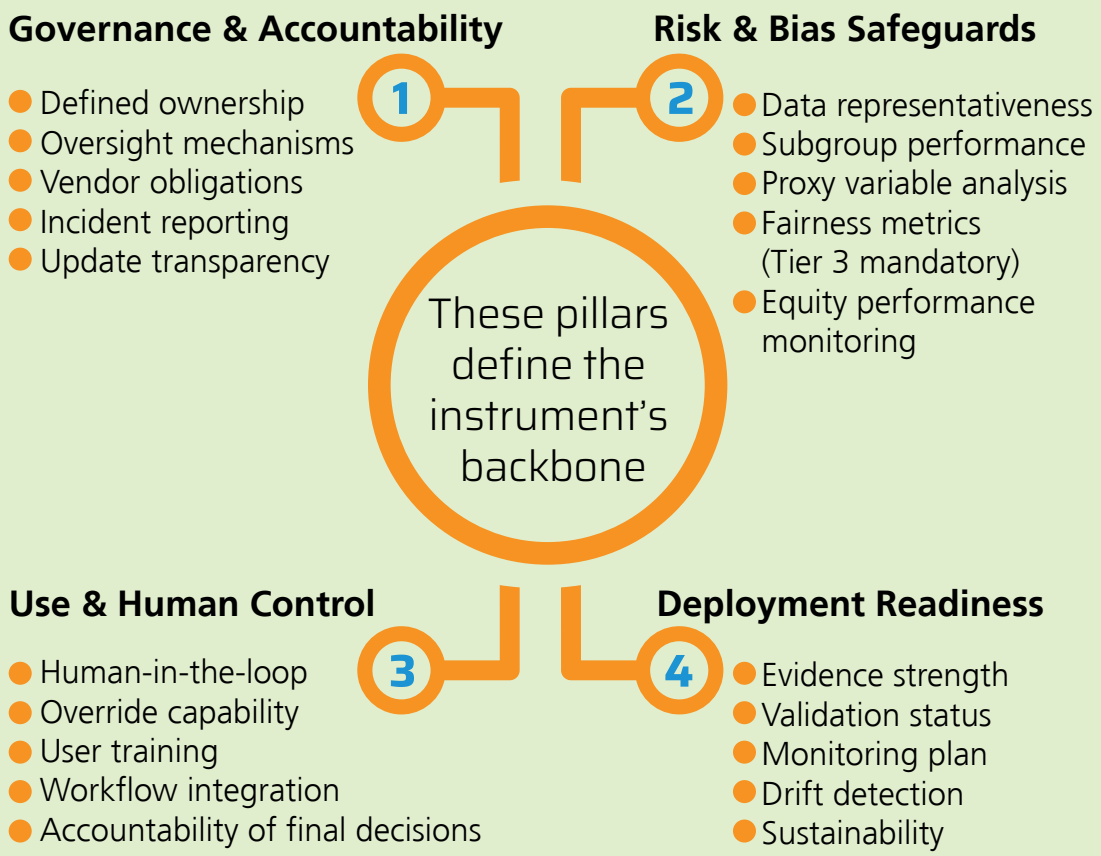
AI-GUARD does not replace national regulatory frameworks, legal review processes, or procurement rules. Rather, it strengthens institutional decision-making processes by providing a structured methodology to assess readiness and safeguards prior to deployment. In doing so, AI-GUARD supports responsible innovation that contributes to health system resilience, protects equity, enhances public trust, and ensures that technological advancement translates into measurable public health benefit.

AI-GUARD General Structure

AI-GUARD structured Decision Path

- **1.** Initiative Definition
- **2.** Public Health Value Assessment
- **3.** Risk Exposure Classification
- **4.** Tier-Based AI-GUARD Pillar Assessment
- **5.** Structured Recommendation & Safeguards

The Four Core AI-GUARD Pillars



AI-GUARD Tier Model

Tier 1	Tier 2	Tier 3
Responsible AI Use	Program-Level AI	High-Impact AI Governance

- **Tier 1.** Responsible AI Use for:
  - Generative AI-type usage
  - Administrative automation
  - Internal workflow tools

Purpose: Enable safe everyday AI usage without bureaucratic overload.

- **Tier 2.** Program-Level AI for:
  - Risk scoring tools
  - Public-facing apps
  - Clinical support systems
  - Surveillance dashboards

Purpose: Provide structured readiness screening before adoption or pilot.

- **Tier 3.** High-Impact AI Governance for:
  - National outbreak detection
  - Diagnostic AI
  - Resource allocation systems
  - Eligibility models

Purpose: Ensure responsible governance before high-impact deployment.

Final Recommendation:

- Proceed
- Temporary Conditional Authorization — Pending Verification
- Reconsider / Redesign

Executive Overview

Purpose

Artificial intelligence (AI) is increasingly influencing decision-making across health systems. From administrative automation, research and clinical decision support to national surveillance and resource allocation, AI technologies offer significant opportunities to improve efficiency, access, quality, and public health preparedness.

At the same time, AI initiatives may introduce risks related to governance, accountability, bias, data protection, operational sustainability, and unintended consequences, particularly when deployed at scale or affecting vulnerable populations.

AI-GUARD has been developed to support decision-makers at all levels of health institutions in assessing the readiness and risk profile of any AI initiative before adoption, procurement, development, or scaling.

AI-GUARD is a structured, tiered instrument designed to:

- Evaluate the strategic public health value of AI initiatives.
- Identify risk exposure and potential impact.
- Assess institutional governance and operational readiness
- Detect and mitigate bias and equity-related concerns.
- Strengthen transparency, accountability, and responsible deployment.

The instrument promotes disciplined, evidence-informed decision-making while enabling innovation in alignment with public health priorities.

Scope of Application

AI-GUARD should be applied prior to:

- Procuring AI-based products or platforms.
- Developing AI models internally.
- Scaling pilot AI initiatives.
- Integrating AI into clinical or public health workflows.
- Introducing general AI tools (e.g., large language models) into institutional operations.

The instrument is applicable across:

- Ministries of Health.
- Public health agencies.
- Hospitals and clinical institutions.
- National health programs.
- Digital health and information system units.
- Regulatory and oversight bodies.

AI-GUARD is designed to be adaptable to both low-risk and high-impact AI initiatives through a tiered assessment model.

What AI-GUARD Is Not

AI-GUARD is not:

- A regulatory approval mechanism.
- A certification or accreditation tool.
- A substitute for legal, ethical, or procurement review.
- A barrier to innovation.

Rather, it is a structured advisory instrument intended to strengthen institutional readiness and responsible decision-making before AI deployment.

- **Transparency:** Evidence, limitations, and system updates must be clearly documented.
- **Sustainability:** Long-term operational viability and monitoring must be planned before deployment.

Structure of the Instrument

AI-GUARD operates through a tiered, adaptive framework:

1. **Executive Entry Scan:** A brief assessment to determine strategic value, risk exposure, and preliminary readiness.
2. **Tier Classification:** Automatic assignment to one of three levels based on impact and risk:
    - Tier 1: Responsible AI Use

- Tier 2: Program-Level AI
- Tier 3: High-Impact AI Governance

3. **Four Core Pillar Assessment:** Governance, human control, risk & bias safeguards, and deployment readiness.
4. **AI-GUARD Dashboard Summary:** A structured output including indices and recommended next steps.

This structure ensures consistency across institutions while maintaining flexibility for different levels of complexity.

Expected Outcomes

Application of AI-GUARD supports institutions to:

- Improve the quality of AI-related decisions.
- Identify governance and readiness gaps before deployment.
- Strengthen safeguards for high-impact systems.
- Reduce exposure to operational, reputational, and equity-related risks.
- Promote responsible, transparent, and sustainable innovation.

AI-GUARD Tool

Executive Entry Scan

Strategic Value

Public Health Problem Addressed

What primary objective does the initiative aim to achieve?

(Select one)

- ☐ Administrative efficiency or workflow improvement (1)
- ☐ Reduction of waiting times or improved access to services (2)
- ☐ Improved use of health data for planning, policy decisions, or service organization (3)
- ☐ Strengthened surveillance or early detection (3)
- ☐ Strategic resource allocation or system optimization (4)
- ☐ Exploratory or undefined objective (4)

Justification for AI Use

Why is artificial intelligence considered necessary for this initiative?

(Select one)

- ☐ Conventional digital tools are insufficient for the intended objective (1)
- ☐ AI may improve speed, scale, prediction, or analytical capacity (2)
- ☐ AI is proposed mainly by an external partner, donor, or vendor (3)
- ☐ AI has been considered without clear comparative justification (4)

IInitiative Identification

What type of AI initiative is being considered?

(Select one)

- ☐ Use of a generative AI tool for limited operational support (e.g., large language model, drafting, summarization, administrative automation) (1)
- ☐ Procurement of an AI-based product or platform from an external provider (2)
- ☐ Scaling, institutional expansion, or broader deployment of an existing AI initiative (3)
- ☐ Internal development of an AI model or system (4)

Decision Impact

What level of decision impact does the initiative have?

(Select one)

- ☐ No direct impact on patient care, programmatic decisions, or resource allocation (1)
- ☐ Supports decision-making but does not automate or determine final decisions (2)
- ☐ Influences prioritization, planning, or allocation of resources, or interventions (3)
- ☐ Directly influences diagnosis, eligibility, benefit determination, or critical decisions (4)

Population Impact

Who is affected by this initiative?

(Select one)

- ☐ Internal staff only or limited internal operational users (1)
- ☐ A defined program population or specific beneficiary group (2)
- ☐ General population or multiple population groups (3)
- ☐ Vulnerable, underserved, or disproportionately affected populations (4)

Preliminary Readiness

Evidence Base

What level of evidence supports the effectiveness of this initiative?

(Select one)

- ☐ Strong validation evidence (independent, peer-reviewed, or externally validated in comparable settings) (1)
- ☐ Pilot evidence or limited validation (2)
- ☐ Vendor claims without independent validation (3)
- ☐ No clear supporting evidence (4)

Governance Readiness

Are governance and accountability clearly defined for this initiative?

(Select one)

- ☐ Clear ownership, formal oversight structure, and decision responsibility defined (1)
- ☐ Ownership defined; oversight informal or developing (2)

☐ Governance roles unclear or only partially assigned (3)

☐ Governance and accountability not yet defined (4)

Executive Scan Scoring Guidance

The Executive Entry Scan provides three preliminary outputs:

- Strategic Value (Low / Moderate / High)
- Risk Exposure (Low / Moderate / High)
- Preliminary Readiness (Weak / Developing / Strong)

Strategic Value Estimation

High Strategic Value typically includes initiatives that:

- Address clearly defined public health priorities.
- Demonstrate measurable impact.
- Improve access, surveillance, or system performance.

Limited Strategic Value may be indicated when objectives are exploratory or not clearly aligned with institutional priorities.

Risk Exposure Estimation

Risk Exposure is elevated when:

- The initiative directly influences diagnosis or eligibility decisions.
- It affects the allocation of public resources.
- It impacts vulnerable or underserved populations.
- It automates human decision-making.
- It operates at large or national scale.

Preliminary Governance Readiness Estimation

Governance Readiness is stronger when:

- Clear ownership is defined.
- Oversight mechanisms are in place.
- Accountability and reporting processes are established

Executive Scan Summary Table

Complete this summary after calculating the average score for each dimension.

Dimension	Sections included	Calculation	Average score	Interpretation
Strategic Value	3.1 + 3.2 + 3.3	(3 scores ÷ 3)	____	Low / Moderate / High
Risk Exposure	3.4 + 3.5	(2 scores ÷ 2)	____	Low / Moderate / High
Preliminary Readiness	3.6 + 3.7	(2 scores ÷ 2)	____	Weak / Developing / Strong

Final Tier Assignment

Final Classification	Assessment – Check Section 3.10
Assigned Tier	☐ Tier 1 – Responsible AI Use ☐ Tier 2 – Program-Level AI ☐ Tier 3 – High-Impact AI Governance

Executive Conclusion

Item	Assessment
Initiative Type	_____ (concise free text description of the AI initiative being assessed)
Overall Recommendation (complete after index calculation; see Sec. 4 for Tier 1, Sec. 5 for Tier 2, Sec. 6 for Tier 3)	☐ Proceed ☐ Proceed with Conditions ☐ Reconsider / Redesign

Tier Determination

Tier 1 – Responsible AI Use

Tier 1 applies to low-risk AI initiatives that do not directly influence clinical decisions, eligibility determinations, or allocation of public resources.

Examples include:

- Use of large language models for drafting or summarizing documents.
- Administrative automation tools.
- Internal workflow optimization systems.
- Non-clinical data analysis tools.
- Translation or transcription applications.

Tier 1 ensures that basic governance, accountability, and data protection safeguards are in place without imposing unnecessary administrative burden.

Estimated completion time: 5–7 minutes.

Governance & Accountability

(Tier 1 Scope)

Confirm the following:

Item	Yes	Partial	No
Defined ownership for AI tool use	☐	☐	☐
Clear internal guidance on acceptable use	☐	☐	☐
Vendor or provider terms reviewed	☐	☐	☐
Process for reporting misuse or incidents	☐	☐	☐
Update or version transparency understood	☐	☐	☐

Guidance

- Ownership should identify a responsible unit or officer.
- Even low-risk tools require clear accountability.
- Staff should know whom to contact in case of error or concern.

Use & Human Control

(Tier 1 Scope)

Confirm the following:

Item	Yes	Partial	No
Human verification required before output use	☐	☐	☐
AI outputs not treated as final decisions	☐	☐	☐
Users informed of limitations and hallucination risk	☐	☐	☐
AI use does not replace professional judgment	☐	☐	☐
Staff training or awareness provided	☐	☐	☐

Guidance

- AI-generated outputs must always be reviewed by a human.
- Staff must understand that AI tools may produce inaccurate or fabricated content.
- Professional accountability remains with the human user.

Risk & Bias Safeguards

(Tier 1 Scope)

Confirm the following:

Item	Yes	Partial	No
No identifiable patient data entered into public AI tools	☐	☐	☐
Sensitive data protection guidance available	☐	☐	☐
Data sharing policies respected	☐	☐	☐
Awareness of potential content bias	☐	☐	☐

Guidance

- Tier 1 (Responsible AI Use) tools must not be used to process identifiable health data unless approved secure environments are in place.
- Staff should avoid entering confidential, proprietary, or legally protected information into external systems.

Deployment Readiness

(Tier 1 Scope)

Confirm the following:

Item	Yes	Partial	No
Tool evaluated for operational relevance	☐	☐	☐
Use case clearly defined	☐	☐	☐
Basic monitoring or feedback mechanism exists	☐	☐	☐
Sustainability or subscription implications considered	☐	☐	☐

Guidance

- Even low-risk AI use should serve a defined operational purpose.
- Institutions should consider licensing, cost and continuity implications.

Tier 1 - Scoring Summary

(Responsible AI Use)

Assign scores as follows:

Yes = 2 points	Partial = 1 point	No = 0 points

Calculate total possible points across all Tier 1 items (and normalize the score as a percentage of total possible points).

Tier 1 Interpretation

Proceed

80-100 % of total points	PHVI ≥ **50**	Composite Institutional Readiness ≥ **60**

Temporary Conditional Authorization – Pending Verification

60-79 % of total points

Reconsiderar / Rediseñar

Below **60%** of total points

Tier 1 Recommendation Template

Assessment Result:

☐ Proceed

☐ Temporary Conditional Authorization — Pending Verification

☐ Strengthen Safeguards Before Use

If conditions are required, specify corrective actions below:

1. .....

2. .....

Tier 2 – Program-Level AI

Tier 2 applies to AI initiatives that:

- Support or influence decision-making.
- Operate at program, institutional, or regional level.
- Affect identifiable patient populations.
- Involve public-facing tools or service delivery components.
- Require structured governance and monitoring.

Examples include:

- Risk scoring systems.
- Clinical decision support tools.
- Public-facing health applications.
- Surveillance dashboards.
- Triage assistance systems.
- Program-level predictive analytics.

Tier 2 ensures structured governance, bias safeguards, and monitoring mechanisms before deployment.

Estimated completion time: 15–20 minutes.

Governance & Accountability

Confirm the following:

Item	Yes	Partial	No
Defined ownership and responsible authority	☐	☐	☐
Formal oversight mechanism established	☐	☐	☐
Vendor obligations clearly documented	☐	☐	☐
Incident reporting process defined	☐	☐	☐
Update notification procedures defined	☐	☐	☐
Clear documentation of system limitations	☐	☐	☐

Guidance

- Governance should include a named accountable authority.
- Vendor contracts should address transparency, updates, and performance reporting.
- Known limitations must be documented prior to deployment.

Use & Human Control

Confirm the following:

Item	Yes	Partial	No
Human review required before action	☐	☐	☐
Override capability available	☐	☐	☐
Override actions logged and reviewable	☐	☐	☐
Users trained in system capabilities and limitations	☐	☐	☐
Clear delineation between AI recommendation and final decision	☐	☐	☐

Guidance

- AI should not support professional judgment.
- Override mechanisms must be technically feasible and operationally clear.
- Training should include system limitations and error scenarios.

Risk & Bias Safeguards

Confirm the following:

Item	Yes	Partial	No
Training data sources documented	☐	☐	☐
Population representativeness assessed	☐	☐	☐
Subgroup performance evaluated	☐	☐	☐
Proxy variables reviewed for unintended bias	☐	☐	☐

Item	Yes	Partial	No
Fairness or equity considerations documented	☐	☐	☐
Plan for equity monitoring post-deployment	☐	☐	☐

Guidance

- Assess whether the training data reflects the intended deployment population.
- Subgroup performance should be examined where feasible.
- Proxies (e.g., geographic location, socioeconomic indicators) should be reviewed carefully.
- Equity impacts must be considered before scale-up.

Deployment Readiness

Confirm the following:

Item	Yes	Partial	No
Evidence supporting effectiveness documented	☐	☐	☐
External or independent validation available	☐	☐	☐
Monitoring plan defined	☐	☐	☐
Performance metrics clearly specified	☐	☐	☐
Drift detection or recalibration plan defined	☐	☐	☐
Sustainability and operational capacity assessed	☐	☐	☐
Recurring operational cost estimate over 24 months and identified funding source	☐	☐	☐
Human oversight staffing sustainability	☐	☐	☐

Guidance

- Validation evidence should be reviewed independently of vendor marketing claims.
- Monitoring plans should include frequency, responsibility, and reporting mechanisms.
- Institutions must confirm operational capacity before implementation.

Tier 2 - Scoring Summary

(Program-Level AI)

Assign scores as follows:

Yes = 2 points	Partial = 1 point	No = 0 points

Calculate total possible points across all Tier 2 items.

Then calculate (see Section 7 for calculation and normalization details):

- Governance Readiness Index (GRI)
- Use & Human Control Index
- Bias & Equity Risk Index (BERI)
- Evidence & Transparency Index (ETI)

Normalize each to a 0–100 scale according to the scoring framework in Section 7.

Tier 2 Interpretation

For Tier 3 initiatives:

Proceed

- Public Health Value Index ≥ 60
- Composite Institutional Readiness ≥ 65
- Bias & Equity Risk Index (BERI) ≥ 60

Temporary Conditional Authorization — Pending Verification

- Strategic value present, but readiness gaps identified.
- Safeguards are required prior to implementation.

Reconsider / Redesign

- Limited strategic value.
- Significant governance or bias deficiencies.
- Insufficient evidence base.

Tier 2 Recommendation Template

Assessment Result:

☐ Proceed

☐ Temporary Conditional Authorization — Pending Verification

☐ Reconsider / Redesign

Required Safeguards (if applicable):

1. .....

2. .....

3. .....

Tier 3 – High-Impact AI Governance

Tier 3 applies to AI initiatives that:

- Directly influence diagnosis, eligibility, or benefit determination.
- Influence allocation of public resources.
- Operate at national or large-scale level.
- Affect vulnerable or underserved populations.
- Automate or substantially replace human decision-making.
- Process identifiable health data at scale.

Examples include:

- National outbreak detection systems.
- AI-assisted diagnostic tools.
- Resource allocation or prioritization models.
- Eligibility determination systems.
- Population-level predictive risk models.

Tier 3 requires comprehensive governance structures, formal oversight, bias mitigation safeguards, and continuous monitoring before deployment.

Estimated completion time: 25–35 minutes.

Governance & Accountability

For Tier 3 initiatives, governance structures must be formal and documented.

Confirm the following:

Item	Yes	Partial	No
Defined institutional ownership at senior level	☐	☐	☐
Formal oversight committee or review body established	☐	☐	☐
Clear legal and regulatory review completed	☐	☐	☐
Vendor contracts include transparency and audit clauses	☐	☐	☐
Incident reporting and escalation protocol defined	☐	☐	☐
Version control and update notification procedures documented	☐	☐	☐
Independent evaluation permitted by contract	☐	☐	☐

● Mandatory Requirement

For Tier 3, absence of formal oversight or defined ownership significantly limits readiness and may prevent deployment. The AI-GUARD assessment must be conducted by a review team that does not include the project lead or primary development team. This internal cross-review must be documented and include a conflict-of-interest declaration.

External review may be conducted by academic institutions, peer health agencies, national regulatory bodies, or qualified independent experts. Technical cooperation may be requested to support this review where applicable.

Use & Human Control

Tier 3 systems must preserve meaningful human oversight.

Definitions Box — Section 6.2

- **Human-in-the-Loop (HITL):** A human reviews and approves the AI recommendation before any action is taken.

- **Human-on-the-Loop (HOTL):** The system acts by default but a human can intervene and override within a defined window. Higher risk than HITL.
- **Human-in-Command (HIC):** A human retains full operational control and can suspend or deactivate the system at any time. Required for Tier 3 deployment.

(Note: This oversight taxonomy aligns with the "Human Agency and Oversight" requirements defined by the European Commission's High-Level Expert Group on AI, and supports the WHO guiding principle of "Protecting human autonomy" in healthcare).

Confirm the following:

Item	Yes	Partial	No
Human review required before final decision (HITL implemented)	☐	☐	☐
Individual override capability technically implemented	☐	☐	☐
System suspension or emergency deactivation capability (HIC) implemented	☐	☐	☐
Override decisions logged and auditable	☐	☐	☐
User training formally documented	☐	☐	☐
Failure and error scenarios defined	☐	☐	☐

● **Mandatory Requirement**

Tier 3 systems must operate with **Human-in-the-Loop (HITL**) oversight unless a shift to **Human-on-the-Loop (HOTL)** or fully automated execution is explicitly justified, documented, and approved under legal and ethical review.

Regardless of the operational mode (HITL, HOTL, or autonomous), all Tier 3 systems **must maintain Human-in-Command (HIC)** capabilities, ensuring a designated authority can safely deactivate the system in the event of a Level 3 incident (see Annex E).

Risk & Bias Safeguards

Tier 3 (High-Impact AI Governance) requires structured fairness and representativeness assessment.

Confirm the following:

Item	Yes	Partial	No
Training data sources fully documented	☐	☐	☐
Population representativeness formally assessed	☐	☐	☐
Subgroup performance metrics evaluated	☐	☐	☐
Proxy variables analyzed for bias risk	☐	☐	☐
Fairness metrics applied and documented	☐	☐	☐
Equity impact assessment conducted	☐	☐	☐
Plan for ongoing equity monitoring defined	☐	☐	☐

Mandatory Requirements

For Tier 3:

- Subgroup performance evaluation is required.
- Fairness metrics must be documented.
- Absence of subgroup testing significantly reduces readiness.

Deployment Readiness

Tier 3 systems require robust validation and monitoring before deployment.

Confirm the following:

Item	Yes	Partial	No
External or independent validation completed	☐	☐	☐
Model card or equivalent documentation available	☐	☐	☐
Performance metrics clearly defined	☐	☐	☐
Monitoring and evaluation plan documented	☐	☐	☐
Drift detection or recalibration plan defined	☐	☐	☐
Adverse impact monitoring plan defined	☐	☐	☐

Item	Yes	Partial	No
Operational sustainability assessed	☐	☐	☐
Recurring operational cost estimate over 24 months and identified funding source	☐	☐	☐
Recurring operational cost estimate over 24 months Human oversight staffing sustainability	☐	☐	☐

Mandatory Requirement

For Tier 3:

- External validation is strongly recommended.
- A continuous monitoring plan must be documented prior to deployment.
- An unsustainable financial plan should be treated as a readiness gap requiring 'Temporary Conditional Authorization — Pending Verification' at minimum.

Tier 3 - Scoring Summary

(High-Impact AI Gobernance)

Assign scores as follows::

Yes = 2 points	Partial = 1 point	No = 0 points

Calculate sub-scores for:

- Governance Readiness Index (GRI)
- Use & Human Control Index
- Bias & Equity Risk Index (BERI)
- Evidence & Transparency Index (ETI)

Normalize each to a 0–100 scale according to the scoring framework in Section 7.

Tier 3 Interpretation

For Tier 3 initiatives:

- **Proceed**
  - Public Health Value Index (PHVI) ≥ 70
  - Composite Institutional Readiness ≥ 70
  - Bias & Equity Risk Index (BERI) ≥ 70
- **Temporary Conditional Authorization — Pending Verification**
  - If remediable gaps can be identified.
- **Reconsider / Redesign**
  - If structural deficiencies are present.

Tier 3 Recommendation Template

- **Assessment Result:**

☐ Proceed

☐ Temporary Conditional Authorization — Pending Verification

☐ Reconsider / Redesign

Mandatory Safeguards Prior to Deployment:

1. .....

2. .....

3. .....

Oversight Body Responsible:

.....

Monitoring Frequency:

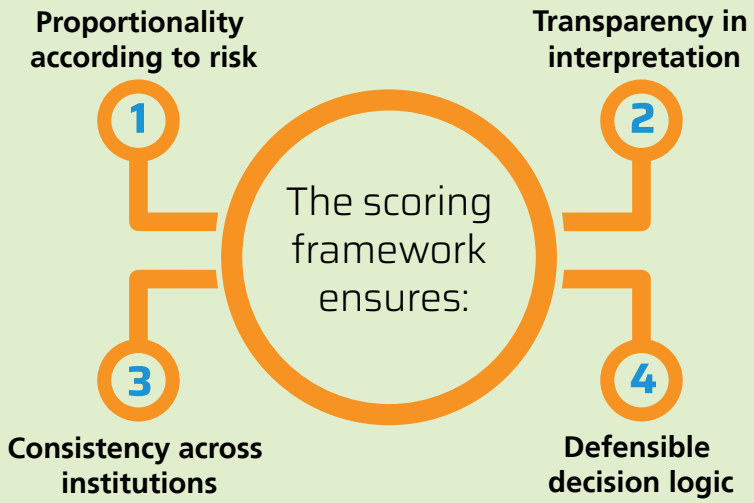
.....

Reassessment Schedule:

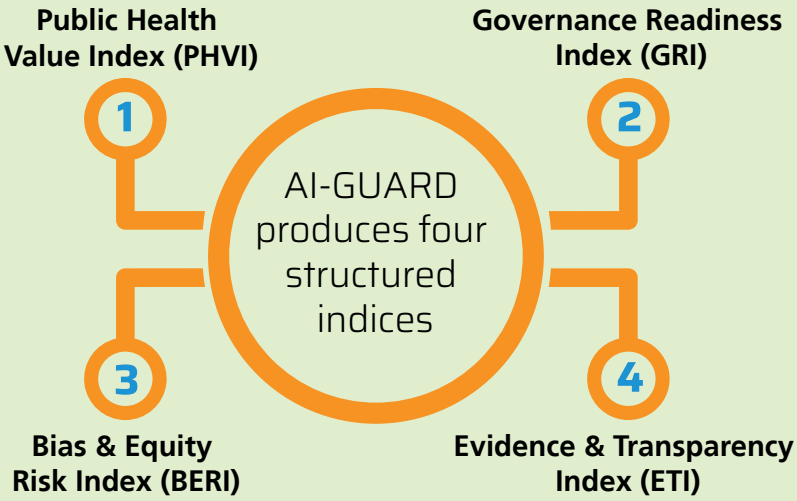
.....

AI-GUARD Scoring Framework and Decision Rules

This section describes how AI-GUARD assessment results are calculated and how final recommendations are determined.



Overview of AI-GUARD Indices



Additionally, Tier classification reflects the level of risk exposure.

Final recommendations are based on the combined interpretation of:

- Tier level
- PHVI
- Composite Institutional Readiness
- BERI

Public Health Value Index (PHVI)

The Public Health Value Index evaluates the strategic justification of the AI initiative.

PHVI is calculated based on:

- Problem clarity
- Measurable outcomes
- Alignment with national or institutional strategy
- Expected impact on access, efficiency, surveillance, or system performance
- Sustainability and integration capacity

Each component is scored on a standardized scale and normalized to 0–100.

Interpretation of PHVI

PHVI Score	Interpretation
80–100	High Strategic Value
60–79	Moderate Strategic Value
40–59	Limited Strategic Value
Below 40	Weak Justification

Initiatives with weak justification should be reconsidered, particularly in Tier 2 or Tier 3 contexts.

Governance Readiness Index (GRI)

GRI reflects the presence and strength of:

- Defined ownership
- Oversight mechanisms
- Vendor transparency obligations
- Incident reporting structures
- Version control and update procedures

Scores are derived from Tier-specific assessment items and normalized to 0–100.

Higher-tier initiatives require more formal governance structures to achieve full readiness.

Bias & Equity Risk Index (BERI)

BERI reflects the degree to which bias and equity risks are mitigated.

It evaluates:

- Data representativeness
- Subgroup performance assessment
- Proxy variable analysis
- Fairness metrics (mandatory in Tier 3)
- Equity impact monitoring

Scores are normalized to 0–100.

Higher scores indicate stronger bias mitigation safeguards.

In Tier 3 initiatives:

- Absence of subgroup evaluation significantly reduces BERI.
- Absence of fairness metrics may prevent progression to full approval.

Evidence & Transparency Index (ETI)

ETI evaluates:

- Validation evidence
- External or independent evaluation

- Model documentation (e.g., model card)
- Defined monitoring plan
- Drift detection mechanisms
- Operational sustainability planning

Scores are normalized to 0–100.

Higher-tier initiatives require stronger evidence to meet readiness thresholds.

Use & Human Control Index

It evaluates:

- The presence of human review prior to action and clarity of the human role in AI supported decisions.
- The availability and effectiveness of override mechanisms allowing users to intervene or reject AI outputs.
- The logging and auditability of override actions and human interventions.
- The existence of system suspension or emergency deactivation capabilities, where applicable.
- User training and awareness of system capabilities, limitations, and failure scenarios.
- Clear delineation between AI recommendations and final human decisions, including accountability assignment.

Scores are normalized to a 0–100 scale.

Higher tier initiatives require stronger and more formalized human control mechanisms to meet readiness thresholds. In Tier 3 initiatives, absence of technically implemented human override or emergency deactivation capabilities may prevent progression to full approval.

Composite Institutional Readiness

Composite Institutional Readiness is calculated as the average of:

- Governance Readiness Index (GRI)
- Use & Human Control Index
- Evidence & Transparency Index (ETI)

This composite reflects whether the institution is structurally prepared to deploy the AI initiative responsibly.

Tier-Based Thresholds

AI-GUARD applies progressively stricter thresholds according to tier level.

Tier 1 – Responsible AI Use	
Proceed when:	
PHVI ≥ **50**	Composite Institutional Readiness ≥ **60**

Tier 1 emphasizes safe operational use rather than structural governance.

Tier 2 – Program-Level AI		
Proceed when:		
PHVI ≥ **60**	Composite Institutional Readiness ≥ **65**	BERI ≥ **60**

Tier 2 (Program-Level AI) requires structured governance and bias safeguards.

Nivel 3: Gobernanza de la IA de alto impacto		
Proceed when:		
PHVI ≥ **70**	Composite Institutional Readiness ≥ **70**	BERI ≥ **70**

Tier 3 requires robust governance, bias mitigation, and evidence validation prior to deployment.

Final Recommendation Categories

AI-GUARD produces one of three outcomes:

- **Proceed**

The initiative demonstrates adequate strategic value, governance readiness, bias safeguards, and evidence.

- **Temporary Conditional Authorization — Pending Verification**

The initiative shows strategic value but requires corrective actions for readiness gaps before implementation and full deployment.

Conditions may include:

- Strengthening governance structures
- Completing subgroup evaluation
- Conducting external validation
- Bias mitigation
- Establishing monitoring plans

There are important procedures to be considered:

- A Compliance Deadline must be established, not to exceed 90 days from the date of assessment for Tier 2, and 60 days for Tier 3.
- Each required condition must be assigned to a named responsible party.
- The designated oversight authority must formally verify compliance and sign a Compliance Verification Statement before the system transitions to full deployment.

● Reconsider / Redesign

Significant deficiencies exist in strategic justification, governance, bias mitigation, or evidence strength.

Redesign or further institutional preparation is required before reassessment.

Reassessment and Continuous Review

For Tier 2 (Program-Level AI) and Tier 3 (High-Impact AI Governance) initiatives:

- Reassessment is recommended upon major system updates.
- Reassessment is recommended prior to scaling.
- Periodic review should be conducted at intervals appropriate to system impact.

Hard Fail Rules – Non-compensable requirements

The following rules apply independently of composite index scores. When any Hard Fail criterion is triggered, the overall AI-GUARD recommendation is automatically set to Reconsider / Redesign, regardless of PHVI, GRI, BERI, or ETI values. No scoring average can override a Hard Fail determination.

#	Hard Fail Criterion	Applicable Tier(s)	Automatic Outcome
HF-1	Absence of subgroup performance evaluation	Tier 3 (mandatory) Tier 2 (recommended)	Reconsider / Redesign
HF-2	Absence of documented external or independent validation	Tier 3 (mandatory)	Reconsider / Redesign
HF-3	Absence of technically implemented and documented human override mechanism	Tier 2 & 3	Reconsider / Redesign
HF-4	No defined institutional ownership at senior level	Tier 2 & 3	Reconsider / Redesign
HF-5	Absence of a continuous monitoring plan prior to deployment	Tier 3	Reconsider / Redesign

AI-GUARD Assessment Summary Template

The AI-GUARD Assessment Summary provides a structured overview of the evaluation results. It should be completed after the full assessment and retained as part of institutional documentation.

This summary may be used in internal decision-making meetings, procurement discussions, oversight reviews, or funding deliberations.

AI-GUARD Assessment Summary

Initiative Title:

.....

Institution / Program:

.....

Assessment Date:

.....

Responsible Authority:

.....

Tier Classification:

☐ Tier 1 – Responsible AI Use

☐ Tier 2 – Program-Level AI

☐ Tier 3 – High-Impact AI Governance

AI-GUARD Indices

Index	Score (0–100)	Interpretation
Public Health Value Index (PHVI)	_____	☐ High ☐ Moderate ☐ Limited ☐ Weak
Governance Readiness Index (GRI)	_____	☐ Strong ☐ Adequate ☐ Developing ☐ Weak
Bias & Equity Risk Index (BERI)	_____	☐ Strong Safeguards ☐ Moderate ☐ Limited ☐ High Risk
Evidence & Transparency Index (ETI)	_____	☐ Strong ☐ Adequate ☐ Limited ☐ Insufficient

Composite Institutional Readiness: _____

Overall Recommendation

☐ Proceed

☐ Temporary Conditional Authorization — Pending Verification

☐ Reconsider / Redesign

Required Safeguards

(if applicable)

List corrective actions required prior to deployment or scaling:

1. .....

2. .....

3. .....

4. .....

Governance Oversight

- **Tier 2 (Program-Level AI)**
- **Tier 3 (High-Impact AI Governance)**

Oversight Body Responsible:

……..

Monitoring Frequency:

……..

Reassessment Schedule:

……..

……..

Summary Narrative

Provide a brief narrative (3–5 sentences) summarizing the rationale for the decision:

……. .

……. .

……. .

……. .

……. .

Final Approval

Name and Title of Reviewing Authority:

……. .

Signature: ……..

Date: ……. .

Annexes

Annex A: How to Use AI-GUARD

AI-GUARD is designed to be applied in a structured and sequential manner. The instrument can be completed by an individual decision-maker, a technical team, or through a facilitated institutional review process, depending on the scale and complexity of the AI initiative under consideration.

The assessment consists of four main stages.

Step 1 – Complete the Executive Entry Scan

All AI initiatives should begin with the Executive Entry Scan.

This brief assessment:

- Clarifies the type and scope of the initiative.
- Identifies the intended impact and affected population.
- Estimates strategic value.
- Assesses preliminary governance readiness.
- Determines the appropriate AI-GUARD tier level.

The Executive Entry Scan typically requires approximately five minutes to complete.

Upon completion, the initiative will be classified into one of three tiers:

- **Tier 1** – Responsible AI Use
- **Tier 2** – Program-Level AI
- **Tier 3** – High-Impact AI Governance

The assigned tier determines the depth of assessment required in subsequent steps.

The Executive Entry Scan is the initial decision-entry stage of the AI-GUARD assessment process. It provides a rapid appraisal of the strategic relevance, expected public health value, potential risk exposure, and preliminary governance readiness of a proposed artificial intelligence initiative before any detailed technical assessment is undertaken. This section should be completed before proceeding to tier-specific assessment.

Scoring approach for the Executive Entry Scan

Each response selected in sections **3.1 to 3.6** receives a score from **1 to 4**, reflecting the level of governance relevance, potential impact, and institutional responsibility associated with that answer.

The scoring scale is interpreted as follows:

● 1 = Lowest governance concern / simplest condition
● 2 = Moderate-low
● 3 = Moderate-high
● 4 = Highest governance concern / strongest governance need

A higher score does **not** indicate better quality. It indicates that the initiative may require stronger governance attention, deeper assessment, or additional safeguards.

All questions use the same **1–4 scale**, even when a question includes more than four answer choices. In such cases, scores are assigned according to the governance significance of each answer rather than the order in which options appear.

For each question, only **one option should be selected**: the option that best represents the highest applicable condition at the current stage of the initiative. If more than one option appears applicable, the option with the **highest governance relevance** should be chosen.

After completing sections **3.1 to 3.6**, add all selected scores and divide the total by the number of answered questions to obtain an **average score**.

The average score is interpreted as follows:

● 1.0 a 1.9 = Low	● 2.0 a 2.9 = Moderate	● 3.0 a 4.0 = High

This average score supports the preliminary estimation of:

● Strategic Value	● Risk Exposure	● Preliminary Readiness

These three dimensions are summarized in section **3.7** and used to support **tier determination in section 3.8**.

The average score should not be interpreted mechanically. Certain responses, particularly those involving diagnosis, eligibility, resource allocation, vulnerable populations, or large -scale deployment, may require a higher tier even when the overall average is moderate.

A high score does **not automatically indicate readiness to proceed**. It may also indicate that the initiative has significant governance implications and requires stronger safeguards before implementation.

Step 2 –Tier Determination

Tier classification is based on the level of impact and risk exposure associated with the AI initiative.

Key determinants include:

- Whether the system influences clinical or eligibility decisions.
- Whether it affects the allocation of public resources.
- Whether it targets or disproportionately affects vulnerable populations.
- Whether it automates or replaces human decision-making.
- Whether it operates at national or large-scale program level.
- Whether it processes identifiable health data.

The tier level ensures that assessment depth is proportional to potential impact.

Institutions should not manually downgrade tier classification. If uncertainty exists, the higher tier should be applied as a precautionary approach.

Based on the entry scan responses, assign the initiative to one of the following tiers. Use the qualitative definitions below together with the consolidated interpretation of Strategic Value, Risk Exposure and Preliminary Readiness from Section 3.9. When uncertainty exists between tiers, the higher tier should be applied.

Definición cualitativa

Tier 1: Responsible AI Use

Low decision impact, minimal risk exposure, and limited population impact.

Tier 2: Program-Level AI

Moderate impact, supports or influences decisions, requires structured governance review.

● **Tier 3:** High-Impact AI Governance

Direct influence on diagnosis, eligibility, or resource allocation; affects vulnerable populations; or operates at large scale.

If uncertainty exists between tiers, the higher tier should be applied.

Decision Rule

Use the following rule to translate the Executive Entry Scan interpretation into a **preliminary** Tier assignment. This rule is **supportive, not mechanical**, and must be applied together with the qualitative criteria in this section.

● **Tier 1 – Responsible AI Use:**

If the Executive Entry Scan yields **three "Low"** across the dimensions, **or two "Low" plus one "Moderate"**, classify as **Tier 1**.

● **Tier 2 – Program-Level AI:**

If the Executive Entry Scan yields **two "Moderate"** (or more), **or three "Moderate"**, classify as **Tier 2**, unless Tier 3 qualitative triggers apply.

● **Tier 3 – High-Impact AI Governance:**

If the Executive Entry Scan indicates **two or more "Strong/High"** implications (e.g., high Risk Exposure and high Strategic Value in national/large-scale contexts), classify as Tier 3. Any **Tier 3** qualitative trigger below also sets Tier 3.

● Step 3 – Complete the Tier-Specific Assessment

Following tier confirmation, the initiative must undergo evaluation under the Four Core AI-GUARD Pillars:

● 1 Governance & Accountability
● 2 Use & Human Control
● 3 Risk & Bias Safeguards
● 4 Deployment Readiness

The depth and number of assessment items increase according to tier level.

● **Tier 1 - Responsible AI Use:** focuses on basic safeguards for low-risk AI use.

- **Tier 2 - Program-Level AI:** requires structured review of governance, bias mitigation, and monitoring plans.

- **Tier 3 - High-Impact AI Governance:** requires comprehensive governance structures, validation evidence, and continuous monitoring mechanisms.

Assessment items are scored according to the AI-GUARD scoring framework described in Section 5.

All responses should be documented with supporting evidence where available.

Step 4 – Review the AI-GUARD Dashboard Summary

Upon completion of the assessment, results are summarized in the AI-GUARD Dashboard.

The Dashboard includes four indices:

- 1 Public Health Value Index (PHVI)
- 2 Bias & Equity Risk Index (BERI)
- 3 Governance Readiness Index (GRI)
- 4 Evidence & Transparency Index (ETI)

Based on the combined results, the initiative will receive one of the following recommendations:

- Proceed
- Temporary Conditional Authorization — Pending Verification
- Reconsider / Redesign

For initiatives receiving "Temporary Conditional Authorization — Pending Verification" or "Reconsider / Redesign," AI-GUARD will identify required safeguards and readiness gaps that must be addressed prior to implementation.

Institutional Responsibilities

AI-GUARD results should be:

- Reviewed by the designated responsible authority.

- Documented and retained as part of project records.
- Reassessed if the AI system undergoes substantial modification, scaling, or functional change.

For Tier 3 (High-Impact AI Governance) initiatives, periodic reassessment is strongly recommended.

Frequency of Application

AI-GUARD should be applied:

- Before procurement decisions.
- Before pilot implementation.
- Before national or program-scale deployment.
- When significant model updates occur.
- When governance or operational context changes.

The instrument is intended to support continuous institutional learning and responsible AI governance. AI models, particularly those used in clinical or epidemiological contexts, are subject to performance degradation over time due to data drift, population change, and evolving clinical practice. Because of that it is necessary to have a reassessment periodically.

AI-GUARD assessments carry a defined validity period. Institutions must conduct a full reassessment before the applicable expiration date or upon any mandatory early reassessment trigger, whichever occurs first.

Tier	Standard Expiration	Mandatory Early Reassessment Triggers
Tier 1	24 months	Significant change in tool use scope or data handling practices
Tier 2	18 months	Major model update; change in target population; expansion to new program areas; adverse event reports
Tier 3	12 months	Any update to the underlying model; scaling to new geographic areas; change in deployment context; adverse impact detected; legal or regulatory changes

Annex B: Simulation – Tier 1: Responsible AI Use

Case: Use of a Large Language Model for Policy Drafting

Initiative Description

A Ministry of Health department proposes authorizing the use of an enterprise large language model (LLM), such as ChatGPT or Gemini, to support internal administrative work, including:

- drafting policy briefs
- summarizing technical reports
- translating internal documents
- improving writing clarity

No identifiable patient data will be entered into the system. Use is limited to internal administrative functions and outputs remain subject to human review before institutional use.

Executive Entry Scan – Section Scoring

Section	Selected score
3.1 Public Health Problem Addressed	1
3.2 Justification for AI Use	2
3.3 Initiative Identification	1
3.4 Decision Impact	1
3.5 Population Impact	1
3.6 Evidence Base	2
3.7 Governance Readiness	2

Executive Scan Summary Table

Dimension	Sections included	Calculation	Average score	Interpretation
Strategic Value	3,1 + 3,2 + 3,3	(1+2+1) ÷ 3	1,3	Low
Risk Exposure	3,4 + 3,5	(1+1) ÷ 2	1,0	Low
Preliminary Readiness	3,6 + 3,7	(2+2) ÷ 2	2,0	Developing

Final Tier Assignment

Tier 1 – Responsible AI Use

Executive Conclusion

Item	Assessment
Initiative Type	Use of generative AI tool for limited operational support
Overall Recommendation	Temporary Conditional Authorization – Pending Verification

Rationale

- The initiative improves internal operational efficiency.
- It does not directly influence clinical decisions, eligibility, or resource allocation.
- Population exposure is minimal because use remains internal.
- Governance requirements remain moderate but still require institutional guidance.

Tier 1 Assessment Summary

Governance & Accountability

- Defined ownership: Yes (2)
- Internal usage guidance: Partial (1)

- Vendor terms reviewed: Yes (2)
- Incident reporting defined: Limited (0)

Use & Human Control

- Human verification required: Yes (2)
- AI output not treated as final decision: Yes (2)
- User training provided: Partial (1)

Risk & Bias Safeguards

- No identifiable patient data entered: Yes (2)
- Confidentiality guidance provided: Yes (2)
- Awareness of hallucination risk: Partial (1)

Deployment Readiness

- Use case clearly defined: Yes (2)
- Operational sustainability considered: Yes (2)
- Feedback mechanism defined: Partial (1)

Total: 76,9 ▸ Pending Verification

Recommended Safeguards

- Formalize internal institutional guidance for generative AI use.
- Provide structured user training on limitations and output verification.
- Establish periodic monitoring of usage patterns and recurring risks.

Tier 1 – Scoring Walkthrough

This walkthrough illustrates how to translate **Yes / Partial / No** selections into a **percentage score** for Tier 1, consistent with Section **4.5 (Yes=2; Partial=1; No=0)**. First, sum the points across all Tier 1 items (Sections **4.1–4.4**). Then **normalize the total as a percentage of the maximum possible points**, and interpret the result using Section **4.6**.

Step 1 — Point assignment (example layout):

- Governance & Accountability (4 items) ▸ [5 / 8]
- Use & Human Control (3 items) ▸ [5 / 6]

- Risk & Bias Safeguards (3 items) ▸ [5 / 6]
- Deployment Readiness (3 items) ▸ [5 / 6]

Step 2 — Normalization, PHVI and Composite Instituional Readiness:

Normalization:

$$\text{Tier 1 Score (\%)} \frac{\text{Sum of points}}{\text{Max points (36)}} \text{x100} = \frac{20}{26} \text{x100} = 76{,}9\%$$

PHVI:

55, Interpretation: Limited Strategic Value (40–59)

- Problem clarity: The initiative has a clear and bounded objective (internal administrative support), but it does not address a defined public health problem directly.
- Measurable outcomes: Efficiency gains (time saved, productivity) are measurable, but indirectly related to public health outcomes.
- Alignment with institutional strategy: The use aligns with general digital transformation and administrative efficiency goals, but not with core service delivery, surveillance, or population health priorities.
- Expected impact: Impact is internal and operational, with no direct effect on access, quality of care, surveillance, or system performance at population level.
- Sustainability and integration capacity: High—enterprise LLMs are relatively easy to sustain and integrate for administrative use.

GRI: 62,5

Derived from Tier 1 Governance & Accountability items: 5/8 points ▸ (5/8) × 100.

UHCI: 83,3

Derived from Tier 1 Use & Human Control items: 5/6 points ▸ (5/6) × 100.

ETI: 83,3

Derived from Tier 1Deployment Readiness items: 5/6 ▸ (5/6) × 100.

Composite Institutional Readiness:

GRI: 62.5%; UHCI: 83.3%, ETI: 83.3%

▸ Composite = 76.4 %

Step 3 — Interpretation (Section 4.6):

- **Proceed:**
  - 80 – 100 %
  - PHVI ≥ 50
  - Composite Institutional Readiness ≥ 60
- **Pending Verification:** 60 – 79% (FALLS HERE)
- **Redisegn:** Below 60 %

Annex C: Simulation – Tier 2: Program-Level AI

Case: AI-Based Emergency Department Triage Support Tool

Initiative Description

A regional hospital proposes adoption of an AI-supported triage tool to help predict patient deterioration risk in the emergency department.

The system:

- supports clinician prioritization
- does not replace final clinical decisions
- uses historical electronic health record data
- affects patient care workflows and operational prioritization

The tool is intended to improve responsiveness in emergency care while maintaining human clinical oversight.

Executive Entry Scan – Section Scoring

Section	Selected score
3.1 Public Health Problem Addressed	2
3.2 Justification for AI Use	2
3.3 Initiative Identification	2
3.4 Decision Impact	3
3.5 Population Impact	2
3.6 Evidence Base	2
3.7 Governance Readiness	2

Executive Scan Summary Table

Dimension	Sections included	Calculation	Average score	Interpretation
Strategic Value	3,1 + 3,2 + 3,3	(2+2+2) ÷ 3	2,0	Moderate
Risk Exposure	3,4 + 3,5	(3+2) ÷ 2	2,5	Moderate
Preliminary Readiness	3,6 + 3,7	(2+2) ÷ 2	2,0	Developing

Final Tier Assignment

Tier 2 – Program-Level AI

Executive Conclusion

Item	Assessment
Initiative Type	Procurement of an AI-based product supporting emergency department triage
Overall Recommendation	Proceed with Conditions

Rationale

- The initiative supports clinical prioritization but does not automate final decisions.
- It affects identifiable patients and operational care flow.
- Moderate governance safeguards are required because decisions influence prioritization within a clinical environment.

Tier 2 Assessment Summary

Governance & Accountability

- Defined ownership: Yes
- Formal oversight mechanism: Partial

- Vendor transparency clauses: Yes
- Incident reporting defined: Partial

Use & Human Control

- Human review required: Yes
- Override capability available: Yes
- Override logging implemented: Partial
- User training conducted: Partial

Risk & Bias Safeguards

- Training data documented: Yes
- Population representativeness assessed: Partial
- Subgroup performance evaluated: Partial
- Proxy bias analysis conducted: No
- Equity monitoring plan defined: Partial

Deployment Readiness

- Pilot validation available: Yes
- External validation: No
- Monitoring plan documented: Partial
- Drift detection plan defined: No

Recommended Safeguards

- Conduct subgroup performance evaluation before wider deployment.
- Perform proxy variable bias analysis.
- Establish structured monitoring and drift detection procedures.
- Formalize periodic oversight review.

Reassessment Recommendation

A reassessment is recommended before broader institutional or regional expansion.

Tier 2 – Index Calculation Walkthrough

This walkthrough illustrates how to translate **Yes / Partial / No** selections into **normalized indices** for Tier 2, consistent with **Section 5.5** (Yes = 2; Partial = 1; No = 0) and the **AI GUARD Scoring Framework (Section 7)**. First, assign points to the Tier specific items (Sections **5.1–5.4**). Then normalize each pillar index to **0–100**, compute the **Composite Institutional Readiness** (Section **7.7**), and interpret the result using **Tier 2 thresholds** (Sections **5.6 / 7.8**).

● Step 1 — Point assignment (example layout)

- Governance & Accountability (selected items used in this example: 4) ▸ [6 / 8]
  - Yes (Defined ownership, Vendor transparency) = 2 + 2
  - Partial (Formal oversight, Incident reporting) = 1 + 1
- Use & Human Control (selected items used in this example: 4) ▸ [6 / 8]
  - Yes (Human review, Override capability) = 2 + 2
  - Partial (Override logging, User training) = 1 + 1
- • Risk & Bias Safeguards (selected items used in this example: 5) ▸ [5 / 10]
  - Yes (Training data) = 2
  - Partial (Representativeness, Subgroup performance, Equity monitoring) = 1 + 1 + 1
  - No (Proxy bias analysis) = 0
- Deployment Readiness (selected items used in this example: 4) ▸ [3 / 8]
  - Yes (Pilot validation) = 2
  - Partial (Monitoring plan) = 1
  - No (External validation, Drift detection) = 0 + 0

● Step 2 — Normalization, PHVI and Composite Institutional Readiness

Public Health Value Index (PHVI):

Derived from the five PHVI components (problem clarity, measurable outcomes, strategic alignment, expected impact, sustainability/integration). For this use case (ED triage support improving responsiveness with human oversight), the component level appraisal yields: PHVI = 65 ▸ Moderate Strategic Value.

Normalization to a 0–100 scale (per pillar):

- **GRI** = (6/8) × 100 = **75**
- **UHCI** = (6/8) × 100 = **75**
- **BERI** = (5/10) × 100 = **50**
- **ETI** = (3/8) × 100 = **37,5**

Composite Institutional Readiness (per 7.7):

$$\text{"Composite"} = \frac{\text{"GRI"} + \text{"UHCI"} + \text{"ETI"}}{3} = \frac{75+75+37.5}{3} = 62{,}5\%$$

Step 3 — Interpretation (Sections 5.6 / 7.8)

Tier 2 – Proceed when all are met:

- PHVI ≥ 60 ▸ Met (65)
- Composite Institutional Readiness ≥ 65 ▸ No met (62,5)
- BERI ≥ 60 ▸ No met (50)

Overall Recommendation: Temporary Conditional Authorization — Pending Verification (falls here)

What must be addressed prior to full deployment (see B.8):

- Bias & Equity: complete subgroup performance evaluation, proxy bias analysis, and formalize equity monitoring.
- Evidence & Transparency: obtain external/independent validation; define performance thresholds (incl. calibration); formalize monitoring & drift detection.
- Governance & Human Control: formal oversight mechanism; override logging/audit trail; structured user training and role boundaries (AI recommendation vs final decision).

Annex D: Simulation – Tier 3: High-Impact AI Governance

Case: National AI-Based Early Outbreak Detection System

Initiative Description

A Ministry of Health proposes deployment of a national AI-supported surveillance system to detect early signals of infectious disease outbreaks using:

- electronic health records
- laboratory reporting systems
- syndromic surveillance data

The system is intended to support:

- allocation of emergency resources
- prioritization of public health response strategies
- early identification of population-level risks

The initiative processes identifiable health data at national scale and may influence decisions affecting vulnerable populations.

Executive Entry Scan – Section Scoring

Section	Selected score
3.1 Public Health Problem Addressed	3
3.2 Justification for AI Use	1
3.3 Initiative Identification	4
3.4 Decision Impact	4
3.5 Population Impact	4
3.6 Evidence Base	3
3.7 Governance Readiness	4

Executive Scan Summary Table

Dimension	Sections included	Calculation	Average score	Interpretation
Strategic Value	3,1 + 3,2 + 3,3	(3+1+4) ÷ 3	2,7	Moderate–High
Risk Exposure	3,4 + 3,5	(4+4) ÷ 2	4,0	High
Preliminary Readiness	3,6 + 3,7	(3+4) ÷ 2	3,5	Weak

Final Tier Assignment

Tier 3 – High-Impact AI Governance

Executive Conclusion

Item	Assessment
Initiative Type	Internal development of a national AI-supported surveillance system
Overall Recommendation	Reconsider / Redesign

Rationale

- The initiative operates at national scale.
- It influences prioritization of emergency public health action and resource allocation.
- It affects populations with unequal vulnerability.
- Governance and evidence remain insufficient for immediate deployment.

Tier 3 Assessment Summary

Governance & Accountability

- Senior ownership defined: Yes
- Formal oversight committee: Not yet established
- Vendor audit clauses: Partial
- Incident reporting protocol: Limited

- **Use & Human Control**
  - Human oversight defined: Partial
  - Override logging implemented: No
  - Formal training program implemented: No
- **Risk & Bias Safeguards**
  - Training data documented: Yes
  - Subgroup performance evaluated: No
  - Fairness metrics applied: No
  - Equity impact assessment conducted: No
- **Deployment Readiness**
  - External validation completed: No
  - Model card available: Draft only
  - Monitoring plan defined: Partial
  - Drift detection plan defined: No

Mandatory Safeguards Prior to Pilot Deployment

- Establish a formal multidisciplinary oversight committee.
- Conduct subgroup performance and fairness evaluation.
- Complete independent validation studies.
- Define continuous monitoring and drift detection procedures.
- Consolidate governance documentation and accountability lines.

Reassessment Requirement

A full reassessment is required before pilot implementation.

Tier 3 – Index Calculation Walkthrough

This walkthrough details how to compute **Tier 3 indices** from **C.7** checklist using **Yes = 2; Partial = 1; No = 0**, then **normalize to 0–100**. First, assign points to the Tier specific items (Sections **6.1–6.4**). Then **normalize** each pillar index to **0–100**, compute the **Composite Institutional Readiness**, and interpret against **Tier 3 thresholds** and **Hard Fail rules**.

Step 1 — Point assignment (example layout)

- **Governance & Accountability ▸ [4 / 8]**
  - **Yes:** Senior ownership defined = **2**
  - **No:** Formal oversight committee not yet established = **0**
  - **Partial:** Vendor audit clauses = **1**
  - **Partial:** Incident reporting protocol (limited) = **1**
- **Use & Human Control ▸ [1 / 6]**
  - **Partial:** Human oversight defined = **1**
  - **No:** Override logging implemented = **0**
  - **No:** Formal training program implemented = **0**
- **Risk & Bias Safeguards ▸ [2 / 8]**
  - **Yes:** Training data documented = **2**
  - **No:** Subgroup performance evaluated = **0**
  - **No:** Fairness metrics applied = **0**
  - **No:** Equity impact assessment conducted = **0**
- **Deployment Readiness ▸ [2 / 8]**
  - **No:** External/independent validation completed = **0**
  - **Partial:** Model card available (draft) = **1**
  - **Partial:** Monitoring plan defined = **1**
  - **No:** Drift detection plan defined = **0**

Step 2 — Normalization, PHVI and Composite Institutional Readiness

Public Health Value Index (PHVI) (per 7.2):

Derived from the five PHVI components (problem clarity, measurable outcomes, strategic alignment, expected impact, sustainability/integration). For a **national early outbreak detection system**, the component level appraisal is typically high on public health value:

- **Problem clarity** (early detection of outbreaks at population level) = **85**
- **Measurable outcomes** (timely detection, earlier response, resource targeting) = **70**
- **Alignment with strategy** (core national surveillance priority) = **85**

- **Expected impact** (resource allocation, emergency response prioritization) = **80**
- **Sustainability & integration** (complex multi system integration; high O&M requirements) = **60**

$$\text{"PHVI"} = \frac{85+70+85+80+60}{5} = 76 \text{ ("High Strategic Value")}$$

Normalization to a 0–100 scale (per pillar):

- **GRI** = (4/8) × 100 = **50,0**
- **UHCI** = (1/6) × 100 = **16,7**
- **BERI** = (2/8) × 100 = **25,0**
- **ETI** = (2/8) × 100 = **25,0**

Composite Institutional Readiness (per 7.7):

$$\text{"Composite"} = \frac{\text{"GRI"}+\text{"UHCI"}+\text{"ETI"}}{3} = \frac{50.0+16.7+25.0}{3} = 30,6\%$$

Step 3 — Interpretation

Tier 3 – Proceed thresholds (all must be met):

- PHVI ≥ 70 ▸ Met (76)
- • Composite Institutional Readiness ≥ 70 ▸ No met (30,6)
- BERI ≥ 70 ▸ No met (25,0)

Hard Fail Rules (per 7.10):

- **HF-1:** Absence of **subgroup performance evaluation ▸ Triggered** (C.7 = No).
- **HF-2:** Absence of **documented external/independent validation ▸ Triggered** (C.7 = No).
- **HF-3:** Absence of **technically implemented human override ▸ Likely** (not documented; if confirmed **No**, triggers HF 3).

Overall Recommendation: Reconsider / Redesign

What must be addressed prior to any pilot (see C.8):

- **Governance:** establish a formal multidisciplinary oversight committee, define legal/regulatory review, document versioning & update notification, and allow independent evaluation by contract.
- **Human Control:** implement HITL (or justify HOTL), ensure HIC emergency deactivation, and override logging; deliver formal user training and define failure scenarios.
- **Bias & Equity:** perform subgroup performance analysis, apply fairness metrics, assess equity impact, and plan ongoing equity monitoring.
- **Evidence & Transparency:** complete external/independent validation, finalize model card, formalize monitoring & drift detection, define performance metrics (incl. thresholds & alerting), and document operational sustainability & resources.

Annex E: Vendor Evidence & Model Card Requirements

This annex provides a structured template to request essential documentation and evidence from vendors or developers of AI systems intended for deployment in health settings.

For Tier 2 (Program-Level AI) and Tier 3 (High-Impact AI Governance) initiatives, this annex should be formally transmitted to vendors as part of due diligence or procurement processes.

Minimum Vendor Documentation Requirements

The following documentation should be requested prior to procurement or deployment:

1. General System Information

- System name and version
- Developer organization
- Contact point for technical accountability
- Intended use and target population
- Deployment context assumptions

2. Model Description

- Type of AI model (e.g., machine learning, deep learning, NLP)
- Description of input data sources
- Description of output format
- Decision-support vs automated decision function
- Known system limitations

3. Training Data Transparency

Vendors should provide:

- Description of training data sources
- Geographic origin of training data
- Time period covered
- Population characteristics (age, sex, region, relevant demographics)
- Data inclusion and exclusion criteria

For Tier 3 initiatives, subgroup representation disclosure is strongly recommended.

4. Performance Metrics

Vendors should provide:

- Primary performance metric(s)
- Sensitivity and specificity (where applicable)
- False positive and false negative rates
- Calibration performance (if risk scoring)
- Subgroup performance (where applicable)
- Validation dataset characteristics

For Tier 3 initiatives, subgroup performance evaluation is expected.

5. Fairness & Bias Mitigation Measures

Vendors should disclose:

- Whether bias testing was conducted
- Methodology used for bias evaluation
- Fairness metrics applied
- Identified disparities (if any)
- Mitigation measures implemented
- Ongoing fairness monitoring plans

For Tier 3 (High-Impact AI Governance) initiatives, absence of fairness documentation should be considered a major readiness gap.

6. Validation & Evaluation

- Internal validation methods
- External or independent validation studies
- Peer-reviewed publications (if available)
- Real-world pilot results
- Regulatory approvals (if applicable)

Tier 3 systems should demonstrate external validation where feasible.

7. Governance & Update Transparency

Vendors should specify:

- Frequency of model updates
- Change notification procedures

- Version control documentation
- Audit log availability
- Incident reporting collaboration procedures

8. Monitoring & Drift Management

- Post-deployment monitoring strategy
- Performance monitoring indicators
- Drift detection methods
- Recalibration procedures
- Responsibilities for performance degradation management

Model Card Template (Minimum Fields)

Institutions may request that vendors complete the following structured model card.

Model Card – Minimum Required Information

1. Model Overview

- Model name and version
- Developer
- Date of release
- Intended use
- Intended users

2. Model Inputs and Outputs

- Description of input variables
- Output format and interpretation
- Threshold definitions (if applicable)

3. Training Data Summary

- Data sources
- Population coverage
- Data time frame
- Data preprocessing steps

4. Performance Metrics

- Overall performance
- Subgroup performance (if applicable)
- Calibration metrics (if applicable)

5. Fairness Assessment

- Fairness metrics applied
- Identified disparities
- Mitigation strategies

6. Limitations and Warnings

- Known limitations
- Situations where the model should not be used
- Population groups where performance may vary

7. Monitoring Plan

- Post-deployment monitoring frequency
- Drift detection mechanisms
- Contact point for reporting issues

Interpretation Guidance

Absence of documentation in any of the above domains should be reflected in:

- Governance Readiness Index (GRI)
- Bias & Equity Risk Index (BERI)
- Evidence & Transparency Index (ETI)

For Tier 3 initiatives, absence of subgroup performance evaluation or validation evidence may prevent progression to "Proceed" status.

Strategic Impact of Annex D

This annex:

- Protects institutions during procurement
- Elevates vendor accountability
- Encourages transparency

- Aligns with global best practices in AI governance
- Strengthens AI-GUARD's credibility

Internally Developed and Open-Source Adapted Models

D.4.1 Scope of Application

This section applies when an institution or team:

- Develops an AI model internally using institutional or national health data.
- Fine-tunes an open-source foundation model using local, regional, or national health datasets.
- Deploys an open-source model in a health system context without material modification but assumes operational responsibility for its outputs.

D.4.2 Principle of Equivalent Accountability

The internal development team or responsible institution assumes all documentation, bias testing, and monitoring obligations that Sections D.1 and D.2 assign to external commercial vendors. The absence of a third-party vendor does not reduce these obligations.

D.4.3 Additional Requirements Specific to Open-Source Adaptation

In addition to all requirements in D.1 and D.2, the following must be documented:

- Identity and version of the foundation model used (e.g., Llama 3.1 8B, Mistral 7B v0.3)
- Known limitations, biases, and documented risks of the foundation model as reported by its original developers.
- Description of the fine-tuning dataset: origin, size, demographic coverage, time period, and preprocessing steps.
- Bias introduced or inherited through fine-tuning, including subgroup performance on the local dataset.
- Named institutional authority responsible for ongoing maintenance, monitoring, and incident response.

For Tier 3 applications, internal development teams must complete the full Model Card template (Section D.2) prior to any deployment, including in pilot phases.

Annex F - Incident Response and Emergency Deactivation Protocol

Purpose

This annex defines the minimum requirements for incident identification, classification, escalation, and emergency deactivation for AI systems deployed under AI-GUARD Tier 2 and Tier 3. All Tier 3 systems must complete this protocol prior to deployment. Tier 2 systems are strongly encouraged to adopt it.

Incident Classification

The following event types constitute reportable incidents:

- **Level 1 - Minor:** Isolated output errors with no patient or population impact; corrected by user override.
- **Level 2 - Moderate:** Systematic errors affecting a defined population group, workflow disruption, or evidence of emerging performance degradation (data drift).
- **Level 3 - Critical:** Evidence of significant patient harm, mass false negatives/positives affecting public health response, discriminatory outputs across population subgroups, or loss of human oversight capability.

Escalation Pathway

Prior to deployment, the responsible institution must define and document:

- **Named responsible authority** for each incident level.
- **Maximum response time** by level (suggested: L1 = 5 business days; L2 = 48 hours; L3 = immediate).
- **Communication pathway** to affected stakeholders and, where applicable, to national regulatory authorities.

Emergency Deactivation Criteria (Rollback Protocol)

The following conditions must trigger immediate system suspension pending investigation and remediation:

- Detection of a Level 3 incident as defined in E.2.
- Loss of human override capability.
- Performance degradation below pre-defined thresholds (to be established during Deployment Readiness assessment).
- Unresolved Level 2 incidents persisting beyond the established response window.

Deactivation must include a documented return-to-manual-process plan to ensure continuity of care or public health function during the suspension period.

Post-Incident Review Requirements

Following any Level 2 or Level 3 incident:

- Root cause analysis must be completed and documented.
- AI-GUARD reassessment is required before redeployment.
- Oversight authority must formally approve redeployment..

PAHO
Pan American Health Organization
World Health Organization
Americas Region
IDB
www.paho.org