Annex B: Background on Singapore's AI Governance Work

Overview

Artificial Intelligence ("**AI**") has been identified as a key step of Singapore's Smart Nation journey. While AI bring about benefits, its risks have been an on-going discussion in international fora and among governments, international organisations, industry, academia, and civil society. The increasing interest in AI ethics and governance is due to public concerns with issues relating to AI systems' transparency, explainability, safety, robustness, unintended bias, and accountability. At the heart of these discussions is the goal to foster public trust in AI technologies while allowing innovation to flourish.

Singapore has been contributing to the international discourse on AI ethics and governance since 2018[2]. Our approach to balancing the need for technological innovation and building trust in AI is one of voluntary adoption of government's detailed guidance on AI deployment. The Infocomm Media Development Authority Singapore (IMDA) and Personal Data Protection Commission (PDPC) have published the following detailed guidance, namely:

a. Model AI Governance Framework (Model Framework). The First Edition was issued in January 2019, and the Second Edition was published in January 2020. The Model Framework converts high level AI ethics principles into implementable measures for organisations. Read Model Framework at: go.gov.sg/ai-gov-mf-2/

b. Implementation and Self-Assessment Guide (ISAGO). ISAGO was published in January 2020 to help organisations assess the alignment of their internal processes with the Model Framework. Read ISAGO at: https://go.gov.sg/isago

c. Compendium of Use Cases. Two volumes of Use Cases were published in 2020 to showcase successful implementation of measures in Model Framework by organisations of different sizes, from different sectors, locally and internationally. Read Volume 1 at: https://go.gov.sg/ai-gov-use-cases; Volume 2 at: https://go.gov.sg/ai-gov-use-cases-2

Development of *A.I. Verify* - AI Governance Testing Framework and Toolkit

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.

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

[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.

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. IMDA is working with regulators and standards organisations to map the MVP 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.

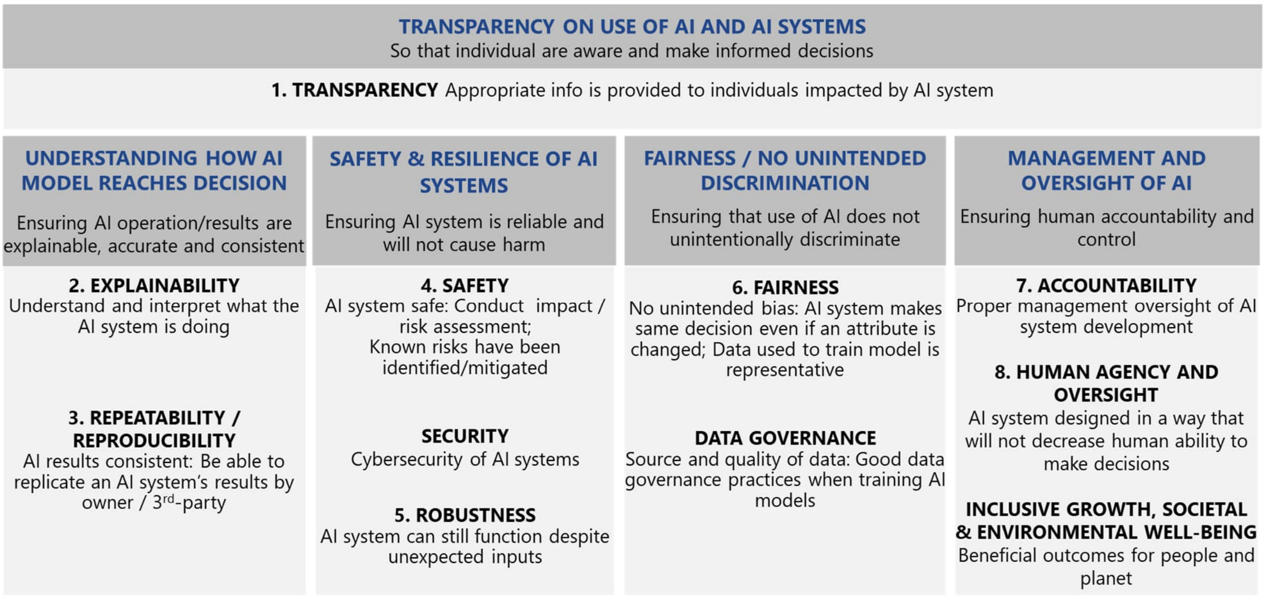
In developing the AI Governance Testing Framework and Toolkit, IMDA aligned it with internationally accepted AI ethics principles, guidelines, and frameworks, such as those from the EU and OECD. Countries are generally coalescing around 11 key AI ethics principles, grouped into 5 pillars (See Figure 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 and not reinventing the wheel.

Figure 1: Initial set of AI ethics principles for MVP for AI Governance Testing



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

The 8 AI ethics principles selected can be assessed by a combination of technical tests and/or process checks.

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.

Scope and limitations of the MVP
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.