INFOCOMM
MEDIA
DEVELOPMENT
AUTHORITY

# INVITATION TO PILOT

# A·I·VERIFY

# AI GOVERNANCE TESTING FRAMEWORK & TOOLKIT

25 May 2022

Supported by:

PERSONAL DATA
PROTECTION COMMISSION
SINGAPORE

In support of:

SG:DIGITAL

## PART I: WHY DO WE NEED TESTING FOR AI GOVERNANCE?

1. With more products and services employing AI to provide greater personalisation or to make autonomous predictions, the public needs to be assured that AI systems are fair, explainable, and safe, and companies that deploy them are transparent and accountable. The goal is to foster public trust in AI technologies while supporting the increasing use of AI.

2. Voluntary AI governance frameworks and guidelines have been published to help system owners and developers implement trustworthy AI products and services.[1] Singapore has been at the forefront of international discourse on AI ethics and governance, and guiding industry on responsible development and deployment of AI since 2018.[2] The Infocomm Media Development Authority Singapore (IMDA) and Personal Data Protection Commission (PDPC) published the Model AI Governance Framework (now in its 2nd edition), a companion Implementation and Self-Assessment Guide for Organisations, and two volumes of Compendium of Use Cases that provide practical and implementable measures for industry's voluntary adoption.

3. Voluntary self-assessment is a start. With greater maturity and more pervasive adoption of AI, the industry needs to **demonstrate to their stakeholders** their implementation of responsible AI in an **objective and verifiable** way. IMDA and PDPC have taken the first step to develop an AI Governance Testing Framework and Toolkit to enable industry to demonstrate their deployment of responsible AI. This is currently available as a **Minimum Viable Product (MVP)** for system developers and owners who want to be more transparent about the performance of their AI systems through a combination of technical tests and process checks.

4. With the MVP, Singapore hopes to achieve the following objectives:

   a. **Enable businesses to build trust with their stakeholders.** The MVP allows businesses to determine their own benchmarks and demonstrate the **claimed performance** of their AI systems to their stakeholders, thereby enhancing stakeholders' trust in the AI systems.

   b. **Facilitate interoperability of AI governance frameworks.** The MVP addresses common principles of trustworthy AI and can potentially help businesses bridge different AI governance frameworks and regulations.[3] IMDA is working with regulators and standards organisations to map the MVP

[1] Jurisdictions including Australia, Hong Kong, Japan, Korea, Singapore, UAE have published AI ethics principles and/or guidelines for industry's voluntary adoption.

[2] In June 2018, Singapore published a *Discussion Paper on Artificial Intelligence and Personal Data Protection and Personal Data*, which formed the basis for the subsequent publication of the Model AI Governance Framework.

[3] For example, in April 2021, European Commission proposed an Artificial Intelligence Act, establishing rules for the development and placement of trustworthy AI systems on the EU market, and use of AI. Another example is China's new Personal Information Protection Law (PIPL) that includes rules on "automated decision-making", which encompasses use of AI technologies.

to established AI frameworks. This helps businesses that offer AI-enabled products and services in multiple markets.

c. **Contribute to development of international standards on AI.** Singapore participates as a member in ISO/IEC JTC1/SC 42 on Artificial Intelligence. Through industry adoption of the MVP, Singapore aims to work with AI system owners/developers globally to collate industry practices and build benchmarks that can help develop international standards on AI governance.

5. As AI governance testing is still nascent, IMDA aims to create an **AI testing community** comprising:

   a. AI developers and system owners seeking to test their AI systems,
   b. technology providers developing AI governance implementation and testing solutions,
   c. advisory service providers specialising in testing and certification support, and
   d. researchers developing testing technologies.

   This community will enable the sharing of experiences and development of best practices, and foster collaboration to build benchmarks, thereby, catalysing the development of AI governance testing.

**Why participate in piloting the MVP?**

6. In developing this MVP, IMDA engaged a small group of industry partners to conduct early-stage testing to obtain their feedback so as to ensure that the testable criteria in the Testing Framework are implementable and the Testing Toolkit is able to produce insightful results. The IMDA is now **inviting participants from the broader industry to participate in the pilot phase of the MVP**. Participants will have the unique opportunity to:

   a. Have early and full access to an internationally-aligned AI Governance Testing Framework and Toolkit MVP and use it to **conduct self-testing on their AI systems/models**;

   b. Produce reports to **demonstrate transparency and build trust** with their stakeholders;

   c. Provide feedback to IMDA to help **shape the MVP** so that it can **reflect industry's needs** and benefit the industry; and

   d. **Join the AI testing community to network, share and collaborate** with other participating companies to **build industry benchmarks and contribute to international standards development**.

### Roadmap Beyond the MVP

7. Participating companies in the pilot phase will be using the MVP to conduct self-testing in their **own environment**. Feedback received from the pilot participants will be used to further enhance the MVP. IMDA targets to release an updated AI Governance Testing Framework and Toolkit Version at the end of the pilot.

8. As AI governance testing technologies are still emerging, there will be **research and development opportunities** for technology providers and researchers to enhance and build new testing tools. For instance, this could include novel algorithms for robustness testing or methods to test unsupervised AI models. IMDA will work with the AI testing community to identify these development opportunities and engage **industry and research institutions** to address these technological gaps.

9. IMDA will also work with the industry and the AI testing community to develop the ecosystem beyond self-testing. This could include, in the longer term, **testing and certification** by independent third parties such as audit firms, and testing and certification service providers.

## PART II: WHAT IS THE AI GOVERNANCE TESTING FRAMEWORK AND TOOLKIT?

10. In developing the AI Governance Testing Framework and Toolkit, IMDA aligned it with internationally accepted AI ethics principles, guidelines, and frameworks, such as those from the EU and OECD. Countries are generally coalescing around 11 key AI ethics principles, grouped into 5 pillars (See Figure 1). The 11 principles are transparency, explainability, repeatability/reproducibility, safety, security, robustness, fairness (*i.e.,* mitigation of unintended discrimination), data governance, accountability, human agency & oversight, and inclusive growth, societal & environmental well-being. For a start, **an initial set of 8 principles** were selected for the MVP based on the following practical considerations:

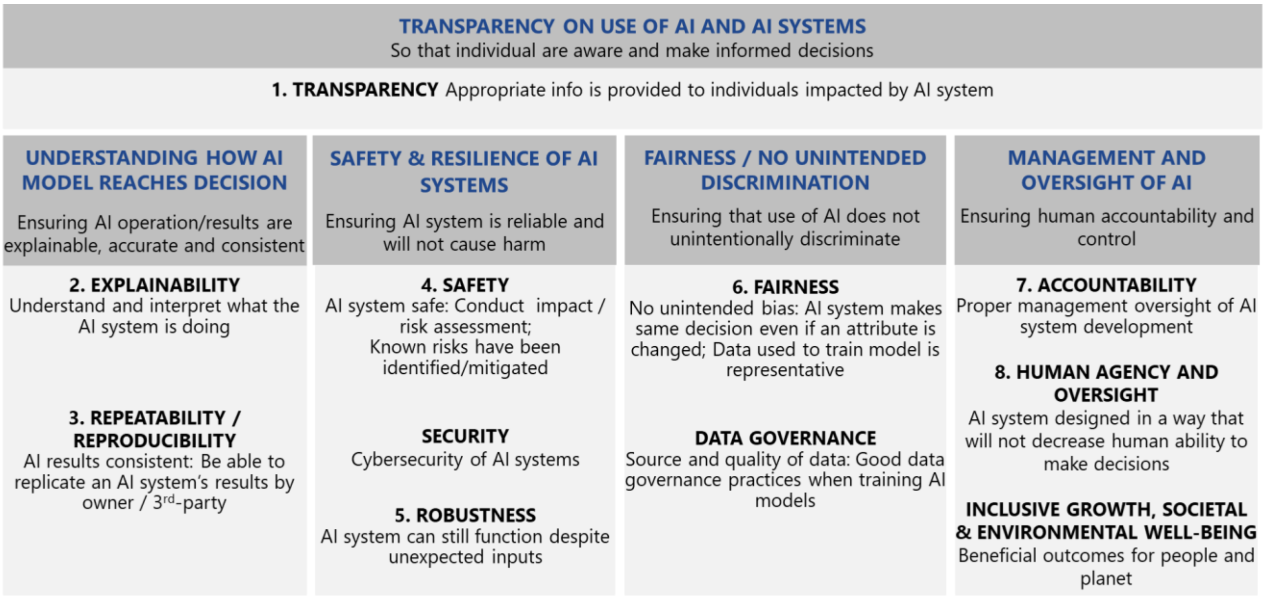
    a. At least one principle chosen from each of the 5 pillars for comprehensiveness;

    b. Availability of open-source tools or established methodologies that can be packaged and used to carry out testing against chosen principles; and

    c. Leverage existing testing and certification regimes and efforts, *i.e.,* cybersecurity and data governance (including data protection/privacy) and not reinventing the wheel[4].

[4] Example includes ISO/IEC 27001 Information Security Management Certification, ISO/IEC JTC1/SC27's efforts on AI security, Singapore's Data Protection Trustmark (DPTM) Commission and Singapore's Technical Committee on Artificial Intelligence's work on AI security.

*Figure 1 Initial set of AI ethics principles for MVP*



**TRANSPARENCY ON USE OF AI AND AI SYSTEMS**
So that individual are aware and make informed decisions

**1. TRANSPARENCY** Appropriate info is provided to individuals impacted by AI system

| **UNDERSTANDING HOW AI MODEL REACHES DECISION**<br>Ensuring AI operation/results are explainable, accurate and consistent | **SAFETY & RESILIENCE OF AI SYSTEMS**<br>Ensuring AI system is reliable and will not cause harm | **FAIRNESS / NO UNINTENDED DISCRIMINATION**<br>Ensuring that use of AI does not unintentionally discriminate | **MANAGEMENT AND OVERSIGHT OF AI**<br>Ensuring human accountability and control |
|---|---|---|---|
| **2. EXPLAINABILITY**<br>Understand and interpret what the AI system is doing | **4. SAFETY**<br>AI system safe: Conduct impact / risk assessment;<br>Known risks have been identified/mitigated | **6. FAIRNESS**<br>No unintended bias: AI system makes same decision even if an attribute is changed; Data used to train model is representative | **7. ACCOUNTABILITY**<br>Proper management oversight of AI system development |
| **3. REPEATABILITY / REPRODUCIBILITY**<br>AI results consistent: Be able to replicate an AI system's results by owner / 3rd-party | **SECURITY**<br>Cybersecurity of AI systems | **DATA GOVERNANCE**<br>Source and quality of data: Good data governance practices when training AI models | **8. HUMAN AGENCY AND OVERSIGHT**<br>AI system designed in a way that will not decrease human ability to make decisions |
|  | **5. ROBUSTNESS**<br>AI system can still function despite unexpected inputs |  | **INCLUSIVE GROWTH, SOCIETAL & ENVIRONMENTAL WELL-BEING**<br>Beneficial outcomes for people and planet |

11. The 5 pillars describe how system owners and developers can build trust with customers and consumers by demonstrating the following:

    a. **Transparency on Use of AI & AI systems.** By disclosing to individuals that AI is used in the system, individuals will become aware and can make an informed choice of whether to use the AI-enabled system.

    b. **Understanding how an AI model reaches a decision.** This allows individuals to know the factors contributing to the AI model's output, which can be a decision or a recommendation. Individuals will also know that the AI model's output will be consistent and performs at the level of claimed accuracy given similar conditions.

    c. **Ensuring safety and resilience of AI system.** Individuals know that the AI system will not cause harm, is reliable and will perform according to intended purpose even when encountering unexpected inputs.

    d. **Ensuring fairness i.e., no unintended discrimination.** Individuals know that the data used to train the AI model is sufficiently representative, and that the AI system does not unintentionally discriminate.

    e. **Ensuring proper management and oversight of AI system.** Individuals know that there is human accountability and control in the development and/or deployment of AI systems and the AI system is for the good of humans and society.

12. The 8 AI ethics principles selected can be assessed by a combination of technical tests and/or process checks. The following principles can potentially be assessed using both technical tests and process checks:

    a. **Explainability** – Assessed through a combination of technical tests and process checks. Technical tests are conducted to identify factors contributing to AI model's output. Process checks include verifying documentary evidence of considerations given to the choice of models, such as rationale, risk assessments, and trade-offs of the AI model.

    b. **Robustness** – Assessed through a combination of technical tests and process checks. Technical tests attempt to assess if a model performs as expected even when provided with unexpected inputs. Process checks include verifying documentary evidence, review of factors that may affect the performance of AI model, including adversarial attacks.

    c. **Fairness (Mitigation of unintended discrimination)** – Assessed through a combination of technical tests and process checks. Technical tests check that an AI model is not biased on protected or sensitive attributes specified by the AI system owner, by checking the model output against the ground truth. Process checks include verifying documentary evidence of having a strategy for the selection of fairness metrics that are aligned with the desired outcomes of the AI system's intended application; and the definition of sensitive attributes are consistent with the legislation and corporate values.

13. The following principles are assessed through process checks:

    a. **Transparency** – Assessed through process checks of documentary evidence (*e.g.,* company policy and communication collaterals) of providing appropriate information to individuals who may be impacted by the AI system. The information includes, under the condition of not compromising IP, safety, and system integrity, use of AI in the system, intended use, limitations, and risk assessment.

    b. **Repeatability/Reproducibility** – Assessed through process checks of documentary evidence including evidence of AI model provenance, data provenance and use of versioning tools.

    c. **Safety** – Assessed through process checks of documentary evidence of materiality assessment and risk assessment, including how known risks of the AI system have been identified and mitigated.

    d. **Accountability** – Assessed through process checks of documentary evidence, including evidence of clear internal governance mechanisms for proper management oversight of the AI system's development and deployment.

e. **Human agency and oversight** – Assessed through process checks of documentary evidence that AI system is designed in a way that will not reduce human's ability to make decisions or to take control of the system. This includes defining role of human in its oversight and control of the AI system such as human-in-the-loop, human-over-the-loop, or human-out-of-the-loop.

14. We would like to emphasise that the MVP:

    a. **Does not define ethical standards.** It aims to provide a way for AI system developers and owners to demonstrate their claims about the performance of their AI systems vis-à-vis the 8 selected AI ethics principles.

    b. **Does not guarantee** that any AI system tested under this Framework will be **free from risks or biases or is completely safe**; and

    c. Is used by AI system developers/owners to conduct **self-testing** so that data and models remain in the company's operating environment.

**Components of the MVP**

15. The MVP consists of a Testing Framework and a Toolkit. The Testing Framework specifies the testable criteria relevant to the selected AI ethics principles. The Toolkit is used to execute technical tests and record process checks described in the Testing Framework.

**Testing Framework**

16. The structure of the Testing Framework comprises the following key components:

    a. **Definitions of AI ethics principles.** The Testing Framework provides definitions for each of the AI ethics principles.

    b. **Testable criteria.** For every principle, a set of testable criteria will be ascribed. Testable criteria are a combination of technical and non-technical (e.g., processes and organisational structure) factors contributing to the achievement of the desired outcomes of that governance principle.

    c. **Testing process.** Testing processes are actionable steps to be carried out in order to ascertain if each testable criterion has been satisfied. The testing processes could be quantitative such as statistical tests and technical tests. They can also be qualitative such as producing documented evidence during process checks.

    d. **Metrics.** These are well-defined quantitative or qualitative parameters that can be measured, or the presence of evidence can be demonstrated for each testable criterion.

e. **Thresholds (where applicable).** As AI technologies are rapidly evolving, thresholds that define acceptable values or benchmarks for the selected metrics (whether defined by industry or by regulators) often do not exist. Hence, thresholds are not available in the current version of Testing Framework. However, we aim to collate and develop meaningful and context-specific metrics and thresholds as industry test their AI systems against the Testing Framework.

**Toolkit**

17. As a start, this Toolkit covers technical testing for three principles: Fairness, Explainability and Robustness. The Toolkit provides a "one-stop" tool for technical tests to be conducted by identifying and packaging widely used open-source libraries into a single Toolkit. These tools include SHAP (SHapley Additive exPlanations) for explainability, Adversarial Robustness Toolkit for adversarial robustness, and AIF360 and Fairlearn for fairness testing[5].

18. In terms of user experience, the Toolkit:

    a. Provides a user interface to guide users step by step in the testing process, including a guided fairness tree to guide users to the fairness metrics relevant for their use case;

    b. Supports certain binary classification and regression models that use tabular data, such as decision trees and random forest algorithms;

    c. Produces a basic summary report to help system developers and owners interpret the results of the tests;

    d. Is intended to be deployed in the user's environment and is packaged into a Docker® container which allows for easy deployment.

    For the process checks, the report will be in the form of a checklist, stating the presence or absence of documentary evidence specified in the Testing Framework.

[5] **SHAP**: Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
**AIF360**: Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.
**Fairlearn**: Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., ... & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020-32.
**Adversarial Robustness Toolbox**: Nicolae, M. I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., ... & Edwards, B. (2018). Adversarial Robustness Toolbox v1. 0.0. arXiv preprint arXiv:1807.01069.

## Scope and limitations of the MVP

19. As we are in the early stages of development and iteration, the Toolkit currently has the following features and limitations:

    a. **Works with a certain subset of common AI models**, such as binary classification, and regression algorithms from common frameworks such as scikit-learn, Tensorflow, and XGBoost. The toolkit does not support unsupervised models at this time;

    b. Can **handle tabular datasets** for most principles, with certain limitations (e.g., robustness tests cannot yet be executed on regression models). The toolkit has limited support for image datasets;

    c. **Supports small-to-medium scale models** (~2GB) which can be fully imported to the toolkit using a web interface. Larger models and AI pipelines may not work at this time;

    d. Over the course of the industry pilot, more functionalities that will gradually be made available with industry contribution and feedback.

## Illustration

20. The following is an illustration of how an AI model can be assessed against the explainability principle through technical testing and process checks.

| **EXPLAINABILITY** - Ability to assess the factors that led to AI system’s decision, its overall behaviour, outcomes, and implications |
|---|
| *Explainability is about ensuring AI driven decisions can be explained and understood by those directly using the system to enable or carry out a decision, to the extent possible. The degree to which explainability is needed also depends on the aims of the explanation, including the context, the needs of stakeholders, types of understanding sought, mode of explanation, as well as the severity of the consequences of erroneous or inaccurate output on human beings. Explainability is an important component of a transparent AI system. The testable criteria in this section focuses on system-enabled explainability. However, it may not be possible to provide an explanation for how a black box model generated a particular output or decision (and what combination of input factors contributed to that). In these circumstances, other explainability measures may be required (e.g., accountability and transparent communication). As state-of-the-art approaches to explainability become available, users should refine the process, metrics and/or thresholds accordingly.* |

| No. | Testable Criteria | Testing Process | Metric | Threshold | Technical Tool/Process checks |
|---|---|---|---|---|---|
| 1 | For each model being developed, run explainability methods to help users understand the drivers of the AI model. | Perform analysis to determine feature contributions. | Features contributing to model output as obtained from technical tool | Not applicable | IMDA Toolkit (comprising SHAP and LIME tools) |
| 2 | Lean towards a preference for developing AI models that can explain their decisions or that are interpretable by default. | If choosing a less explainable modelling approach, document the rationale, risk assessments, and trade-offs of the AI model. | Documented evidence of considerations given to choice of final model, which include rationale, risk assessments, and trad-offs. | Not applicable | Process checks of presence of documented evidence |

**PART III: INVITATION TO PILOT THE MVP**

21. The IMDA is inviting the following to participate in the pilot of the MVP:

a. AI system owners and developers who wish to verify their AI systems against internationally accepted AI ethics principles;

b. Technology solution providers who wish to contribute to the development of AI governance implementation and testing tools; and

c. Other testing framework owners and developers who wish to have early discussions on compatibility and interoperability with Singapore’s AI Governance Testing Framework and Toolkit.

22. The objectives of this pilot are to:

a. Validate that the MVP can be implemented by owners and developers for a wider range and variety of AI systems;

b. Identify research and development opportunities for testing tools and engage research institutions and technology solution providers for collaboration;

c. Begin collating industry consensus on acceptable performances of AI systems in terms of metrics and thresholds;

d. Begin collating industry best practices on implementing trustworthy AI systems; and

e. Explore compatibility and interoperability with like-minded owners and developers of AI systems testing frameworks.

23. Please contact the following if you are interested to participate in the pilot, or if you require more information.

| **Name** | **Email** |
|---|---|
| Cyrus Chng,<br>Assistant Manager (AI Governance) | Cyrus_CHNG@imda.gov.sg |
| Tan Wen Rui,<br>Manager (AI Governance) | TAN_Wen_Rui@pdpc.gov.sg |
| Chung Sang Hao,<br>Deputy Director (AI Governance) | CHUNG_Sang_Hao@pdpc.gov.sg |
| Lee Wan Sie,<br>Director (Development of Data-Driven Technologies) | LEE_Wan_Sie@imda.gov.sg |

INFOCOMM
MEDIA
DEVELOPMENT
AUTHORITY

## Acknowledgement

The Info-communications Media Development Authority and Personal Data Protection Commission express their sincere appreciation to the following for their valuable feedback and/or participation in the early-stage testing of the MVP:

1. AWS
2. DBS
3. Google
4. Meta
5. Microsoft
6. Singapore Airlines
7. NCS (Part of Singtel Group)/Land Transport Authority
8. Standard Chartered Bank
9. UCARE.AI
10. X0PA.AI

Supported by:

pdpc PERSONAL DATA PROTECTION COMMISSION SINGAPORE

In support of:

SG :DIGITAL