DEVELOPING THE MVP FOR AI GOVERNANCE TESTING FRAMEWORK

Issued 14 July 2021

TABLE OF CONTENTS

PART I: INTRODUCTION

1. The baseline or Minimum Viable Product (MVP) testing framework for AI governance, also known as the Testing Framework 1.0, continues our efforts to promote responsible AI adoption. It aims to help AI system-owners and/or developers test and verify the performance of their AI solutions. This will be done through a mix of technical/statistical tests and process checks. The Testing Framework translates AI ethical principles into tangible results and is the practical next step for organisations.

PART II: AI GOVERNANCE TESTING FRAMEWORK

2. The Testing Framework identifies 12 ethical principles that describes four central aspects of a trustworthy AI system (please refer to **Annex**). These are common to prominent frameworks on trustworthy AI, including from international bodies such as the Organisation for Economic Co-operation and Development (OECD) and the European Union.

   (a) **Understanding how an AI model reaches a decision**: For users to know what the AI model does and that its results are consistent. The relevant principles are *explainability*, *repeatability* and *reproducibility*.

   (b) **Ensuring safety and resilience of AI system**: For users to know that the system will not cause harm and is reliable. The relevant principles are *safety*, *security* and *robustness (including accuracy)*.

   (c) **Ensuring fairness and no unintended discrimination**: To ensure that the AI system does not unintentionally discriminate. The relevant principles are *fairness* and *data governance*; and

   (d) **Ensuring management and oversight of AI system**: To ensure 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 relevant principles are *accountability, transparency, human agency and oversight*, and *inclusive growth, societal and environmental well-being*.

PART III: TESTING FRAMEWORK 1.0

3. The Testing Framework allows AI system owners to objectively assess and verify their claim(s) regarding their AI systems with respect to internationally accepted AI ethics and governance principles. The primary target audience of the Testing Framework 1.0 are AI system owners, i.e., those who implement AI systems to offer products/services to their end-users. AI developers who provide solutions to AI system owners will also find this

Testing Framework relevant as AI system owners often seek technical support from their solution providers.

4. The structure of the Testing Framework 1.0 comprises the following key components:

(a) **Definitions of AI governance principles**

The Testing Framework provides the definition of each governance principle.

(b) **Testable criteria**

For every governance 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 carried out 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 showing documented evidence.

(d) **Metrics**

These are well-defined quantitative or qualitative parameters that can be measured or has presence of evidence that can be demonstrated.

(e) **Thresholds**

These are acceptable values or benchmarks for the selected metrics. As AI technologies are still nascent and rapidly evolving, thresholds (whether defined by industry or by regulators) often do not exist. As we develop new versions of the Testing Framework, we aim to develop meaningful and context-specific metrics and thresholds.

PART IV: WHAT'S NEXT

5. We plan to work with companies and organisations to enhance the Testing Framework so that it will be relevant, useful and add value to industry. Please contact the following if you would like to be part of our community and receive updates, or if you require more information:

Name	**Email**
Tan Wen Rui (Ms) Manager (AI Governance) Trusted AI and Data	Tan_Wen_Rui@pdpc.gov.sg
Chung Sang Hao (Mr) Deputy Director (AI Governance) Trusted AI and Data	Chung_Sang_Hao@pdpc.gov.sg

END OF DOCUMENT

ANNEX: 12 AI ETHICAL PRINCIPLES ORGANISED INTO FOUR KEY AREAS

UNDERSTANDING HOW AI MODEL REACHES DECISION	SAFETY AND RESILIENCE OF AI SYSTEM	FAIRNESS / NO UNINTENDED DISCRIMINATION	MANAGEMENT AND OVERSIGHT OF AI SYSTEM
To know what it does and that results are consistent	AI system is reliable and will not cause harm	AI system does not unintentionally discriminate	Human accountability and control
EXPLAINABILITY Ability to understand and interpret what the AI system is doing **REPEATABILITY** Check that it's consistent: Be able to replicate an AI system's results **REPRODUCIBILITY** Ability to replicate an AI system's results (by independent third-party)	**SAFETY** Check that it's safe: Known risks have been identified/mitigated **SECURITY** Cybersecurity of AI systems **ROBUSTNESS** Ensuring that AI system can still function despite unexpected inputs	**FAIRNESS** Check that there is no unintended bias: AI systems makes same decision even if an attribute is changed **DATA GOVERNANCE** Know the source and quality of data: Good data governance practices when training AI models	**ACCOUNTABILITY** Proper management and oversight of AI system development **TRANSPARENCY** Appropriate information is provided to individuals impacted by AI system **HUMAN AGENCY AND OVERSIGHT** AI system designed in a way that will not decrease human ability to make decisions **INCLUSIVE GROWTH, SOCIETAL AND ENVIRONMENTAL WELL-BEING** Beneficial outcomes for people and planet