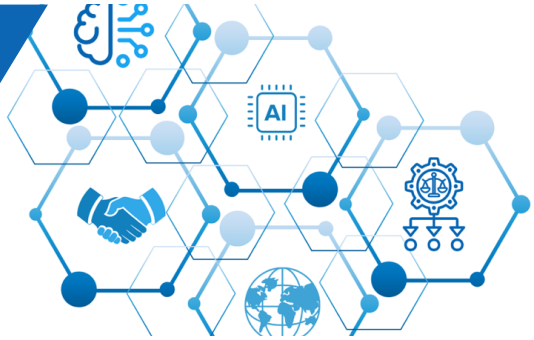
BRIEFING NOTE

ESCAP | APCICT
Asian and Pacific Training Centre for Information and Communication Technology for Development

Responsible AI Engineering and Security



Key Topics

Why the module was developed

Manila-to-implementation journey

AIDDE development approach

Course objective

Learning outcomes

Target audience and entry profile

Key benefits

OECD lifecycle backbone

Nine-session learning and evidence chain

Secure AI Development Evidence Pack

Emerging technologies and way forward

Pilot experience and learner feedback

From governance principles to engineered controls and evidence

Responsible AI Engineering and Security translates approved governance, risk and security obligations into implementable controls, verification plans and operational safeguards across the AI system lifecycle.

This briefing note introduces Specialist Module 4 of the AI Engineering and Security Package. The two-day, nine-session learning flow progresses from requirements and architecture through data, modern AI controls, verification, operations, suppliers and an evidence-bounded readiness recommendation.

The AIDDE methodology governs courseware analysis, design, development, delivery and continuous enhancement. The OECD AI system lifecycle is the subject-matter backbone, complemented by ISO/IEC 42001, secure SDLC, MLOps/LLMOps/AgentOps, OWASP/AISVS, MITRE ATLAS and selected AI security resources.

Applying responsible and secure AI engineering across the AI system lifecycle

1 | WHY THIS MODULE WAS DEVELOPED

From AI governance awareness to implementation capability

The Capacity Building Programme on Data and AI Governance for Philippine Government Officials, organised by the Department of Information and Communications Technology (DICT) and APCICT/ESCAP in Manila on 3-6 June 2025, strengthened policy, ethical and governance understanding among mid- to senior-level officials.

The programme also exposed a practical implementation gap. Policy and governance teams could articulate principles, risks and oversight expectations, while technical and project officers needed a method for converting those expectations into system requirements, engineered controls, verification evidence and operational safeguards.

The engineering question

> **PRACTICAL DESIGN QUESTION**
> How can an AI system be designed, developed, deployed, operated, changed and retired so that approved governance, risk, Responsible AI and security obligations become enforceable controls and auditable evidence?

Why principles alone are insufficient

- **Principles** do not define system boundaries, assets, identities, trust relationships or attack paths.
- **Policies and risk treatments** must become requirements, control owners, acceptance criteria and evidence.
- **Modern AI** adds retrieval, runtime, identity, tool-use and autonomy risks that conventional software controls may not fully address.
- **Safe operation** requires release gates, monitoring, incident response, recovery, change control, revalidation and secure retirement.

2 | THE IMPLEMENTATION JOURNEY

From Manila need to a refined specialist module

➢ June 2025 - capability need identified

The Manila programme established a foundation in data governance, AI fundamentals, ethics, governance, law, risk management and auditing. Follow-up discussions highlighted the need for practitioner-focused implementation training.

➢ July-September 2025 - analysis, design and development

AIDDE translated the identified need into a two-day technical and project-officer module centred on functional roles, implementation tasks, a coherent learning sequence, practical activities and cumulative evidence outputs.

➢ 1-3 October 2025 - pilot experience

The wider Trustworthy AI workshop ran in Manila from 1 to 3 October 2025. Its two-day practitioner track for project managers and technical staff provided pilot experience and learner feedback that informed the subsequent refinement Responsible AI Engineering and Security Module

➢ Post-pilot - evaluation and enhancement

Learner feedback and technical review improved the learning flow, reduced duplication, strengthened the case-based evidence chain and expanded coverage of secure AI engineering, GenAI, RAG, agentic AI, verification and operational readiness.

RESULT
A nine-session Responsible AI Engineering and Security module that connects approved obligations to engineered controls, verification plans, operational safeguards and an evidence-bounded readiness recommendation.

3 | DESIGN APPROACH

AIDDE: the courseware development pathway

AIDDE describes how the module moved from an observed capability need to a designed, developed, delivered and continuously improved curriculum. It is a courseware-development method; the OECD AI system lifecycle is the organising backbone for the subject matter.

The five phases

A. Analysis. Identify learner roles, job tasks, capability gaps, prior knowledge, delivery constraints and public-sector implementation use cases.

ID. Instructional Design. Define outcomes, boundaries, learning sequence, timing, methods, assessment evidence and specialist hand-offs.

D. Development. Create source-grounded lessons, demonstrations, learner activities, templates, the Project AMANAH case and evidence criteria.

D&I. Delivery and Implementation. Facilitate the two-day programme, apply the cumulative case, observe performance and capture implementation challenges.

E&E. Evaluation and Enhancement. Use feedback, technical review and emerging-risk analysis to revise content, activities, references and learning flow.

AIDDE outputs for the module

1. **Learner-needs and role map.** Connects the capability gap to the people who build, operate, secure, challenge, own and source AI systems.

2. **Curriculum blueprint.** Specifies nine sessions, two learning days, measurable outcomes, timing, content boundaries and evidence hand-offs.

3. **Grounded learning assets.** Defines lessons, demonstrations, learner builds, reference extracts, templates and the cumulative case.

4. **Pilot and delivery evidence.** Captures participation, engagement, applicability, facilitation observations and learner feedback.

5. **Revision record.** Documents how feedback, duplicated content, new threats and changing technologies are addressed.

Quality gates

- Each session is traceable to functional roles and implementation tasks.
- Each session produces an output that is consumed or checked later.
- Claims and activities are grounded in approved sources; unsupported details remain assumptions or evidence gaps.
- Timing allows concept input, demonstration, application and reflection.
- Evaluation leads to a documented revision, not merely a satisfaction score.

DESIGN PRINCIPLE
The module teaches one connected implementation workflow rather than nine isolated technical topics.

4 | COURSE OBJECTIVE

Translate obligations into controls, evidence and readiness

Enable cross-functional AI delivery and assurance teams to translate approved Responsible AI, risk and security obligations into testable requirements, defensible architectures, implemented-control designs, verification evidence and operational-readiness plans across the AI system lifecycle.

Using the Project AMANAH case, participants build a traceable evidence pack that supports - but does not replace - management decision-making, specialist validation or independent assurance.

What the course enables learners to do

- Define the AI system, intended purpose, lifecycle boundary, principal actors, providers and deployment responsibilities.
- Convert governance decisions and risk treatments into testable engineering requirements, control designs, release gates and acceptance criteria.
- Design Responsible AI and security together across architecture, data, model protection, identity, applications, infrastructure and operations.
- Plan verification, red teaming, monitoring, incident response, recovery, change control, revalidation and secure retirement before production.
- Integrate evidence into a bounded technical-readiness recommendation while preserving management accountability for deployment and risk acceptance.

COURSE FORMAT
Two-day specialist programme; nine linked sessions; 390 learning minutes per day; demonstrations, learner builds, a cumulative case and structured evidence hand-offs.

5 | LEARNING OUTCOMES

A coherent progression from system context to decision evidence

By the end of the module, learners should be able to:

1. **Translate obligations.** Turn approved governance obligations and risk treatments into testable requirements, release gates, acceptance criteria and evidence.
2. **Decompose AI systems.** Map ML, foundation-model, RAG and agentic systems into assets, data flows, identities, dependencies, boundaries, tools and human decision points.
3. **Engineer controls.** Design data, privacy, cryptographic, model-CIA, application, infrastructure and lifecycle-IAM controls.
4. **Bound modern AI behaviour.** Apply verified identity, scoped authority, orchestration controls, a runtime control harness, guardrails, tool gateways, sandboxing, human confirmation, safe stop and rollback.
5. **Plan verification and operations.** Create integrated testing, red-team, deployment, monitoring, incident-recovery and supplier-assessment plans.
6. **Integrate evidence.** Produce a traceable Secure AI Development Evidence Pack and an evidence-bounded readiness recommendation.

Assessment emphasis

BOUNDARY
The module assesses the quality, traceability and defensibility of learner artefacts. It does not imply that planned controls have been implemented, tests have been executed, or management risk acceptance has been granted.

6 | TARGET AUDIENCE AND ENTRY REQUIREMENTS

Cross-functional implementation roles with shared accountability

1. **Build and architect.**
   AI/ML, GenAI, RAG and agent engineers; developers; solution, security and enterprise architects.
2. **Data and model assurance.**
   Data engineers and stewards; model validators; privacy engineering and data-protection specialists.
3. **Platform and operate.**
   MLOps, LLMOps, AgentOps, cloud, platform, site-reliability and infrastructure engineers.
4. **Secure and respond.**
   Application and cloud security, penetration testing, red team, SOC, threat hunting, digital forensics and incident response.
5. **Govern and challenge.**
   Responsible AI, AI governance, technology risk, compliance and internal audit roles requiring implementation literacy.
6. **Own and source.**
   Product, system and service owners; procurement; third-party risk; and supplier-management roles.
7. **Entry requirements.**
   Participants should have foundation-to-intermediate knowledge of AI delivery, risk, data, privacy, secure development or cybersecurity. No coding is required for course exercises; technical participants may optionally implement selected controls in a lab environment.
8. **Delivery requirement.**
   AI-assisted activities require facilitator-supplied sources, the Project AMANAH case and an approved enterprise AI assistant. Organisational authorisation, data-handling rules and access controls apply.

7 | KEY BENEFITS

What the module adds to organisational capability

1. **One traceable method.** Connects requirement to threat, control, test, telemetry, recovery evidence and recommendation.
2. **Evidence by design.** Defines what must be recorded during build and operation before management or assurance asks for proof.
3. **Modern AI coverage.** Differentiates controls for ML, foundation models, multimodal AI, RAG, tools, agents and multi-agent patterns.
4. **Operational readiness.** Plans release, monitoring, incident response, recovery, change, revalidation and retirement before production.
5. **Clear hand-offs.** Separates engineering readiness, management decision and risk acceptance, specialist validation and independent assurance.
6. **AI-assisted, evidence-grounded learning.** Uses structured RACE prompts with an approved enterprise AI assistant to accelerate analysis, develop first drafts and explore engineering scenarios.

> **LEARNER RESPONSIBILITY**
> All AI-generated outputs must be critically reviewed, validated against facilitator-supplied sources and clearly marked where evidence is missing or an assumption has been made.

Recommended pre-reading

- Primer on ML, foundation and multimodal models, RAG, agents and AI system components.
- The organisation's AI governance, security, privacy, risk and secure-development requirements.
- Project AMANAH case and supplied governance/risk hand-off evidence.

A simple end-to-end frame for responsible and secure AI

The module adapts the OECD AI system lifecycle into seven learner-friendly delivery stages. The lifecycle is not a waterfall: monitoring, incidents, material changes and retirement decisions can return the team to requirements, threat modelling, control design and revalidation.

Seven learner-friendly stages

1. **Plan and design.** Define intended purpose, stakeholders, system boundary, obligations, risk treatments and acceptance criteria.
2. **Data.** Establish authority, provenance, quality, representativeness, privacy, security, lineage, retention and recovery controls.
3. **Build or adapt.** Develop, configure, fine-tune or integrate models, RAG, agents, applications, platforms and controls.
4. **Test and verify.** Evaluate functionality, performance, Responsible AI, privacy, security, resilience and operational readiness.
5. **Deploy.** Approve a controlled baseline, apply release gates, stage or canary the change, and preserve rollback evidence.
6. **Operate and monitor.** Observe behaviour, detect threats and drift, manage vulnerabilities, respond to incidents, recover and revalidate changes.
7. **Retire.** Decommission models, data, RAG stores, tools, identities and suppliers securely, retaining required records and deletion evidence.

Seven stages, continued

Four complementary technical scaffolds

1. **Management and accountability.** ISO/IEC 42001 A.6-A.10 anchors lifecycle expectations, owners, governance controls and evidence hand-offs.
2. **Engineering and operations.** Secure SDLC plus MLOps, LLMOps and AgentOps translate requirements into implementation, release and runtime control.
3. **Threats and verification.** OWASP GenAI and agent resources, AISVS and MITRE ATLAS support threat knowledge, selectable requirements and tests.
4. **Practitioner completeness.** Selected AI security resources extend infrastructure, identity, cryptography, SecOps, incident response and lifecycle coverage.

> **CORE MESSAGE**
> The lifecycle tells learners where they are; the scaffolds tell them what must be governed, engineered, secured and evidenced at that point.

9 | NINE-SESSION LEARNING AND EVIDENCE CHAIN

Day 1 - frame, threat-model and engineer the control baseline

Day 1 moves from the approved purpose and lifecycle boundary to a defensible architecture, data-control and identity baseline. Each session owns a distinct decision and produces a named artefact.

1. 75 minutes - From Governance Requirements to Secure AI Engineering

Introduce the Course Learning Journey, Define the system and lifecycle boundary, clarify accountability, introduce the AI concepts that materially change attack surface, and create the traceable engineering backlog.

Output: *M4-01 Requirements and System/Security Context*

2. 90 minutes - AI Architecture Decomposition and Formal Threat Modelling

Map components, flows, assets, identities and trust boundaries; apply complementary threat-modelling lenses; prioritise adversarial ML, GenAI, RAG, agent and resource-exhaustion scenarios.

Output: *M4-02 Architecture, Trust Boundaries and Threat Model*

3. 105 minutes - Secure AI Data, Privacy and Cryptographic Engineering

Design preventive, detective and recovery controls for training, validation, operational, telemetry, RAG/vector and synthetic-data flows.

Output: *M4-03 Data, Privacy and Cryptographic Control Matrix*

Day 1, continued

4. 120 minutes - Responsible and Secure AI Architecture, Infrastructure, Model Protection and IAM

Build defence in depth across applications, models, RAG and agents, pipelines and platforms; protect model confidentiality, integrity and availability; control human and non-human identities.

Output: *M4-04 Architecture, Model CIA and IAM Register*

Day 1 control logic

1. **Start with.** Approved purpose, obligations, risk treatments and supplied case evidence.
2. **Build.** Requirements, threat model, data controls, architecture and identity controls.
3. **Preserve.** Unique IDs, owners, acceptance criteria, assumptions, evidence gaps and evidence references.
4. **Hand off to Day 2.** A controlled platform, model-CIA and IAM baseline ready to be extended, verified and operationalised.

> **OVERLAP CONTROL**
> Session 2 identifies threats; Sessions 3-5 design controls; Session 6 designs verification. A later session consumes an earlier artefact but does not reteach its core content.

Day 2 - secure modern AI, verify, operate, recover and recommend

5. 75 minutes - Engineering GenAI, RAG and Agentic AI Safely

Address prompt and context attacks, RAG poisoning and excessive agency. Explain agents, orchestration, the runtime control harness, guardrails and tool gateways; apply scoped authority, sandboxing, human confirmation, safe stop and rollback.

Output: *M4-05 GenAI, RAG and Agent Control Design*

6. 75 minutes - Integrated AI Verification, Red Teaming and Findings Triage

Turn requirements and controls into repeatable functional, Responsible AI, privacy and security tests; define evidence and acceptance criteria; distinguish vulnerabilities, control failures, Responsible AI failures and model limitations.

Output: *M4-06 Verification and Findings Management Plan*

7. 90 minutes - Secure MLOps, LLMOps and AgentOps; AI SecOps, Incident Response and Recovery

Plan gated release, controlled baselines, telemetry, runtime detection, posture management, containment, forensics, trusted recovery, change, revalidation and secure retirement.

Output: *M4-07 Deployment, AI SecOps, Incident Response and Recovery Plan*

Day 2, continued

8. 75 minutes - AI Supply Chain, Provenance and Third-Party Security

Assess external models, datasets, APIs, cloud services, libraries and tools; record provenance, shared responsibility, AIBOM/SBOM, supplier evidence, concentration risk, contingency and exit readiness.

Output: *M4-08 Supply Chain and Supplier Assessment*

9. 75 minutes - Capstone: Secure AI Deployment Evidence and Readiness Recommendation

Integrate readiness gates and traceability; use only supplied evidence; record open threats, failed or not-run tests, conditions, owners, due dates and revalidation triggers.

Output: *M4-09 Final Evidence Pack and Readiness Recommendation*

Recommendation options

> **EVIDENCE-BOUNDED CONCLUSION**
> Recommend Go | Recommend Conditional Go | Recommend Controlled Pilot Only | Recommend No-Go

Management retains the formal deployment decision and risk acceptance. Specialist validation and independent assurance remain separate from the learner team's technical-readiness recommendation.

Signature learning sequence

Frame and require -> threat-model -> engineer controls -> verify -> plan release, operations and recovery -> assess dependencies -> integrate evidence -> recommend.

> **DAY 2 TOTAL**
> 75 + 75 + 90 + 75 + 75 = 390 learning minutes.

10 | EMERGING TECHNOLOGIES AND WAY FORWARD

A controlled refresh path for a fast-moving discipline

The module will be periodically reviewed so that its customer promise and evidence chain remain stable while technical examples, threats, control patterns and references evolve.

Priority review areas

- Foundation and multimodal models; RAG and retrieval architectures; agentic and multi-agent systems.
- Tool protocols, orchestration patterns, runtime control harnesses, guardrails and tool gateways.
- Attack paths, evaluation methods, assurance practices, operational telemetry and relevant standards.

> **CHANGE CONTROL**
> Updates should preserve session ownership, cumulative evidence hand-offs and management/assurance boundaries. New content replaces or refines existing material rather than creating duplication.

11 | PILOT EXPERIENCE AND LEARNER FEEDBACK

What the October 2025 practitioner track showed

The wider Trustworthy AI workshop ran from 1 to 3 October 2025. Day 1 was an executive briefing; Days 2-3 formed the practitioner track for project managers and technical staff. That track provided pilot experience and learner feedback that informed the subsequent refinement of Specialist Module 4.

Feedback themes

1. **Engagement.** Learners showed high levels of attention and active involvement across lectures and workshops.
2. **Relevance.** Content was viewed as well structured, holistic and directly applicable to participants' roles.
3. **Interactive delivery.** Lectures, case studies and hands-on activities supported deeper understanding and practical skill-building.
4. **Practical value.** Learners appreciated the clarity, relevance and usefulness of the materials for Philippine AI governance and digital transformation.
5. **Multi-stakeholder value.** Participation from government bodies and representatives of UNDP, UNICEF, ILO and UN Women broadened the institutional perspective.

Implication

The feedback supported a case-based, interactive delivery model while strengthening the implementation pathway for technical and project officers.

How feedback shaped the revised module

- Optimised the flow from governance requirements to architecture, controls, verification, operations, dependencies and decision evidence.
- Reduced duplication by assigning each session a distinct decision and artefact.
- Expanded practical coverage of GenAI, RAG, agents, secure operations, incident recovery and the AI supply chain.
- Made outputs cumulative so learners build one working file rather than isolated exercises.

- Strengthened evidence discipline and the distinction between technical readiness, management decision and independent assurance.

Project AMANAH

> **FICTIONAL CASE**
> Project AMANAH is a fictional, cross-functional AI implementation case used to connect all nine sessions and evidence outputs. It does not describe a real organisation, system, incident or control environment.

Primary source anchors

- OECD AI system lifecycle materials;
- ISO/IEC 42001:2023;
- OWASP AI Security Requirements and GenAI/agent resources;
- AI engineering and AI security resources;
- Project AMANAH case materials
- final Manila workshop report.

This briefing note is a curriculum and implementation summary. It is not legal advice, regulatory approval, management risk acceptance or independent assurance.