# Singapore's Approach to AI Governance

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

# Our Approach to AI Governance

**AI for Public Good**

- Building a trusted AI ecosystem is not just about safeguards but also ensuring that good societal and economic benefits can be accrued to all
- While it is necessary to protect broader public interest and build consumer trust, we should also allow space for innovation

**Practical & Risk-based**

- AI is a nascent and fast developing space, so it is premature to impose hard regulations. For now, to focus on supporting industry voluntary adoption
- Policies should be proportional to risks, and move in tandem with AI developments and beyond concepts

**Multi-stakeholder & Global**

- Always in partnership – with industry, academia, other governments
- Collaboration helps to crowd in different expertise, to collectively experiment, build and improve
- Goal is to contribute to global development of norms and knowledge

# Overview of AI Governance Initiatives

- Model AI Governance Framework
- Self-Assessment Guide
- Use Cases


2019 - 2021

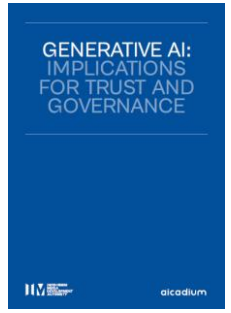
- AI Verify Testing Framework & Toolkit
- Open-Source Foundation

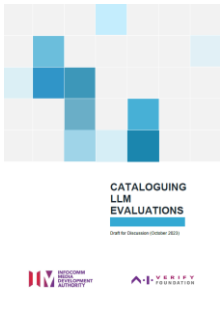
AI VERIFY

AI VERIFY FOUNDATION

2022 – 2023

- Discussion Paper for Gen AI governance
- Gen AI Evals Sandbox & LLM Eval Catalogue




June 2023

Oct 2023

Detailed guidelines for organisations to implement trustworthy AI

Accountability and risk-based approach, takes ethical principles into corporate governance, risk management, operations and stakeholder engagement

➡

Practical AI testing tool for organisations to demonstrate trustworthy AI

➡

Generative AI

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

# AI Verify

## Why did we build it?

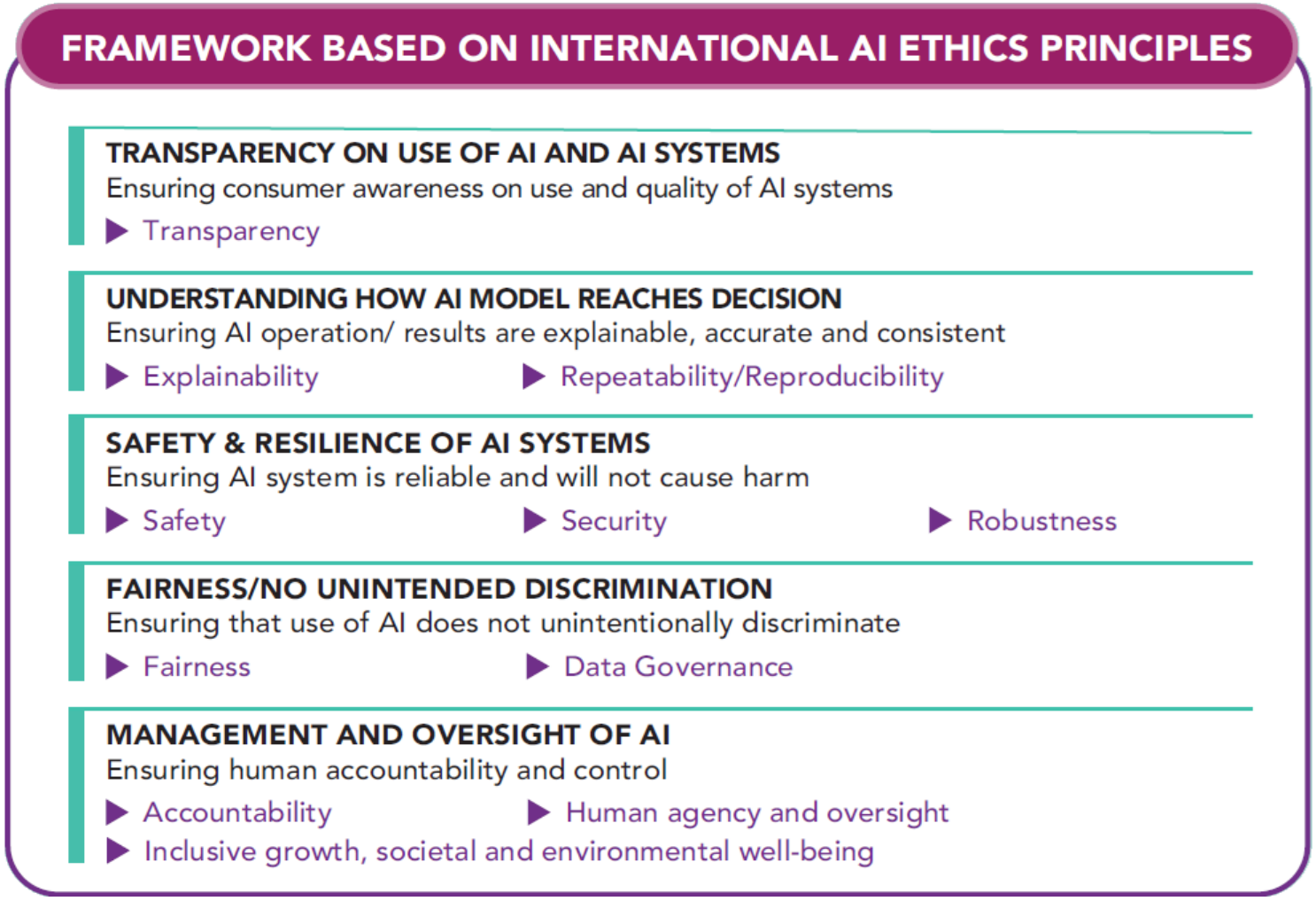
- Need for a one-stop tool for organisations to easily **demonstrate that their AI system are trustworthy** or compliant to defined requirements/standards
- So that they can be **transparent** about their AI implementation with…
  - Stakeholders like customers, partners, shareholders
  - Regulators
  - Consumers

## What is it?

- A **testing framework** that defines process checks based on key global ethical AI principles and requirements
- A **software toolkit** to operationalise the process checks, and to technically test a subset of the principles
- Does not define “pass/fail”, nor certify that AI systems have met standards - but to demonstrate organisations’ own claims

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

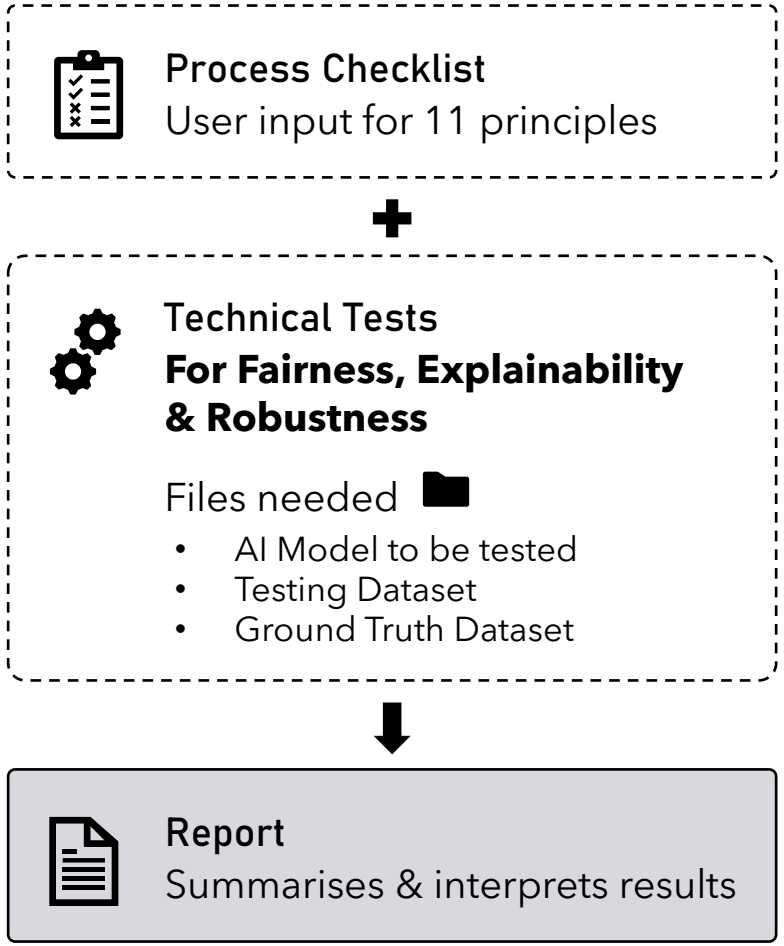
# AI Verify Testing Framework




- **Consists of >80 process checks**
  - Spanning 11 global AI principles
  - Addresses 5 major areas of concern for AI systems
- **Mapped to NIST AI Risk Management Framework** (Crosswalk)
- Can be extended to include other standards

# AI Verify Software Toolkit (Discriminative AI)

- **Guided workflow** to operationalise the Framework
  - Checklist + technical tests
- **Blackbox testing**: similar to cyber pen test, geared towards post-development testing
- **Reports**: customizable to support different needs e.g. different stakeholders, frameworks and use-cases
- **Currently supports common supervised learning classification & regression models** for tabular & image data, and does not test for Gen AI
  - Extensible - include new models and testing algorithms as testing sciences improve



# Customizable & Extensible testing tool which works for different use cases

## Open Source on GitHub (Apache 2.0 license)

- Facilitate rapid innovation and extension of functionalities
- Can be used in 3 ways:
  - Organisations' self test
  - Third party testing for assurance or certification
  - Embedded in testing or audit products or services

## Extensible Architecture

- 3rd parties can build plugins to extend functionalities

**Types of Plugins:**

**Display Components**
Custom graph modules to display test results

**Workflow Components**
Sector specific checklists or workflows

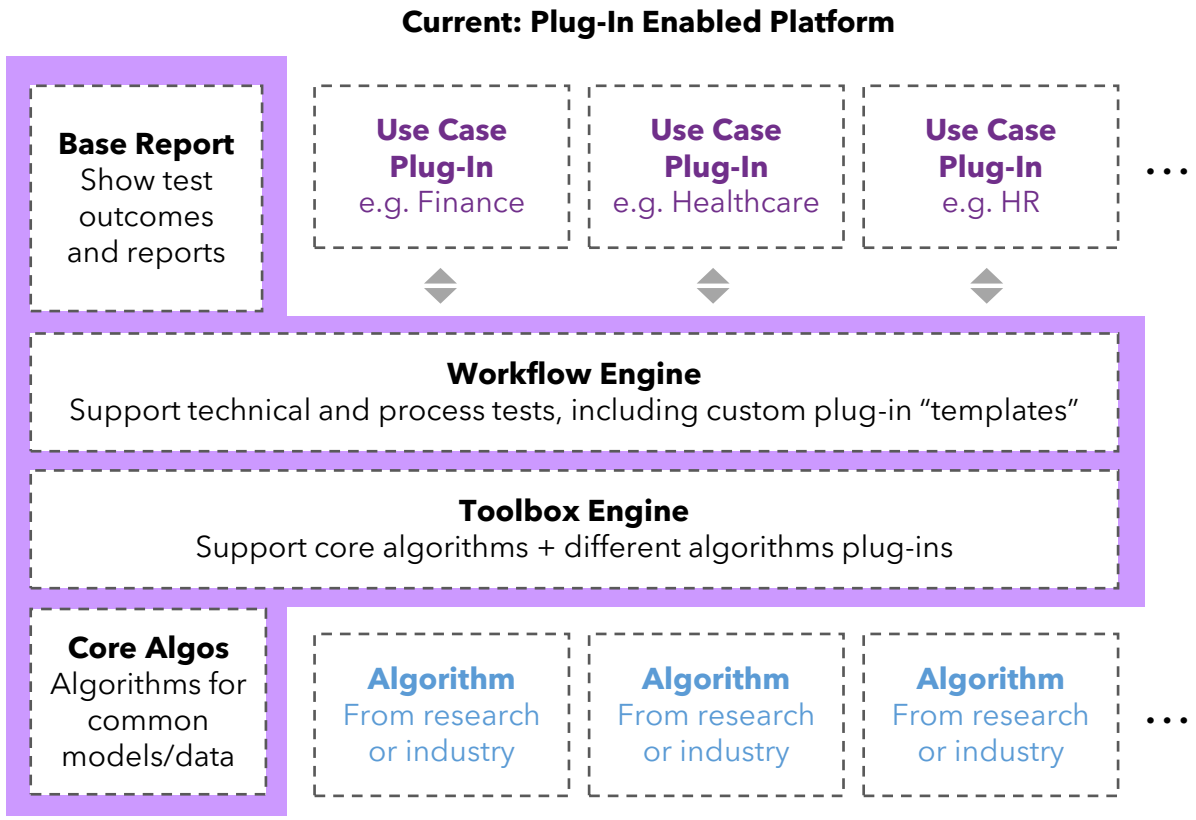
**Testing Algorithms**
Introduce new test approaches

**Report Templates**
Sectoral reports designed for specific use cases

**AI Verify Dev Tools**
Available on GitHub to facilitate plug-in development

# Standards and AI Verify

- **AI Verify is extensible and can be updated to incorporate international standards (e.g. ISO)**




- **AI Verify has >80 process checks for 11 Governance principles, which can serve as input to relevant ISO SC42 work items**

**Process Checks for Fairness**

- Assess within-group fairness
- Processes to test for potential biases during the entire lifecycle of the AI system, so that practitioners can act to mitigate biases based on feedback
- Establish a strategy for the selection of fairness metrics that are aligned with the desired outcomes of the AI system's intended application
- Define sensitive features for the organisation that are consistent with the legislation and corporate values
- Establish a process for identifying and selecting sub-populations between which the AI system should produce fair outcomes
- …

# Foundation set up to crowd-in global experts to grow sciences and capabilities for AI testing

Made available under **Apache 2.0** license and hosted with the **AI Verify Foundation:**

Not for profit entity to **promote the use and development of AI testing frameworks, standards, and best practices** by

- fostering a community
- creating a neutral platform for collaboration
- educating and advocating for AI testing

**Open source enables the community to contribute and audit the code**

### New Frameworks
Companies can add new frameworks or sectoral reports designed for specific use cases

### New Components
Developers can add custom interpretation and graphs to display results in the way they prefer

### New Algorithms
Researchers can add new algorithms which support new models, data types, or ML frameworks

**Meetups and community workshops**

# Encouraging Traction from Industry

Organic growth to more than 100 members in 5 months

**Premium Members**

aicadium, Google, IBM, IMDA (Infocomm Media Development Authority), Microsoft, Red Hat, salesforce

**General Members**

accenture, Adobe, ADVANCE.AI, AiDA, AIDX, Allytics, amdocs, 蚂蚁集团 ANT GROUP, AsiaVerify, Asia Internet Coalition, asurion, aws, bbp, betterdata, Bowen Asia Group, Bright Raven, BSA, CSDA, Citadel AI, citydata, CrimsonLogic, credo ai, data iku, DataRobot, DBS, DELL Technologies, Deloitte., ELGO, EngageRocket, Envision digital, enzai, evco, FAIRLY, fairnow, FairPriceGroup, finbots.ai, Handshakes, Hewlett Packard Enterprise, HITACHI Inspire the Next, HUAWEI, impress, innovBtif, INTERACT AI, INVIGILO, Lazada, LIODS., mediacorp, MEIYUME, Meta, MAIEI, ncs, NERVOTEC, OCBC Bank, OPPi, QUALGRO, Regtank, sas, SCANTIST, 商汤 sensetime, SINGAPORE AIRLINES, SigmaRed, squirro, ST Engineering, standard chartered, storewise, STRIDES, TAU EXPRESS, TeamSolve, Humanity, Temus, IC AG, TISANE LABS, TRIBE, truera, TrustedAI, UCARE.AI, UOB, FLOW TECH, Vidreous, virtusa, whale, workday., XOPA AI, ZOIOZ, zoom, 2021.AI



As of 1 Nov 2023

# Gen AI Testing & Evaluation – Expanding AI Verify

## Impetus

- **Need to grow capacity** – evaluation necessary to validate if models are safe to use, but capabilities mostly reside with model developers and research labs
- **Need more evals** to address increasing context-specific use, e.g. culture, domains
- **Need standards**
  - For consistent eval across models and applications
  - To ensure that evals are reliable

## Focus Areas

- **LLM Evaluation Catalogue**
  - First stop reference for LLM evals, first step towards defining basic standards, e.g. taxonomy, baseline for safety
  - Compile and organise evals that are currently available – will continue to grow
- **Evaluation Sandbox**
  - Build benchmarks and codify methods – focusing on context and use cases
  - Different parts of the ecosystem to partner and build practical knowledge
- **Improving evaluation standards & science**
  - **Methods to build standard benchmarks and assess existing ones** so test results are reliable, and testing can be done more consistently and easily across different models
    - Quality of testing datasets: sufficiency and validity of questions, etc.
    - "Testing pedagogy": pre/post prompting set up, ordering of multiple choice questions, regurgitation/memorisation, etc.
    - Technical interface: tools to provide interfaces to benchmarks and to models
  - **Methods for red-teaming**
  - **Methods for testing advanced model capabilities**

# LLM Evals Catalogue

**Common reference and starting point for LLM Evals**

## TAXONOMY + CATALOGUE

- Surveyed the landscape of LLM evals and formulated a systematic framework to organize existing tests
  a) General Capabilities
  b) Domain Specific Capabilities
  c) Safety and Trustworthiness
  d) Extreme Risks
  e) Undesirable Use Cases.
- Compiled a catalogue of evals (benchmarks and other methods) based on these five categories

## OBSERVATIONS + FUTURE WORK

- Provided a gap analysis of the LLM evaluation landscape and recommendations for future work
  a) Context-specific evals lacking, e.g. domains, cultures
  b) Frontier model evals – very few such tests
  c) Standardized assessment frameworks to improve eval reliability

## BASELINE SAFETY EVALUATIONS

- Based on AI Verify principles, recommend a baseline evaluation that LLMs should minimally be tested on before deployment
  a) Bias
  b) Factuality
  c) Toxicity Generation
  d) Robustness
  e) Data Governance

Needs continuous updating to be relevant – invite community to feedback and add on as new benchmarks, methods are developed

# Gen AI Eval Sandbox

## Collaboratively build practical evaluation capabilities and grow body of knowledge for Gen AI testing

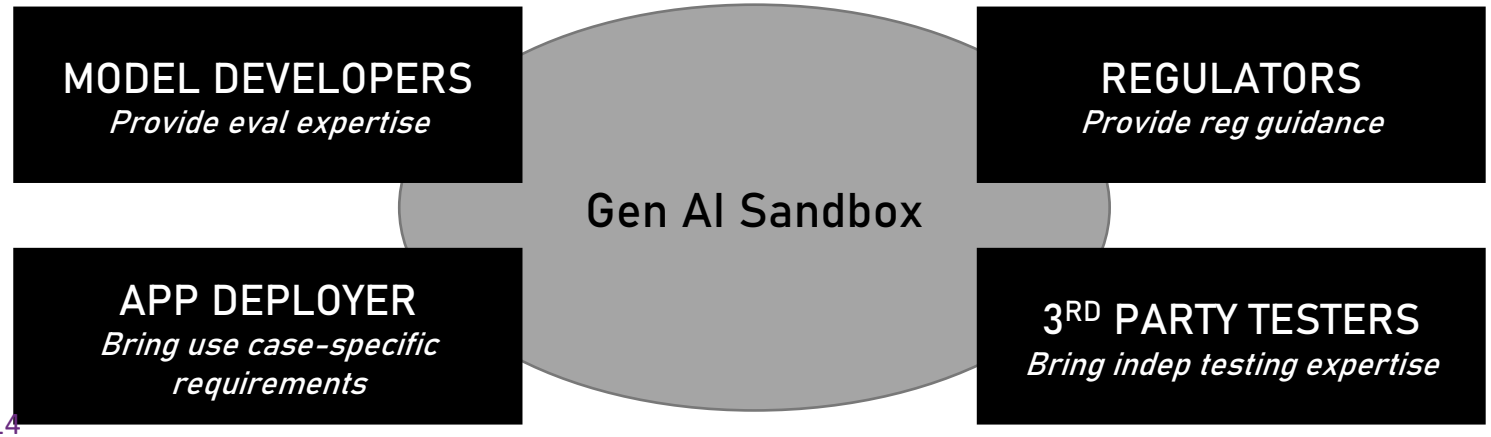
### OUTPUT

| | |
|---|---|
| 1. EVAL METHODOLOGIES | e.g. red teaming standards, methods for testing images |
| 2. EVAL BENCHMARKS | e.g. domain-specific, use case centric, cultural representation |
| 3. CASE STUDIES | e.g. end-to-end testing, third party testing |
| 4. AREAS FOR FURTHER DEVELOPMENT | e.g. gaps that require more investment to plug |

# Gen AI Eval Sandbox

## Ways to participate

- **Identify your needs for evaluation**, e.g. use cases, involving 3rd party validation
- **Offer your expertise** – as model developer, app deployer, etc
- **Co-opt your partners** – we can also help match-make, e.g. companies in the same domain co-building benchmarks
- **Jointly experiment**, and develop output




### Example: IMDA-Anthropic Cultural Red Teaming

- Aim: Develop cultural red-team methodology so that models can be safer for different cultures
- Anthropic: Provide access to model, red-teaming platform and experience
- IMDA: Provide cultural considerations and conduct red-teaming

Google Microsoft DataRobot aws XOPA AI IBM Deloitte.

… and more

# Welcome to collaborate

## AI Verify & Foundation

1. **Join the Foundation as a member & be part of a global testing community**
   - Collaboratively grow testing capabilities
   - Co-develop evaluation standards
2. **Demonstrate responsible AI implementation**
   - Profile your use-cases of responsible AI implementations on Foundation website
3. **Avenue for developers to actively participate in the technical community and contribute to the AI Verify open-source project**
   - E.g. Use-case specific plugins (testing algos, report templates, etc) and enhancing the core code base with new functions

## Gen AI

1. **Identify suitable LLM apps to participate in Gen AI Evaluation Sandbox**
   - Co-create evaluation benchmarks and methodologies through experimental projects
2. **Contribute feedback to LLM Evaluation Catalogue**
   - Baseline set of safety evaluations can serve as starting point for App evaluation & deployment

# Links to Resources

1. LLM Evaluation Catalogue (PDF), (Github)
2. Generative AI Evaluation Sandbox
3. Generative AI: Implications for Trust and Governance
4. AI Verify Foundation
5. AI Verify (Github)
6. AI Verify (SG) x NIST AI Risk Management Framework (US) Crosswalk
7. Model AI Governance Framework
8. Compendium of Use Cases
9. Implementation & Self Assessment Guide for Organisations

Please contact info@aiverify.sg if you have questions.

# Thank you

INFOCOMM MEDIA DEVELOPMENT AUTHORITY