BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

NATIONAL AI CAPACITY PROGRAMME

A Tiered National Capability Programme — Safety, Trust, and Resilience

Programme Overview

Artificial intelligence is being adopted across Malaysian organisations at remarkable speed — but the capacity to use it safely, to preserve trust in a world of synthetic media, and to remain resilient against AI-enabled threats is not keeping pace. A single awareness session cannot close that gap. Building genuine capability requires a structured, tiered curriculum that takes people from understanding, to hands-on skill, to governance leadership.

Building AI Safety Capacity & Digital Trust Resilience is that curriculum. It is a three-tier national capability programme built on three enduring pillars — AI Safety, Digital Trust, and Resilience — that deepen progressively across each tier. Everyone gains awareness; technical staff gain practitioner skills; leaders gain the ability to govern and lead. Together they build a workforce, and ultimately a nation, that can adopt AI with confidence.

The programme is designed to support the capability-building mandate of the National AI Office (NAIO) and the cybersecurity mission of CyberSecurity Malaysia, and is flexible enough to be deployed as a national initiative, an agency-wide public-sector programme, or an enterprise and critical-infrastructure capability build.

The Three Pillars

Three connected pillars run through every tier of the programme, each explored more deeply as participants progress.

- **AI Safety —** Ensuring AI systems behave reliably and do not cause harm — managing bias, hallucination, and unsafe or unreliable outputs, and keeping meaningful human oversight and control.
- **Digital Trust —** Preserving trust in identity, content, and transactions in an era of deepfakes and synthetic media — through verification, provenance, and data integrity.
- **Resilience —** The operational capacity to secure AI systems and to withstand, detect, respond to, and recover from AI-related threats and failures.

BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

Programme Architecture

The programme is structured as three stackable tiers. The three pillars run through all of them, deepening at each level, and completion builds a clear certification ladder from awareness to governance leadership.

Tier	Duration	Audience	Focus
Tier 1: Awareness & Literacy	1 day	All staff, general officers, and leaders	Understand safety, trust, and resilience and apply safe everyday practices
Tier 2: Practitioner	3 days	IT, security, risk, data, and development professionals	Test, secure, and operate AI systems safely and resiliently
Tier 3: Governance & Leadership	2 days	Decision-makers, governance officers, policy and risk leads	Govern and lead organisational AI safety, trust, and resilience

Certification ladder: Certified AI Aware (Tier 1) → Certified AI Safety & Security Practitioner (Tier 2) → Certified AI Governance & Leadership (Tier 3). Approximately 42 hours of structured capability development across the full programme.

Deployment & Positioning

The programme is designed for flexible deployment across three buyer contexts, using the same core architecture.

- **National Programme —** A vehicle for a national AI-capability and AI-literacy target, delivered across the public service.
- **Public Sector —** An agency-wide capability build taking officers, technical teams, and leadership through the tiers appropriate to their roles.
- **Enterprise & Critical Infrastructure —** An organisation-wide programme for enterprises and critical-infrastructure operators, from all-staff awareness to board-level governance.

BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

Tier 1 — Awareness & Literacy

The universal foundation — everyone understands AI safety, trust, and resilience, and applies safe practices every day. (1 day, 7 contact hours)

Tier 1 is designed for the entire workforce, from frontline officers to senior leaders. It requires no prior AI experience and builds a shared language and baseline of safe, responsible AI practice across the organisation.

Agenda

Time	Module & Content
9:00 – 9:30	**Welcome & National Context** • Why AI safety, trust, and resilience matter now — for you, your organisation, and the nation • Malaysia's national direction: NAIO, the AIGE guidelines,etc.
9:30 – 10:45	**Pillar 1 — AI Safety: When AI Gets It Wrong** • How AI fails: bias, hallucination, and unreliable or unsafe outputs, with real examples • Why human oversight and judgment remain essential • Recognising when to trust — and when to question — an AI output
10:45 – 11:00	*Morning Break*
11:00 – 12:30	**Pillar 2 — Digital Trust: Deepfakes, Identity & What to Believe** • Deepfakes, voice cloning, and synthetic identity explained • Real threats in Malaysia: deepfake investment scams and executive impersonation • Verifying what you see and hear: provenance and simple verification habits
12:30 – 1:30	*Lunch*
1:30 – 3:00	**Pillar 3 — Resilience: Securing AI & Staying Safe** • AI-powered attacks and the risk of unmanaged 'shadow AI' at work • Protecting data and privacy when using AI tools • Spotting, reporting, and responding to AI-related incidents
3:00 – 3:15	*Afternoon Break*

BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

3:15 – 4:15	**Applied — Safe & Responsible AI in Your Daily Work** • Practical safe-use habits aligned to responsible-AI principles • Your organisation's AI acceptable-use basics • Hands-on scenario: spotting the safety, trust, and resilience issues
4:15 – 5:00	**Closing — Your Safe-AI Commitments & National Alignment** • Personal and team commitments to safe, trusted AI use • How everyday practice supports the national AI agenda • Certification and the pathway to the Practitioner tier

Tier Objective	Learning Outcomes	Certification
Give every member of the organisation a working understanding of AI safety, digital trust, and resilience, and the habits to use AI safely and responsibly.	Recognise AI safety risks and when to question outputs Spot deepfake and identity threats Understand AI security and responsible use Apply safe practices in daily work	Certified AI Aware Certificate of completion and a personal safe-AI commitment

BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

Tier 2 — Practitioner

Hands-on capability to test, secure, and operate AI systems safely and resiliently. (3 days, ~21 contact hours)

Tier 2 is designed for technical professionals — IT, security, risk, data, and development staff. It is intensively practical, with a hands-on lab closing each day. Basic technical literacy is assumed; Tier 1 completion is recommended.

Day 1 — AI Safety Engineering

Time	Module & Content
9:00 – 10:45	**AI Failure Modes & Output Evaluation** • Understanding how AI systems fail: bias, hallucination, drift, and robustness gaps • Methods for evaluating and benchmarking AI outputs systematically • Establishing acceptance criteria and human-in-the-loop checkpoints
10:45 – 11:00	*Morning Break*
11:00 – 1:00	**AI Red-Teaming Fundamentals** • Adversarial prompting and jailbreak testing techniques • Probing models for unsafe, biased, or leaking behaviour • Documenting findings and severity in a structured way
1:00 – 2:00	*Lunch*
2:00 – 3:30	**Guardrails & Safe-by-Design Controls** • Designing input/output guardrails and content filters • Human oversight and escalation patterns for higher-risk use • Safety testing as part of the development lifecycle
3:30 – 3:45	*Afternoon Break*
3:45 – 5:00	**Hands-On Lab: Red-Team & Evaluate an AI System** • Teams red-team a provided AI application and evaluate its safety • Produce a findings report with recommended guardrails

BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

Day 2 — Digital Trust & AI Security

Time	Module & Content
9:00 – 10:45	**Securing the AI Stack** • Protecting data, models, and pipelines across the AI lifecycle • Access control, secrets, and the AI supply chain • Threats unique to AI systems versus traditional software
10:45 – 11:00	*Morning Break*
11:00 – 1:00	**OWASP Top 10 for LLMs in Practice** • Prompt injection, insecure output handling, and training-data poisoning • Model denial-of-service, supply-chain, and sensitive-information disclosure • Hands-on: exploit and then defend a vulnerable AI application
1:00 – 2:00	*Lunch*
2:00 – 3:30	**Digital Trust: Deepfake & Synthetic-Media Defence** • Deepfake and synthetic-media detection techniques and tools • Content provenance, watermarking, and authentication standards • Defending identity verification and e-KYC processes
3:30 – 3:45	*Afternoon Break*
3:45 – 5:00	**Hands-On Lab: Attack & Defend** • Teams exploit AI vulnerabilities, then implement and test defences • Debrief against the OWASP LLM reference

BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

Day 3 — Resilience Engineering & Operations

Time	Module & Content
9:00 – 10:45	**Threat Modelling for AI Systems** • Mapping the attack surface of an AI-enabled service • Prioritising risks and designing mitigations • Integrating AI risk into existing security processes
10:45 – 11:00	*Morning Break*
11:00 – 1:00	**Monitoring, Detection & Response** • Logging, monitoring, and anomaly detection for AI systems • Detecting AI-specific incidents: abuse, drift, data leakage, and compromise • Building an AI incident-response process
1:00 – 2:00	*Lunch*
2:00 – 3:30	**Secure Deployment & AI Security Operations** • Secure deployment patterns and continuous monitoring • Recovery and continuity for AI-related failures • Building an ongoing AI security operations capability
3:30 – 3:45	*Afternoon Break*
3:45 – 5:00	**Hands-On Lab: AI Incident-Response Simulation** • Teams run a simulated AI-security incident end to end • Produce a reusable AI incident-response playbook

Tier Objective	Learning Outcomes	Certification
Equip technical staff to test AI systems for safety, secure them against AI-specific threats, defend digital trust, and operate them resiliently.	Test AI systems for safety and robustness Conduct AI red-teaming Secure AI against OWASP LLM threats Detect deepfakes and defend identity Build AI incident detection and response	Certified AI Safety & Security Practitioner Assessed through the hands-on labs and a practitioner portfolio

BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

Tier 3 — Governance & Leadership

Equipping leaders to own and lead organisational AI safety, trust, and resilience. (2 days, ~14 contact hours)

Tier 3 is designed for decision-makers, AI governance officers, and policy, risk, and compliance leads. It translates the safety and security disciplines of the lower tiers into organisational governance, accountability, and leadership.

Day 1 — Governance Foundations & National Alignment

Time	Module & Content
9:00 – 10:45	**The AI Governance Landscape** • Malaysia's framework: the AIGE guidelines, NAIO direction, and PDPA developments • International references: NIST AI Risk Management Framework, ISO/IEC 42001, and EU AI Act awareness • How the pieces fit into a coherent governance approach
10:45 – 11:00	*Morning Break*
11:00 – 1:00	**Risk-Based AI Governance** • Classifying AI use cases by risk and impact • Matching controls and oversight to the level of risk • Balancing innovation with safety, trust, and accountability
1:00 – 2:00	*Lunch*
2:00 – 3:30	**Roles, Accountability & Operating Model** • The AI governance operating model and who owns what • Accountability, escalation, and decision rights • Building the governance function within the organisation
3:30 – 3:45	*Afternoon Break*
3:45 – 5:00	**Hands-On: Map Your AI Risk Landscape** • Participants classify their organisation's AI use against a risk-based framework • Identify priority governance gaps

BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

Day 2 — Building & Leading the AI Governance Posture

Time	Module & Content
9:00 – 10:45	**Policies & Assurance** • Core AI policies: acceptable use, safety, incident, and procurement • AI assurance: risk assessment, impact assessment, and audit • Embedding safety and trust requirements into policy
10:45 – 11:00	*Morning Break*
11:00 – 1:00	**Managing Third-Party & Vendor AI Risk** • Assessing and contracting for AI in the supply chain • Due diligence on vendor safety, security, and data practices • Ongoing oversight of third-party AI
1:00 – 2:00	*Lunch*
2:00 – 3:30	**Leadership, Reporting & Culture** • Board and stakeholder reporting on AI risk and posture • Building a culture of responsible, safe AI adoption • Leading change and sustaining the capability
3:30 – 3:45	*Afternoon Break*
3:45 – 5:00	**Hands-On: Draft Your AI Governance Roadmap** • Participants build a prioritised AI governance roadmap for their organisation • Present for peer and facilitator feedback

Tier Objective	Learning Outcomes	Certification
Enable leaders and governance officers to build, run, and lead an organisational AI governance posture aligned to Malaysia's national framework.	Build an AI governance framework Apply AIGE and international standards Establish risk assessment and assurance Manage third-party AI risk Lead accountable AI adoption	Certified AI Governance & Leadership Assessed through the governance roadmap deliverable

BUILDING AI SAFETY CAPACITY & DIGITAL TRUST RESILLENCE

Delivery, Frameworks & Certification

The programme is modular and stackable: tiers may be delivered as a full pathway or taken independently by the appropriate audience. Delivery is instructor-led with hands-on labs at the Practitioner tier, available in classroom or hybrid formats, and anchored throughout by real Malaysian examples and current threats. The programme is grounded in national frameworks and recognised international standards.

Framework / Reference	Role in the Programme
National Guidelines on AI Governance & Ethics (AIGE)	Core responsible-AI principles applied across all tiers
NAIO & AI Technology Action Plan 2026–2030	National direction and capability-building context
CyberSecurity Malaysia guidance	AI security best practice and national cyber readiness
Personal Data Protection Act (PDPA)	Data protection, profiling, and automated decision-making
NIST AI Risk Management Framework & ISO/IEC 42001	International references for AI risk and management systems
OWASP Top 10 for LLM Applications	Practical reference for AI application security (Practitioner tier)

Certification & National Impact

Completion builds a clear, stackable certification ladder — Certified AI Aware, Certified AI Safety & Security Practitioner, and Certified AI Governance & Leadership — totalling approximately 42 hours of structured capability development. Delivered at scale, the programme provides a measurable, tiered, and certifiable vehicle for national AI-capability and AI-literacy targets, and a ready pathway for organisations to build safe, trusted, and resilient AI adoption from the ground up.