
Malaysia AI Safety Institute (MY-AISafe)

AI safety enables Malaysia to move faster with AI while keeping people protected.

AI safety

Evaluation

Standards Building

Technical Assurance

STRATEGIC MANDATE & ALIGNMENT

Contributing to **E12: AI Trust Function** under the **National AI Action Plan 2026–2030**, establishing national capabilities for testing, risk response, and trustworthy AI adoption.

Malaysia AI Safety Institute (MY-AISafe)

National AI Ecosystem • Kementerian Digital • National AI Action Plan 2026–2030

The Malaysia AI Safety Institute (MY-AISafe) brings together capabilities for testing and evaluating AI systems, monitoring AI incidents, and developing the standards and technical guidance needed for safer AI deployment.

Announced on 28 July 2026 as part of the launch of AI Malaysia Berhad, MY-AISafe forms part of Malaysia’s national AI ecosystem under Kementerian Digital. Its work contributes to E12: AI Trust Function in the National AI Action Plan 2026–2030, supporting greater trust, assurance and accountability as AI adoption grows.

FROM RESPONSIBLE AI PRINCIPLES TO REAL-WORLD EVIDENCE

Malaysia set out seven principles for responsible AI in the National Guidelines on AI Governance and Ethics (AIGE), published in 2024. The National AI Action Plan 2026–2030 builds on these principles to support responsible AI adoption.

MY-AISafe puts these principles into practice through testing, evaluation and technical evidence. The institute develops tests, benchmarks and evaluation reports to assess AI systems against key areas of responsible AI.

Fairness Bias tests across groups of people	Reliability, Safety and Control Robustness testing and red teaming
Privacy and Security Data protection and security tests	Inclusiveness Testing with the full range of intended users
Transparency System documentation and disclosure	Accountability A named owner and incident reports
Pursuit of Human Benefit and Happiness Impact assessment on the people affected	

Principles as named in the National Guidelines on AI Governance and Ethics (AIGE), 2024.

MY-AISAFE DELIVERS E12 AI TRUST FUNCTION, ONE OF THE PLAN’S 28 INITIATIVES

The Action Plan carries 14 impact engines that put AI to work in the economy and society, and 14 foundational enablers that make that work possible. E12 is the enabler that gives Malaysia its own capability to evaluate, monitor and respond to AI risks.

E12 AI Trust Function

“Create an AI evaluation, assessment & response capability to ensure Malaysia can adequately pre-empt and manage AI risks”

Target: “Establish a rapid and reliable national capability to detect, assess and contain critical AI-related incidents, with continuous improvement in response readiness, coordination and effectiveness across government and key sectors.”

The national AI Trust Function brings together:

Standards and Policy AI safety and security standards, risk classifications and guidance on safe use	Evaluation and Certification Testing of AI models and applications against those standards
Incident Monitoring and Forensics Tracking AI-related incidents and establishing why they happened	International Collaboration and Capacity Building Working with partners abroad and building specialist skills in Malaysia

THREE AREAS OF WORK, FROM THE TEST LAB TO THE RULEBOOK

The institute draws on expertise from government, industry, academia and research institutions to support the safe adoption of AI in Malaysia.

AI RDCI (Research, Development, Commercialisation and Innovation)

The institute examines how models and AI applications behave, including how they fail, before people come to depend on them. This covers safety evaluations, testing and AI red teaming.

Strategic collaboration

Partnerships in Malaysia and with international partners build the skills to test AI systems in Malaysia and keep the institute's methods in step with work elsewhere. This covers talent development, expertise sharing and technology transfer.

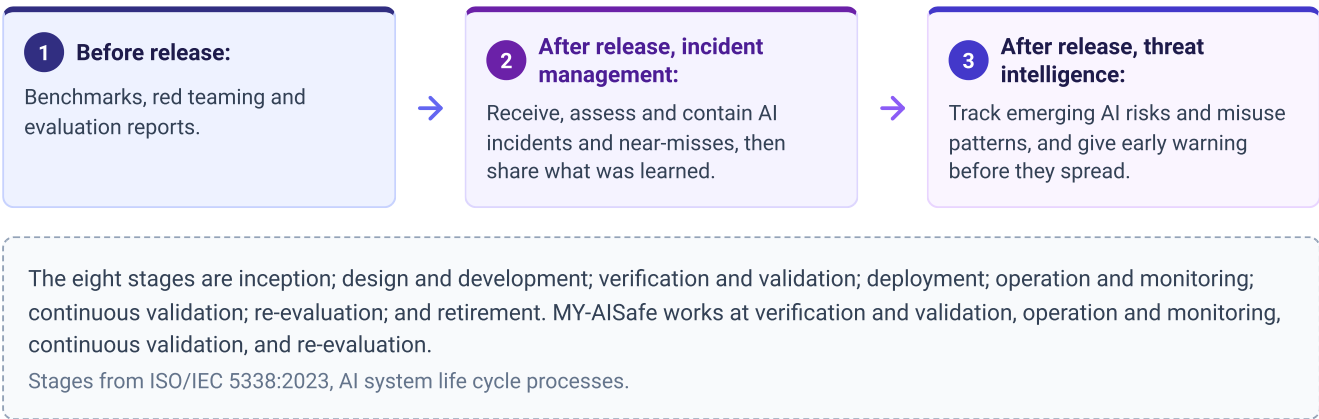
AI Assurance, Standards & Evaluation

The institute develops the technical approaches needed to assess whether AI systems meet safety and reliability expectations. This includes evaluation methodologies, benchmarks, testing protocols, standards and assurance mechanisms.

Areas of work as announced on 28 July 2026.

A SYSTEM IS TESTED BEFORE RELEASE, AND ITS INCIDENTS AND THREATS ARE MANAGED FOR AS LONG AS IT RUNS

Many AI risks only show up after launch, when systems meet real users, new data and people who try to misuse them. The international standard for the AI life cycle sets out eight stages. MY-AISafe tests systems before release, then manages incidents and tracks emerging threats once they are in use.



WHO IT SERVES

Turning technical evidence into action

MY-AISafe's work is designed to be useful to the people building, deploying, evaluating and overseeing AI.

Regulators and agencies: Technical evidence and advice for decisions in their own sectors.	Developers and deployers: Clear test methods and benchmarks to build towards, and a place to raise questions early.
Researchers and universities: Joint work on evaluation methods and benchmarks.	International partners: Shared methods, joint research and capacity building with AI safety institutes in other countries.

FROM ETHICAL PRINCIPLES IN 2024 TO A NATIONAL INSTITUTE IN 2026

●	Sep 2024 National Guidelines on AI Governance and Ethics (AIGE) published
●	Dec 2024 National AI Office launched
●	Jul 2026 Public consultation on the AI Governance Bill, 10 to 31 July
●	28 Jul 2026 AI Malaysia launched with the National AI Action Plan 2026-2030 and the Malaysia AI Safety Institute
●	2027 AI Governance Bill targeted for tabling in Parliament