Digital Policy Office

Ethical Artificial Intelligence Framework

Version: 2.0

December 2025

© The Government of the Hong Kong Special Administrative Region of the People's Republic of China

Amendment History				
Change Number	Revision Description	Pages Affected	Revision Number	Date
1	1. Added PCPD's "Guidance on the Ethical Development and Use of Artificial intelligence" as a reference in Section 4.1.1.2 of the Framework 2. Added "Personal Information Protection Law" of the People's Republic of China as an example in Section 4.1.4.4 of the Framework. 3. Updated Appendix B 4. Supplemented Appendix E	4-7, 4-30, 6-4 to 6-6 Appendix E	1.1	June 2023
2	Updated the references in Appendix E	Appendix E	1.2	July 2023
3	Updated the references in Appendix E	Appendix E	1.3	August 2023
4	Change "Office of the Government Chief Information Officer" to "Digital Policy Office"		1.4	July 2024
5	1. Updated Section 1-5 2. Updated Appendix A-D 3. Removed Appendix E	Section 1-5, Appendix A-D	2.0	December 2025

Note: In the event of any discrepancies between the Chinese and English versions of this document, please refer to the English version and notify the Digital Policy Office.

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY 1-2

1.1 INTRODUCTION 1-2
1.2 WHAT IS THE ETHICAL AI FRAMEWORK? 1-3
1.3 HOW CAN THE ETHICAL AI FRAMEWORK BE USED? 1-12

2. PURPOSE 2-2

3. OVERVIEW OF THE ETHICAL AI FRAMEWORK 3-2

3.1 VISION FOR THE ETHICAL AI FRAMEWORK 3-2
3.2 OBJECTIVES 3-2
3.3 BENEFITS 3-2
3.4 INTRODUCTION TO AI AND DATA ETHICS 3-3
3.5 KEY COMPONENTS AND RELATIONSHIPS 3-5
3.5.1 Ethical AI Principles 3-8
3.5.2 AI Governance 3-23
3.5.3 AI Lifecycle 3-29

4. AI PRACTICE GUIDE 4-2

4.1 AI LIFECYCLE AND PRACTICES 4-2
4.1.1 Project Strategy 4-3
4.1.2 Project Planning 4-8
4.1.3 Project Ecosystem 4-13
4.1.4 Project Development 4-19
4.1.5 System Deployment 4-42
4.1.6 System Operation and Monitoring 4-51

5. AI ASSESSMENT 5-2

5.1 AI APPLICATION IMPACT ASSESSMENT 5-2
5.1.1 Process for Completing AI Application Impact Assessment 5-4
5.2 FREQUENCY OF AI ASSESSMENT 5-5
5.3 RECOMMENDATION 5-7

6. APPENDIX 6-2

APPENDIX A – GLOSSARY 6-2

APPENDIX B – REFERENCE 6-4

APPENDIX C – AI APPLICATION IMPACT ASSESSMENT TEMPLATE 6-11

APPENDIX D – AI STRATEGY TEMPLATE 6-38

SECTION 1

EXECUTIVE SUMMARY

1. EXECUTIVE SUMMARY

1.1 INTRODUCTION

Artificial intelligence ("**AI**") and big data analytics have the potential to enhance social well-being and are increasingly being applied to various business areas to improve operational efficiency and to provide new services, but at the same time this can bring about different challenges. It is important for organisations to consider AI and data ethics when implementing Information Technology ("**IT**") projects and providing services.

When organisations are considering the application of AI and big data analytics, they need to consider a range of factors such as the requirements of relevant legislation and stakeholder expectations on the applicable ethical standards of data and technology that appropriately reflect the value and culture of the local community.

The Ethical Artificial Intelligence Framework (called the "**Ethical AI Framework**" hereunder) document consists of:

- A **Tailored Framework** for ethical use of AI and big data analytics when implementing IT projects; and
- An assessment template (used to complete "**AI Assessment**") for AI and big data analytics to assess the impact of AI applications or projects.

In this document, the term "AI" is used to refer to analytic operations involving big data analytics, advanced analytics and machine learning that use massive data sets and processing capabilities to find correlations and make predictions. **The term "AI applications" has been used to refer to a collective set of applications whose actions, decisions or predictions are empowered by AI models.** Examples of AI applications are IT projects which have prediction functionality and/or model development involving training data. For IT projects that have AI applications, organisations can make reference to the requirements of the Ethical AI Framework.

The adoption of Ethical AI Framework is a step that establishes a common approach and structure to govern the development and deployment of AI applications with the intention to maximise the benefits of the application of AI in IT projects based on the following guiding principles:

- Facilitate organisations to understand the application of AI and big data analytics in their respective business areas;
- Complement other operating guidelines (e.g. privacy, security and data management);
- Foster and guide the ethical use of AI and big data analytics in the organisation;

- Facilitate organisations to consider requirements on AI ethics relevant to their business and technical domains, govern and assess the compliance of the IT projects; and
- Assist organisations in identifying and managing potential risks in adopting AI and big data analytics in IT projects by conducting AI Assessment.

1.2 WHAT IS THE ETHICAL AI FRAMEWORK?

The Ethical AI Framework was originally developed to assist government bureaux/departments ("**B/Ds**") in planning, designing and implementing AI and big data analytics in their IT projects and services. It consists of guiding principles, leading practices and AI Assessment that should be adopted for B/Ds' AI-powered IT projects. The AI Assessment includes assessment template for B/Ds to utilise at application level which would support B/Ds on the management of benefits, impact and risks. Nonetheless this framework, including guiding principles, practices and assessment template, is also applicable to other organisations in general and this version of framework is suitably revised (e.g. removal or adjustment of government specific terms) for general reference by organsiations when adopting AI and big data analytics in their IT projects.

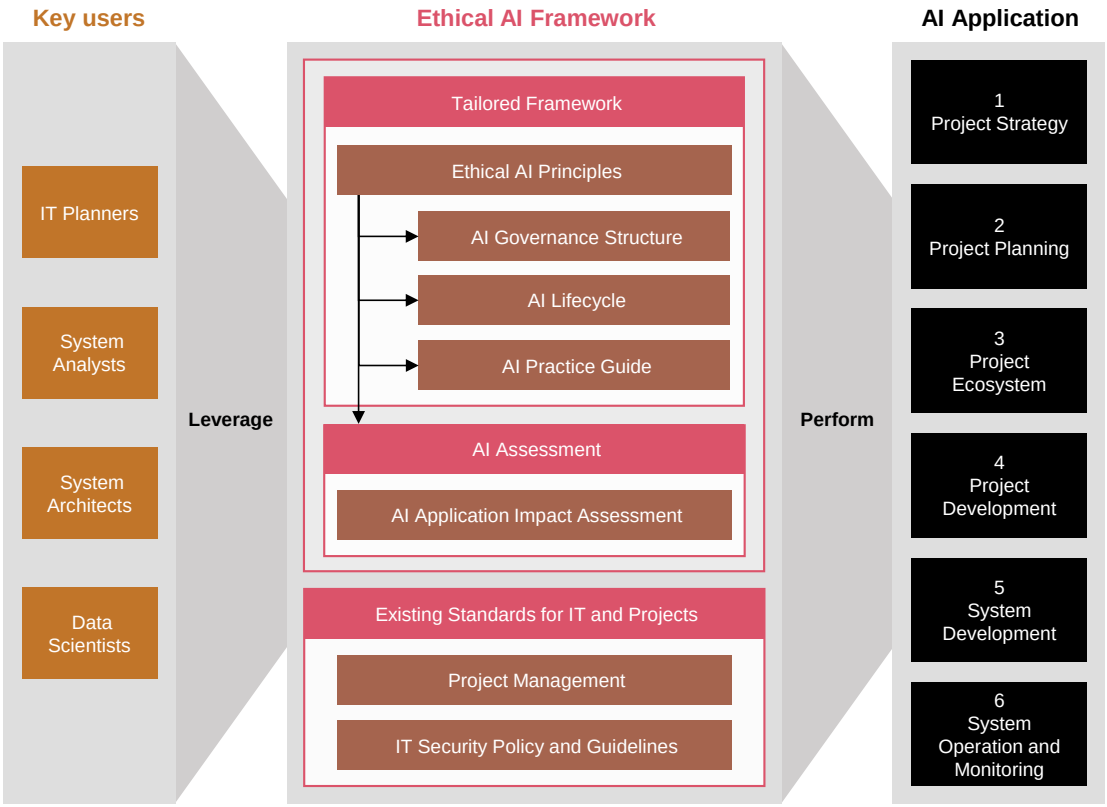


Figure 1: Overview of the Ethical AI Framework

The Ethical AI Framework consists of the following key components:

- The ***Tailored Framework*** provides a set of Ethical AI Principles, an AI Governance Structure, an AI Lifecycle and an AI Practice Guide. Please refer to Section 3.5 "Key Components and Relationships" for details.

- The ***AI Assessment*** includes an assessment template for organisations to assess the benefits, impact and risks of AI applications. An **AI Application Impact Assessment** is used to assess AI applications to ensure that the impact across the AI Lifecycle is managed and that related Ethical AI Principles have been considered. Please refer to Section 5 "AI Assessment" for details.
- **Existing standards and practices** (e.g. project management, IT security standards) should be used as part of the overall framework for managing IT projects. AI applications would be a subset of these projects. Please refer to "Appendix B – Reference" for details.

Ethical AI Principles

Twelve Ethical AI Principles should be observed for all AI projects. Two out of the twelve principles (1) **Transparency and Explainability** and (2) **Reliability, Robustness and Security** are "Performance Principles". These fundamental principles must be achieved to create a foundation for the execution of other principles. For example, without achieving the Reliability, Robustness and Security principle, it would be impossible to accurately verify that other Ethical AI Principles have always been followed.

The other principles are categorised as "General Principles", including (1) **Fairness,** (2) **Diversity and Inclusion**, (3) **Human Oversight**, (4) **Lawfulness and Compliance**, (5) **Data Privacy**, (6) **Safety**, (7) **Accountability**, (8) **Beneficial AI**, (9) **Cooperation and Openness** and (10) **Sustainability and Just Transition**. They are derived from the United Nations' Universal Declaration of Human Rights and Hong Kong's relevant ordinances:

Principle	Definition
Transparency and Explainability	Organisations should be able to explain the decision-making processes of the AI applications to humans in a clear and comprehensible manner.
Reliability, Robustness and Security	Like other IT applications, AI applications should be developed such that they will operate reliably over long periods of time using the right models and datasets while ensuring they are both robust (i.e. providing consistent results and capable to handle errors) and remain secure against cyber-attacks as required by the relevant legal and industry frameworks.
Fairness	The recommendation/result from the AI applications should treat individuals within similar groups in a fair manner, without favouritism or discrimination and without causing or resulting in harm.
Diversity and Inclusion	Inclusion and diverse usership through the AI application should be promoted by understanding and respect the interests of all stakeholders impacted.
Human Oversight	The degree of human intervention required as part of AI application's decision-making or operations should be dictated by the level of the severity of ethical issues involved.
Lawfulness and Compliance	Organisations responsible for an AI application should always act in accordance with the law and regulations and relevant regulatory regimes.

Principle	Definition
Data Privacy	Individuals should have the right to: (a) be informed of the purpose of collection and potential transferees of their personal data and that personal data shall only be collected for a lawful purpose, by using lawful and fair means, and that the amount of personal data collected should not be excessive in relation to the purpose. Please refer to the Data Protection Principles ("**DPP**")**1** "Purpose and Manner of Collection" of the Personal Data (Privacy) Ordinance (the "**PD(P)O**")[1]. (b) be assured that data users take all practicable steps to ensure that personal data is accurate and is not kept longer than is necessary. Please refer to the **DPP2** "Accuracy and Duration of Retention" of the PD(P)O. (c) require that personal data shall only be used for the original purpose of collection and any directly related purposes. Otherwise, express and voluntary consent of the individuals is required. Please refer to the **DPP3** "Use of Personal Data" of the PD(P)O. (d) be assured that data users take all practicable steps to protect the personal data they hold against unauthorised or accidental access, processing, erasure, loss or use. Please refer to the **DPP4** "Security of Personal Data" of the PD(P)O. (e) be provided with information on (i) its policies and practices in relation to personal data, (ii) the kinds of personal data held, and (iii) the main purposes for which the personal data is to be used. Please refer to the **DPP5** "Information to Be Generally Available" of the PD(P)O.
Safety	Throughout their operational lifetime, AI applications should not compromise the physical safety or mental integrity of mankind.
Accountability	Organisations are responsible for the moral implications arising from their AI applications, including their use and misuse. Besides, there should be a clearly identifiable accountable party, be it an individual or an organisational entity.
Beneficial AI	The development of AI should promote the common good.
Cooperation and Openness	A culture of multi-stakeholder open cooperation in the AI ecosystem should be fostered.
Sustainability and Just Transition	The AI development should ensure that mitigation strategies are in place to manage any potential societal and environmental system impacts.

Table 1: Ethical AI Principles and Definitions

[1] https://www.pcpd.org.hk/english/data_privacy_law/ordinance_at_a_Glance/ordinance.html

AI Governance

AI governance refers to the practices and direction by which AI projects and applications are managed and controlled. The three lines of defence is a well-established governance concept in many organisations. Figure 2 shows the different defence lines and their roles.

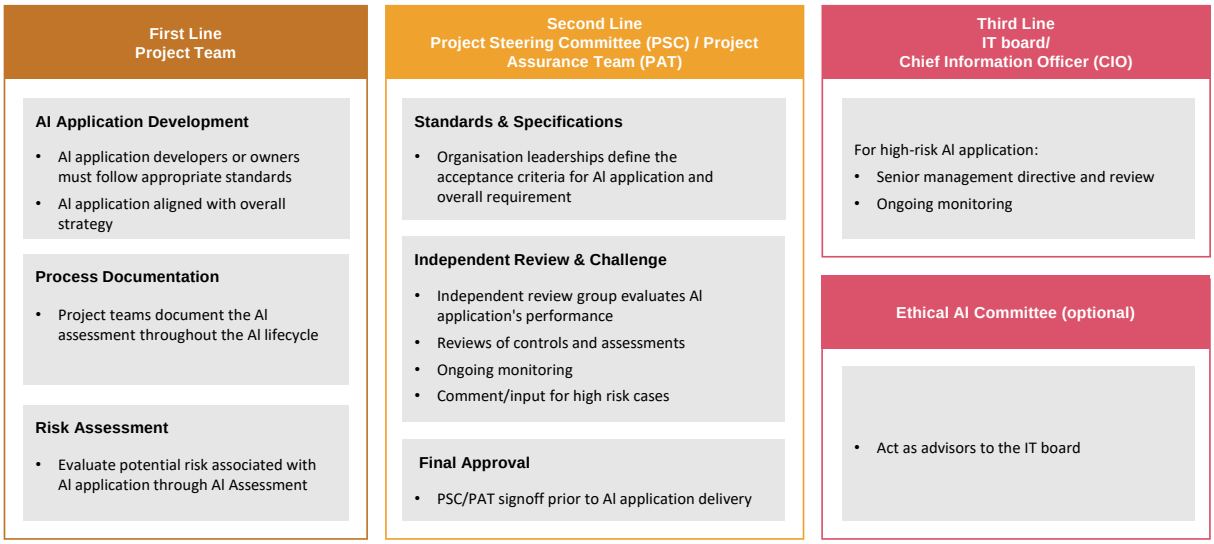


Figure 2: Lines of Defence Model

The governance structure consists of the following.

- The **first line of defence** is the Project Team who is responsible for AI application development, risk evaluation, execution of actions to mitigate identified risks and documentation of AI Assessment.
- The **second line of defence** is comprised of the Project Steering Committee ("**PSC**") and Project Assurance Team ("**PAT**") who are responsible for ensuring project quality, defining acceptance criteria for AI applications, providing independent review and approving AI applications. The Ethical AI Principles should be addressed through the completion of AI Assessment before approval of the AI application.
- The **third line of defence** involves the IT Board, or Chief Information Officer ("**CIO**") if the IT Board is not in place, and is optionally supported by an Ethical AI Committee, which may consist of external advisors. The purpose of the Ethical AI Committee is to provide advice on ethical AI and strengthen organisations' existing competency on AI adoption. The third line of defence is responsible for reviewing, advising and monitoring of high-risk AI applications.

AI Lifecycle

In order to structure the practices for organisations to follow when executing AI projects/creating AI applications, practices in different stages of the AI Lifecycle have been detailed in the AI Practice Guide (Please refer to Section 4 "AI Practice Guide" in the Ethical AI Framework for further details). A way to conceptualise the AI Lifecycle appears in the following 6-step schematic.

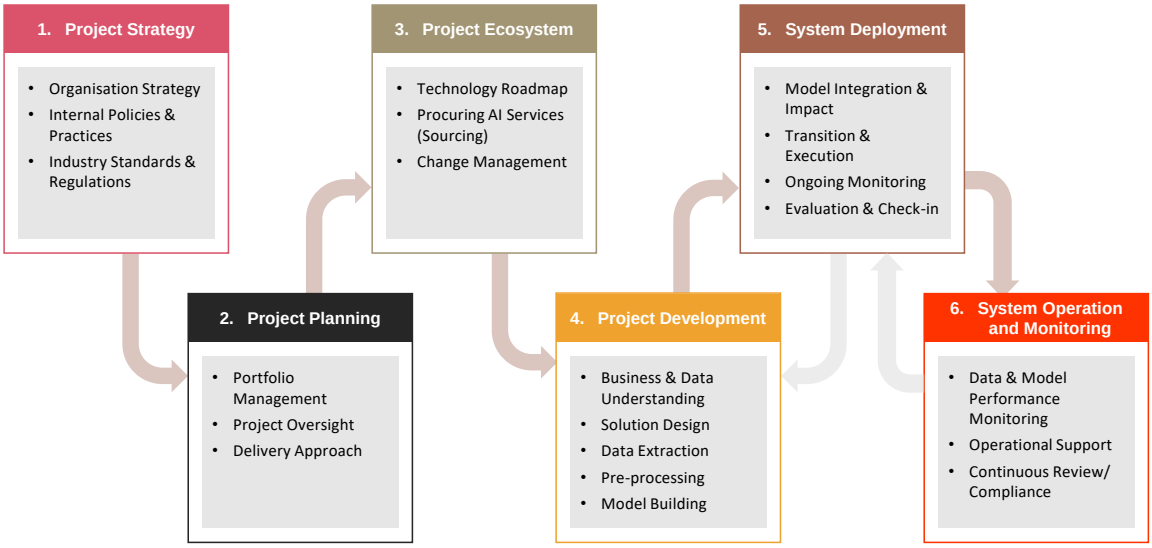


Figure 3: Overview of the AI Lifecycle

The AI Lifecycle shows the different stages involved in AI projects that can

- Guide organisations to understand the different stages and requirements involved; and
- Serve as a reference for the development of AI practices to align with actual stages of how AI is typically developed.

The AI Lifecycle is used to align the practices in the AI Practice Guide. The AI Lifecycle also aligns with a traditional System Development Lifecycle ("**SDLC**") model as depicted in Figure 4.

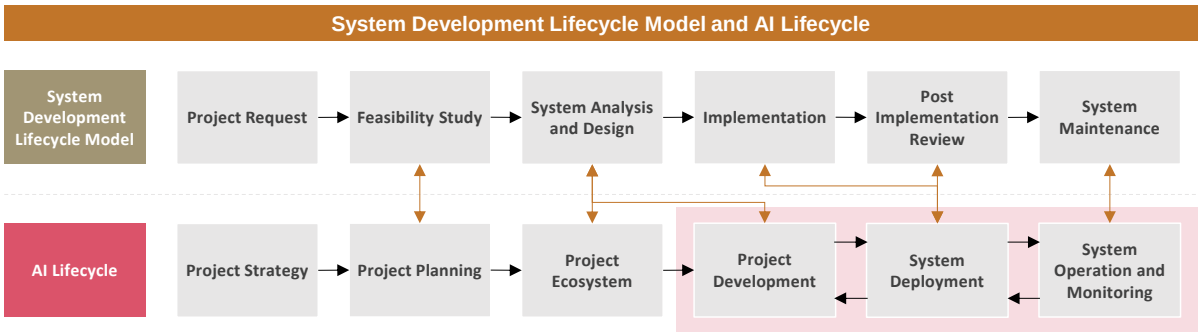


Figure 4: AI Lifecycle Aligned with a System Development Lifecycle Model

In a typical development process of an AI application, great emphasis is placed on data because good data are often required for creating a good AI model. Data sourcing and preparation which is part of the project development can be a continuous exercise. This is because an AI model can often benefit from better or more data for iterative model training throughout the development

process. The approach of conventional software lifecycle is to program the IT application with a set of instructions for a pre-defined set of events. Thereafter, the IT application will exploit its computing capabilities and other resources to process the data fed into the system. This is different from an AI application where a huge amount of data are fed into the application, which in turn processes all the data resulting in a trained model or AI solution. This trained model is then used to solve new problems.

There is often a continual feedback loop between the development and deployment stages, as well as system operation and monitoring of the AI Lifecycle for iterative improvements making this distinct from a traditional system development lifecycle.

AI Practice Guide

Section 4 "AI Practice Guide" contains detailed practices to be followed for a number of practice areas. Such practice areas are assessed as part of the AI Application Impact Assessment. A summary of the practice areas in the AI Practice Guide is listed below.

AI Lifecycle	Practice Area	Definition
Project Strategy	**Organisations Strategy, Internal Policies and Practices**	Organisations should be able to explain decision-making processes of the AI applications to humans in a clear and comprehensible manner.
	Industry Standards and Regulations	Relevant regulations and standards require an assessment to ensure that AI and related processes adhere to any relevant laws or standards.
Project Planning	**Portfolio Management**	Portfolio management is performed to ensure that the IT investments embedded in the organisation's processes, people and technology are on course. Assessment of AI projects to ensure that they gainfully address business requirements and objectives.
	Project Oversight and Delivery Approach	Quality control not only monitors the quality of deliverables; it involves monitoring various aspects of the project as defined in the Project Management Plan ("**PMP**").
Project Ecosystem	**Technology Roadmap for AI and Data Usage**	A technology roadmap should enable the organisations to plan and strategise which, when and what technologies will be procured for AI and big data analytics.
	Procuring AI Services	Off-the-shelf products and data can be procured for AI projects. In conducting such procurement exercises, organisations should duly consider the related ethical considerations.
Project Development	**Business & Data Understanding**	Organisations should determine the objectives of using AI and weigh and balance the benefits and risks of using AI in the decision-making process.

AI Lifecycle	Practice Area	Definition
	Solution Design	Organisations should assess the AI model within an AI application for suitability compared to the organisation's objectives as well as the appropriate level of human intervention required.
	Data Extraction	Organisations should ensure the data quality, validity, reliability and consistency of information from various sources, within or outside of the organisations.
	Pre-processing	Sensitive data containing an individual's information require extra care during solution development to prevent data leakage as well as breaches of privacy and security.
	Model Building	Model building should aim at mitigating common AI application errors such as inaccurate model assumptions, model overfitting and adversarial attacks.
System Deployment	**Model Integration & Impact**	Verification, validation and testing is the process of ensuring the AI applications perform as intended based on the requirements outlined at the beginning of the project. This would ensure proper integration of the AI application.
	Transition & Execution	Assuming the AI application would fail, mitigation steps should be incorporated to minimise damages in the case of failure prior to deployment of the AI application.
	Ongoing Monitoring	Provide feedback to monitor/maintain model performance and robustness of the AI application.
	Evaluation & Check-in	Traceability, Repeatability and Reproducibility of the AI application are required to ensure the AI application is operating correctly and to help build trust from the public and stakeholders.
System Operation and Monitoring	**Data and Model Performance Monitoring**	AI models (as part of AI applications) should be continuously monitored and reviewed due to the likelihood of the AI models becoming less accurate and less relevant.
	Operational Support	Upon AI applications deployment, ongoing operational support should be established to ensure that the AI applications performance remains consistent, reliable and robust.

AI Lifecycle	Practice Area	Definition
	Continuous Review/Compliance	The Project Manager (or Maintenance Team), PSC/PAT (or Maintenance Board) and IT Board/CIO (or its delegates) are responsible to monitor risks of AI such as non-compliance with applicable new/revised laws and regulations.

Table 2: Practice areas covered in the AI Practice Guide within the Ethical AI Framework

AI Application Impact Assessment

The AI Application Impact Assessment should be conducted on an AI application at different stages of the AI Lifecycle. The AI Application Impact Assessment introduces a systematic thinking process for organisations to go through different aspects of considerations of individual applications for their associated benefits and risks whilst highlighting the need for additional governance activities and identifying follow-up actions to ensure necessary measures and controls required for implementing ethical AI.

The AI Application Impact Assessment template used for this assessment is provided in Appendix C "AI Application Impact Assessment Template" in the Ethical AI Framework document for reference.

The AI Application Impact Assessment has the following components:

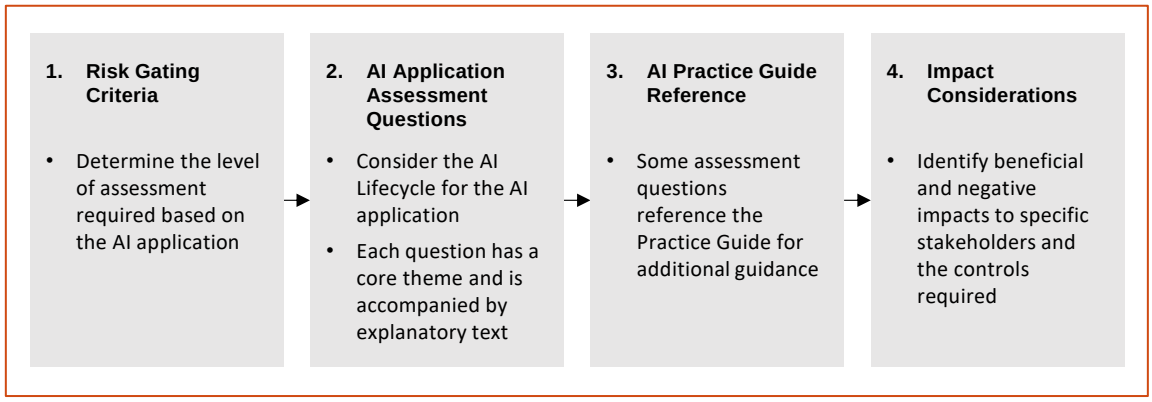


Figure 5: AI Application Impact Assessment Components

An AI Application Impact Assessment should be conducted regularly (e.g. annually) or when major changes take place as AI projects progress and when the AI application is being operated.

The stages of the AI Lifecycle where AI Application Impact Assessment should be reviewed are shown in Figure 6.

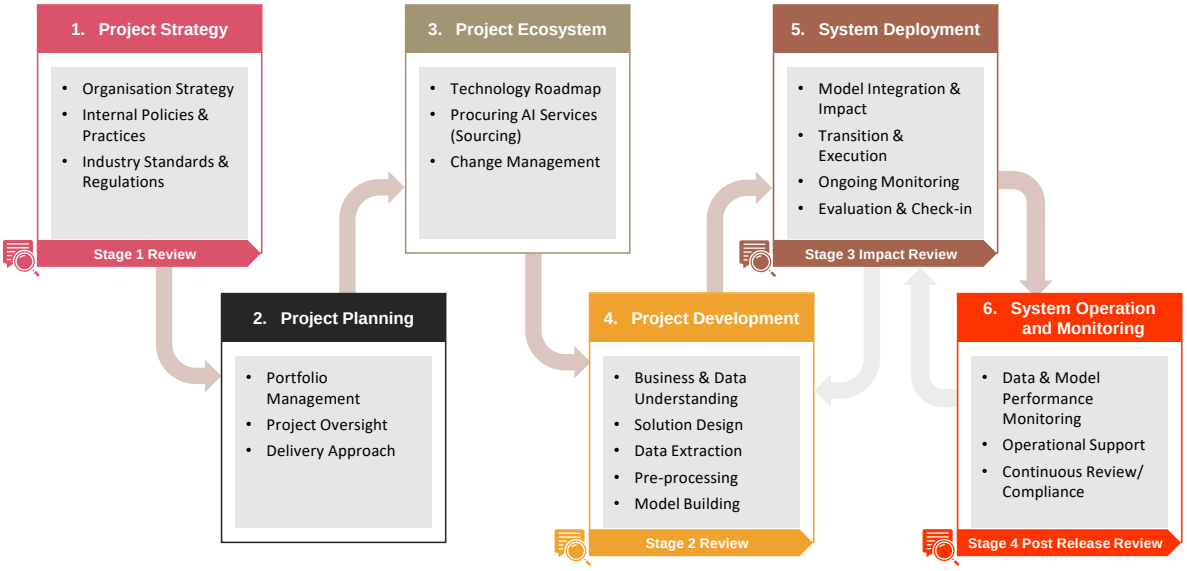


Figure 6: Stages for AI Application Impact Assessment

The AI Application Impact Assessment can be used as a 'live' document throughout the AI Lifecycle, but the associated AI Application Impact Assessment should be reviewed at 4 key stages of the AI Lifecycle with a copy of the AI Application Impact Assessment being retained for historical records. The mapping of these reviews to the system development lifecycle with responsible parties and actions to be performed is summarised below.

Lifecycle stage	Responsible party	Actions to be performed
SDLC: Project Request **AI Lifecycle:** Project Strategy	Project Team	Categorise the project as "high-risk" or "non-high-risk" based on answers to the "Risk Gating questions" Conduct AI Application Impact Assessment • Answer questions 1-8, 9(i), 41-49, 57-58
	PSC/PAT	Review and endorse the assessment for non-high-risk project
	IT Board/CIO	Review and endorse the assessment for high-risk project
SDLC: System Analysis and Design **AI Lifecycle: Project Ecosystem** Project Development	Project Team	Conduct AI Application Impact Assessment • Answer questions 9(ii), 10-14, 18-19, 22-29, 30(i)(ii), 31-34, 37(i), 52-56 • Review and update the answers of questions 1-8, 9(i), 41-49, 57-58 according to the latest position • If third-party technology or data is used, complete and review questions 15-17, 48 before procurement
	PSC/PAT	• Review and endorse the assessment

Lifecycle stage	Responsible party	Actions to be performed
SDLC: Implementation (before rollout) **AI Lifecycle:** System Deployment	Project Team	Conduct AI Application Impact Assessment • Answer questions 20, 21, 30(iii), 35, 36, 37(ii)(iii), 38-40, 50, 51 • Review and update the AI Application Impact Assessment according to the latest position (questions 1-14, 18-19, 22-29, 30(i)(ii), 31-34, 37(i), 41-49, 52-58) • If third-party technology or data is used, review questions 15-17, 48
	PSC/PAT	• Review and endorse the assessment
SDLC: System Maintenance (annually or when major changes take place) **AI Lifecycle:** System Operation and Monitoring	Maintenance Team	• Review and update the "AI Application Impact Assessment" according to the latest position • Escalate any significant issues
	Maintenance Board IT Board/CIO (or its delegates)	• Handle escalation • Monitor high-risk projects

Table 3: Actions to be performed for completing and reviewing the AI Application Impact Assessment

1.3 HOW CAN THE ETHICAL AI FRAMEWORK BE USED?

Organisations can make use of the Ethical AI Framework when adopting AI and big data analytics in their IT projects or services. Ethical AI Framework is defined not only to serve as a reference/guide for the IT project team during the development and maintenance process. It also defines the governance structure to enable organisations to demonstrate accountability in building trust upon adoption of AI by evaluating the impact, ensuring safe and responsible use of AI and facilitating innovation.

Furthermore, IT Planners and Executives can make reference to the Ethical AI Framework to embed appropriate ethical AI considerations starting from the strategy formulation, planning and establishment of the ecosystem. For details of the relevant section to different roles, please refer to Section 3.5 "Key Components and Relationships".

SECTION 2

PURPOSE

2. PURPOSE

This document is intended to provide readers with an understanding of the Ethical AI Framework and procedures that should be carried out to embed ethical elements in organisations' planning, design and implementation of AI applications in IT projects or services. The major sections of this document comprise of:

- **Section 1 - Executive Summary** provides an outline of the Ethical AI Framework components, usage and governance structure;
- **Section 2 - Purpose** outlines the objectives of every section in this document;
- **Section 3 - Overview of the Ethical AI Framework** introduces the Ethical AI Framework, Ethical AI Principles, vision statement, roles and responsibilities and objectives;
- **Section 4 - AI Practice Guide** provides practical guidance with references to the Ethical AI Principles as part of the Ethical AI adoption process by the organisations; and
- **Section 5 - AI Assessment** refers to a template on the ethical AI aspects and considerations that require completion by the organisations to assess AI Application Impact.

The intended audience and recommended sections are listed in the table below:

Audience	Recommended Sections
Executives of organisations	Section 1 - Executive Summary
Chief Information Officers ("CIOs"), **IT Planners/IT Board,** **Project Steering Committee ("PSC"),** **Project Assurance Team ("PAT"), and** **Business Users**	Section 1 - Executive Summary Section 3 - Overview Section 4.1.1 - Project Strategy Section 4.1.2 - Project Planning Section 4.1.3 - Project Ecosystem Section 5 - AI Assessment
Project Managers	All Sections
Project Team (including System Analysts, System Architects and Data Scientists)	Section 4.1.4 - Project Development Section 4.1.5 - System Deployment Section 4.1.6 - System Operation and Monitoring Section 5 - AI Assessment

SECTION 3

OVERVIEW OF THE ETHICAL AI FRAMEWORK

3. OVERVIEW OF THE ETHICAL AI FRAMEWORK

3.1 VISION FOR THE ETHICAL AI FRAMEWORK

The vision for the Ethical AI Framework is to enable organisations to manage potential ethical issues and implications through assessing their AI capabilities and applications. This enables delivery of ethical AI whilst managing the potential impact of AI applications.

The purpose of this section is to provide an overview of the Ethical AI Framework for organisations. Examples used are taken from different sources and are purely for illustrative purposes only.

3.2 OBJECTIVES

The objectives of the Ethical AI Framework are:

- To support organisations to understand Ethical AI and its applications by:
  - establishing Ethical AI Principles with definitions, practices and guidelines; and
  - defining roles and responsibilities for organisations when implementing Ethical AI.
- To guide the ethical use of AI in organisations by:
  - providing a description of activities to be covered throughout all stages of the AI Lifecycle and the corresponding capabilities required to apply ethical AI; and
  - providing practical guidelines for organisations to observe and apply when they incorporate AI in IT projects to ensure ethical adoption.
- To assist organisations to govern the compliance of AI applications by:
  - providing an AI Application Impact Assessment template that enables organisations to assess their AI application over a set of practical considerations for implementing ethical AI.

3.3 BENEFITS

The adoption of Ethical AI Framework is a foundation step that establishes a common approach and structure to govern the subsequent development and deployment of AI applications. Benefits of having an Ethical AI Framework include:

- Establishing common best practices to ensure organisations have guidance and references to adopt AI in IT projects with appropriate ethical considerations;
- Identifying the benefits, risks and impacts of an AI application to enable better risk mitigation decisions that maximise benefits; and
- Acting as a bridge between the strategy and execution which helps ensure the AI application is aligned with organisations' vision and needs.

3.4 INTRODUCTION TO AI AND DATA ETHICS

In the age of big data, enormous quantities of data are being generated, collected and analysed to identify insights and support decisions making. Big data are often described in terms of the 'five Vs' where:

- volume refers to the vast quantity of the data available;
- velocity refers to the speed at which data must be stored and/or analysed to provide the right information at the right time to make appropriate management decisions;
- variety refers to a huge variation in types and sources of data including both structure and unstructured data (e.g. file objects, social media feeds, tags, data from sensors, audio, image and video);
- veracity refers to the trustworthiness of the data over its accuracy and quality; and
- value refers to the ability to transform data to improve outcomes/values.

Analytical techniques that are used to analyse big data are often being described as advanced analytics, machine learning and AI. These technical terms have similar meanings and overlap with each other. They all refer to analytic operations that take advantage of large volume data describing the past situations (i.e. historical data), and massive processing capabilities and advanced algorithms, and that use them to find correlations and make predictions with acceptable accuracy.

In a broad definition, AI is a collective term for computer systems that can sense their environment, think, learn and take actions in response to the gathered data, with the ultimate goal of fulfilling their design objectives. AI systems are a collection of interrelated technologies used to help solve problems autonomously and perform tasks to achieve defined objectives without explicit guidance from a human being. We can distinguish the four main categories of AI (see Figure 7):

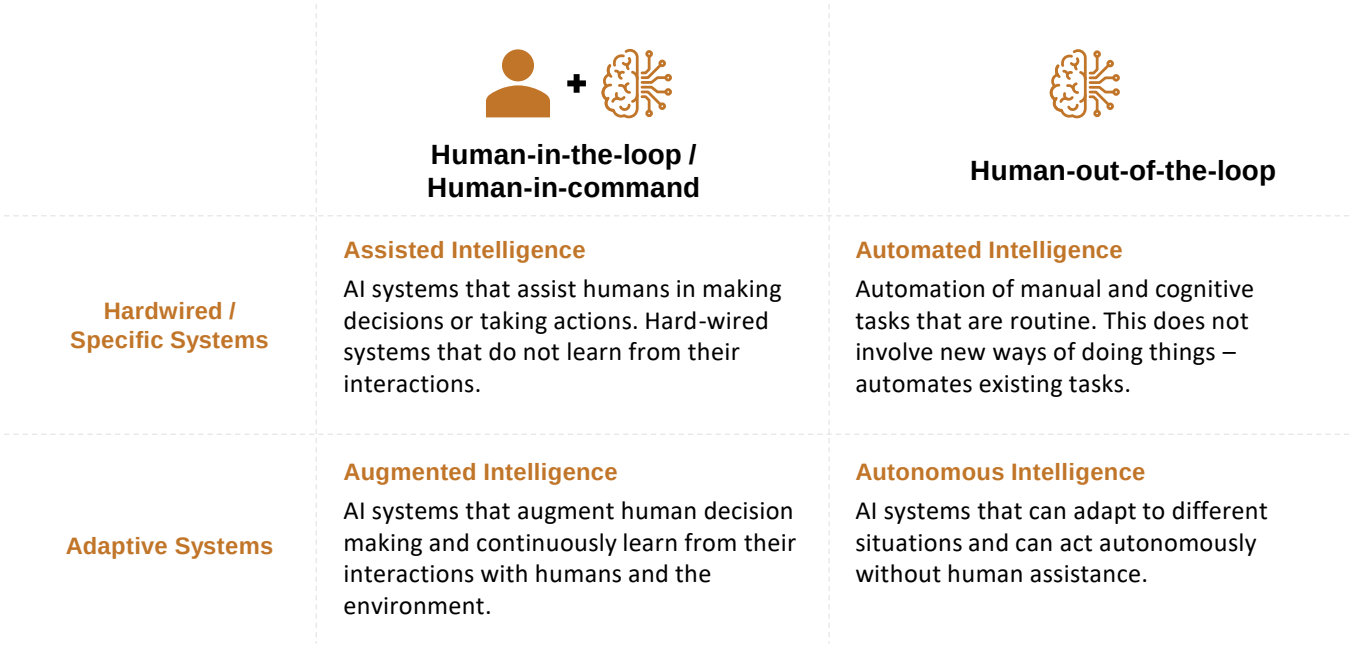


Figure 7: Categorisation of Artificial Intelligence[2]

- **Assisted Intelligence:** AI applications in this category assist humans in making decisions or taking actions. These applications do not necessarily learn from their interactions with the environment, and the final decision maker is eventually a human agent. Their objective is to facilitate human beings in performing cognitive tasks faster and better, improving what human beings and organisations are already doing. A simple example, prevalent in cars today, is the Global Positioning System ("**GPS**") navigation program that offers directions to drivers with adjustments to road conditions.
- **Augmented Intelligence:** This category includes AI applications that augment human decision making and continuously learn from their interactions with human beings and the environment they operate in. They do not make decisions autonomously and their objective is to assist humankind in making better decisions, enabling them and organisations to do things they could not otherwise do.
- **Automated Intelligence:** This category refers to the automation of manual/cognitive and routine/non-routine tasks. However, it does not involve new ways of functioning since the objective is simply to automate existing tasks. Examples of Automated Intelligence are Robotic Process Automation applications used to automate back-office processes. A key characteristic of these applications is that there is "no human-in-the-loop" meaning the system makes decisions in a completely autonomous manner.
- **Autonomous Intelligence**: The goal of these AI applications is to automate decision making processes without human intervention. To achieve this goal, they learn from their interactions

[2] https://www.pwc.no/no/publikasjoner/Digitalisering/pwc-ai-analysis-sizing-the-prize-report.pdf

with human beings, as well as with their environment. Examples of Autonomous Intelligence solutions are autonomous vehicles, robots, chatbots and virtual assistants.

Definitions of some key ethical AI terms are as follows:

- **AI Ethics:** The consideration of ethical implications as they relate to the use, development, implementation or outcome of AI. AI Ethics can be considered a subset of Data Ethics.
- **Ethical AI Principles:** A set of behavioural principles related to the use, development, implementation or outcome of AI.
- **Data Ethics:** Branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including AI/machine learning models, artificial agents and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes) in order to formulate and support morally good solutions (e.g. right conducts or right values).

3.5 KEY COMPONENTS AND RELATIONSHIPS

The key components of the Ethical AI Framework are depicted in Figure 8. Details of the components are described in the subsequent sections (i.e. Section 4 “AI Practice Guide” and Section 5 “AI Assessment”). The Ethical AI Framework should be read in conjunction with existing standards and practices for IT and project management.

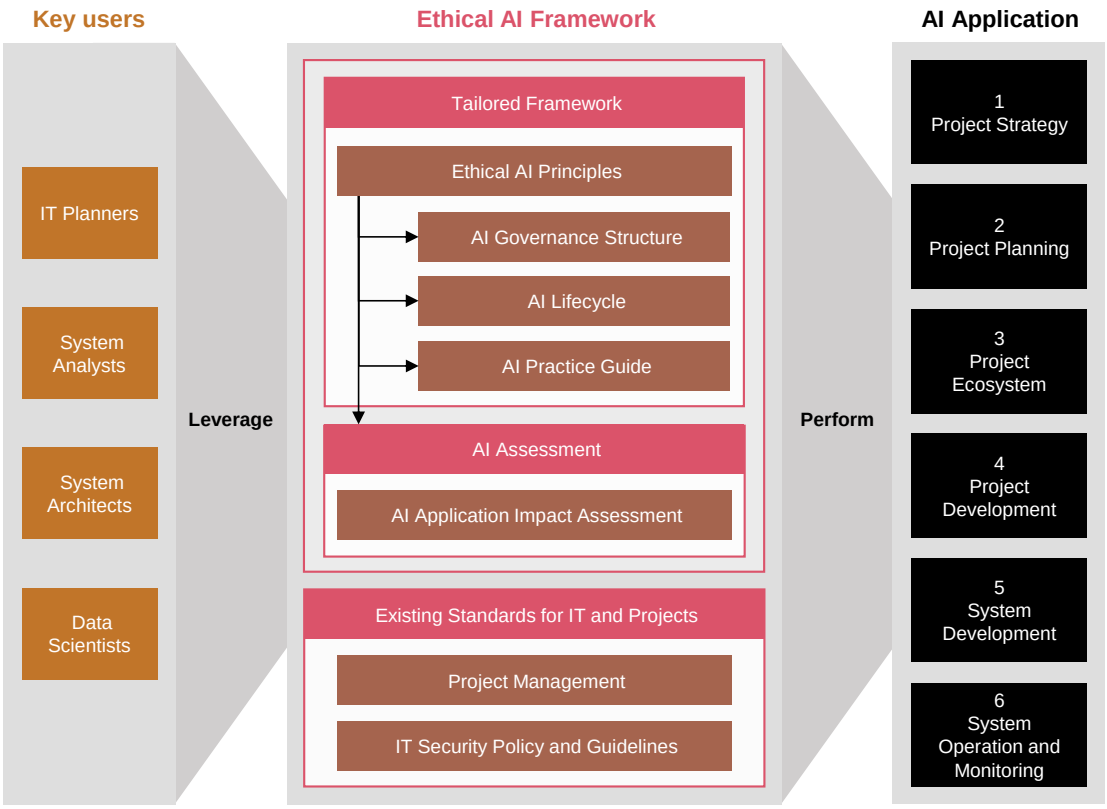


Figure 8: Key Components of the Ethical AI Framework

The Ethical AI Framework consists of the following components:

- The **Tailored Framework**, which consists of the following sub-components:
  - The **Ethical AI Principles** define principles to be followed when designing and developing AI applications. The Ethical AI Principles are applicable to all roles as defined in the AI Governance Structure;
  - The **AI Governance Structure** defines standard structures and roles and responsibilities over the adoption process of AI against practices set out in the Ethical AI Framework;
  - The **AI Lifecycle** defines an AI lifecycle model that is used to structure the layout of practices in the AI Practice Guide and the questions in the AI Assessment; and
  - The **AI Practice Guide** defines a set of practices for all stages in the AI Lifecycle. The guiding practices are derived from the Ethical AI Principles.
- The **AI Assessment** consists of the **AI Application Impact Assessment** template which defines questions to be answered by target readers across the stages as part of the AI Lifecycle to assess the impact of AI applications or projects and to ensure that Ethical AI Principles have been considered.

Target readers of the Ethical AI Framework include IT Planners, System Analysts, System Architects and Data Scientists. Data Scientists encompass a role including development, deployment and monitoring AI for the Ethical AI Framework. These areas can also be separate

roles depending on the individual setup of the organisations. These target readers should leverage the Ethical AI Framework to:

- understand Ethical AI Principles and practices;
- initiate discussions on the impact of AI;
- adopt standardised of practices and terminology; and
- perform AI Assessment.

For IT Planners and Executives

- ***Project Strategy***

  IT Planners and Executives can refer to Section 4.1.1 "Project Strategy" of the AI Practice Guide to establish organisational AI/data strategy, and to ensure Ethical AI Principles are embedded and relevant regulations and ordinances are considered.

- ***Project Planning***

  IT Planners in organisations can refer to Section 4.1.2 "Project Planning" of the AI Practice Guide along with the risk gating criteria within the AI Application Impact Assessment to ensure ethical AI requirements are met, impacts are effectively assessed, and to determine which AI projects require further review at senior level. Ethics, roles and responsibilities for AI should be taken into consideration.

- ***Project Ecosystem***

  IT Planners can refer to Section 4.1.3 "Project Ecosystem" of the AI Practice Guide when evaluating the existing technology landscape, business needs and sourcing (procurement) options to identify technology gaps. IT planners who work with the sourcing team and the Project Team should refer to this section for considerations when deploying third-party AI applications. Existing change management procedures are to be followed when making changes to existing systems/AI applications.

For System Analysts, System Architects and Data Scientists

- ***Project Development***

  System Analysts and System Architects (who are responsible for developing data plumbing, which transforms and feeds data to the AI applications) and Data Scientists (who are responsible for performing development of AI applications) can refer to Section 4.1.4 "Project Development" of the AI Practice Guide to ensure aspects such as data validation, documentation, biased data mitigation, data privacy, training, testing and AI modelling techniques are considered for ethical AI.

- ***System Deployment***

  Data Scientists (who are responsible for managing the integration, scaling and deployment of AI applications, managing post-deployment performance and stability of AI applications) can refer to Section 4.1.5 "System Deployment" of the AI Practice Guide for areas specifically related to deployment of AI applications such as integration, testing, feedback loops, tuning metrics and performance checks.

- ***System Operation and Monitoring***

  Data Scientists (who are responsible for monitoring AI systems and ensuring suitable actions are taken to address feedback from the AI applications) can refer to Section 4.1.6 "System Operation and Monitoring" of the AI Practice Guide to ensure actions including escalation, continuous review and compliance checking on internal requirements and external expectations are considered.

3.5.1 Ethical AI Principles

Twelve Ethical AI Principles should be observed for all AI projects. Two out of the twelve principles (1) **Transparency and Explainability** and (2) **Reliability, Robustness and Security** are "Performance Principles". These fundamental principles must be achieved to create a foundation for the execution of other principles. For example, without achieving the Reliability, Robustness and Security principle, it would be impossible to accurately verify that other Ethical AI Principles have always been followed.

The other principles are categorised as "General Principles", including (1) **Fairness,** (2) **Diversity and Inclusion**, (3) **Human Oversight**, (4) **Lawfulness and Compliance**, (5) **Data Privacy**, (6) **Safety**, (7) **Accountability**, (8) **Beneficial AI**, (9) **Cooperation and Openness** and (10) **Sustainability and Just Transition**. They are derived from the United Nations' Universal Declaration of Human Rights and the Hong Kong's relevant ordinances.

Definitions for the principles are listed in Table 4 with further details provided in subsequent subsections.

Principle	Definition
Transparency and Explainability	Organisations should be able to explain decision-making processes of the AI applications to humans in a clear and comprehensible manner.

Principle	Definition
Reliability, Robustness and Security	Like other IT applications, AI applications should be developed such that they will operate reliably over long periods of time using the right models and datasets while ensuring they are both robust (i.e. providing consistent results and capable to handle errors) and remain secure against cyber-attacks as required by the relevant legal and industry frameworks.
Fairness	The recommendation/result from the AI applications should treat individuals within similar groups in a fair manner, without favouritism or discrimination and without causing or resulting in harm.
Diversity and Inclusion	Inclusion and diverse usership through the AI application should be promoted by understanding and respect the interests of all stakeholders impacted.
Human Oversight	The degree of human intervention required as part of AI application's decision-making or operations should be dictated by the level of the perceived severity of ethical issues.
Lawfulness and Compliance	Organisations responsible for an AI application should always act in accordance with the law and regulations and relevant regulatory regimes.

Principle	Definition
Data Privacy	Individuals should have the right to: (a) be informed of the purpose of collection and potential transferees of their personal data and that personal data shall only be collected for a lawful purpose, by using lawful and fair means, and that the amount of personal data collected should not be excessive in relation to the purpose. Please refer to the Data Protection Principles ("**DPP**")**1** "Purpose and Manner of Collection" of the Personal Data (Privacy) Ordinance (the "**PD(P)O**")[3]. (b) be assured that data users take all practicable steps to ensure that personal data is accurate and is not kept longer than is necessary. Please refer to the **DPP2** "Accuracy and Duration of Retention" of the PD(P)O. (c) require that personal data shall only be used for the original purpose of collection and any directly related purposes. Otherwise, express and voluntary consent of the individuals is required. Please refer to the **DPP3** "Use of Personal Data" of the PD(P)O. (d) be assured that data users take all practicable steps to protect the personal data they hold against unauthorised or accidental access, processing, erasure, loss or use. Please refer to the **DPP4** "Security of Personal Data" of the PD(P)O. (e) be provided with information on (i) its policies and practices in relation to personal data, (ii) the kinds of personal data held, and (iii) the main purposes for which the personal data is to be used. Please refer to the **DPP5** "Information to Be Generally Available" of the PD(P)O.
Safety	Throughout their operational lifetime, AI applications should not compromise the physical safety or mental integrity of mankind.
Accountability	Organisations are responsible for the moral implications arising from their AI applications, including their use and misuse. Besides, there should be a clearly identifiable accountable party, be it an individual or an organisational entity.
Beneficial AI	The development of AI should promote the common good.
Cooperation and Openness	A culture of multi-stakeholder open cooperation in the AI ecosystem should be fostered.

[3] https://www.pcpd.org.hk/english/data_privacy_law/ordinance_at_a_Glance/ordinance.html

Principle	Definition
Sustainability and Just Transition	The AI development should ensure that mitigation strategies are in place to manage any potential societal and environmental system impacts.

Table 4: Ethical AI Principles and Definitions

3.5.1.1 Transparency and Explainability

Principle: Organisations should be able to explain decision-making processes of the AI applications to humans in a clear and comprehensible manner.

Explainability

Explainability is the notion of creating human-understandable explanations of an AI model's characteristics (e.g. features and parameters), as well as the decision-making pathways for each individual prediction. Such explanations may be of interest to various stakeholders of AI and should be constantly improved. Related areas include understanding AI, ensuring models are explainable, transparency of processes, reproducibility and trust.

Across both global and local literature, it is recognised there is a need to be able to explain what AI actually does in a clear way. To create robust and trusted AI applications, an understanding of inner functioning of their models, as well as the data used at training and decision-making times is imperative.

Organisations should be interested in understanding the inner workings of AI models to evaluate and establish trust in utilising AI solutions and make more confident decisions; Data scientists should carefully consider whether a model is ready for deployment, or whether it is required to improve communication and collaboration with their peers, system analysts, system architects and managers for enhancing the explainability; and governance bodies may require comprehensive and immutable reports for auditing purposes.

Whatever the case may be, this necessitates an ethical principle that obliges AI developers to consider building explainability capabilities into AI applications. Explainability methods can be used to generate textual or visual explanations of models from two different perspectives, depending on the use case and stakeholders. Please refer to Section 4.1.4.1 "Business and Data Understanding" for further details. For those complicated machine learning algorithms (e.g. neural networks), hyper-parameters of algorithms should be properly documented. For example, for AI applications involving the use of neural networks, hyper-parameters such as number of layers, activation functions and error functions should be properly documented.

Depending on whether the data, algorithms and configurations of an AI model are available at the time of explanation, two approaches can be taken to generate human-readable explanations:

- *Built-in Explanation:* Some models inherently have the ability to explain their behaviour. For example, decision trees are essentially a cascade of questions and work comparably to the way humans think. When executing against a data point (using its features and values), the pathways to reach a decision can be simply reported back to human users. Similarly, linear models like Logistic Regression are fairly intuitive and easy to explain to a non-expert in data science, as the data points can be visualised in a plot against the learned probability line, and Feature Importance metrics can be calculated, and absolute values reported back to data scientists.

- *Post-hoc Explanation:* Post-hoc explanation of an AI model is an effort beyond exploratory data analysis techniques to understand its behaviour after it has finished training, or when the model is in production and only accessible through an Application Programming Interface ("**API**"). While post-hoc explanation is less likely to elucidate the exact inner workings of a model, it can help end-users to approximate the behaviour of the model by way of examples, surrogate models or visualisation of Feature Importance and Partial Dependence plots. Depending on which stakeholder is on the receiving end of the explanation, natural language explanations of a model's behaviour can be generated using techniques from the Natural Language Generation ("**NLG**") domain to create.

Note, there is an inherent trade-off between explainability and accuracy of machine learning algorithms. Nonparametric machine learning algorithms or models with a very large number of parameters (e.g. deep neural networks or ensemble models) can learn complex, latent patterns from large-scale data and offer flexibility and superior performance, at the cost of being 'opaque' in nature. In contrast, parametric models, such as Naive Bayesian or Simple Neural Networks, that can be trained with a set of fixed parameters are easier to explain but often are less performant though this is not always the case. In addition to the type, size and complexity of training data that usually dictate the learning models, the safety-critical nature of the AI application and desired level of explanation should be considered at design time.

Transparency

Transparency is notably the most critical characteristic of building trust into AI models. Trust needs to be developed and strengthened in a gradual manner. It is realised through carefully designing a process to minimise risk, and therefore, plays a crucial role in the widespread adoption of new, disruptive technologies like AI. Trust is hard to come by and builds upon several factors including purpose and performance of the AI applications, as well as the technology provider (e.g. organisation's brand reputation, level of transparency in design, operations, reliability and being able to explain the rationale behind the AI models' decisions are crucial for building and maintaining trusts. The Transparency principle calls for the adoption of a clear, honest communication channel with all relevant stakeholders.

The stochastic nature of AI applications makes them fundamentally different from conventional software development paradigms. In what follows, we iterate possible ways of embedding Transparency into the design of AI applications based on Institute of Electrical and Electronics Engineers ("**IEEE**") standards:

- *Transparency as Traceability:* The transparency of the software engineering process during implementation that should allow technical inspection of which norms have been implemented, explicitly and implicitly, for which contexts, and how possible conflicts are resolved.
- *Transparency as Verifiability:* This concerns how normative reasoning is approached in the implementation and how it is possible to verify that the normative decisions the system makes match the required norms and values using possibly explicit and exact representations with the objective to provide the basis for a range of strong mathematical techniques, such as formal verification. Verifiability is also important in case of failure, in the sense that if an AI application causes harm, it should be possible to ascertain why, and in case of judicial decision-making involving an AI model, it must be possible to provide a satisfactory explanation auditable by a competent human authority[4].
- *Transparency as Non-deception and Honest Design:* Any stakeholder of the global AI community has a strong and shared interest in avoiding AI applications autonomously creating any information or content in any form (text, audio, video) that intentionally represent states of the world that do not correspond partially or entirely to specific states of fact with the intent of deceiving any third-party (human beings or machines). Of course, in certain use cases of AI application deception may be necessary in serving the core functionality of the system (e.g. in case of gaming, entertainment and art applications), but those actions are no longer norm violations because they are justified by context and user consent[5].
- *Transparency as Personal Privacy:* People should have the right to access and correct their own personal data that can be used by AI applications.

A common misconception regarding Transparency is a belief that organisations must reveal their AI-related intellectual property (e.g. source code, models' configurations, comprehensive views of their data sets) to provide for a "transparent AI". On the contrary, the Transparency principle calls for the adoption of a clear, honest communication channel between an organisation and its end-users, as well as regulators when needed, and its indispensable nature. Otherwise, the "for-profit" nature of many AI applications, as well as third-party AI vendors' business models would never allow organisations to adhere to this principle. In fact, having a very detailed transparency level with end-users of an AI application would possibly make it susceptible to being "gamed" by malicious users. At the same time, auditors and regulators may require accurate, exhaustive views into the organisation's AI-driven processes. In such cases, the level of transparency may be decided by the organisation as seen fit.

[4] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf
[5] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Provability

Provability refers to the mathematical certainty behind AI models' decisions. It mandates a higher level of formalism in explaining an AI application's behaviour. This type of explainability is geared more towards data scientists to place an AI model under scrutiny to ensure the decision-making policies of the model can be mathematically proved and remain consistent as changes in data and environment take place. Equitably, in many safety-critical AI applications, where the use of AI must be approved or certified by regulators, provability becomes indispensable.

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- Project Team should aim to embed elements of transparency into the AI applications and translate them into human language (e.g. using decision tree/illustration to explain to users how decisions are made). This allows human to understand whether the decision made by AI has errors and fix the errors.

Provide examples to explain the AI decision-making process in narrative terms or graphics (for example, drawing a workflow with decision trees) whenever possible for a non-technical audience to understand and visualise the AI operations better. Organisations should make explicit how different factors and data determine the outcomes and conclusions of their AI models.

3.5.1.2 Reliability, Robustness and Security

Principle: Like other IT applications, AI applications should be developed such that they will operate reliably over long periods of time using the right models and datasets while ensuring they are both robust (i.e. providing consistent results and capable to handle errors) and remain secure against cyber-attacks as required by the relevant legal and industry frameworks.

The overarching aim of the *Reliability, Robustness and Security* principle is to ensure that AI applications behave as intended, from training data to final output, over prolonged periods of time. Reliability is about increasing the likelihood of the system being fault-free. Robustness is the ability of a model to continue to perform effectively when assumptions and variables change. Security is to protect the data and model itself.

Across both global and local literature, it is widely recognised that there is a need to ensure AI applications behave as intended to maintain reliability, robustness and security. Areas in the literature include awareness of misuse, integrity, robustness, resilience, effectiveness, quality, appropriateness, accuracy and security, with a particular emphasis on preventing harm.

It is an indispensable requirement for AI applications to be designed and developed in a way that takes into consideration that environment, data and processes on which they rely, change over time. Malevolent actors of AI applications can exploit such drifts in the AI operating environment with a variety of techniques to penetrate and *fool* AI models to make incorrect predictions with high confidence. The adoption of the Reliability, Robustness and Security

principle assists organisations to identify potential weaknesses in an AI application, improve overall performance, withstand adversarial attacks, and monitor long term performance of AI models throughout their operational lifetime, and verifiably so where applicable and feasible.

Reliability

Fundamental to creating reliable software applications, Reliability is an engineering effort to maximise the probability that a system will perform its required functions fault-free within a specified time period and the environment[6]. As the complexity of AI algorithms and systems built upon them increases, the role of disciplined software engineering practices, such as relevant standards and software tests, become more prominent to ensure two measures encompassed in reliability:

- *Service Availability:* The degree to which a system or component is operational and accessible when required, often expressed as a probability[7]. In safety-critical or real-time missions, the service availability of AI applications is of high significance and the absence thereof can lead to potentially fatal consequences.

- *Serviceability:* Sometimes also referred to as Maintainability, is the degree of ease with which a system can be maintained or repaired. Changes to the infrastructure where AI systems are running on, the requirements that AI models have to fulfil, as well as issues and bugs that surface over time, require the AI developers to periodically correct defects, analyse their root causes and prevent future failures.

While individual practitioners should uphold quality and integrity by adhering to industry best practices and organisational codes of conduct, it is also important that organisations closely monitor and manage the quality and integrity of the data being used to develop AI applications. 'Data quality' determines the reliability of the information to serve an intended purpose and attributes that define the usability of information, whereas data integrity refers to the reliability of information in terms of its physical and logical validity, which is based on accuracy and consistency of the data across its lifecycle, and calling for no unintended change between successive updates. The key considerations relating to data quality and integrity are consistency, accuracy, validity, timeliness, uniqueness and completeness.

Robustness

Robustness is a characteristic describing a model's ability to effectively perform while its variables or assumptions are altered. In order to ensure a model is robust, validations and error handling must be incorporated at every step of the data science pipeline, from data preparation and ingestion through to prediction.

Robust models must perform consistently while being exposed to new and independent (but similar) datasets and be able to deal with the errors and corner cases that occur at execution time. Robustness also entails placing effective error handling measures in place to protect AI

[6] http://www.mit.jyu.fi/ope/kurssit/TIES462/Materiaalit/IEEE_SoftwareEngGlossary.pdf
[7] http://www.mit.jyu.fi/ope/kurssit/TIES462/Materiaalit/IEEE_SoftwareEngGlossary.pdf

models when exposed to malicious inputs and parameters.

Security

Many of existing security practices conventional to software development efforts are also applicable for AI and machine learning models. As they largely rely on data curated and integrated with public or third-party sources, they must be able to discern between malicious input and benign anomalous data. During security design, extra care should be taken to cleanse, secure and encrypt data, as well as designing access controls to the trained model.

Adversarial attacks, i.e., the act of introducing small, intentional perturbations to data used to compel AI models to make incorrect predictions with high confidence, is one of such security risks. Additionally, even for organisations that do not provide direct access to their AI models, but expose them as application services, it is still necessary to design and implement security measures to prevent attackers from reverse engineering their model.

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- AI systems should be technically robust, and they should be sufficiently protected from any malicious use. Malicious actors can try to deceive AI models with inputs designed to learn about the AI and deceive it. Controls are necessary to ensure that the AI application will not be exposed to potential hackers who will train the AI model to perform tasks it was not intended to perform. Otherwise, when such cases happen, this will lead to liability and reputational risk to concerned organisations using the AI application.
- Continuous monitoring which includes validation, verification of accuracy and maintenance of the model should be performed regularly to improve the security and robustness of AI. This is because hackers usually look for outdated software with security flaws which are more vulnerable to cyberattacks. By constantly updating the AI application, this will minimise such security risks.

3.5.1.3 Fairness

Principle: The recommendation/result from the AI applications should treat individuals within similar groups in a fair manner, without favouritism or discrimination and without causing or resulting in harm.

With regards to Fairness, Article 2 of the Universal Declaration of Human Rights states that "*Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinions, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing or under any other limitation of sovereignty*". It is therefore important for AI applications to limit discrimination or bias relative to the factors reported in the Article. The same principle against

discrimination is included in the Convention on the Elimination of All Forms of Discrimination against Women.

The following is a list of key legal terminology for consideration when understanding Fairness and how it applies to organisations or AI applications implementation:[8]

- **Prejudice:** Prejudice means to injure or harm a person's rights.
- **Discrimination:** Discrimination means the adverse, unfair or detrimental treatment, preference, exclusion or distinction of a person because of the person's race, colour, sex, sexual orientation, age, physical or mental disability, marital status, family or carer's responsibilities, pregnancy, religion, political opinion, national extraction or social origin.
- **Impartiality:** Impartiality means to act or make a decision based on merit and according to the law without bias, influence, preconception or unreasonableness.
- **Positive Discrimination:** Positive Discrimination means the:
  - treatment of a person;
  - taking of an action affecting a person; and/or
  - making of a decision affecting a person because of that person's race, colour, sex, sexual orientation, age, physical or mental disability, marital status, family or carer's responsibilities, pregnancy, religion, political opinion, national extraction or social origin that is done with the intention of achieving:
    - substantive equality;
    - equal enjoyment;
    - equal exercise of human rights; and/or
    - equal exercise of fundamental freedoms

    for that person.

It is paramount to acknowledge that fairness is a social construct. There are many mathematical definitions of fairness in the context of AI, and when we choose one, we inevitably overlook some aspects of the others. In other words, achieving "universal fairness" in every AI decision may not be entirely possible due to inherent limitations and contextual dependencies. However, AI applications can be designed to meet specific fairness goals, thus mitigating some of the perceived unfairness and creating a more responsible system overall. AI models, which learn from data, present a higher risk of unintended biases due to their complexity and data that may contain incomplete representations or over-representation of certain groups. By contrast, traditional IT systems typically follow predefined rules. Therefore, the unique characteristics of AI application requires greater attention to address fairness issues.

<u>Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):</u>

- Training data should be free from any bias characteristics such as sample size disparity (where there is significantly less data for minority groups), selection bias (where certain groups are less likely to be selected), bias in model design and bias in model use and feedback (e.g. by checking for stereotyping certain groups when relying on data and algorithms). If the AI application is biased, the decisions made will show preference towards certain groups of individuals.

[8] Please note that often there is no universal legal meaning for these words/phrases and that particular definitions can differ based on their purpose and the circumstances of that situation. Therefore, in this context the definitions are purposefully broad and relatively simple, taking into account the context in which they are to be used (i.e. a broad-based decision-making framework).

- Ensure integrity of source data obtained to help ensure a fair outcome from the AI application. Data which are invalid or inaccurate, when used to train an AI model, will lead to biased decisions and affected users will be discriminated unintentionally due to the flaw in datasets.

3.5.1.4 Diversity and Inclusion

Principle: Inclusion and diverse usership through the AI application should be promoted by understanding and respecting the interests of all stakeholders impacted.

This principle ensures AI applications should respect local cultural norms, while aiming to avoid conflicts with any general or performance principles. Given AI can be used globally by a diverse range of people, their interpretation of the way an AI application behaves can vary significantly. To achieve Diversity and Inclusion, it is important to involve the largest possible number of AI users representing the broadest variety of cultures, interests, lifestyles and disciplines.

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- AI application developers should always consider people of different backgrounds, races and gender. If such differences were not considered, the AI application might behave differently towards certain groups of individuals causing inconvenience. One such example is AI recruiting tool showing preference towards Caucasians because the tool was developed by Caucasians.

3.5.1.5 Human Oversight

Principle: The degree of human intervention required as part of AI application's decision-making or operations should be dictated by the level of the severity of ethical issues.

AI applications are regarded as autonomous systems to various degrees and they become prevalent in our lives. Under this principle, it is important to maintain human oversight in AI, which is the capability of humans in making choices that is to think and determine an outcome and consequently enact upon it in the context of AI applications.

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- Ensure an appropriate level of human intervention based on multiple factors such as the benefits and risks of the AI application, impacts of the AI decision, operational cost and evolving societal norms and values.

- Controls should be implemented that allow for human intervention in the event of system failure especially when the system failure will have an impact on human safety. One such example is autonomous vehicle where human should be given the option to prevent the vehicle from causing accidents if it failed to detect human on the road.

3.5.1.6 Lawfulness and Compliance

Principle: Organisations responsible for an AI application should always act in accordance with the law and regulations and relevant regulatory regimes.

In all cases, the principles that AI applications must adhere to are all applicable laws, regulations, standards and guidelines. It is therefore indispensable for any organisation dealing with the development or implementation of AI applications to master and apply consistently all relevant laws, regulations, standards and guidelines.

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- AI system developed must be in compliance with regulations and laws.
- Compliance review processes for AI applications should be defined to keep track of regulatory changes and to ensure that the policies and processes are compliant.

3.5.1.7 Data Privacy

Principle: Individuals should have the right to:

(a) be informed of the purpose of collection and potential transferees of their personal data and that personal data shall only be collected for a lawful purpose, by using lawful and fair means, and that the amount of personal data collected should not be excessive in relation to the purpose. Please refer to the Data Protection Principles ("DPP")1 "Purpose and Manner of Collection" of the Personal Data (Privacy) Ordinance (the "PD(P)O").

(b) be assured that data users take all practicable steps to ensure that personal data are accurate and is not kept longer than is necessary. Please refer to the DPP2 "Accuracy and Duration of Retention" of the PD(P)O.

(c) require that personal data shall only be used for the original purpose of collection and any directly related purposes. Otherwise, express and voluntary consent of the individuals is required. Please refer to the DPP3 "Use of Personal Data" of the PD(P)O.

(d) be assured that data users take all practicable steps to protect the personal data they hold against unauthorised or accidental access, processing, erasure, loss or use. Please refer to the DPP4 "Security of Personal Data" of the PD(P)O.

(e) be provided with information on (i) its policies and practices in relation to personal data, (ii) the kinds of personal data held, and (iii) the main purposes for which the personal

data are to be used. Please refer to the DPP5 "Information to Be Generally Available" of the PD(P)O.

Individuals[9] should have the right to expect that organisations will process data that pertains to them in a manner that creates benefits for the individual or for a broader community of people. In cases where the organisations receive most of the benefit, a demonstrable vetting process should determine there is minimal impact on an individual. Individuals should have the right to control data uses that are highly consequential to them. This should be facilitated through an appropriate level and contextual application of consent and access where possible. Where consent is not possible, unsuitable or less impactful, they have the right to know that accountability processes assure the data uses are fair and responsible[10].

Individuals should have the right to be informed with respect to the use of their personally identifiable data. This is particularly important in the case of AI applications where datasets possibly containing personal and sensitive information are used to train machine learning algorithms. Additionally, people need to understand their online digital personas, as well as the way they interact in digital environments, are profoundly different from real life. This helps them better protect their privacy and rights.

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- Communicate clearly how, why, where, and when customer data are used in AI systems.
- Ensure minimal collection and processing of personal data and retention policies are followed which take into account privacy regulations such as PD(P)O and the related PD(P)O guidance notes.

3.5.1.8 Safety

Principle: Throughout their operational lifetime, AI applications should not compromise the physical safety or mental integrity of mankind.

For safety, unintended risks of harm should be minimised inclusive of physical, emotional and environmental safety. With the evolution and improvement of AI application performance in terms of both cognitive capabilities and level of autonomy, the risk of unanticipated or unintended behaviours increases correspondingly. Different and possibly dangerous scenarios could arise in which AI applications attempt to take control over their own reward systems or where the learning system fails with unpredictable consequences.

[9] By individuals we include individuals and groups of individuals
[10] Not all uses of data are suitable for control – e.g. data used for security analysis

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- Where AI tools are used to augment human in decision-making, they should be safe, trustworthy, and aligned with the ethics and preferences of people who are influenced by their actions. If robots are deployed to provide care for the elderly, the robot should not cause physical or mental harm to the elderly.

3.5.1.9 Accountability

Principle: Organisations are responsible for the moral implications arising from their AI applications, including their use and misuse. Besides, there should be a clearly identifiable accountable party, be it an individual or an organisational entity.

The Accountability principle ensures the responsibilities and liability of stakeholders are made clear and that people can be held accountable. This includes ensuring that responsibilities are being fulfilled from planning through to record-keeping.

Accountability is a cornerstone principle in most legislation and/or privacy frameworks. It is necessary to determine the ethical and legal responsibility of each stakeholders of the AI system of their use, misuse and actions.

Accountability is put into action through a comprehensive, end-to-end governance process. Governance typically refers to the collective set of policies, procedures and oversight internally and externally that manages the risk of systems and meets required obligations. This will help management of ethical responsibilities and assist to track and mitigate risks related to big data and AI projects. By introducing an appropriate governance framework, this can balance the need for innovation within the organisation and the need to safeguard the public interest.

3.5.1.10 Beneficial AI

Principle: The development of AI should promote the common good.

AI applications should not cause harm to humanity and should instead positively promote the common good and well-being. Technologies can be created with the best intentions, but without considering well-being factors, can still have dramatic negative consequences on people's mental health, emotions, sense of themselves, their autonomy, their ability to achieve their goals and other dimensions of well-being.

AI applications can be ethical, legal, profitable and safe in their usage and yet not positively contribute to human well-being. Well-being considerations that include psychological, social, economic fairness, and environmental factors could enable a better evaluation of the

technological progress benefits while being able to assess unintended negative consequences of AI to impact and diminish human well-being.

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- Development of AI should promote the progress of society and human civilisation, create smarter working methods and lifestyles, and enhance people's livelihood and welfare. If the AI application does not show any benefits to the human civilisation, people may opt out of using AI application because it does not benefit them.
- AI that does not promote common good may hurt and destroy the well-being of the society. One such example includes the autonomous weapon which will cause great danger to human race.

3.5.1.11 Cooperation and Openness

Principle: A culture of multi-stakeholder open cooperation in the AI ecosystem should be fostered.

Cooperation and Openness are about building trust. It includes different stakeholder collaborating and communicating with end-users and other impacted groups on risks and plans to handle these risks. This principle is emphasised for educating the public about AI to help build trust[11].

The endeavour of developing or implementing AI applications should be an open and collaborative process involving a diverse internal team and engagement with diverse end-user groups within the community. In this regard, organisations should proactively collaborate with any type of stakeholders, ranging from end-user to universities, research centres, governments and professional associations in order to help to mitigate the risk of exclusion and inherent biases within the AI application.

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- Organisations should actively seek to develop and enhance AI application cross domain, cross sector and cross organisation. If there is no cooperation between different teams, the AI model may make biased decisions due to lack of diverse opinions/user testing. Project Managers can work together with the Project Team as well as end-users to perform testing before deployment.

[11] https://www.pwc.ch/en/publications/2017/pwc_responsible_artificial_intelligence_2017_en.pdf

3.5.1.12 Sustainability and Just Transition

Principle: The AI development should ensure that mitigation strategies are in place to manage any potential societal and environmental system impacts.

AI can help transform traditional sectors to address societal and environmental challenges as well as bolster human well-being. Although AI presents transformative opportunities to address some of the societal and environmental challenges, left unguided, it also has the capability to accelerate the society and environment's degradation. The deployment of AI applications potentially carries profound societal and environmental impacts as in the case of worker displacement or posing challenges and changes in the areas of caring for the elderly, sick and disabled, education and planet preservation, etc. To develop a sustainable AI, the ultimate goal is to ensure that it becomes values-aligned with humanity, promising safe application of technology for humankind. In practice, this means checks and balances developed to ensure that evolving AI applications remain sustainable. This means that there is a need to guarantee that AI applications (with possible societal and environmental impacts) are implemented with an appropriate mitigation plan in a way that is as fair and inclusive as possible to everyone concerned and leaving no one behind.

Examples (please refer to Section 4 "AI Practice Guide" for further examples and practices):

- AI system can be used to contribute to lesser carbon footprint. For example, AI technology can be used to optimise power utilisation at data centres and help save energy.
- AI supporting infrastructure should be scalable to sustain for long-term enhancements. Otherwise, if the AI applications were developed on infrastructure that could not be scaled-up, this may lead to limitations in the future for fine-tuning of the AI model (and hence the AI applications) in a sustainable way.

3.5.2 AI Governance

AI governance refers to the practices and direction by which AI projects and applications are managed and controlled. The following are the important elements associated with the acceptance and success of developing and maintaining AI applications:

- Establishing governance structure to oversee the implementation of AI projects and AI Assessment.
- Defining roles and responsibilities that affect the use and maintenance of the Ethical AI Framework. Please refer to Section 3.5.2.1 "AI Governance Structure" for details.
- Specifying a set of practices to guide and support organisations to plan, develop, deploy and monitor AI applications. Please refer to Section 4 "AI Practice Guide" for details.
- Assessing the adoption of those practices in terms of application impact. Please refer to Section 5 "AI Assessment" for details.

Effective AI governance should not be a purely technology-led effort as it only solves for concerns that the technical stakeholders may have; it does nothing to assuage concerns posed by the public and requirements for end-to-end governance that integrates with the second and third lines of defence. The three lines of defence is a well-established governance concept in many organisations.

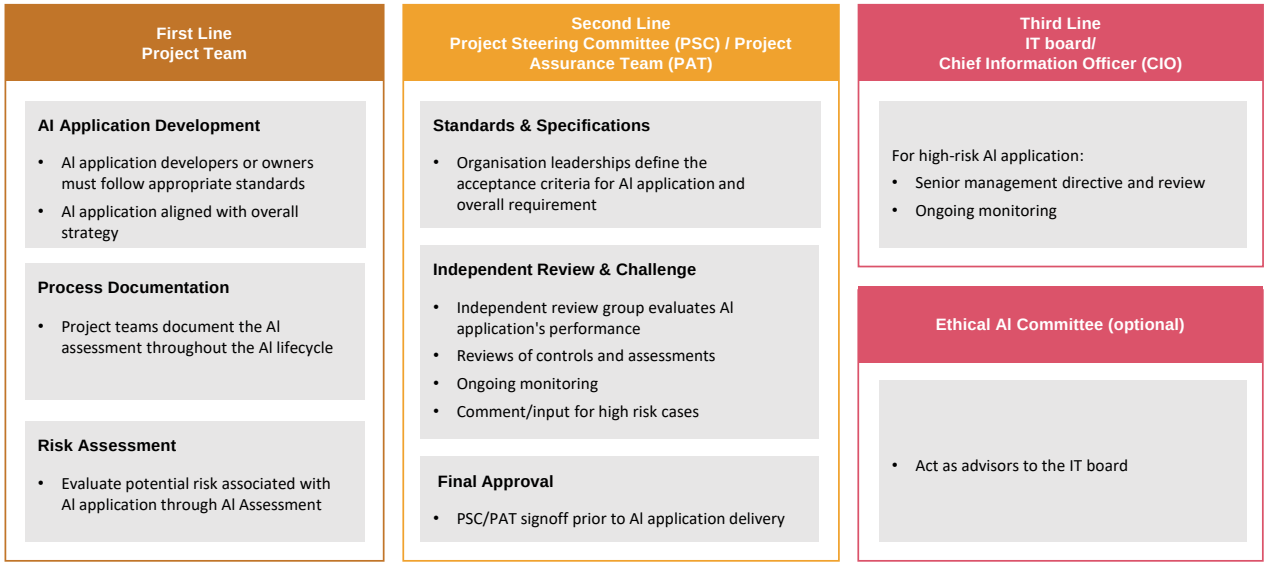


Figure 9: Lines of Defence Model

Figure 9 shows the different defence lines and their roles. The governance structure is a board structure with the following setup.

- The **first line of defence** is the Project Team who is responsible for AI application development, risk evaluation, execution of actions to mitigate identified risks and documentation of AI Assessment.
- The **second line of defence** comprises of the Project Steering Committee ("**PSC**") and Project Assurance Team ("**PAT**") who are responsible for ensuring project quality, defining acceptance criteria for AI applications, providing independent review and approving AI applications. The Ethical AI Principles should be addressed through the use of AI Assessment before approval of the AI application.
- The **third line of defence** involves the **IT Board**, or the Chief Information Officer ("**CIO**") if the IT Board is not in place, and is optionally supported by an Ethical AI Committee, which may consist of external advisors. The purpose of the Ethical AI Committee is to provide advice on ethical AI and strengthen organisations' existing competency on AI adoption. The third line of defence is responsible for reviewing, advising and monitoring of high-risk AI applications.

3.5.2.1 AI Governance Structure

The AI Governance Structure describes the key activities for different roles/functions and defines their corresponding responsibilities.

Organisations make reference to the Ethical AI Framework to plan, implement, and maintain their AI applications. The Project Team can be sourced from existing staff who are familiar with the organisations' business, IT project management and AI development process. The Project Team can involve System Analysts, System Architects and Data Scientists, etc., who execute the development of AI models or applications; manage deployment and post deployment performance. The size of the Project Team will depend on the organisational structure, operating model and scope of AI application being developed by the organisation.

Roles and responsibilities for the AI Governance Structure are listed in the table below.

Roles	Responsibilities
IT Board, or CIO if the IT Board is not in place	The IT Board is responsible for overseeing an organisation's IT applications, including AI applications and reviewing AI Application Impact Assessment (please refer to Section 5 "AI Assessment" for details) for high-risk AI applications before commencement of projects. The Terms of Reference ("**TOR**") for the IT Board outlines high-level responsibilities including steering the use of IT in the organisation, integrating the Information Strategy direction with the business objectives and ensuring that IT initiatives support the organisation's direction and policies. Any AI projects that trigger one of the risk gating criteria as mentioned in Section 5.1 "AI Application Impact Assessment" and Appendix C "AI Application Impact Assessment Template" within an AI Application Impact Assessment are defined as high-risk AI applications and would require the IT Board's approval. The Project Manager is responsible for verification of answers to the risk gating criteria to determine if the AI application is of high risk with support from the Project Team as needed. The definition of high-risk AI applications can be further refined by the IT Board as the experience and AI capabilities of an organisation increases. The IT Board has the responsibilities to: • Review AI Application Impact Assessment for high-risk AI applications; and • Review the recommendations and remediation actions provided by the Project Team based on the reviewed AI Application Impact Assessment and provide comments to the PSC/PAT and the Project Team to ensure appropriate considerations of risks, ethical implications and benefits.
Ethical AI Committee (optional)	The Ethical AI Committee is responsible for advising the IT Board/CIO on ethical issues related to AI applications and AI Assessment. The Ethical AI Committee may consist of external advisors with expertise in technical aspects (e.g. Data, AI Analytics), ethics, legal, risk/benefits

Roles	Responsibilities
	assessment and security. Committee members should be well versed in assessing ethical type issues for AI projects and be aware of the criteria for approving an AI project. For example: • Risks are determined to be reasonable and have been mitigated in relation to the anticipated benefits that may reasonably be expected to result; • Risks to the population making up the data subjects (e.g. children, prisoners, educationally disadvantaged persons, mentally disabled, as well as other vulnerable groups) are considered; • The ethical (permissible) basis for the collection and use of the personal data are appropriately documented (e.g. it is within the scope of expected use, consent was obtained); • There are adequate provisions to protect the privacy of individuals involved in the project; and • The related AI Assessment and decisions reached by the Project Team and the IT Board/CIO. The Ethical AI Committee has the responsibilities to: • Provide recommendations and act as advisors towards the development of ethical AI to support organisations with high-risk AI applications; • Advise on the organisation's decisions and AI Assessment to moderate and increase transparency around the use of AI applications; and • Bring oversight and external knowledge to assist organisations when trying to provide trust and transparency over the AI application for the public. The number of reviewers for an AI Application Impact Assessment review is dependent on the nature and complexity of the AI projects. Reviewers can include specialists in: • Privacy; • Data security/information security; • Engineering/technology/human factors (i.e. the psychological and physiological principles to engineering and design of products, processes and systems); • Data analytics/data science; • Legal; • Public relations; and

Roles	Responsibilities
	• General organisation expertise/knowledge of best practices. An Ethical AI Committee can be used when a project is sufficiently large, high-impact and high-profile that its ethical values may be challenged and therefore there is a need for an independent review. Organisations can assess this need based on the potential negative impacts of AI applications when planning for AI application projects. The impact section of the AI Application Impact Assessment can be used to document this impact. The objective of the Ethical AI committee is to provide some measure of assurance that the activity is ethical, appropriate and defensible under public scrutiny. Examples of where an Ethical AI Committee could be used include health care services which have significant impacts on the welfare of patients, and AI credit scoring systems where sensitive personally identifiable information ("**PII**") is processed in large scale. It should be noted that compositions may vary depending on the needs from the specific organisation and the industry that the organisation relates to on a case-by-case basis. This also includes the channels for the appointment. Organisations may follow their existing practices of setting up expert groups or advisory committees, in which external experts or industry representatives are invited to provide advice to organisations.
Project Steering Committee ("PSC") **Project Assurance Team ("PAT")**	PSC/PAT have the responsibilities to: • Define acceptance criteria for AI applications and overall requirements; • Review AI Assessment to ensure impacts of the AI application are managed; • Perform ongoing monitoring throughout the AI Lifecycle; • Provide comments for high-risk AI applications; • Provide signoff prior to AI application delivery; • Communicate with the Office of the Privacy Commissioner for Personal Data ("**PCPD**") for high-risk AI projects that have potential data privacy issues as appropriate; and • Seek advice from your legal department or lawyers for high-risk AI projects that have potential legal issues as appropriate. Project approval is always required through existing project management structures.
Project Team	The Project Team is responsible for the delivery of AI projects and AI Assessment. Members of the Project Team can be internal or contractor resources who are assigned to complete the project tasks as directed by

Roles	Responsibilities
	a Project Manager. Examples of key roles for the Project Team include System Analysts, System Architects and Data Scientists. Business users can also be included. The Project Team has the responsibilities to: • Assist the Project Management to ensure that the AI application complies with the quality standards throughout the AI Lifecycle (e.g. AI and data ethics, quality procedures, industry standards and government regulations); • Complete AI Assessment and deliver AI applications. Each AI application being developed by the Project Team should have an AI Application Impact Assessment completed; • Recommend AI projects and provide AI Application Impact Assessment (via the Project Manager) for the PSC/PAT's review; • Develop organisation specific AI standards and guidelines if necessary, leveraging the Ethical AI Framework; • Communicate with the PCPD for high-risk AI projects that have potential data privacy issues as appropriate; and • Seek advice from your legal department or lawyers for high-risk AI projects that have potential legal issues.
Project Manager **(Project Team)**	The Project Manager has the responsibilities to: • Ensure that the AI application complies with the quality standards throughout the AI Lifecycle (e.g. AI and data ethics, quality procedures, industry standards and government regulations); • Ensure that reports/checks/assessments are performed on the AI Project including any data governance checks.; • Qualify the use case, developing the end-to-end vision and subsequent design of the AI application; • Ensure that relevant trainings on AI are conducted to upskill existing staff in relation to the AI model; and • Provide administrative support to the IT Board/CIO in arranging meetings, preparing minutes, drafting documents and deliverables, circulating materials to respective parties for comment and approval, triaging inquiries and coordinating with different stakeholders over AI initiatives. Even for Proof of Concepts ("**POCs**") projects, a Project Manager with similar responsibilities should be assigned.
System Analysts, System Architects (Project Team)	System Analysts and System Architects have the responsibilities to: • Assign and track ownership of data sets used in AI models;

Roles	Responsibilities
	• Ensuring licences for any purchased data are in place; • Handle Extract, Transform and Load ("**ETL**") activities that prepare and transform the data for Data Scientists; and • Ensure existing processes such as archival, backup, security, privacy and retention are adhered to.
Data Scientists (Project Team)	Data Scientists have the responsibilities to: • Determine the capabilities of the AI application and training procedures to suffice that capability defined. This includes POCs; • Follow controls and procedures defined for model testing and validation; • Notify project management who subsequently notifies the IT Board/CIO of changes to systems or infrastructure that impact governance and control of AI applications; • Manage, integrate, scale and deploy AI applications; • Transfer AI applications to production code and perform model training at scale; • Mange post deployment performance and stability of AI applications; and • Manage infrastructure and platforms for AI application development, training, deployment, monitoring and testing/validation as well as managing, monitoring and troubleshooting AI applications.

3.5.3 AI Lifecycle

In order to structure the practices for organisations to follow when executing AI projects/creating AI applications, practices in different stages of the AI Lifecycle will be described in Section 4.

A way to conceptualise the AI Lifecycle appears in the following 6-step schematic.

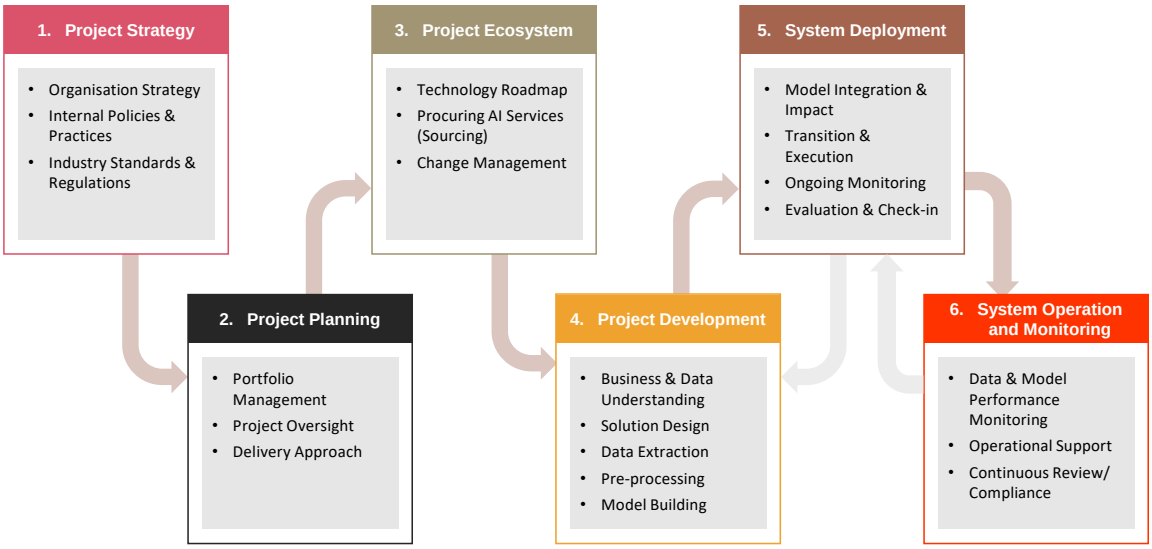


Figure 10: Overview of the AI Lifecycle

The AI Lifecycle shows the different stages involved in AI projects that can

- Guide organisations to understand the different stages and requirements involved; and
- Serve as a reference for the development of AI practices to align to actual stages of how AI is typically developed.

From this AI Lifecycle, key competencies and capabilities can be implemented that serve as the ingredients of an effective, comprehensive programmatic governance system. Further descriptions of the corresponding coverage and capabilities for each AI Lifecycle stage are provided below.

1. Project Strategy

- The AI and Data strategy for individual organisations should be formulated with alignment to its strategic goals and the Ethical AI Principles. Such strategy should be documented and effectively communicated within the organisations;
- The key leadership roles accountable for the implementation of the integrated strategy and key roles who have responsibility for execution should be formally assigned. Subsequently, a comprehensive and effective process should be established to routinely monitor and analyse external changes in AI and data use expectations. These changes should be analysed and used to routinely adopt the organisation's strategy, policies and governance;
- The organisational strategy, policies and procedures should be reviewed and ensure that they have no conflicts with the Ethical AI Principles; and
- The overall AI application governance process includes a well-embedded assessment process to ensure all policies are effectively applied and the application should be evaluated against benefits, impacts and risks with respect to all stakeholders.

2. Project Planning

- A defined AI portfolio management process should be established to ensure alignment around strategic business goals. These are prioritised against short-term and long-term goals and include achieving a positive impact on society;
- Key governance processes should exist with executive responsibility and oversight to ensure risk and ethical based decisions guide on how the portfolio of applications is applied;
- A defined project management structure should be in place to ensure the portfolio of AI applications can facilitate scaling and adoption;
- Key responsibilities should be established with board or senior level sponsorship;
- Management processes should be in place to ensure effective data governance throughout the AI portfolio;
- Clear processes should be in place to establish that all AI initiatives are assessed as to their negative impact on individuals, groups or society and are evaluated against a broad range of benefits and risks. Public expectations should be routinely evaluated for change. The assessment process ensures auditability, including traceability and logging of the AI system's processes and outcomes; and
- A senior level decision making body/group should be formally established to address AI applications of higher risks.

3. Project Ecosystem

- A defined process or procedure with clear responsibilities should be defined to account for possible future technology requirements, necessary model updates, etc. This includes a formalised process to evaluate the existing technology landscape, needs and sourcing options to identify gaps and to adjust the AI roadmap as required;
- Formalised responsibility and processes should be established to evaluate and ensure all related staff are equipped with the skills and knowledge they need to take on the goals and responsibility for AI objectives;
- The sourcing team should be routinely evaluated and allocated the right expertise to perform the change management delivery, training and transition to business as usual; and
- Third-party vendor tools, data and techniques should be evaluated to ensure alignment with Ethical AI Principles, data use and/or AI governance.

4. Project Development

- A defined and programmatic project management plan and system, aligned with the organisation's ethical values, should be developed to address all required organisational processes, functional and non-functional requirements, business value alignment and risk

and testing assessment as part of development process for all new AI initiatives. This includes formal integration to the enterprise master data management requirements;

- The integrated development of data, analytics/AI, automation/software with Ethical AI Principles should be embedded across all the organisational dimensions during design and development with appropriate validation and verification;
- AI application and data suitability should be matched to the business objectives and technology required;
- Where third-party data or technology is used, all required organisational process, risk and testing requirements should be formally assessed. Concurrently, requirements should be defined, and project management should incorporate strategy;
- Plans and/or systems, or a set of procedures should exist to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design;
- Formalised processes should be established to test and monitor for potential biases during the development, deployment and operation of the AI application;
- The AI application should be evaluated for reliability, model sensitivity and model performance against the formalised selected definition of fairness; and
- Bias trade-offs with respect to performance, trade-offs between explainability and performance should be routinely evaluated.

5. System Deployment

- A formalised process with assigned responsibility should be in place to ensure all the AI applications are assessed across all dimensions for impact on all stakeholders at deployment. This includes assessing whether an appropriate balance of benefits and mitigated risks supports the AI application and data processing activity, achieves a goal of ethical AI and that effective mitigating controls are established to reduce risk;
- A formalised decision-making structure and risk-based escalation path should be established to resolve issues, assess the usage of high-risk data and make decisions as well as formally approve the deployment of AI applications;
- A defined project management system should be established to assess and develop a comprehensible plan to roll out AI applications and management systems and processes should be in place to continuously identify, review and mitigate risks of using the identified AI applications post deployment; and
- The project management system and process should ensure the project continues to add value to the services, benefits are measured, have been defined and are actively tracked and reported on.

6. System Operation and Monitoring

- A defined systematic monitoring capability, including customer redress processes should be established to capture issues/incidents related to AI applications;

- An escalation process should be established to escalate significant issues/incidents related to AI applications to the IT Board/CIO for their awareness and action; and
- Clear responsibilities and service level agreements should be defined for all AI application support groups coupled with continuous learning (both machine learning and human/organisational learning) which have been built to ensure integration with other processes such as security event management.

The AI Lifecycle aligns with a traditional System Development Lifecycle ("**SDLC**") model as depicted in Figure 11.

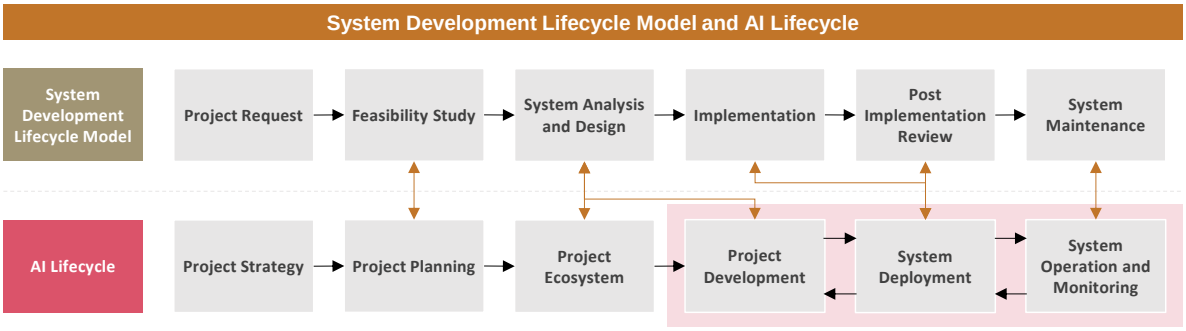


Figure 11: AI Lifecycle Aligned with a System Development Lifecycle Model

In a typical development process of an AI application, great emphasis is placed on data because good data are often required for creating a good AI model. Data sourcing and preparation which is part of the project development can be a continuous exercise. This is because an AI model can often benefit from better or more data for iterative model training throughout the development process. The approach of conventional software lifecycle is to program the IT application with a set of instructions for a pre-defined set of events. Thereafter, the IT application will exploit its computing capabilities and other resources to process the data fed into the system. This is different from an AI application where a huge amount of data are fed into the application, which in turn processes all the data resulting in a trained model or AI solution. This trained model is then used to solve new problems.

There is often a continual feedback loop between the development and deployment stages, as well as system operation and monitoring of the AI Lifecycle for iterative improvements making this distinct from a traditional system development lifecycle.

SECTION 4

AI PRACTICE GUIDE

4. AI PRACTICE GUIDE

Organisations can refer to this AI Practice Guide throughout various project stages of AI Lifecycle starting from strategy, planning, ecosystem, development, deployment and ongoing monitoring of AI applications. Organisations should continue to follow the related existing ordinances, policies, and guidelines for IT projects which are relevant to their respective business domain or industry and current practices. Practices listed in the AI practice guide are additional, ethical AI related guidance that organisations are advised to adopt.

This section describes good practices for organisations to consider and adopt as they progress along the AI Lifecycle when developing AI applications. Organisations can follow the AI Practice Guide at an early stage when conducting the planning stage of the AI Lifecycle for their projects. Gradually, they can then use the AI Practice Guide for other AI Lifecycle areas based on organisational needs and progress with AI.

The AI Practice Guide contains sections that require technical data and AI knowledge. The Project Team Manager should work with the Project Team to ensure that the appropriate data analysis has been done and evidenced in assessment. Training for Project Teams involved in AI is fundamental to ensure that these teams can achieve AI capabilities.

Examples used in the AI Practice Guide are taken from different sources and are purely for illustrative purposes only. They do not imply that the specific examples have to be followed by organisations.

Note: Roles mentioned in Section 3.5.2 “AI Governance” are related to the AI Governance structure, while the intended audience of the Ethical AI Framework (as listed under Section 2 “Purpose”) includes a larger group of audience. Other roles such as procurement team exist and are referenced in the AI Practice Guide as they will be involved in certain practices/regulations (e.g. organisational procurement practices would be involved when a project team is looking to purchase AI solutions) however these existing roles would not be part of the AI Governance structure and hence are not included in Section 3.5.2 “AI Governance”.

4.1 AI LIFECYCLE AND PRACTICES

The AI Lifecycle stages (please refer to Section 3.5.3 “AI Lifecycle”) are used to group the different practices with examples. Subject to the situation and context of individual organisations, each organisation can decide the action party among the suggested candidates.

4.1.1 Project Strategy

4.1.1.1 Organisation Strategy, Internal Policies and Practices

Definition:

An organisation strategy for deployment of AI projects should be aligned with any existing organisation goals and IT strategy. The Ethical AI Principles should be considered or included in formulating the organisational strategy and its decision-making process.

Who are involved at this stage?

IT Planners/Executives, Business Users

Practices and examples:

Practice	Practice detail
Form an AI strategy or plan with a review performed	• Organisations should formulate an AI strategy to ensure that the AI outcome conforms to the organisations' vision and mission. This can be accomplished by explaining the intended outcomes of the AI development and ways of achieving this. Examples of key components for an AI strategy include a vision, objectives, roadmap, and alignment of AI goals with organisation goals, and considerations for governance. The governance roles, people and technology resources for deployment of AI applications constitute the AI governance. • The AI strategy can be documented using standard strategy document formats in line with existing organisational strategy documents. • The timeframe for completion of an AI strategy would be subject to the timeframe for planning and developing AI applications. There is no set timeframe for the strategy and each organisation would complete this as they need it for their AI plan when following the Ethical AI Framework. AI strategy could be included in organisational IT plan. • The committee or group forming and reviewing the strategy can have a diverse team including the public relations, risk, legal, technology, human resources and other teams as deemed necessary. • Organisations should include the Ethical AI Principles in their AI strategy to ensure there is an awareness for all stakeholders and to show endorsement at an organisational level. • If high-risk AI applications are anticipated to deploy, an IT Board/CIO review is required prior to the initiation and production deployment of the project. An Ethical AI Committee can be set up if external advisors are needed to enhance the competency of organisations in performing oversight/monitoring on high-risk AI applications or use cases.

Practice	Practice detail
	High-risk AI applications should be identified using the risk gating criteria outlined in the AI Application Impact Assessment. Please refer to Section 3.5.2 "AI Governance" and Section 5 "AI Assessment" for details on the governance structure and review. • A current-state assessment of the organisations' AI readiness or maturity can be documented across various functions and is completed to ensure there is an achievable roadmap for organisational goals and investments. Gaps in process, people, systems and data may be considered when formulating the organisational strategy. • The AI strategy (including values and goals for the ethical use of AI) for the organisation should have a documented review including any key considerations for the organisation. One example of an AI strategy goal for an organisation would be having AI transformation to relieve manual workloads and increase the utilisation of AI skills and resources. Key considerations to achieve this goal should include the identification of areas which AI can be leveraged, the benefits of the AI application towards end-users, success rates or other measurable goals of AI application adoption and expected outcomes from the AI applications. • A strategic assessment as part of setting up the AI strategy (e.g. performance of SWOT analysis) can be performed as part of the IT planning process. This can be performed by an organisational project team, IT team or via the use of external consultants. • The AI strategy should be well communicated (e.g. through internal email) and understood across the organisation. • Refer to Appendix D "AI Strategy Template" for an illustrative template. Organisations can use other formats and templates for their AI strategy.

Form an organisational view on AI solutions	• Organisations should complete the AI Application Impact Assessment process (please refer to Section 5 "AI Assessment" for details) to identify, vet and resource new ideas supported by AI applications (e.g. facial recognition, video analytics).
	  ○ The AI application should be assessed to identify the benefits to organisation and impacts on individuals or society; and
	  ○ Negative impacts or risks which should include the sensitivity of information required, potential social discrimination towards people, impacts on the environment and potential costs from a legal perspective should be identified. The strategy and priorities should be aligned, with the benefits and risks considered. This should be standardised for identification and approval of use cases.
	• The Project Team should periodically scrutinise the academic and business landscape to identify potential applications or research of AI that can be applied to achieve business objectives.

Practice	Practice detail
	Organisations need to carry out market research areas specific to them (e.g. healthcare AI applications, law enforcement). • Market research should include an assessment of the opportunities and risks associated with the use of AI. This assessment should contain any existing examples (where available) to identify benefits, risks and lessons learned from others who have implemented similar projects to ensure that these lessons are considered when the organisation is assessing potential AI applications.
Develop and establish metrics	• Organisations can have quantifiable or qualitative metrics to measure the impact of the AI applications (e.g. cost savings and reduction of man-days achieved or the impact on organisations' reputation based on the public's perception with either positive or negative sentiments on social media and the ability to respond to unexpected disaster events). • For organisations that are using advanced AI applications, organisations should establish metrics to measure feedback, performance and prioritise AI applications. Examples of metrics include inspecting the success and failure of predictions to verify that the AI application is performing in an acceptable manner. Please refer to Section 4.1.6.1 "Data and Model Performance Monitoring" for further examples of metrics that can be established.
Adhere to the right internal policies	• Internal policies and practices (including any AI guidelines) should be easily accessible within the organisation to create awareness and ensure that the AI and Project Teams fully comprehend the policies and practices. • Organisations should disseminate internal policies and practices to remind all staff to exercise the policies and practices in their projects. • Organisations should ensure that there are processes and reviews to ascertain that the Project Team is using AI applications in a manner consistent with the organisation's core values, legal requirements and/or societal expectations (e.g. through completion of AI Application Impact Assessment). This is to ensure that the AI projects' execution is in accordance with internal policies and practices.
Update policies	• The organisational strategy, policies and procedures should be reviewed and ensure they comply with the Ethical AI Principles (refer to Section 3.5.1 "Ethical AI Principles"). • The translation of standards into IT standards can be performed by IT team and is important to ensure organisations are aware of compliance requirements. Periodic policy reviews should be performed by IT team to evaluate whether new IT standards are required. Examples of industry standards under development are listed in Appendix B "Examples of Relevant Industry Standards".

Practice	Practice detail
	• Organisations should assess the coverage of internal policies and identify compliance gaps or areas of concern regarding data ethics and AI in existing policies. For example, organisations should evaluate whether local standards or regulations such as data protection and privacy law have been integrated. • The AI Application Impact Assessment (please refer to Section 5 "AI Assessment" for details) should be utilised to identify whether the AI application might harm the rights of individuals or groups, and ensure that the AI application is evaluated against a broad range of benefits and risks. This is to identify any developing social or cultural norms that can impact internal policies and procedures relevant to AI.
Equip staff with necessary skills and knowledge	• Provide adequate training to relevant personnel to equip them with the necessary knowledge, skills, and awareness to work effectively under the organisations' AI strategy. The training should cover a range of areas, including: o AI governance and compliance: provide training courses covering topics such as bias and fairness, explainability, accountability, security, privacy, and other relevant aspects of AI governance, ensuring compliance with applicable laws, regulations, and policies, and fostering awareness of importance of fact-checking to maintain accuracy and reliability in AI systems; o General AI technology: provide training courses covering practical applications, common AI use cases, and limitations, etc., as and where necessary; o Use of AI Tools: appropriate training should be provided regarding AI tools developed or adopted by organisations; and o IT security: provide training courses covering topics like data poisoning, adversarial attacks, and other aspects of IT security related to AI systems. • Review and update training programs regularly to ensure that they remain relevant to the latest advancements in best practices and the rapidly evolving field of AI technology.
Example: To adhere to the principle "Cooperation and Openness", an organisation may formulate a strategy of data sharing. By leveraging the central data repositories which have data sharing features built-in, the organisation can conduct analytics and AI application development which fulfil this strategy.	

Related Ethical AI Principles:

1. All Ethical AI Principles should be considered and referenced as part of an organisation strategy for AI. This can evidence endorsement of the Ethical AI Principles by the organisation.

2. Accountability – Roles and responsibilities instituted in policies determine and demonstrate who is accountable for different areas of an AI project.
3. Data Privacy – Ensuring that there is a data privacy internal policy being observed for AI projects.

4.1.1.2 Industry Standards and Regulations

Definition:

Relevant regulations and standards require an assessment to ensure that AI and related processes adhere to any relevant laws or standards.

Who are involved at this stage?

IT Planners/Executives

Practices and examples:

Practice	Practice detail
Perform continual assessment of standards/ regulations	• Development of international standards for AI is currently underway by various standards bodies and professional associations such as the International Organisation for Standardisation ("ISO") and IEEE. These standards are expected to serve as references once finalised. Please refer to Section B.1 of "Appendix B - Reference" for details. The IT team plays a supportive role in studying and incorporating elements of these standards, as necessary.
Embed standards or regulations with strategy	• Organisations should address relevant standards related to data ethics and AI from an operational perspective by embedding these in the organisation strategy and ensuring requirements from standards and regulations are considered and fulfilled in AI projects. • A good standard or regulatory compliance strategy should take a top-down approach. For example, the CIOs, who are responsible for overseeing ICT development in their organisations, and the PSC, should set clear expectations on standards that the Project Team should comply (e.g. PD(P)O) with when executing AI projects and emphasise the impact of being incompliant. This is to ascertain that the Project Team understands the importance of industry standards and regulations.
Example: An AI Application Impact Assessment can be used to assure standards and regulations are being considered. The AI Assessment will assist organisations to develop and deploy AI applications ethically because the assessment questions will allow organisations to assess their compliance towards all the Ethical AI Principles outlined.	

Related Ethical AI Principles:

1. Accountability – Standards and regulations should necessitate organisations and individuals to be accountable.
2. Data Privacy – To ensure personal data are collected on a fully informed basis and in fair manner, with due considerations towards minimising the amount of data collected, the PD(P)O and related Data Protection Principles are required to comply with. This includes ensuring that the use of personal data in AI shall be the same as or directly related to the original collection purpose of the personal data.
3. Lawfulness and Compliance – Ensuring industry standards and regulations are considered in the project so that the organisation and AI application users are aware and can comply with.

4.1.2 Project Planning

4.1.2.1 Portfolio Management

Definition:

A portfolio is defined as a collection of projects. Portfolio management is performed to ensure that the individual IT investments embedded in the organisation's processes, people and technology are on course.

Who are involved at this stage?

IT Planners/Executives, Business Users

Practices and examples:

Practice	Practice detail
Create an inventory of existing AI projects	• An AI projects inventory of existing, potential and Proof of Concepts ("**POC**") being planned by the organisation should be created. • A POC is an experimental activity intended to demonstrate the feasibility, benefits and risk that the AI may introduce for evaluation of the appropriateness of the solution towards the real-world problem. As such, a POC is usually a smaller scale activity served as a test example to support the case for a larger project.
Map AI projects to the organisation objectives	• Categorise the AI projects based on the status of each project and determine whether these projects support the current organisation's strategy and objectives.
Select and prioritise AI projects	• Organisations should prioritise AI projects that have a favourable balance of risks and benefits, that bring positive impacts, high probability of success and should assist the organisation to achieve

	its objectives. The AI Application Impact Assessment in Section 5 "AI Assessment" can assist organisations to document and determine which AI projects to prioritise. Identify whether some existing IT projects can be combined or enhanced through the adoption of big data and AI and subsequently merit being implemented for the benefit of the organisation and end-users. • Define the intended scope of the AI project. This is important to identify and choose a suitable AI application that would fit the intended scope. • Assess the long-term and potential impacts of applying AI technology in the intended scope to avoid technology abuse.
Example: Implementing an AI application such as a chatbot on websites could be mapped to an organisation's objective to improve services. Organisations should map each AI project to its objective and subsequently select the AI project to prioritise. Organisations can prioritise AI projects that will benefit themselves and end-users to adhere to the "Beneficial AI" principle.	

Related Ethical AI Principles:

1. Beneficial AI – Projects that promote common good can be prioritised with higher priority.
2. Fairness – Potential impacts on groups can be factored in at project level based on project goals.
3. Diversity and Inclusion – Stakeholders impacted can be mapped across the range of AI projects being performed.

4.1.2.2 Project Oversight and Delivery Approach

Definition:

It is important to decide whether an AI project will need a review from the IT Board/CIO at the outset of the project. Each project has its characteristics and features as well as business values. Roles and responsibilities are required to operationalise AI and ensure accountability. Please refer to Section 3.5.2 "AI Governance" for details. The Project Manager should define the quality standard, quality control and assurance activities and the acceptance criteria for major deliverables of the project in the Quality Management Plan. Quality control not only monitors the quality of deliverables, but it also involves monitoring various aspects of the project as defined in the Project Management Plan ("**PMP**") to ensure that the AI application complies with the quality standards throughout the AI Lifecycle (e.g. AI and data ethics, quality procedures, industry standards and government regulations).

Who are involved at this stage?

IT Planners/Executives, Business Users

Practices and examples:

Practice	Practice detail

Assess the general criteria that qualify an AI project for a review by the IT Board/CIO	Review and approval from the IT Board/CIO are required if the AI application meets any one of the following criteria. The criteria are based on General Data Protection Regulation ("**GDPR**") guidance for identifying situations "likely to result in high risk". These situations will evolve over time and organisations may tailor these criteria based on their own experience with AI applications and their needs for IT Board/CIO's review. Please refer to Section 5 "AI Assessment" for more details on the risk gating criteria.
	
	• The AI application has a high degree of autonomy. This means the AI application is highly capable of accomplishing tasks and interacting with its surrounding objects or humans.
	• The AI application is used in a complex environment. This means that many variables from the environment are being considered throughout the execution of the AI application. For example, using the AI application in an open environment such as a public park where moving surrounding objects and other living things have to be factored in by the AI applications.
	• Sensitive personal data are used in the AI application. This includes any information that could be used to relate and identify a living person. Examples of personal data with high sensitivity include identity card numbers, bank account data and health data.
	• Personal data are processed on a large scale and/or are combined data sets taking into account:
	  ○ The number of individuals concerned, either as a specific number or as a proportion of the relevant population;
	  ○ The volume of data and/or the range of different data items being processed;
	  ○ The duration, or permanence, of the data processing activity; and
	  ○ The geographical extent of the processing activity.
	• The AI application can result in a potentially sensitive impact on human beings (e.g. to make decisions about individuals or determine their emotions).
	• The AI application involves the evaluation or scoring of individuals. This can be done by profiling and predicting a living person, based on aspects such as the data subject's performance at work, economic situation, health, personal preference, reliability, behaviour and location.
	• The AI application makes automated decisions/complex decisions that have a significant impact on persons or entities that have legal consequences for them where the decisions were independently made without human intervention.
	• The AI application involves systemic observation or monitoring. This is a means of processing used to observe, monitor or control

Practice	Practice detail
	individuals, including data collected through networks or of a monitoring system in a publicly accessible area.
Perform Quality Control	• Quality assurance activities include reviewing and performing ongoing monitoring based on the lines of defence model of the AI governance. Please refer to Section 3.5.2 "AI Governance" of this document for the AI governance model. Organisations should evaluate whether the AI governance structure and processes are consistent with the changing AI requirements periodically. For example, a new AI technology may require the development of new roles or skills. Protocols, processes and procedures should be established to manage and ensure proper data governance throughout the AI project with existing data governance processes and procedures observed where available.
Consider all Ethical AI Principles throughout the AI Lifecycle	• This document provides end-to-end practices throughout the AI Lifecycle to consider. • The Ethical AI Framework should be leveraged to consider the ethical impacts (with reference to the Ethical AI Principles) and document related considerations in the AI Application Impact Assessment questionnaire. Please refer to Section 5 "AI Assessment" for details. • Ensure that the project plan considers ethics over development and implementation of AI (e.g. creating requirements for fairness in the proposed AI model and listing the methods in place to measure and test for achieving this). Please refer to Section 4.1.4.3 "Data Extraction" for more details on defining fairness. The project plan should account for each AI Lifecycle stage. All AI projects (including POC) should follow existing project management practices.
Ensure appropriate resourcing needs are met	• Determine the required resources for the implementation of AI. This includes considerations for suitable technology and services as well as resources with a mixture of functional knowledge from the organisation in addition to technical expertise. • Please refer to Section 4.1.3.1 "Technology Roadmap for AI and Data Usage" for considerations of technology. • Please refer to Section 4.1.3.2 "Procuring AI Services" for considerations of third-party AI services.
Ensure oversight exists for key processes	• Ensure all AI applications meet internal policy and process requirements including performing the AI Application Impact Assessment. • Ensure that effective and robust processes to assure testing of AI models for fairness, bias, the risk of adversarial attacks and explainability standards are established. Please refer to Section 4.1.4.1 "Business and Data Understanding", Section 4.1.4.3

Practice	Practice detail
	"Data Extraction", Section 4.1.4.5.4 "Model Overfitting", Section 4.1.4.5.6 "Adversarial Attacks" and Section 4.1.5.1 "Model Integration & Impact". • Assess emerging risks from changing market expectations on compliance of AI applications (e.g. changing regulatory or standards surrounding the use of AI). Please refer to Section B.1 of "Appendix B - Reference". • Ensure the Data Scientists are fully trained on AI application development and deployment as well as ongoing testing and validation standards to allow them to oversee compliance with organisation policies and strategic goals.
Engage and communicate with stakeholders	• Maintain clear and honest communication and engage effectively with all stakeholders, including staff, AI vendors and end-users, to foster trust and promote ethical AI adoption and do not exaggerate AI capabilities. • Establish proactive communication channels with AI vendors at an early stage to ensure clear understanding of product capabilities, limitations, and requirements. Please refer to Section 4.1.3.2 "Procuring AI Services (Sourcing)" for details. • Disclose the use of AI technology and provide clear information about the purposes, limitations, and potential risks associated with the AI application. Please refer to Section 4.1.5.2 "Provide information to users" for details. • When required, organisations should maintain transparent communication with regulators.
Define roles and responsibilities	• Organisations remain accountable and liable for the moral implication arising from the AI applications. Organisations are responsible for the actions, decisions, and outputs of AI systems they develop, deploy, or use. • Where applicable, define clear Terms of Use with users to state the chain of responsibility for AI services. • Responsibilities of the various roles should be defined to the level appropriate for the individuals to perform their duties. • The project management plan should contain clearly articulated roles, responsibilities and timelines to facilitate AI adoption. • Proper organisation of roles and responsibilities is essential to ensure that the ultimate decision-making authority for the AI and data activity has been assigned. • A three line of defence model is recommended for organisations where feasible. Please refer to Section 3.5.2 "AI Governance" for more details on roles and responsibilities. The roles include

Practice	Practice detail
	Project Managers, System Analysts, System Architects and Data Scientists that form part of the Project Teams.
Example: If videos are being used for an AI application, staff resources may require experience in video analytics in terms of technical expertise as well as ethical sense in video data handling and considering all related ethical AI principles.	

Related Ethical AI Principles:

1. Accountability – Ensuring accountability through AI Assessment, assigned responsibilities for quality control and consideration of ethics in the project.
2. Human Oversight – Human intervention can be reviewed as part of the project and oversight.
3. Sustainability and Just Transition – Mitigation strategies can be verified alongside schedule or budget if any identified adverse impacts are to be mitigated.
4. Cooperation and Openness – Organisations should proactively engage both internal and external stakeholders to foster trust and ethical AI adoption.

4.1.3 Project Ecosystem

4.1.3.1 Technology Roadmap for AI and Data Usage

Definition:

A technology roadmap should enable the organisations to plan and strategise which, when and what technologies will be procured for AI and big data analytics. An effective technology roadmap should outline a strategy to achieve the digital transformation goals.

Who are involved at this stage?

IT Planners/Executives, Data Scientists, IT Security Management, System Analysts, System Architects, Business Users

Practices and examples:

Practice	Practice detail
Establish appropriate technology	• Organisations should check with their IT teams (or equivalent) when choosing an infrastructure that fits the organisation's objectives, compliance needs and budget. This includes the use of existing infrastructure that may be available centrally in the organisation such as a centralised data analytics platform.

Practice	Practice detail
	• Organisations should check with their IT teams (or equivalent) when considering a suitable scalable technical infrastructure required for data collection and storage, data processing, data modelling, execution and deployment. This is to ensure that the solution can scale and have resources to use and process the data required. As the number of requests and data increases, the technical infrastructure should be able to respond automatically to maintain the speed and reliability. All AI and big data applications should be built with scalability and service availability in mind.
Conduct data assessments	• Organisations can assess what data are needed for the project and whether relevant data will be available for the project. Availability and quality of relevant data are the prerequisites for any AI application as they are required to power the AI application for training or fine-tuning the AI model (within the AI application) and to make predictions. • Organisations should identify the specific types of data and sources of data that will be collected, tracked, transferred, used, stored or processed as part of the AI application and whether the data involved are sensitive. • Organisations should document the data lineage to understand the source, path, license or other obligations and transformations of data which would be utilised in the AI application.
Example: The use of video analytics can be from recorded footage or a live camera. It is important to define how and what data are collected, stored and processed (e.g. biometric information) as part of data assessments. This is to help adherence to the "Data Privacy" principle during the entire data lifecycle stages from data collection to data disposal.	

Related Ethical AI Principles:

1. Reliability, Robustness and Security – The security of technology leveraged is required to ensure that AI applications are secure.
2. Sustainability and Just Transition – Technology used should be scalable and have mitigation for failure which could impact AI applications.

4.1.3.2 Procuring AI Services (Sourcing)

Definition:

AI projects can be conducted through outsourcing arrangements. Off-the-shelf products or even external data can be procured for AI projects. In conducting such procurement exercises, organisations should duly consider the related ethical considerations.

Who are involved at this stage?

Sourcing Team (i.e. Procurement), Project Manager, Business Users

Practices and examples:

Practice	Practice detail

Define requirements of the AI application for procurement	**<u>Plan for AI Procurement</u>**
	• Organisations should evaluate that they have the right sourcing team (procurement and Project Team with diverse expertise for opinions). For example, the Project Team can have different expertise including domain experts, system and data engineering experts and security experts. This would ensure that the right vendor with the right technology and skillsets requirements is chosen.
	• Organisations should allocate the right expertise to perform the change management delivery, training and handover upon procurement of the AI application. Training and education can be provided to IT to enhance capabilities. This will ensure that all relevant staff are equipped with the skills and knowledge that they should have to take on the goals and responsibility for AI (including understanding the probabilistic nature of AI algorithms).
	• Organisations may consider suitably including the requirements on the experience of the tenderers and/or the personnel engaged by the tenderers in AI development in the brief or specifications for procuring AI applications.
	• Organisations should define a clear problem statement on the objectives of the AI application.
	• Requirements for services can include consideration of model and ethical risks, importance to business objectives or strategy and model explainability as appropriate.
	
	<u>Understand Third-party's Products and Services</u>
	• As AI systems involve typically complex algorithms and modelling techniques, organisations should pay extra attention when outsourcing the implementation of AI projects. Organisations should request and review the documentation including the algorithm's design specification, coding and techniques the AI application is based on, its outcomes, ongoing support and monitoring or maintenance of the proposed AI application. Areas to be checked when using third parties include:
	  ○ Operation and intended use;
	  ○ Limitations and usage prohibitions;
	  ○ Training procedures;
	  ○ Testing procedures;

Practice	Practice detail
	○ Capabilities and performance metrics; ○ Checks performed to evaluate ethical values such as fairness; reliability, robustness and security; ○ Access to model configuration; and ○ Potential vulnerabilities, risks or biases in the AI application. • Understand the data processing, safety, security and privacy protection measures implemented by AI products and services before adoption. • Review the Terms of Use or responsibility statements provided by the AI vendors to understand the chain of responsibility for the AI services, where applicable. **<u>Ensure Third-party's Compliance with Existing Standards</u>** • The Project Team should observe that the AI vendor is compliant with local and international standards where available and applicable. Development for standards of AI is ongoing at the international standards body or professional associations such as the ISO and IEEE which can be referenced in the future. • For example, there are currently existing standards specific to medical devices with machine learning capabilities such as: ○ China Centre for Medical Device's "Technical Guiding Principles of Real-World Data for Clinical Evaluation of Medical Devices" [12] ○ United States Food and Drug Administration's "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)" [13] ○ ISO standards such as ISO 13485:2016 Medical devices – A practical guide • The International Electrotechnical Commission ("**IEC**") standards such as IEC 62304. **<u>Assess Applicability of the Third-party's AI Application</u>** • The applicability of the vendor AI application to the organisation's intended portfolio for use in the given context should be examined, tested and assessed. • For example, when buying an external advanced analytic AI application such as video analytics, ensure that the AI model (e.g. face recognition) within the AI application fits the organisation objective and intent as well as privacy and security regulations.

[12] https://www.cmde.org.cn/xwdt/zxyw/20240115162832179.html
[13] https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

Practice	Practice detail
	Ensure that AI application and data usage are compliant and align with Ethical AI Principles such as privacy (e.g. PD(P)O). **<u>Avoid Vendor Lock-in</u>** • Some vendor AI applications can be 'black boxes' and it is often unclear how the AI applications analyse data and subsequently arrive at results or decisions. Emphasise the need for explainability and request documentation on this (e.g. algorithms design specification, coding and techniques the AI model is based on and its outcomes) for review of the AI models within AI applications. Doing so will enable the organisations to engage with other vendors in the future to continue or enhance the initial AI application rather than being locked into one vendor. • Where vendors will not release documentation on AI algorithms due to trade secrets in practice, there can be a business decision for the organisation to take on where the risk of such non-disclosure should be considered with the possible business consequences. Organisations (using their individual project governance bodies) would consider case by case with their own justifications, considerations and decision documented. • Ensure the data models and data formats for the AI application are usable across a variety of platforms, rather than formats that are specific to a given vendor where possible. This can be checked with the individual vendors. **<u>Assess Supply Chain Risks and Diversity</u>** • The AI industry relies on a highly globalised supply chain. However, certain countries or regions may impose unilateral coercive measures, such as technology barriers and export restrictions, potentially creating development obstacles and disruptions to the global AI supply chain. As a result, there may lead to supply disruptions for hardware, software, and tools. • Organisation should thoroughly evaluate the supply chain diversity across key components, including chips, software, tools, computing infrastructure, and data sources used in AI applications. It is imperative to monitor for the vulnerabilities and defects in both software and hardware products and to promptly address them to ensure system security. **<u>Agree on Intellectual Property</u>** • Some third-party vendors may not disclose the algorithm and technique used in the development of the AI application to

Practice	Practice detail
	protect their intellectual property (IP) rights. When the third-party vendor designs and develops a new AI application for the use of organisation, the new intellectual property owner should be mutually agreed in line with organisation procurement procedures. • Organisation should agree with the third-party AI provider on who should be the controller of the AI application and the data being processed by the AI solution upon procurement of the solution. In such cases, this should be explicitly stated and stipulated in a contract or service level agreement.
Perform risk mitigation procedures	• The credibility and reliability of the third-party provider indirectly affect the reliability and robustness of the AI application. • Conduct an independent evaluation of the third-party provider as part of a due diligence process. This applies to custom AI applications as well as off-the-shelf or turnkey AI applications. • This can include basic company information checks such as business license, official references, global sanctions list, law enforcement list, past experiences/credentials and qualifications. Where these checks are not passed there can be a business decision for the organisation to take on where the risk of reliability of the vendor should be considered with the possible business consequences. Organisations (using their individual project governance bodies) should consider this on a case by case basis with their own justifications, considerations and decision documented. • Organisations can maintain documented processes for dealing with third parties and evidence their due diligence completed. This includes maintaining evidence of documentation obtained and reviewed related to the third-party provider. • Risk considerations such as any third-party interactions with the AI application, access to model configuration files or endpoints, access to the models and how they are used should all be considered and documented.
Example: When procuring a video analytics solution, organisations need to understand the third-party's approach. Including requirements in the tender that the contractor should comply with related security standards and ensure the proposed design takes into consideration security issues such as the use of masking, anonymising and maintaining compliance to data privacy standards assists with adherence to the "Data Privacy" principle.	

Related Ethical AI Principles:

1. Fairness – Third parties should be able to demonstrate why their solutions are considered 'fair'.
2. Explainability – 'Black box' AI applications should be avoided where possible.

3. Reliability, Robustness and Security – AI Applications from third parties should be reliable, robust and secure.

4.1.4 Project Development

4.1.4.1 Business and Data Understanding

Definition:

When developing AI applications, organisations should determine the objectives of using AI and weigh and balance the benefits and risks of using AI in the decision-making process.

Who are involved at this stage?

Project Managers, Business Users

Practices and examples:

Practice	Practice detail
Define business requirements	• A process (such as the AI Application Impact Assessment, please refer to Section 5 "AI Assessment" for details) is defined and should be followed through to ensure the potential application impact and the data usage is considered alongside Ethical AI Principles, strategy and any applicable legal requirements. • A standard process exists to define and validate the performance and business success criteria for a given AI project and application. Examples of these criteria include measuring performance metrics such as true positive rates or false negative rates or false positive rates for classification models. High true positive rates can indicate that the AI application performs as intended and provides results with high accuracy whereas high false positive rates denote that negative cases are incorrectly predicted as positive and the AI application may require further fine-tuning to produce accurate results. High false negative rates denote that positive cases are not correctly included in the prediction result. Given there is no single set of metrics that can fit all AI applications (given higher false negative rates may trigger a lower false positive rate and vice versa), this definition should be tailored by Project Team for individual AI application. • Organisations should engage both internal and external stakeholders to determine the functional requirements, non-

Practice	Practice detail
	functional requirements and quality assurance procedures needed for the AI application. • Organisations should consider potential alternatives to the current or proposed AI applications and check the performance compared to the AI application in terms of accuracy metrics and any business, legal, economic or social impacts. For example, is the alternative method more accurate and economical from an operational perspective than an AI application? Should the model minimise false negative or false positive? This would assist the organisation to gauge whether the alternative method would impose less risk than an AI application comparing to the benefits of the proposed AI application. • The overall AI application development process should include an iterative impact assessment (please refer to AI Application Impact Assessment in Section 5 "AI Assessment") that includes testing requirements and ensures that the AI project is aligned with the organisations approach to overall risk management (for example, alignment with all internal policies and Ethical AI Principles).

Related Ethical AI Principles:

1. Cooperation and Openness – organisations should engage both internal and external stakeholders to determine the functional requirements, non-functional requirements and quality assurance procedures needed for the AI application.

4.1.4.2 Solution Design

Definition:

When developing AI applications, organisations should determine the design requirements. This includes assessing various aspects such as the suitability of data and technology and the degree of human intervention required. Riskier decisions should incorporate a higher level of human intervention in the process.

Who are involved at this stage?

Project Managers, Business Users

Practices and examples:

Practice	Practice detail
Assess AI model's suitability	• An AI model within an AI application should be assessed for suitability compared to the organisation's objectives. There should be a process in place to produce reasons for the decisions or recommendations that an AI model reaches when required. Factors affecting the decision-making of the model can include the: ○ business objective; ○ accuracy and explainability of the model; ○ model development and training time; ○ scalability of the model; ○ complexity of the model; and ○ IT architecture or infrastructure. • A process should be in place to assure that a systematic review is conducted to determine the suitability of organisation's existing technology architecture for the AI model's design. To identify suitability, the following activities should be performed: ○ Understand the datasets being consumed by AI models - Certain AI models can only operate with certain types of data; some algorithms can function with smaller datasets while others do not. For example, Naïve Bayes or linear regression algorithms may perform well for small datasets. If the training dataset is large, it may be more suitable to consider using decision tree algorithms [14]. A sufficient amount of training data should be allocated to fine-tune the reliability of an AI model to a satisfactory level. Refer to section 4.1.4.3 "Data Extraction" for details on the portion of training data compared to testing data and validation data. ○ Identify the types of learning algorithm required which can be categorised as: ▪ Supervised learning - If the training dataset has an input and a corresponding output pair, then a supervised learning algorithm should be applied. Supervised learning is the process of training AI models using training datasets that contain an input and a correct output pair which allow the model to learn. Examples of supervised learning algorithms are regression and anomaly detection. ▪ Unsupervised learning - If the dataset does not have any corresponding output pair, then an unsupervised learning algorithm should be used. Unsupervised

[14] https://www.kdnuggets.com/2020/05/guide-choose-right-machine-learning-algorithm.html

Practice	Practice detail
	learning is a learning algorithm that will attempt to identify the pattern and hidden structures within the data. Examples of unsupervised learning algorithms are clustering and dimension reduction. ▪ Reinforcement learning - Reinforcement learning is a reward and punishment mechanism used to train AI models where the reward function determines the rewards for the actions of the AI model and vice versa. The reward function allows the AI application to derive conclusions by learning from experience instead of making a prediction. Reinforcement learning may not be applicable for certain AI applications. This reinforcement learning technique can train the AI model to respond based on existing reward information. o Understand the constraints – Several key constraints to be understood include the scalability of the AI model, storage capacity and model development time.
Determine the appropriate level of human intervention	• As part of completing an AI Application Impact Assessment, evaluate whether the AI application requires human intervention. Apart from the severity of impact to individuals or organisations as a result of the decisions to be made by the AI applications, organisations should consider factors such as: o Probability of the AI Application causing harm or affecting human safety during the decision-making process; o The operational cost of executing the decision making with human intervention in comparison to executing it without human intervention; o Legal implications or similar effects for automated decisions when there is no "human" involvement in the decision making; and o Evolving societal norms and values. Considerations should include whether it would be generally acceptable to society when the decision-making process is fully automated by an AI application. • Organisations should determine the level of human intervention required from the AI application. In general, there are three different approaches when determining the appropriate level of human intervention. These are 'human-out-of-the-loop', 'human-in-the-loop' and 'human-in-command'. 'Human-in-the-loop' and 'human-in-command' approaches can be considered assisted intelligence or augmented intelligence while human-out-of-the-loop can be considered automated intelligence or

Practice	Practice detail
	autonomous intelligence. The extent of human intervention should be proportionate to the impact level associated with the AI application. A higher impact would have a higher level of human intervention expected. 1. Consider a human-out-of-the-loop approach in decision-making when decisions being made do not have a significant impact on the economy, society, law or when it is impractical to involve human beings in the process. Impact assessment can be performed using the AI Application Impact Assessment. As an example, a human-out-of-the-loop approach would be suitable when swift decisions are required within a limited timeframe and there is limited impact. It is still crucial that the AI application's decision is interpretable by human beings even when the decision making itself will not involve human beings. 2. Consider a human-in-the-loop approach when there is limited data available and humans are able to make a better judgement. 3. Consider a human-in-command when the human-in-the-loop or human-out-of-the-loop approach is not suitable. This allows human beings to oversee and control the overall activity of the AI application with the ability to decide when and how to intervene. This approach can enable human beings to adjust variables that may alter the outcome or decision making of the AI application. • Identify the automated decisions that the AI application is required to perform. When the AI automates decisions that have legal implications or similar effects, and there is no human intervention in the decision, processes should be introduced to ensure the individuals affected have been informed. Users should be made aware of their interactions with the AI application upfront. This is to increase transparency to users and ensure that user consent has been obtained (for example, provide a disclosure that states the automated decisions are solely performed by the AI application and users should only proceed if they agree with such usage). • AI robots serve as an auxiliary role in the workplace, primarily for assisting users and the user organisations in accomplishing their assigned tasks. In principle, AI robots are prohibited from making independent decisions in high-risk applications (such as medical diagnostic systems, autonomous driving, decision making in public services, etc.). All decision making must be carried out under user supervision and guidance. For further

Practice	Practice detail
	reference, please refer to the "Hong Kong Generative Artificial Intelligence Technology and Application Guideline".
Define requirements for transparency and explainability of AI models	• Create a communications plan on AI model explainability to achieve transparency and explainability to impacted stakeholders for exhibiting the relative importance of each variable to explain predictions. • Provide examples to explain the AI decision-making process in narrative terms or graphics (for example, drawing a workflow with decision trees) whenever possible for a non-technical audience to understand and visualise the AI operations better. Organisations should make explicit how different factors and data determine the outcomes and conclusions of their AI models. The reasoning that is usually acceptable from a human point of view can include aspects such as: ○ logic; ○ societal norm, practices and beliefs; and ○ moral justifications.
Example: It is important to determine the level of appropriate human involvement for adherence to the "Human Oversight" principle. Human-in-the-loop is often used for the automation of security threat detection for computer systems. The changing threat landscape in today's cybersecurity leads to new forms of attack which are not detected by AI applications. A security specialist would still be required to perform security threat detection alongside the AI application and to check the accuracy of the results from the AI application. Predictive railway maintenance that triggers work orders or orders maintenance parts can adopt human-out-of-the-loop approaches although this may be limited to a notification, prioritisation and order placement as the maintenance itself may still require in-person involvement). A human-in-command approach should be adopted in situations when broader economic, societal, legal or ethical impacts are involved (for example, medical operations).	

Related Ethical AI Principles:

1. Human Oversight – Ensure an appropriate level of human intervention based on multiple factors such as the benefits and risks of the AI application, impacts of the AI decision, operational cost and evolving societal norms and values.
2. Accountability – Ensure that there is accountability for high-risk AI applications including having human beings involved when decisions are being made.

4.1.4.3 Data Extraction

Definition:

AI models rely on information from various sources, within or outside of the organisations, which can possess high risks associated with data quality, validity, reliability and consistency. An Extract, Transform and Load ("**ETL**") tool is typically used to extract huge volumes of data from various sources and to transform and load the data based on the AI model's needs. Complex data cleansing and transformation steps can be prone to unintended user errors that are difficult to identify and may lead to erroneous modelling results. Data integrity is a necessary component to ensure data fairness. Data integrity ensures that the results generated from the AI application are not generated by biased or skewed datasets.

Who are involved at this stage?

System Analysts, System Architects

Practices and examples:

Practice	Practice detail
Use or establish a centralised, standardised data layer	• Utilise the central data repositories in organisations as a common data registry for AI applications where appropriate. The repository should have access controls in place that are regularly maintained. • Leverage data repositories such as data warehouses, data lakes or data marts to act as a 'single source of truth'. The data repository can then be used to provide a cleansed data source for the AI application development.
Perform data validation	• Validate consistency of data formats when data are obtained from multiple sources. • Perform evaluations to ensure data quality and characteristics align with the business and AI application objectives. Some common examples of data validation rules include validating data types or formats, unique key checks, duplicate data checks, consistency of acronyms and preventing unexpected null values. • Identify whether personal information exists in the dataset and complies with data usage policies for personal data instituted by regulators. • Select and sanitise training data carefully to prevent infringement of intellectual property rights, exclude sensitive data, and mitigate risks of data poisoning attacks.

Practice	Practice detail
Obtain assurance for public or third-party data	• Verify reliability, representativeness and relevance of the data acquired. An example of confirming data reliability is verifying the information obtained against its source. Further verification can be performed if an independent reliable data source exists. • Examine the content of the source data and determine if it fits the needs of the AI model by identifying its relevance to the project objectives and whether it addresses the subject matter. • Assess the reliability level of the vendor and/or institutions of the data source. Please refer to Section 4.1.3.2 "Procuring AI Services" for details. Data reliability is often contextual e.g. how critical is the decision that the data needed is to be relied on. An example would be the difference between the data used in a simple marketing application with limited impact compared to the data used to decide who cannot get on a plane. The data for the latter case would be expected to typically be much more reliable than a simple marketing application.
Standardise training data annotation	• Provide comprehensive data annotation guidelines and adequate training to annotators. Insufficient annotation guidelines, inadequately trained annotators, and annotation errors can adversely affect the accuracy, reliability, and effectiveness of models and algorithms. Moreover, these issues may introduce training biases, amplify discrimination, reduce generalisation abilities, and result in incorrect outputs.
Complete documentation on data sources	• There should be a standard requirement or process to define and document the specific types of data which will be collected, tracked, transferred, used, stored or processed as part of AI model and application development. • Tracking and maintaining information on the data source, owner, licenses and key assumptions records should be performed as a good data governance practice. System analysts and system architects are responsible for governing the collection or acquisition of all sources of training, testing and operational data related to the AI model.
Enhance transparency and traceability	• Ensure ETL processes are transparent (e.g. through scripting) by documenting or using a workflow approach that visualises data lineage for the organisation, System Analysts and System Architects to understand the data flow.

Practice	Practice detail
Enable error handling mechanism	• Develop an error handling mechanism for the ETL solution. The error handling mechanism should capture the ETL's task name, error number and error description. This should then be captured in a separate table or file that enables the Project Team to analyse issues and fix them. For example, an alerting mechanism can be incorporated to alert Project Team members via email when errors in ETL processes occur.
Example: For video analytics, to obtain assurance for public or third-party data, ensure that the videos have been obtained from a valid source based on the metadata of the video and verify the recording location, recording date/time, duration, etc. Additional analysis can be done on the video to ensure source validity by identifying whether the human being's faces have been altered due to adversarial attacks. This would assist adherence to the "Reliability, Robustness and Security" principle. For adherence to the "Reliability, Robustness and Security" principle, organisations should enable the error handling mechanism. An AI model for counting footfall from CCTV should be tested using actual video and re-tested when there are code changes. Data extracted from videos could be broken down into different stages of extraction to identify for errors (for example, assess for exclusions of objects at different stages of extraction).	

Related Ethical AI Principles:

1. Fairness – Integrity of data sources is important to help ensure a fair outcome.
2. Reliability, Robustness and Security – Data collected should be complete and accurate to ensure the reliability of the data for model training purposes.
3. Safety, Reliability, Robustness and Security – ETL tools must be secure and robust when performing data cleansing and transformation to avoid errors in subsequent stages of development.

4.1.4.4 Pre-processing

Definition:

In the data processing procedures for AI modelling purposes, the data processed need to be fair and appropriate in terms of data samples, size and distributions to ensure the AI application formulate meaningful and representative inferences. Representational flaws in datasets such as overrepresentation or underrepresentation of data samples may lead to bias in the outcomes of trained AI models.

Personal data containing an individual's information and sensitive data require extra care during solution development to prevent data leakage as well as breaches of privacy and security policies. The amount of personal data collected, processed and used should be minimised where feasible.

Note: For considerations related to the use of personal data and privacy in IT projects including AI or big data projects, organisations should refer to PD(P)O guidance.

Who are involved at this stage?

Project Managers, System Analysts, System Architects, Data Scientists

Practices and examples:

Practice	Practice detail
Strengthen data security and protection	• Ensure that operations comply with data security and protection requirements such as using encryption technology and other protective measures in particular when processing personal data and classified data. • Implement proper data access controls, develop robust data backup and recovery plans, and check data access records regularly. • Enhance the capabilities in areas such as network security and supply chain risks to reduce exposure to cyber threats or dependency on single sources.
Conduct Privacy Impact Assessments ("PIA")	• Conduct PIA to identify whether it is necessary to collect the data type, amount and extent of personal data for use in AI applications.
Perform data anonymisation	• Use safeguards such as pseudonyms and full anonymisation to prevent the connection of the personal data to an identifiable person. Data anonymisation is the process of protecting sensitive information via encrypting, masking and aggregating any information that links an individual to the stored data. • Reaffirm that the benefits of keeping anonymised data outweigh the potential risk that such data are used to identify individuals, and the impact that such de-identification could have on the individuals. For this reason, the Project Team should review regularly whether anonymised data can be re-identified and adopt appropriate measures to protect personal data. • A similar analysis on benefits and risks should be applied to assess the loss of data utility if the data are being de-identified. For example, blurring faces in video analytics is a form of data anonymisation performed to reduce class bias and protect the identities of the individuals. • When data anonymisation is performed, document the process to achieve this as part of the logging procedures.
Remove illegal or harmful information	• Remove illegal or harmful information including false, biased and intellectual property rights infringing content. Training data

Practice	Practice detail
	is at risk of being poisoned from tampering, error injection, or misleading actions by attackers. This can distort the model's probability distribution, reducing its accuracy and reliability.

Perform data minimisation and aggregation	• In situations where the anonymisation of data is not feasible or required, organisations should minimise the personal data that an AI application has access to. Organisations should identify if there are any less data-intensive methods to achieve the goals of the activity.
	• Data minimisation involves identifying the minimum amount of data needed to fulfil the purpose and objective of the AI application.
	• Project Teams can achieve minimised data by ensuring the data collected is:
	  ○ sufficient to meet the objective;
	  ○ relevant and has a rational connection to achieve the objective; and
	  ○ limited to only what is required.
	• For example, data aggregation, a process of collating and summarising the data can be performed to reduce the level of detail if it is not required.

Create a single analytical data environment	• Process and analyse sensitive data in one secure work area within the same server or cloud environment where the data was collected to avoid the risk of leakage when transferring the data. Additionally, the organisation should avoid using ad-hoc insecure environments where possible to ensure proper handling of data is maintained throughout the project lifecycle. • Implement and adhere to retention policies for all data (e.g. personal data and video data) collected.
Identify outliers and anomalies	• Detect outliers in the source dataset and in the final modelling datasets by program or plotting the data in the form of graphs such as box plots, scatter plots or histograms. Outliers and anomalies can be defined as data points in the dataset being some distance away from the mean of the dataset's population. If the detected outliers and anomalies in the dataset involve errors or do not represent the population accurately, they should be removed or patched with reasonable values if they do not represent the entire population accurately. This is because, in supervised learning AI models, outliers and anomalies can be deceiving, resulting in prolonged training times, or leading to less precise AI models.

Practice	Practice detail
	• When the datasets contain a large number of variables, identifying and removing outliers can be challenging. In such cases, automatic outlier detection algorithms can be used in the modelling pipeline. Examples of automatic outlier detection algorithms are isolation forest, minimum covariance determinant and local outlier factor.
Define and test for fairness and bias	• Fairness can be understood from the perspective of groups, individuals or third-party agents. Organisation should define fairness in the context of their AI application to be able to test for bias and fairness. As an example, a facial recognition system should display statistical (equal) parity in order to classify people from different races accurately. Examples of mathematical fairness definitions are as follows: 1. Statistical (Equal) Parity – Positive classification rates across both groups should be equal. 2. Equal Opportunity – Mathematically, it implies True Positive Rates across all groups should be equal. 3. Overall Accuracy Equality – Both groups should give similar accuracy measures for the given model. 4. Gini Equality – Gini is defined as the ratio area between the Receiver Operating Characteristic ("**ROC**") curve and the diagonal line as well as the area of a related triangle. In essence, the model should do an equally good task of classifying 1's as 1's and 0's as 0's across all groups. • Bias can encroach upon an AI application when bias exists in data being used to train the AI model. Organisations can mitigate inherent bias using bias detection methods to evaluate whether the model discriminates against various groups or individuals, given a test-dataset. Examples of bias detection methods include: ○ Comparing performance metrics – Analyse performance of the AI model across groups, by calculating different rates of correct and incorrect predictions. Different performance metrics such as true positive rate, false positive rate, true negative rate or false negative rate for each possible group in the dataset are then computed. Output such as distribution plots of features can identify skewness across groups. ○ Measuring fairness – This can be done by performing comparison analysis for the disadvantaged against advantaged groups based on fairness definitions normalised treatment quality, statistical parity, etc. Bar charts and heat maps are examples of outputs that could capture disparities across groups. ○ Detecting proxies – This can be done by calculating mutual information of all features with different sensitive attributes (e.g. age, gender and race) to understand how they might act as proxies.

Practice	Practice detail
Fix unbalanced class problems	• Unbalanced classes are a classification issue that arises when the distribution of data across different classes is biased. For example, when implementing an AI system for processing applications, adequate amount of historical data on approved and not approved cases should be made available for training the model. This means that not all the classes or categories in the dataset are represented equally. AI predictive modelling involves predicting a class label for a given observation from the datasets. When an unbalanced class complication appears, this results in AI models that show poor predictive performance, specifically for the classes from minority groups, resulting in class discrimination. • Benchmark data distributions against the full dataset population statistics to quantify the representativeness of the data being tested. This can assist in mitigating bias in the data used for model development. • For example, use standard algorithms such as random forest, logistic regression and decision trees to ensure a good classification model that can predict classes with limited instances and ensures that training data contains equal data samples from different groups. • In the event where unbalanced datasets led to the unbalanced class issue, this could be handled using sampling techniques such as: o random resampling to rebalance the class distribution for the imbalanced dataset; o random undersampling to delete dataset examples from the majority class with the intention that the model loses some information to achieve a balanced dataset comparable to the minority class; and o random oversampling to duplicate dataset examples from the minority class to achieve a balanced dataset comparable to the majority class.
Compare training, validation and test data	• Once the AI model has been tuned to a satisfactory level, a testing dataset can be used to evaluate the learned models to gauge the model's performance on unseen data. Training data are the sample data used to train the AI model. If several models or tuning of hyper-parameters[15] are involved, validation data can be used to evaluate the model fit among different models during model development. Test data can be considered the data that will be held from the AI model for final modelling testing purpose until the very end of the training stage.

[15] Hyperparameter refers to a parameter which its value is used to control the learning process of a machine learning model. For example the maximum level for a random forest model.

Practice	Practice detail
	• Typically, the ratio for the data is as follows: ○ Training data: test data: validation data are commonly set at 60:20:20 or 70:15:15. If validation data are not used, the recommended sample size for training data to test data are 80:20. • Analyse and compare both training and validation/test data to ensure the results from what is used to evaluate AI model are equivalent to what was used to build the model. This is to confirm that the outcome of the AI model is comparable when new and unseen test data are used. For example, if the AI model is able to classify an image of a person as a human being, based on the training data, it should be able to perform comparably based on the test data even though both images presented are two distinct individuals. • Create or use a library containing training and validation/test data which can be utilised to evaluate for potential unintended bias. For example, some Python deep learning libraries[16] [17] [18] can be used to detect prejudicial or discriminatory correlations between features, labels and groups.
Example: One example to practice data anonymisation in the case of video analytics is to blur or obscure faces to ensure individuals cannot be identified in the videos (when identification is not necessary). This is in line with the "Data Privacy" principle. If the objective of the analytics is to measure footfall, limit the data extraction and storage to the total count. Consider blurring of faces at the point of data collection and do not extract information such as the behaviour of individuals and their activities. To adhere to the "Fairness" principle, organisations can analyse the training data used to assess fair representation. For video analytics (to measure footfall), ensure that the AI model measures footfall correctly for different groups of people. Training data can contain equal data samples from individuals in different groups to ensure different types of individuals are recognised. Training data, when used to measure footfall via CCTV should include various scenarios to ensure the outcome of the model meets the objective correctly.	

Related Ethical AI Principles:

1. Data Privacy - Ensure minimal collection and processing of personal data and retention policies are followed which take into account privacy regulations such as PD(P)O and the related PD(P)O guidance notes.
2. Lawfulness and Compliance - Ensure consideration and/or compliance with local and international data privacy and protection standards as appropriate such as the PD(P)O, the

[16] https://github.com/dssg/aequitas
[17] https://aws.amazon.com/sagemaker/clarify/
[18] https://fairlearn.org/

Chinese Mainland's Personal Information Protection Law (《個人信息保護法》) or European Commission's General Data Protection Regulation ("**GDPR**") where applicable when handling data.

3. Fairness – Over representation or under representation of certain groups or categories in the data can lead to biased and unfair outcomes and should be avoided.

4.1.4.5 Model Building

Model building encompasses the following areas which will be discussed in detail in the subsequent subsections:

- Model Assumptions
- Model Objectives and Incentives
- Input Variable Selection
- Model Overfitting
- Model Training Ownership
- Adversarial Attacks

Model customisation refers to the process of adapting a pre-trained AI model, usually developed by a third-party, to meet specific requirements rather than building a model from scratch. It may involve re-engineering, fine-tuning or grounding the AI model to improve its performance and align it with specific business needs. Should security flaws occur in the foundation model, it will lead to risk transmission to downstream models. Organisations should be mindful of the risk transmitted from foundation models and adhere to the practices in this Section for model customisation.

As organisations often outsource their model customisation work to external service contractors, a thorough assessment of potential service providers is recommended when considering such procurement. Please refer to Section 4.1.3.2 "Procuring AI Services" for details on the practices and considerations in procuring AI services or products.

4.1.4.5.1 Model Assumptions

Definition:

Model assumptions are conditions that should be satisfied by the model before performing the relevant modelling analysis. The assumptions underpinning the model must be checked for accurate interpretations and conclusions.

Who are involved at this stage?

Project Managers, Data Scientists

Practices and examples:

Practice	Practice detail

Perform rigorous testing	• Formulate clear rules and methods of the testing and utilise technologies such as sandbox simulations to test and verify AI models.
	• Statistical and other data assumptions must be tested during development to ensure model predictions and insights are valid.
	• Lack of assumptions review and approval of AI models trained leads to erroneous or risky assumptions.
	  ○ The Project Team should conduct testing and document the results of the testing to verify the AI model's assumptions. Examples of testing include algorithm testing and regression testing. Lack of documentation or logging of model results and configurations can make it difficult to manage or maintain AI assets in subsequent stages. The logs should contain information such as:
	  ▪ Jobs/operations performed by the AI application including the algorithm(s) executed; and
	  ▪ Decisions made and outcomes.
	• Monitor the decisions made by AI through the logs and compare them to human decisions or real outcomes and document the action taken. When there is a need to re-train the model (for example, when accuracy levels are too low), include a thorough justification for the change and the features affected.

Example: Organisations should perform rigorous testing to assess model predictions and insights. This is to check that "Fairness" principle is being followed for accuracy. In the case of a weather prediction model, the data may only be applicable for certain locations and therefore not as practical for predicting the weather in different locations. The location assumed in the data should be checked.	

Related Ethical AI Principles:

1. Fairness – To achieve fairness in the outcomes, all data assumptions should be tested and verified for accuracy.
2. Reliability, Robustness and Security – The AI application should be robust and reliable by producing the same results and outcomes when dealing with the same scenarios throughout the testing.

4.1.4.5.2 Model Objective and Incentives

Definition:

AI applications motivated to achieve targets can be misused to satisfy the stated objective but fail to solve the problem and result in bad behaviours.

Who are involved at this stage?

Project Managers, Data Scientists

Practices and examples:

Practice	Practice detail
Create risk-sensitive objectives	• Risk-sensitive objectives aspire to balance the risks and benefits of the AI decisions made by AI applications. The Project Team should incorporate downside risk measures in machine learning objective functions to operate more conservatively to avoid situations in which the AI makes disastrous decisions. For example, if the AI model does not meet the success criteria, other measures should be performed by the AI model that can include escalating to a human operator or aborting the decision-making process.
Example: Reinforcement learning can be used for autonomous vehicles. For example, when AI moves the vehicles in a certain direction, the pattern of the traffic surrounding the vehicles can be used as a reward function (i.e. if the surrounding vehicles are greater than a specified distance away from the autonomous vehicle itself and in line with the highway it is more likely that the decisions being made by the AI are correct). Scenarios should be checked to ensure that traffic patterns recognised are safe for others. This adheres to the "Reliability, Robustness and Security" principle.	

Related Ethical AI Principles:

1. Reliability, Robustness and Security – Rewards functions in AI can reinforce learning models, leading to more reliable and robust AI models.

4.1.4.5.3 Input Variable Selection

Definition:

Input variables or features are values within datasets that are loaded into an AI application for the purpose of training the AI model. A robust AI model relies on these informative inputs to provide an output, often referred to as a target variable (i.e. what the AI model is trying to predict). Selections of input variables should consider both organisation knowledge and causal relationships.

Who are involved at this stage?

System Analysts, System Architects, Data Scientists

Practices and examples:

Practice	Practice detail
Comply with policies and ethical practice to obtain and use input variables	• Define a process to ensure the potential impact and the data being used aligns with Ethical AI Principles and internal policies (including legal requirements). • State the objective for data usage and obtain permission from data holders prior to using data input variables in AI models especially when PII that could be used to identify a specific individual is involved. • Ensure input variables have been collected in a compliant manner (e.g. in compliance with PD(P)O requirements) and that their use is appropriate and permissible.
Avoid target variable leakage	• The Project Team should avoid target variable leakage which is where information that should be beyond the training dataset is used to create the AI model. Target variable leakage arises when the AI model is being trained based on a dataset that includes information that would not normally be available at the time of prediction. This future information can make the AI model's results appear more accurate than the model performance would be in real situations. • Time-series data (if used) should be evaluated if data are not used where the data occurred after the case being predicted by the AI model. A cut-off value on time can be useful to prevent obtaining information after the time of prediction. • Features that are highly correlated with the input variable should be assessed to ensure they are not a result of future information (as these features are more likely to be the target variable leakage if it is occurring).
Example: Organisations should aim to avoid target variable leakage to adhere to the "Fairness" principle. When AI is used in medical diagnosis to detect disease, the training dataset should be limited to data that would be available at the time of prediction. If it has data such as future surgery used, this may indicate that the real results of the model are not as good as expected.	

Related Ethical AI Principles:

1. Fairness – Data used in AI models should be obtained fairly.
2. Reliability, Robustness and Security – To operate reliably and as expected, avoiding target variable leakage means that test results for predictions are more likely to mirror accuracy levels of real predictions.

4.1.4.5.4 Model Overfitting

Definition:

Overfitting is a modelling error that emerges when an AI model is trained to closely fit a limited set of data points. An overfitting model will often exhibit high accuracy on the training dataset but low accuracy on new data. If the AI model does not generalise well from the training data to new or unseen data, the AI model may perform poorly in its prediction.

Who are involved at this stage?

Data Scientists

Practices and examples:

Practice	Practice detail
Perform cross-validation	• Cross-validation is a resampling procedure that would be performed by a development team to evaluate the effectiveness and accuracy of an AI model based on a limited data sample. One example of a resampling procedure is the k-fold technique. • The K-fold technique allows training and testing of the AI model k-times on different subsets of training data to derive a more accurate estimate of the model's prediction performance. • Every resampled dataset should have a similar distribution as the full population's distribution for this approach to be acceptable. For example, if the full population consists of 10 distinct categories, the resampled dataset should have a comparable number of distinct categories. This is because missing categories during the cross-validation process can lead to bias in the AI model's outcome.
Perform regularisation and other forms of model selection	• Consider the use of regularisation. Regularisation is the process to normalise and bring uniformity to AI models. Regularisation is often a technique used in machine learning to fine-tune the AI model function by discouraging the learning of additional and unnecessary complex models. This can reduce the risk of overfitting. • An example of a regularisation technique is pruning. This can be used to remove features that are the least important for the model to learn from. AI models can learn not only basic features of provided datasets but also noise as well as fluctuations that place importance on unimportant features and hence pruning can help. Examples of unimportant features can include backgrounds in an image containing the actual object being detected.

Practice	Practice detail
Example: To avoid overfitting in the case of medical diagnosis, cross-validation can be performed. AI models can be further evaluated by splitting sample data by specific medical conditions and re-evaluating the model's diagnosis. Another method to avoid overfitting is reducing factors unrelated to the condition (e.g. social background or location in some cases to reduce the complexity of the AI model). Such practices would assist organisations to adhere to the "Reliability, Robustness and Security" principle.	

Related Ethical AI Principles:

1. Reliability, Robustness and Security – Overfitting should be avoided to ensure outcomes from AI models when used on unseen data are reliable and accurate.

4.1.4.5.5 Model Training Ownership

Definition:

AI models where the training process is partially or fully outsourced to the public cloud or relies on third-party pre-trained models can introduce new security risks. Please note that if the AI model training process is not outsourced, the practices and examples in this subsection are not applicable.

Who are involved at this stage?

Sourcing Team (i.e. Procurement).

Practices and examples:

Practice	Practice detail
Use trusted data/platform services	• The process of training an AI model involves providing the AI model with training data to learn from. Precise training data are required to help AI models to understand patterns to derive a conclusion. Any third-party vendor used to train an AI model or to provide a cloud training environment for the organisation should be subject to the third-party procurement process. This helps lower the risk of receiving a tampered model that may feature a backdoor. Please refer to Section 4.1.3.2 "Procuring AI Services" for details on third-party procurement process. • Organisations should assess the suitability of storing and using personal data or data belonging to organisations on the public cloud. • Developers should ensure the security of training environment for AI model algorithms, including proper cybersecurity configurations and appropriate data encryption measures. • The trained model should be transferred only through channels that provide surety of integrity in transit. Examples include docker containers and pickle files (for python scripts) which can then be encrypted before being transferred. • Comply with requirements or regulations on cross-boundary data flow where applicable.
Example: When using cloud/third-party services, validate that cloud vendor is compliant with international security standards such as ISO and organisational security practices. Compliance with international standards can help to ensure that Ethical AI Principles such as "Data Privacy" and "Reliability, Robustness and Security" are followed.	

Related Ethical AI Principles:

1. Reliability, Robustness and Security – Third-party's compliance with relevant international standards increases the likelihood that their AI applications remain reliable, robust and secure.

4.1.4.5.6 Adversarial Attacks

Definition

An adversarial attack takes place when malicious actors deceive the AI models to intentionally influence the AI application's outputs without being detected. This is attempted by subtly modifying the input data with changes that are imperceptible to humans, to induce the AI model to make an incorrect prediction. Such attacks can occur in the training phase or the test phase.

Attacks that appear during the training phase are poisoning attacks whereas attacks that exist in the test phase are evasion attacks.

Who are involved at this stage?

Data Scientists, Project Managers

Practices and examples:

Practice	Practice detail
Perform adversarial training	• The Project Team should obtain as many adversarial examples as possible and explicitly train the model not to be misled by these samples. Adversarial training involves enlarging the dataset to include adversarial examples from which the AI models can learn from. Adversarial training is introduced to secure the AI application from disruptive perturbations at different points of vulnerability in an AI model. Disruptive perturbations happen when a small random value is added to existing data points in training data that alter the original features of the data. Such examples can often be seen in digital images where random values are added to the original image which results in the wrong classification by the AI model (e.g. an image of a panda is classified as gibbon due to disruptive perturbations). • Build a new adversarial machine learning model that predicts whether a dataset is from a recent dataset or from the AI training dataset. This could be a method to identify new changing data, leading to new forms of adversarial attacks that would require retraining of the AI model. • For example, there are open-source libraries written in Python programming language such as the Adversarial Robustness Toolbox[19] that can be used to train AI models against adversarial attacks.
Conduct independent assessment and testing	• When conducting a security assessment of the AI application, consider involving an independent third party. For large-scale or high-risk projects, a "red team" approach could be adopted to ensure the preparedness, quality and reliability of the systems. The "red team" will play the antagonistic role in testing the ultimate limits of the system by identifying potential vulnerabilities and putting the system under stress in challenging scenarios, so as to help organisations plug any loopholes and identify any risks and issues before production launch.

[19] https://research.ibm.com/projects/adversarial-robustness-toolbox

Practice	Practice detail

	• Assess the ability to withstand or overcome adverse conditions under faults, attacks, or other anomalies, and prevent unexpected results and behavioural errors, while maintaining a minimum level of operational functionality.
	• Conduct assessments, which covers the following areas:
	  ○ Data - evaluate potential vulnerabilities in data processing, such as data poisoning[20] or manipulation. Assess data integrity, quality, and privacy protection measures;
	  ○ Model - assess model accuracy, fairness, resilience, and robustness of the AI model. Evaluate the risk of model leakage and the potential for unauthorised access or manipulation;
	  ○ Inputs - evaluate the system's resilience to unexpected and malicious inputs such as prompt injection[21]; and
	  ○ Output - evaluate potential for harmful, discriminatory, or objectionable outputs. Organisations should ensure the outputs comply with relevant regulations, including the Safeguarding National Security Ordinance, PD(P)O, Copyright Ordinance, and Anti-discrimination ordinances.
	• Conduct regular assessment and testing to analyse safety and security issues, and propose improvement plans.
	• Perform system audits covering applicable scenarios, safety and reliability of AI systems.

Perform dimension reduction	• Use dimensionality reduction on the training data to enhance the resilience of a classifier as a defence mechanism against evasion attacks. This means there are fewer data dimensions that can be used to falsify data. Dimensionality reduction refers to techniques that reduce the attributes of the data but still retain the meaningful properties of the original data. As an example, personal data attributes can include age, height, weight and race. • Checking the missing values ratio is a method to reduce the dimensionality in data. Data fields with numerous empty or null values are less likely to carry useful information. Therefore, data fields with the number of empty values greater than a certain threshold could be removed. The higher the threshold set, the more data reduction will be performed.
Example: Adversarial training (i.e. training a model using adversarial examples to defense against adversarial attacks) is often required for adherence to the principle of "Reliability,	

[20] https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/data-poisoning.html?context=wx&locale=en
[21] https://www.ibm.com/topics/prompt-injection

Practice	Practice detail
	Robustness and Security". To mitigate adversarial attacks, someone may use standard adversarial library (e.g. CleverHans) to train the AI model for adversarial image data. Alternatively, in an AI model for diagnosis that uses images, these images may be compressed. When the dimension of the images is reduced, whilst this leads to a reduction in image precision, it is harder for an attacker to manipulate the pixel values of medical images due to smaller resolutions. Please note that any libraries used should be subject to evaluation and approval before use in line with other IT programs

Related Ethical AI Principles:

1. Reliability, Robustness and Security – AI applications have to remain robust and secure with capabilities to maintain the integrity of information that constitutes it when confronted with possible adversarial attacks.

4.1.5 System Deployment

4.1.5.1 Model Integration & Impact

Definition:

Verification, validation and testing is the process of ensuring the AI applications perform as intended based on the requirements outlined at the beginning of the project. AI applications should be thoroughly tested before deployment to evaluate if the application breaks down and whether it performs as intended.

Who are involved at this stage?

Data Scientists, Project Managers

Practices and examples:

Practice	Practice detail
Perform integration, system, decision and User Acceptance Testing ("UAT")	• Formal and robust testing requirements for AI applications should be initiated prior to deployment (inclusive of AI models within the AI application) to validate they are appropriate for production (for example, requirements relating to fairness, transparency and explainability, error rates and consistency of performance). Please refer to o Section 4.1.4.3 "Data Extraction" for more details on fairness; and

Practice	Practice detail
	○ Section 4.1.2.2 "Project Oversight and Delivery Approach" for more details on transparency and explainability. • Formulate clear test rules and methods, including manual testing, automated testing, and hybrid testing and utilise technologies such as sandbox simulations to thoroughly test and verify models. • Additional testing to perform on the AI application itself include regression testing in the AI to identify model error. Regression testing can include testing of the AI application based on normal inputs to estimate generalisation error. • End-users (non-technical experts) should be involved in the testing process to ensure the AI application makes meaningful decisions for the organisations and provides better user interactions. • Standardised testing benchmarks or metrics for comparing candidate models or evaluating readiness for deployment can assist in selecting the optimal AI application. Please refer to Section 4.1.6.1 "Data and Model Performance Monitoring" for examples of performance metrics. • Stakeholders including the Project Team, business users and governance functions should review testing results and provide approval for AI assets prior to deployment into the production environment. This includes assessing the AI application's suitability for the objective of the organisation and that there is a process in place to produce reasons for the decisions or recommendations from the AI application when requested.
Enable edge cases and exception handling	• The Project Team should consider using edge cases to train the AI model on how it should respond when encountering edge cases. Edge cases typically occur when an AI application encounters unexpected scenarios where it may not perform as expected. • AI models should be developed and incorporated to handle unseen data, new use cases and potential malicious inputs. • As an example of edge cases, when using AI to measure footfall from videos, different examples should be verified (e.g. when a person is using an object such as a bike or a person is wearing a mask). Precautions should be taken to ensure that these situations are handled correctly. • Estimate the impact of your AI application when it provides inaccurate results and identify ways to minimise it. For example, if the AI application produces erroneous results, what is the likelihood that it causes harm to users? Increased human

Practice	Practice detail
	intervention can support to minimise the impact of the erroneous results.
Example: Integration, system, decision and UAT should be performed with different end-users and Project Team members prior to deployment to further the principle of "Cooperation and Openness" with adequate check-and-balance controls. When deploying a weather forecasting AI application, this application should be tested by a human forecaster to ensure that the output of the AI application is valid. The AI application should be evaluated for its functional requirements including weather prediction and non-functional requirements such as performance and usability.	

Related Ethical AI Principles:

1. Explainability – Project Team can explain the decisions of the AI application and factors that affect these decision outcomes during testing.
2. Reliability, Robustness and Security – AI applications should be tested for integration, user acceptance and handling of errors to ensure reliability and robustness.
3. Cooperation and Openness – Project Managers can work together with the Project Team as well as end-users to perform testing before deployment.

4.1.5.2 Transition & Execution

Definition:

Assuming the AI application would fail, mitigation steps should be incorporated to minimise damages in the case of failure.

Who are involved at this stage?

Data Scientists, Project Managers

Practices and examples:

Practice	Practice detail
Establish multiple layers of mitigation	• Multiple layers of mitigation in code can be established to stop system errors or failures at different levels or modules of the AI application. This can assist in earlier detection of errors to stop these errors affecting other parts of the AI application. For high-risk AI applications, drills should be performed to ensure that everyone is conscious of procedures to handle different stop system errors or failures appropriately. • Evaluate and enhance the ability of AI products and services to withstand or recover from adverse conditions due to faults, attacks or other anomalies. • The Project Team should incorporate defence or controls at the beginning or end of each solution step to prevent flow-on effects that could cause a significant system failure. For example, when data are being pre-processed, the processing could stop when there are errors and flag a warning to users. • Consider controls in place that allow for human intervention or auto-shutdown in the event of system failure and create a business continuity plan to address such undesirable events. The plan should include the response and recovery methods in the event of system failure. The system should incorporate fail-safe capability to enable rapid transition to human-based or traditional methods during system failures. • Organisations should also consider the business continuity plan, which identifies what are the processes to be taken when the model fails, needs to be taken offline, or decommissioned. The decommissioning could be because the model no longer works, is no longer relevant, or a newer and better system has been created to take its place. The business continuity plan ensures there is no impact on a process or function during the transition process. For some systems, this may include a parallel run while a new system is coming online before the old AI application is taken offline.
Apply business rules and handle outputs	• Organisations should implement business rules for quality assurance. For example, certain prediction applications can have additional evaluation performed around results to ensure predictions are sensible.

Practice	Practice detail
Implement model/Application Programming Interface ("API") access security practices	• To identify unexpected, untruthful, and inaccurate outputs via technological means, and handle them in accordance with laws and regulations including Safeguarding National Security Ordinance, PD(P)O, Copyright Ordinance, anti-discrimination ordinances. • AI applications should include core data security principles similar to any other IT systems. Security by design and data privacy should be embedded into the whole AI application design and development processes to protect the data and individual's right to privacy. • Establish service scope according to actual application scenarios and implement controls to prevent features beyond the scope through regular reviews. Promptly disable specific system features that are found to be potentially abused. • Restrict or control inputs to the AI applications by limiting API calls and using a subset of ensemble learning (using multiple detection methods) to serve API's or users can prevent AI model theft or AI model tampering. Ensemble learning is used to combine different detection methods (e.g. AI classifiers, detection rules and anomaly detection) to improve robustness. Attackers trying to manipulate the input system then have to avoid all these detection mechanisms to be successful in providing malicious input to the AI model through the AI application. • All users should be treated equally where possible. If organisation has to prioritise certain information or access to technologies differently for different users, this should be justified and documented as part of logging for access.

Practice	Practice detail

Provide information to users	• Create a disclosure document that could be made available to end-users or the public (e.g. on a website) as applicable. This is to encourage transparency and explainability practices to ensure the AI applications' decision-making process is comprehensible to human beings. The document should outline the details on the AI application's:
	  ○ operation and intended use;
	  ○ target users;
	  ○ data used and processed;
	  ○ model training procedures;
	  ○ testing procedures;
	  ○ capabilities and performance metrics; and
	  ○ checks performed to evaluate Ethical AI Principles (for example, reliability, robustness and security).
	• For third-party AI applications, organisations should request equivalent documentation from third parties to explain the AI application. Please refer to Section 4.1.3.2 "Procuring AI Services" for details.
	• Create disclaimer statement or agreement, together with the Terms of Use, that could be made available to end-users or the public (e.g. on a website) as applicable. The disclaimer should include:
	  ○ precautions;
	  ○ limitations;
	  ○ usage prohibitions;
	  ○ right to take corrective measures or terminate services early upon detecting misuse and abuse not conforming to usage intention and stated limitations; and
	  ○ applicable transparency requirements.
	• Create Personal Information Collection Statement and Privacy Policy to inform end-users about the use, rights, and consequences related to their personal data.
	• Publish policies internally and externally (as applicable) to distribute information about governance practices to disclose information such as third-party engagement, data ownership or compliance to industry standards.
	• Inform users upfront if there are interactions involving AI or the content are generated by AI. For example, notify users on a website or messaging application prior to their use of AI chatbots or adding watermarks, metadata, digital signatures, etc. in AI generated content. This is to ensure that end-users or other subjects are adequately aware that a decision, content, advice or outcome is the result of an AI application or algorithmic decision.

Practice	Practice detail
Example: Organisation should implement API access security practices to adhere to the "Reliability, Robustness and Security" principle. In the case of a video analytics application, inputs by users should be limited and have controls in place to ensure accuracy (whether the input is direct or through APIs).	

Related Ethical AI Principles:

1. Reliability, Robustness and Security – Following security practices and implementing multiple layers of mitigations can ensure security and reliability of the AI application by preventing incorrect outcomes.
2. Human Oversight/Safety – Controls should be implemented that allow for human intervention or auto-shutdown in the event of system failure especially when the system failure will have an impact on human safety.
3. Transparency and Explainability – The description of the AI applications algorithms should be accessible to support building trust in AI usage.
4. Beneficial AI – Provide disclosure statements on the benefits of the AI application to illustrate and promote the common good being achieved by the AI application (where applicable).

4.1.5.3 Ongoing Monitoring

Definition

Provide feedback to increase learning and robustness of the AI application.

Who are involved at this stage?

Data Scientists, Project Managers

Practices and examples:

Practice	Practice detail
Track mistakes	• Deploy an appropriate corrective action tracking mechanism (e.g. flagging certain situations for human review) to understand the mistakes that the AI application is making so that these mistakes can be resolved. • When an AI application has to make a decision for situations which are new and unfamiliar, consider implementing a flagging mechanism that will flag these situations for review. • Trigger points such as input data containing new values that were not available in training data can indicate that older training data are no longer valid. AI models can misinterpret information when trying to determine new unknowns from new data input types. Therefore, utilising these trigger points and putting corrective measures in place can limit these mistakes.
Implement an easy-to-use feedback user interface	• Enable end-users or the public (as applicable to the AI application) to share information about failed predictions or cases by designing user-friendly feedback forms. • Organisations can encourage end -users to communicate quality and timely feedback by offering incentives (for example, small gifts or rewards). • Alternatively, organisation can provide a hotline or email contact of relevant personnel that the end-users can reach out to on the organisation's website. • Establish a process to incorporate user's feedback when AI application maintenance activities are performed. For significant issues, this should be escalated through an existing support channel (e.g. incident management).
Example: For a medical diagnosis prediction application, an audit log can be deployed that captures all the medical diagnosis and predictions made. The prediction is then compared and benchmarked against actual results. The application can include a 'feedback' option for the end-user to input their feedback with messages to encourage users to provide responses. This aligns with the "Reliability, Robustness and Security" principle where continuous improvement to the AI application's reliability and robustness is maintained based on user's feedback.	

Related Ethical AI Principles:

1. Reliability, Robustness and Security – Track mistakes to be remediated and user feedback on a regular basis so that the reliability and robustness of the AI application can continually improve.

4.1.5.4 Evaluation & Check-in

Definition

Traceability, Repeatability and Reproducibility are required to ensure the AI application is operating correctly and to help build trust from the public and key stakeholders.

Who are involved at this stage?

Project Managers, Data Scientists

Practices and examples:

Practice	Practice detail
Ensure repeatable and reproducibility end-to-end workflow	• Ensure that the source-to-insights flow of the AI application is a repeatable process through automation. This reduces ad-hoc processes required when executing the AI application. The Project Team should conduct repeatability tests to ensure end-to-end workflow is replicable. • The Project Team should ensure the AI application's reproducibility. Reproducibility will ensure that the AI model produces the same results when the same datasets or methods of prediction were used. The reproducibility of an AI application is an important gauge to measure trustworthiness.
Enable traceability	• The Project Team should ensure that data are stored appropriately by creating audit trails to document the non-functional or functional requirements, model training and decision-making processes. This is to avoid alteration of data and audit trails should be retained based on applicable retention periods. • Consider establishing a traceability mechanism for the entire data pipeline to enforce accountability and allow tracking of data manipulation activities. An AI application is traceable when its decisions, datasets and processes or algorithm, evaluation parameters, outcomes and error logs are documented and tracked. • Ensure the AI application's logs and records (for example, timestamps of any scheduling tasks, count of data rows, the users involved and modelling interactions) are generated and stored securely. Please refer to Section 4.1.3.1 "Technology Roadmap for AI and Data Usage", Section 4.1.4.2 "Solution Design" and Section 4.1.4.5.1 "Model Assumptions" for contents related to different types of logging.
Ensure auditability	• The Project Team should retain evidence that enables the assessment of algorithms, processes and data lineage related to

Practice	Practice detail
	the AI application. This includes evidence from different stages (e.g. data collection, pre-points processing, model training, testing deploying) which supports auditability. All key decisions of the AI application should be documented. • The Project Team should record process flow and metadata management of the AI application in the logging module within the AI application. • To ensure that all documentation is up to date, a review of the logs should occur on a regular basis (for example, quarterly or yearly). For intermediate documentation, ensure that both module developers and integrators understand the rationale behind design decisions. • Please refer to Section B.1 of "Appendix B - Reference".
Example: Organisation should aim to ensure repeatable and reproducibility end-to-end workflow. For example, the collecting, consolidating, analysing and generating insights process for video analytics can be automated using a test script scheduled to run on a daily basis. Logging how results are produced can act as evidence to support identification of accountable areas to facilitate adherence to the accountability principle.	

Related Ethical AI Principles:

1. Lawfulness and Compliance – Relevant measures are taken to ensure industry standards are considered and that there is auditability.
2. Accountability – Logging can serve as evidence for any incidents that occur within the AI applications to support identification of accountable areas for incidents.

4.1.6 System Operation and Monitoring

4.1.6.1 Data and Model Performance Monitoring

Definition:

AI Models (as part of AI applications) should be continuously monitored and reviewed due to the likelihood of the AI models becoming less accurate and less relevant. This can happen when the data and the environment are continually changing with time.

Who are involved at this stage?

Project Managers, Data Scientists

Practices and examples:

Practice	Practice detail
Set performance metrics	• Define and validate the performance criteria for the AI application. This is to provide assurance to the organisation that the AI application continues to add value. Performance metrics can be communicated in management reporting. • A set of metrics relevant to the issue or decision being made by the AI should be developed with exception-based reporting across different subgroups of data (e.g. different product or end-user groups). • Examples of metrics such as precision, recall, F-score or accuracy can be calculated to measure the AI model's performance. This includes checking on the success of predictions and the level of inaccuracy that the predictions have to ascertain that the AI model is still performing in an acceptable manner. Refer to examples of bias in data in Section 4.1.4.3 "Data Extraction". • Organisations should define AI model indicators to assess whether the AI models are operating as designed to continuously meet their intended goals.
Incorporate quality assessments	• Documenting assumptions (on expected inputs, value distributions and boundaries) and integrating them into the AI model can help flag affected instances which could prevent full execution of the AI model and instances that could result in poor predictions. • Independent validation should be defined and be performed periodically by the PSC/PAT according to the risk of AI models. • Organisations should periodically assess the AI model indicators to determine whether their AI models are operating as designed. As AI technologies are developing, new risks will be introduced; and having a process to continually identify, review and mitigate new risks post deployment will contribute to the reliability, robustness and security of the AI application. • Quality assessments should serve as detection and response mechanisms for undesirable adverse effects of the AI application for the end-user or subject (for example, if an AI model is not performing as intended, this should be detected and resolved by retuning the model)
Create anomaly detection techniques	• The Project Team should create, implement and monitor anomaly detection techniques. Anomaly detection refers to identifying outliers or exceptions where the patterns deviate from the expected behaviour within the dataset. It is useful in identifying

Practice	Practice detail
	outliers that significantly deviate from the rest of the data and is a useful technique that can be used to prevent adversarial inputs.
Re-train AI model	• Organisations should perform regular tuning and re-training of their AI models with new data points. This can be automated or performed manually to retrain the AI model periodically when there are new and changing data points. AI Model tuning will be required as part of model maintenance or operational support and when there is feedback from the users of the AI application. • Organisations should record when the AI model and application are being updated, how it is being updated and how this affects the outputs of the AI application to ensure traceability and auditability. This should be factored into change management procedures for the AI application. • Examples of misclassifications in the older AI model can be included in every model tuning update to learn from the past true or positive errors. • Retraining of the AI model should undergo decision and user testing before deployment. This is to ensure that no new errors were introduced when the model is being retrained and updated. As this is an iterative process, continue gathering feedback from end-users upon deployment of the AI application for a continuous improvement of the AI model and application.
Example: Performance metrics can be set to assist adherence to the principle of "Reliability, Robustness and Security". If AI is assisting doctors to perform a diagnosis, they should record whether the AI was (in their view) accurate. This can be set as a performance metric to assess performance.	

Related Ethical AI Principles:

1. Reliability, Robustness and Security – Quality checks and performance metrics assist the organisation to ensure reliability of the AI applications.
2. Accountability – Ensure there are accountable roles to perform quality checks and monitoring of the AI applications.

4.1.6.2 Operational Support

Definition:

Upon AI applications deployment, ongoing operational support should be established to ensure that the AI applications performance remains consistent, reliable and robust.

Who are involved at this stage?

Project Managers, Data Scientists

Practices and examples:

Practice	Practice detail
Establish service-level-agreement (SLA)	• Define and agree on service levels between end-users and IT support (service availability, time to fix, severity thresholds) taking into account any AI application needs. An SLA defines the level of service expected by the end-user from the Project Team, laying out the metrics by which that service is measured. • The support processes should be integrated with any existing operational support processes such as the incident management process.
Create a feedback channel	• Enable mechanisms for users to provide feedback such as via email, website feedback and form input. Please refer to Section 4.1.5.3 "Ongoing Monitoring" for details.
Create or leverage an incident handling plan for dealing with crises situations	• The IT project team should create a new or leverage supplement on an existing incident handling plan for addressing crises arising from adversarial impacts or AI-related incidents, which refer to any event where the development, use, or malfunction of Ai systems leads to harms such as personal injury, disruption of critical infrastructure, or damage to property. This is especially critical for the AI applications applied in key sectors, such as critical information infrastructure and system directly affecting public safety and people's health and safety. • The incident handling plan should encompass key components such as monitoring mechanism to detect AI-related anomalies, reporting policies for suspected incidents, response protocols with clear roles and escalation path, investigation procedures and post-incident actions for remediation. Existing communication channels used by organisations for other IT applications should be used for this purpose. Organisations should promptly report safety and security incidents and vulnerabilities detected in AI system operations. • If there are any issues or disputes regarding the AI or its outputs, this is highly dependent on the model's functions itself and the organisations discretion on how to resolve. Organisations should always aim to incorporate user feedback in the AI application. This feedback is then taken into future training iterations of the model.

Practice	Practice detail
Example: SLAs can be established to set up accountable parties to act and adhere to the "Accountability" principle. When a weather forecasting AI application has errors and is not operating as intended, it should be restored by IT support within a stipulated timeframe.	

Related Ethical AI Principles:

1. Reliability, Robustness and Security – Ensuring that the AI application remains reliable, robust and secure over a prolonged period through acting on feedback and adhering to SLAs.
2. Accountability – IT support should be made accountable to fix the AI application within stipulated timeframes.

4.1.6.3 Continuous Review/Compliance

Definition:

The Continuous Review/Compliance function should be established to monitor and evaluate the AI application to ensure its adequacy, efficiency and effectiveness. The Project Manager, PSC/PAT and IT Board/CIO are responsible to monitor risks of AI such as noncompliance with applicable laws and regulations.

Who are involved at this stage?

Project Managers, IT Planners/Executives

Practices and examples:

Practice	Practice detail

Perform management/ continuous review	• Consider engaging independent auditors or consultants to perform audits/assessments for compliance.
	• Organisations' management should:
	  ○ Maintain an inventory of organisational policies related to the use of AI. Examples of policies related to AI can include security, data governance and procurement of tools;
	  ○ Perform a review of the policies for compliance issues with external legal and regulatory policies and discuss required changes. The timing of the review is dependent on the business requirements (for example, in some cases, an immediate review is required in order to reflect the new legal requirements). Required changes are determined through discussions and through comparison with industry standards and external regulations;
	  ○ Include subject matter experts in the field of ethics and corporate social responsibility to review and revise AI policies; and
	  ○ Create and implement standard risk assessment and business validation procedures for AI assets. Refer to Section 5 "AI Assessment" of this document.

Terminate AI application	• A termination process should be developed. • Properly handle data when terminating AI products and ensure all government data in external services or facilities are cleared or destroyed according to government security requirements. • Systematically revoke user privileges and data access rights when they become no longer required.
Consult IT Board/CIO	• For high-risk AI applications, there should be a review and challenge on Ethical AI Principles considerations (for example, fairness, explainability) and to identify any issues. • Please refer to Section 3.5.2 "AI Governance" for details.
Example: Organisations can consult IT Board/CIO who are responsible for performing reviews. This adheres to the principle of "Accountability". For high-risk cases such as video analytics and facial recognition, these should be reviewed by the IT Board/CIO.	

Related Ethical AI Principles:

1. Lawfulness and Compliance – Ensure compliance as new industry standards and regulations surrounding AI applications are being developed.
2. Accountability – Ensure management has accountability to perform a continuous review.

SECTION 5

AI ASSESSMENT

5. AI ASSESSMENT

AI Assessment suggested in this section provides a set of targeted questions (aligned to the AI Lifecycle) to assist organisations to assess, identify, analyse and evaluate the benefits and impacts of AI applications, to ensure they are meeting the intent of Ethical AI Principles and to determine the appropriate mitigation measures required to control any negative impacts within an acceptable level.

5.1 AI APPLICATION IMPACT ASSESSMENT

The AI Application Impact Assessment should be conducted on an AI application at different stages of the AI Lifecycle. The AI Application Impact Assessment introduces a systematic thinking process for organisations to go through different aspects of considerations of individual applications for their associated benefits and risks whilst highlighting the need for additional governance activities and identifying follow-up actions to ensure necessary measures and controls required for implementing ethical AI.

The AI Application Impact Assessment process includes the identification and analysis of:

- Processes related to the AI application and Ethical AI Principles that align with the AI Practice Guide requirements;
- Benefits that the AI application will bring;
- Negative impacts on specific stakeholders; and
- Establishment of required controls and compliance/monitoring processes.

The benefits of performing an AI Application Impact Assessment are to:

- enable better risk mitigation decisions that maximise benefits (through discussions on identified benefits, risks and impacts);
- provide a balanced view of existing AI application risks; and
- increase the organisations ability to reassure stakeholders on how AI application impact is managed.

The AI Application Impact Assessment template is used to both document and record how an AI application has met the requirements outlined in the AI Practice Guide internal requirements and to facilitate the review and approval process with the considerations made on the approval decision. The AI Application Impact Assessment template is also used to evaluate the risk that an application poses and assist organisations to map and develop specific governance requirements to mitigate those risks.

The AI Application Impact Assessment template used for this assessment is provided in Appendix C “AI Application Impact Assessment Template” in the Ethical AI Framework document for reference.

The AI Application Impact Assessment has the following components:

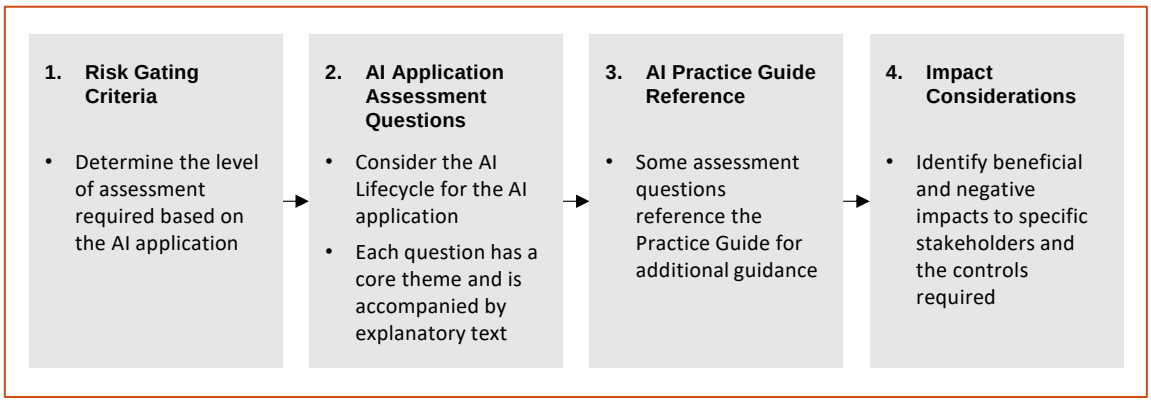


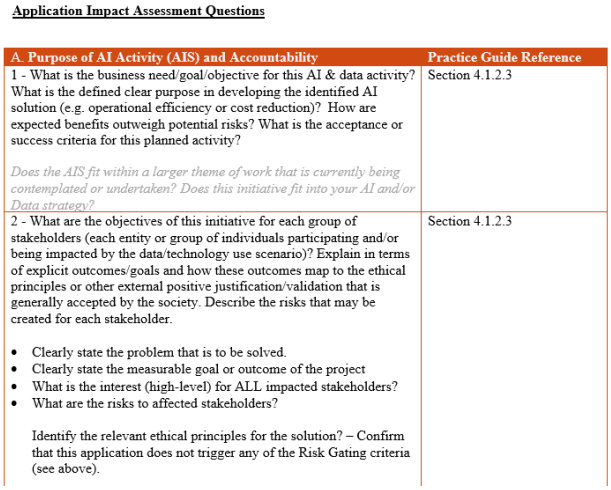
Figure 12: AI Application Impact Assessment Components

1. **Risk Gating Criteria –** A set of questions that are used to distinguish high-risk AI applications. These questions should be completed at the beginning of a proposed AI project or upon conditions of the AI application are changed. AI applications which are considered high-risk would subsequently require review and approval by IT Board/CIO. A sample of the risk gating questions is shown below. Regardless of the answers to risk gating criteria, project team should complete the other parts of the AI Application Impact Assessment.

Question	Context
Is the AI solution within an area of intense public scrutiny (e.g. because of privacy concerns) and/or **frequent** litigation? Or is it a significant new form of technology or involve the combining of different technology and use?	*For example, certain "Internet of Things" applications could have a significant impact on individuals' daily lives and privacy; and therefore, require a comprehensive AIA. The combination of Facial Recognition and a new, sensitive use or potentially controversial use*
Is the AI applied in a new (social) domain? Is the AI applied in a domain where it has not been used before?	*For example, an application that is used for the first time in healthcare while previously, it was only used for marketing purposes. Due to the change of domain, it is possible that the application will rise (new) ethical questions. When the application takes place in a sensitive social area, the risks and the ethical issues are potentially greater. Think of topics such as care, safety, the fight against terrorism or education. Think also of vulnerable groups such as children or the disabled*

Figure 13: Sample of Risk Gating Questions

2. **AI Application Impact Assessment Questions –** The questions are provided to ensure that the impact of the AI application is identified and managed across the AI Lifecycle stages and that related Ethical AI Principles have been considered. The questions consider the impact of the AI applications which includes benefits, risks, the effects on individuals' rights and the balancing of different interests. Answers provided by the Project Team will be assessed qualitatively. A sample of the AI Application Impact Assessment Questions is shown in Figure 14.



Application Impact Assessment Questions

A. Purpose of AI Activity (AIS) and Accountability	Practice Guide Reference
1 - What is the business need/goal/objective for this AI & data activity? What is the defined clear purpose in developing the identified AI solution (e.g. operational efficiency or cost reduction)? How are expected benefits outweigh potential risks? What is the acceptance or success criteria for this planned activity? *Does the AIS fit within a larger theme of work that is currently being contemplated or undertaken? Does this initiative fit into your AI and/or Data strategy?*	Section 4.1.2.3
2 - What are the objectives of this initiative for each group of stakeholders (each entity or group of individuals participating and/or being impacted by the data/technology use scenario)? Explain in terms of explicit outcomes/goals and how these outcomes map to the ethical principles or other external positive justification/validation that is generally accepted by the society. Describe the risks that may be created for each stakeholder. • Clearly state the problem that is to be solved. • Clearly state the measurable goal or outcome of the project • What is the interest (high-level) for ALL impacted stakeholders? • What are the risks to affected stakeholders? Identify the relevant ethical principles for the solution? – Confirm that this application does not trigger any of the Risk Gating criteria (see above).	Section 4.1.2.3

Figure 14: Sample of AI Application Impact Assessment Questions

3. **AI Practice Guide Reference –** Assessment questions have included a reference (where applicable) to relevant sub-sections of Section 4 "AI Practice Guide" that connect the specific area being assessed with the practices.
4. **Impact Considerations –** A set of questions about beneficial and negative impacts on stakeholders are included in the AI Application Impact Assessment template. This guides the evaluation of the need for further mitigating actions from the organisation (e.g. additional controls or processes required to manage the impact).

5.1.1 Process for Completing AI Application Impact Assessment

The process for completing AI Application Impact Assessment is listed below.

1. The Project Team within an organisation should complete an AI Application Impact Assessment template starting from the planning phase of the project (please refer to Section 5.2 "Frequency of AI Assessment" for further details on when the assessment should be completed).
    i. The first stage for the Project Team is to complete answers to the Risk Gating Criteria questions (to assess the applicability of any of the outcomes listed) and to complete the "Purpose of AI Activity and Accountability" section of the AI Application Impact Assessment questions.
    ii. Where the answer to any of the Risk Gating Criteria questions is "Yes", the AI Project is considered high-risk and endorsement should be sought from the IT Board/CIO before the AI project can be commenced.
    iii. Where the answers to all of the Risk Gating Criteria questions are "No" then endorsement from the IT Board/CIO is not required. Endorsement is still required from the PAT/PSC.
    iv. The remainder of the AI Application Impact Assessment questions should be completed by the Project Team as the project progresses through the different stages of the AI Lifecycle. This should occur whether the Risk Gating Criteria have been triggered or not. AI applications can still have risks when not triggering the Risk Gating Criteria and the AI Application Impact Assessment will help the Project Team ensure that impacts are considered.

v. Organisations should respond to every question on the AI Application Impact Assessment. Some questions may not require a detailed response where they are not applicable (e.g. for the assessment section "AI Governance Process – Third-Party Questions" when there are no third parties involved in the AI project).

vi. Where AI practices are not followed (as specified in the AI Practice Guide) and negative impacts are detected, organisations should provide a mitigation plan or document in the assessment why they feel that the project is still appropriate to continue. The decision to be reached by the PSC/PAT is a go-no-go decision at each phase. For example, if there are practices that have not been performed for the areas assessed, or further tasks are required relative to a specific practice requirement (e.g. bias testing), there should be a decision made based on benefits, risks and impacts to determine whether the project can proceed or remediation needs to be performed first.

2. The AI Application Impact Assessment can be used as a 'live' document throughout the AI development lifecycle for capturing risks and mitigations that contribute to the go or no-go decision of the AI development.

3. A detailed review of the full assessment and a collaborative review of the "Impact" section should be completed by the Project Team. Additionally, this review should include a review of whether any other risk gates have been triggered requiring approval escalation.

4. Results of the AI Application Impact Assessment should be reviewed and approved by the PSC/PAT. The PSC/PAT should review the business cases, potential impacts and plan to address if the project is to proceed.

5.2 FREQUENCY OF AI ASSESSMENT

An AI Application Impact Assessment should be conducted regularly (e.g. annually) or when major changes take place as AI projects progress and when the AI application is being operated.

The stages of the AI Lifecycle where AI Application Impact Assessment should be reviewed are shown in Figure 15.

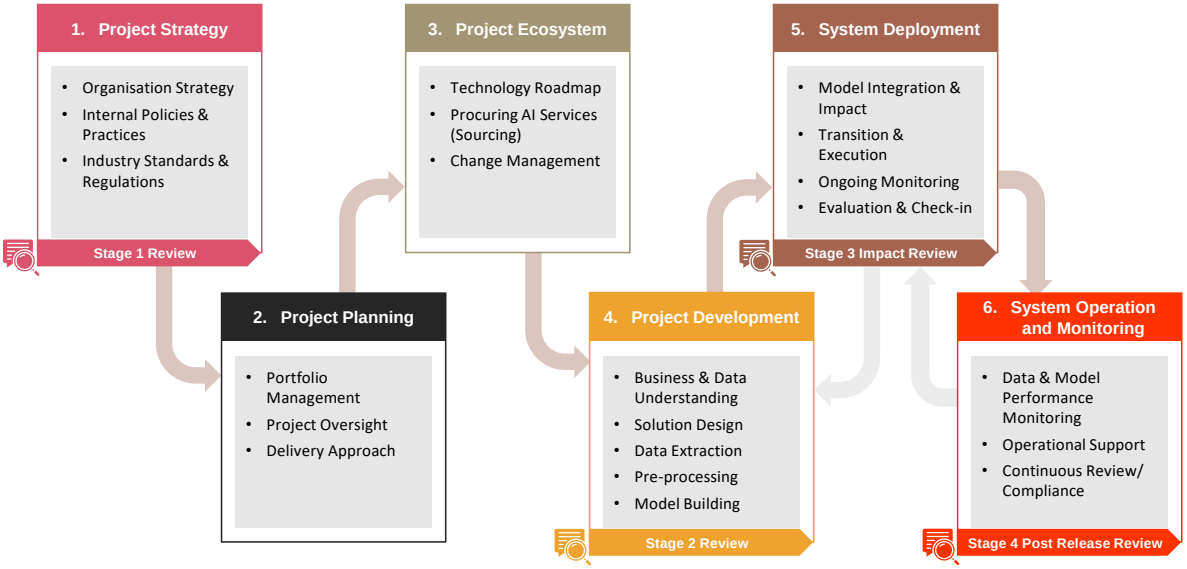


Figure 15: Stages for AI Application Impact Assessment

For new AI applications, organisations should complete the AI Application Impact Assessment and review in the Project Planning, Project Development, System Deployment and 'System Operation and Monitoring' AI Lifecycle stages. These serve as checkpoints to ensure necessary requirements are identified and incorporated in other subsequent AI Lifecycles stages appropriately. The AI Application Impact Assessment can be used as a 'live' document throughout the AI Lifecycle, but the associated AI Application Impact Assessment should be reviewed at 4 key stages of the AI Lifecycle (please refer to Figure 15) with a copy of the AI Application Impact Assessment being retained for historical records.

Lifecycle stage	Responsible party	Actions to be performed
SDLC: Project Request **AI Lifecycle:** Project Strategy	Project Team	Categorise the project as "high-risk" or "non-high-risk" based on answers to the "Risk Gating questions" Conduct AI Application Impact Assessment • Answer questions 1-8, 9(i), 41-49, 57-58
	PSC/PAT	Review and endorse the assessment for non-high-risk project
	IT Board/CIO	Review and endorse the assessment for high-risk project
SDLC: System Analysis and Design **AI Lifecycle: Project Ecosystem** Project Development	Project Team	Conduct AI Application Impact Assessment • Answer questions 9(ii), 10-14, 18-19, 22-29, 30(i)(ii), 31-34, 37(i), 52-56 • Review and update the answers of questions 1-8, 9(i), 41-49, 57-58 according to the latest position • If third-party technology or data is used, complete and review questions 15-17, 48 before procurement
	PSC/PAT	• Review and endorse the assessment
SDLC: Implementation (before rollout) **AI Lifecycle:** System Deployment	Project Team	Conduct AI Application Impact Assessment • Answer questions 20, 21, 30(iii), 35, 36, 37(ii)(iii), 38-40, 50, 51 • Review and update the AI Application Impact Assessment according to the latest position (questions 1-14, 18-19, 22-29, 30(i)(ii), 31-34, 37(i), 41-49, 52-58) • If third-party technology or data is used, review questions 15-17, 48
	PSC/PAT	• Review and endorse the assessment

Lifecycle stage	Responsible party	Actions to be performed
SDLC: System Maintenance (annually or when major changes take place) **AI Lifecycle:** System Operation and Monitoring	Maintenance Team	• Review and update the "AI Application Impact Assessment" according to the latest position • Escalate any significant issues
	Maintenance Board IT Board/CIO (or its delegates)	• Handle escalation • Monitor high-risk projects

Table 5: Actions to be performed for completing and reviewing the AI Application Impact Assessment

5.3 RECOMMENDATION

Operationalising AI in organisations requires establishing a baseline of clear and mutual understanding across several dimensions, from strategy through to continuous review/compliance. The Ethical AI Framework provides structure and best practices across the stages of AI Lifecycle. The practices are aligned with the AI Application Assessment. By virtue of completing the AI Application Impact Assessment, gaps in various processes in the AI Lifecycle can be identified.

As mentioned above, the AI Assessment help identifying the gaps in an organisation's roadmap of adoption and operationalisation of AI and provide guidance for establishing quality processes that can guide organisations to harness the power of AI in a responsible and ethical manner.

A final approval on the go or no-go decision for AI deployment should be provided by the PSC/PAT (with endorsement by the IT Board/CIO as appropriate) based on the consideration of having an appropriate balance of benefits and mitigated risks that the AI applications pose.

SECTION 6

APPENDIX

6. APPENDIX

APPENDIX A – GLOSSARY

List of terms and definitions used in this document

Term	Definition
Class label	Class label is the distinct attribute/feature whose value will be predicted based on the values of another attribute/feature in the dataset. In other words. Class is the category where the data will be classified based on the common property that the data has with other sets of data within a category while label is the outcome of the AI model's classification process.
Clustering	Machine learning algorithm that involves grouping similar data points together.
Data lake	Centralised repository that stores structured and unstructured data.
Data lineage	Data lineage describes the transformation of data over time right from the beginning of its creation.
Data mart	Subset of data warehouse designed for a specific business domain such as finance and operations.
Data warehouse	Centralised and large repository, usually designed for analytics purposes and aggregates data from various system sources
Decision tree	An algorithm that uses the tree representation to solve the problem where each leaf node represents class label and the internal nodes of the tree represent each attribute.
Human-in-the-loop	Human-in-the-loop refers to the capability for human intervention in every decision cycle of the system.
Human-in-command	Human-in-command refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation.
Human-out-of-the-loop	Human-out-of-the-loop refers to the capability of the AI system in making decisions without human intervention
K-fold	Cross validation technique where the original sample is randomly partitioned into equal k sized subsamples.
Logistic regression	A type of classification algorithm used to predict the binary outcome based on a set of independent values.
Personally identifiable information ("PII")	Data or information that can be used to identify an individual such as identification number, biometric and address.

Term	Definition
Random oversampling	Technique that involves duplicating dataset randomly from the minority class and adding them back to the original training dataset.
Random resampling	Technique that involves creating a new version of training dataset that has a different class distribution. This technique aims to achieve a more balanced dataset in the new training dataset.
Random undersampling	Technique that involves selecting dataset randomly from the majority class to delete from the original training dataset.
Random forest	A classification method that operates by constructing decision trees during training stage and output the class that is the mode of the class or mean/average prediction of the individual trees.
Regression	Supervised machine learning technique used to make prediction by estimating the relationship between variables.
Regression testing	Testing performed to confirm the recent code/programme changes does not affect existing AI application's performance negatively.
Reinforcement learning	Reinforcement learning is the training of AI models to make decisions when dealing with problems and learn by reward and punishment based on feedback from its own actions.
Risk gating criteria	A set of questions that are used to distinguish high-risk AI applications. These questions should be completed at the beginning of a proposed AI project or upon conditions of the AI application are changed.
Surrogate model	A surrogate model is an engineering method used when the predictions of an AI model cannot be easily understood or measured, so a model of the outcome is used instead.
Unseen data	Data which are new and have never been 'seen' by the AI model

APPENDIX B – REFERENCE

Section B.1 EXAMPLES OF RELEVANT INDUSTRY STANDARDS

Listed below are published or under development standards relevant to AI.

Organisation/ Institute	Standard	Description
International Organisation for Standardisation (ISO)	ISO/IEC 22989:2022	Information technology - Artificial intelligence – Artificial intelligence concepts and terminology
	ISO/IEC 23053:2022	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
	ISO/IEC 20546:2019	Information technology — Big data — Overview and vocabulary
	ISO/IEC 20547-1:2020	Information technology — Big data reference architecture — Part 1: Framework and application process
	ISO/IEC TR 20547-2:2018	Information technology — Big data reference architecture — Part 2: Use cases and derived requirements
	ISO/IEC 20547-3:2020	Information technology — Big data reference architecture — Part 3: Reference architecture
	ISO/IEC 20547-4:2020	Information technology — Big data reference architecture — Part 4: Security and privacy
	ISO/IEC TR 20547-5:2018	Information technology — Big data reference architecture — Part 5: Standards roadmap
	ISO/IEC 24668:2022	Information technology — Artificial intelligence —Process management framework for Big data analytics
	ISO/IEC TR 24027:2021	Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making
	ISO/IEC TR 24028:2020	Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
	ISO/IEC TR 24029-1:2021	Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
	ISO/IEC TR 24368:2022	Information technology — Artificial intelligence — Overview of ethical and societal concerns

Organisation/ Institute	Standard	Description
	ISO/IEC 23894:2023	Information Technology — Artificial Intelligence — Risk Management
	ISO/IEC TR 24030:2024	Information technology — Artificial Intelligence (AI) — Use cases
	ISO/IEC 38507:2022	Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organisations
	ISO/IEC 24029-2:2023	Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods
	ISO/IEC 42001:2023	Information technology — Artificial intelligence — Management system
	ISO/IEC 5259-1:2024	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples
	ISO/IEC 5259-2:2024	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures
	ISO/IEC 5259-3:2024	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines
	ISO/IEC 5259-4:2024	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework
	ISO/IEC 5259-5:2025	Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance framework
	ISO/IEC TR 24372:2021	Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems
	ISO/IEC TR 27563:2023	Security and privacy in artificial intelligence use cases — Best practices
	ISO/IEC TR 5469:2024	Artificial intelligence — Functional safety and AI systems
	ISO/IEC TS 25058:2024	Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems
	IEEE P7000™-2021	Model Process for Addressing Ethical Concerns During System Design

Organisation/ Institute	Standard	Description
Institute of Electrical and Electronics Engineers (IEEE)	IEEE P7001™-2021	Transparency of Autonomous Systems
	IEEE P7002™-2022	Data Privacy Processes and Methodologies
	IEEE P7003™	Algorithmic Bias Considerations
	IEEE P7004™	Standard on Child and Student Data Governance
	IEEE P7005™-2021	Standard for Transparent Employer Data Governance
	IEEE P7006™	Standard on Personal Data AI Agent Working Group
	IEEE P7007™-2021	Ontological Standard for Ethically driven Robotics and Automation Systems
	IEEE P7008™	Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
	IEEE P7009™-2024	Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
	IEEE Std 7010™-2020	IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being
	IEEE P7011™	Standard for the Process of Identifying & Rating the Trustworthiness of News Sources
	IEEE P7012™	Standard for Machine Readable Personal Privacy Terms
	IEEE P7014™-2024	Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems
	IEEE 3119-2025	Standard for the Procurement of Artificial Intelligence and Automated Decision Systems
China National Standards (GB)	GB 45438-2025	Cybersecurity technology—Labeling method for content generated by artificial intelligence 网络安全技术 人工智能生成合成内容标识方法
	GB/T 45652-2025	Cybersecurity technology—Security specification for generative artificial intelligence pre-training and fine-tuning data 网络安全技术 生成式人工智能预训练和优化训练数据安全规范

Organisation/ Institute	Standard	Description
	GB/T 45654-2025	Cybersecurity technology—Basic security requirements for generative artificial intelligence service 网络安全技术 生成式人工智能服务安全基本要求
	GB/T 45674-2025	Cybersecurity technology—Generative artificial intelligence data annotation security specification 网络安全技术 生成式人工智能数据标注安全规范
	GB/T 45958-2025	Cybersecurity technology—Security framework for artificial intelligence computing platform 网络安全技术 人工智能计算平台安全框架
	GB/T 45081-2024	Artificial intelligence—Management system 人工智能 管理体系
	GB/T 45628-2025	Artificial intelligence—Knowledge graph—Knowledge exchange protocol 人工智能 知识图谱 知识交换协议
	GB/T 45288.1-2025	Artificial intelligence—Large-scale model—Part 1: General requirements 人工智能 大模型 第1部分：通用要求
	GB/T 45288.2-2025	Artificial intelligence—Large-scale model—Part 2: Testing and evaluation for metrics and methods 人工智能 大模型 第2部分：评测指标与方法
	GB/T 45288.3-2025	Artificial intelligence—Large-scale models—Part 3: Service capability maturity assessment 人工智能 大模型 第3部分：服务能力成熟度评估
	GB/T 45280-2025	Artificial intelligence—Unified interfaces for heterogeneous artificial intelligence accelerating units 人工智能 异构人工智能加速器统一接口
	GB/T 45225-2025	Artificial intelligence—Deep learning algorithms evaluation

Organisation/ Institute	Standard	Description
		人工智能 深度学习算法评估
	GB/T 45401.1-2025	Artificial intelligence—Scheduling and cooperation for computing devices—Part 1: Virtualization and scheduling 人工智能 计算设备调度与协同 第1部分：虚拟化与调度
	GB/T 45401.2-2025	Artificial intelligence—Scheduling and cooperation for computing devices—Part 2: Framework for distributed computing 人工智能 计算设备调度与协同 第2部分：分布式计算框架
	GB/T 45079-2024	Artificial intelligence—Technical specification for deep learning framework adaption to multi-hardware platform 人工智能 深度学习框架多硬件平台适配技术规范
	GB/T 45087-2024	Artificial intelligence—Management system 人工智能 服务器系统性能测试方法
	GB/T 46347-2025	Artificial intelligence—Risk management capability assessment 人工智能 风险管理能力评估
	GB/T 41867-2022	Information Technology Artificial Intelligence Terminology 信息技术 人工智能 术语
	GB/T 42018-2022	Information Technology Artificial Intelligence Platform Computing Resource Specifications 信息技术 人工智能 平台计算资源规范
	GB/T 42131-2022	Artificial Intelligence Knowledge Graph Technical Framework 人工智能 知识图谱技术框架
	GB/T 40691-2021	Artificial intelligence—Affective computing user interface—Model 人工智能 情感计算用户界面 模型
	GB/T 43782-2024	Artificial intelligence—Technical requirements for machine learning system 人工智能 机器学习系统技术要求

Organisation/ Institute	Standard	Description
	GB/T 42755-2023	Artificial intelligence—Code of practice for data labeling of machine learning 人工智能 面向机器学习的数据标注规程
	GB/T 41818-2022	Information technology Big data Analysis-oriented data storage and retrieval technical requirements 信息技术 大数据 面向分析的数据存储与检索技术要求
	GB/T 38675-2020	Information technology—General requirements for big data computing systems 信息技术 大数据计算系统通用要求
	GB/T 38643-2020	Information technology—Big data—Functional testing requirements for analytic system 信息技术 大数据 分析系统功能测试要求

Section B.2 Policy / Guidelines / Practices

Listed below are published Regulations / Laws / Policy / Guideline / Practices relevant to AI.

AI projects must also comply with other laws and regulations, including but not limited to:

- Privacy (Cap. 486 Personal Data (Privacy) Ordinance);
- Intellectual Property Rights (Cap. 528 Copyright Ordinance, Cap. 544 Prevention of Copyright Piracy Ordinance, Cap. 559 Trade Marks Ordinance, Cap. 362 Trade Descriptions Ordinance, Cap. 514 Patents Ordinance, Cap. 522 Registered Designs Ordinance);
- Anti-discrimination Ordinances (Cap. 480 Sex Discrimination Ordinance, Cap. 487 Disability Discrimination Ordinance, Cap. 527 Family Status Discrimination Ordinance, Cap. 602 Race Discrimination Ordinance);
- Basic Law of the Hong Kong Special Administrative Region of the People's Republic of China;
- Safeguarding National Security Ordinance; and
- Law of the People's Republic of China on Safeguarding National Security in the Hong Kong Special Administrative Region

Policy / Guidelines / Practices

- 香港生成式人工智能技術及應用指引

  Hong Kong Generative Artificial Intelligence Technical and Application Guideline

Other relevant guidelines:

Office of the Privacy Commissioner for Personal Data

- Guidance on Ethical Development and Use of AI
- 10 TIPS for Users of AI Chatbots
- Artificial Intelligence: Model Personal Data Protection Framework
- Checklist on Guidelines for the Use of Generative AI by Employees

Hong Kong Monetary Authority

- High-level Principles on Artificial Intelligence
- Consumer Protection in respect of the Use of Big Data Analytics and Artificial Intelligence by Authorised Institutions
- Consumer Protection in respect of Use of Generative Artificial Intelligence
- Use of Artificial Intelligence for Monitoring of Suspicious Activities

Judiciary

- Guidelines on the Use of Generative Artificial Intelligence for Judges and Judicial Officers and Support Staff of the Hong Kong Judiciary

Financial Services and the Treasury Bureau

- Policy Statement on Responsible Application of Artificial Intelligence in the Financial Market

Securities and Futures Commission

- Use of generative AI language models

Regulations / Laws for AI practices in the Chinese Mainland

- 生成式人工智能服务管理暂行办法

  Interim Measures for the Management of Generative AI Services
- 人工智能生成合成内容標識辦法

  Measures for Labeling of AI-Generated Synthetic Content

Framework / Guidelines in the Chinese Mainland

- 人工智能安全治理框架

  AI Safety Governance Framework
- 中國關於加強人工智能倫理治理的立場文件

  Position Paper of the People's Republic of China on Strengthening Ethical Governance of AI
- 全球人工智能治理倡議

  Global AI Governance Initiative

APPENDIX C – AI APPLICATION IMPACT ASSESSMENT TEMPLATE

Note

Please refer to Section 5.1.1 "Process for Completing AI Application Impact Assessment" for details on the process to complete the AI Application Impact Assessment

Legend

- Text covers the core questions that are to be addressed in a qualitative manner
- Italicised text is added context to support the core question and should be used as an aid to provide the qualitative answer

Risk Gating Criteria

If one or more of the risk gating questions below are triggered, (i.e. the answer is "yes" for the question), the AI Application Impact Assessment should be subject to IT Board/CIO review before project commencement. Please refer to Section 5 "AI Assessment" for details.

Question	Context
a - Is the AI application within an area of intense public scrutiny (e.g. because of privacy concerns) and/or frequent litigation?	*For example, certain "Internet of Things" applications could have a significant impact on individuals' daily lives and privacy. Examples of such applications include Smart Cities applications:* • *Smart Lighting – intelligent weather adaptive street lighting* • *Smart Traffic Management – Monitoring of vehicles and pedestrian* • *Smart Parking – Monitoring of parking spaces* *Therefore, this requires an IT Board/CIO review of the AI Application Impact Assessment. The combination of facial recognition and a new, sensitive use or potentially controversial use should be considered.*
b - Is the AI application applied in a new (social) domain (i.e. where AI has not been used in Hong Kong)?	*For example, an AI application that is used for the first time in healthcare while previously, it was only used for marketing purposes. Due to the change of domain, it is possible that the AI application will raise (new) ethical questions. When the AI application takes place in a sensitive social area, the risks and the ethical issues are potentially greater. Think of topics such as care, safety, the fight against terrorism or education. Think of vulnerable groups such as children or the disabled.*
c (i) Does the AI application have a high degree of autonomy?	*The more an AI application acts independently with increased freedom to make decisions, the more important it is to properly analyse the consequences of this autonomy. In addition to the freedom to make decisions,*

Question	Context
If the answer to question is 'Yes', please proceed to the question (ii). (ii) Is it used in a complex environment or a critical information infrastructure? If the answer to question (ii) is 'Yes', please proceed to the question (iii). (iii) Does the AI application make automated decisions that have a significant impact on persons or entities or that have legal consequences for them? If the answer to the question (iii) is 'Yes', such application will likely be considered as higher risk application.	*autonomy can also lie in the possibility of selecting data sources autonomously.* *When the AI application is situated in a complex environment, the risks are greater than when the AI application is in a confined environment.* *When the AI application makes decisions automatically (without human intervention) and the decision can lead to someone experiencing legal consequences of that decision or being significantly affected otherwise, the risk is greater. Think of not being able to get a mortgage, losing your job, a wrong medical diagnosis or reputational damage due to a certain categorisation that can lead to the exclusion or discrimination against individuals. Processing with little or no effect on individuals does not match this specific criterion.* *Examples of such AI applications are autonomous vehicle, autonomous military drones and surgical robots. Such AI applications are considered being used in a complex environment because the decisions made by the AI depend on various surrounding environment factors (e.g. surrounding human activities). Applications such as the autonomous vehicle and autonomous military drones are considered to have a high degree of autonomy because decisions were made entirely by the AI.*
d - Is sensitive personally identifiable information used?	*When sensitive personally identifiable information is used in the development and/or deployment of AI applications, the risk is higher. For example, data consisting of racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, biometric data, data concerning health or data concerning a natural person's sex life or sexual orientation.*
e - Does the AI application make complex decisions?	*As the decision-making by the AI application is more complex (for example, more variables or probabilistic estimates based on profiles) the risks increase. Simple AI applications based on a limited number of choices and variables are less risky. If the way in which an AI application has come to its decisions can no longer be (fully) understood or traced back to people, then the risks resulting from the decision are potentially greater.*
f - Does the AI application involve systemic observation or monitoring?	*Processing used to observe, monitor or control individuals, including data collected through networks or "a systematic monitoring of a publicly accessible area". Examples would include widespread video surveillance data and network behavioural tracking. This type of monitoring is a criterion because the personal data may be collected in circumstances where individuals may not be aware of who is collecting their data and how they will*

Question	Context
	be used. Additionally, it may be impossible for individuals to avoid being subject to such processing in public (or publicly accessible) space(s).
g - Does the AI application involve evaluation or scoring of individuals?	*Including profiling and predicting, especially from "aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements".* *Examples of this could include:* • *Financial institution that screens its customers against databases for credit referencing, anti-money laundering ("**AML**"), counterterrorism or fraud checks;* • *Biotechnology company offering genetic tests directly to consumers in order to assess and predict the disease/health risks; or* • *Company building behavioural or marketing profiles based on usage or navigation on its website.*
h - Are personal data processed on a large scale and/or are data sets combined?	*While "large scale" is difficult to define, consider the following factors, when determining whether the processing is carried out on a large scale:* *a. the number of individuals concerned, either as a specific number or as a proportion of the relevant population;* *b. the volume of data and/or the range of different data items being processed;* *c. the duration, or permanence, of the data processing activity; or* *d. the geographical extent of the processing activity.*

AI Application Impact Assessment Questions

A. **Purpose of AI Application and Accountability**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
1 - What is the business need/goal/objective for this AI and data activity? What is the defined clear purpose for developing the identified AI application (e.g. operational efficiency or cost reduction)? How can the expected benefits outweigh potential risks? What are the acceptance or success criteria for this planned activity? *Does the AI application fit within a larger theme of work that is currently being contemplated or undertaken? Does this initiative fit into your AI and/or Data strategy?*	Section 4.1.2.1: • Map AI projects to the organisation objectives • Select and prioritise AI projects IT Planners/Executives, Business Users	
2 - What are the objectives of this initiative for each group of stakeholders (each entity or group of individuals participating and/or being impacted by the data or technology use scenario)? Explain the objectives in terms of the explicit outcomes/goals and how these outcomes map to the Ethical AI Principles or other external positive justification/validation that is generally accepted by society. Describe the risks that may be created for each stakeholder. • Clearly state the problem that is to be solved. • Clearly state the measurable goal or outcome of the project • Who are the stakeholders who are impacted by this AI application? • What is the interest (high-level) for ALL impacted stakeholders? • What are the risks to affected stakeholders? (To be analysed and documented in question 3) Identify the relevant Ethical AI Principles for the AI application – Confirm that this AI Application does not trigger any of the risk gating criteria (see above).	Section 4.1.2.1: • Map AI projects to the organisation's objectives • Select and prioritise AI projects IT Planners/Executives, Business Users	

A. **Purpose of AI Application and Accountability**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
Who are the stakeholders? Identify and list each group. (Stakeholder examples: External stakeholders: users, indirectly affected public, data providers, AI vendors, regulators; Internal stakeholders: employees, project steering committees, IT Board, CIO, business units). What is the initial project impact or implication to each of the Ethical AI Principles: (1) Fairness, (2) Diversity and Inclusion, (3) Human Oversight, (4) Lawfulness and Compliance, (5) Data Privacy, (6) Safety, (7) Accountability, (8) Beneficial AI, (9) 'Cooperation and Openness' and (10) 'Sustainability and Just Transition'. *Objectives should be explicitly describable for each stakeholder and map to some externally validated objective (e.g. broader public policy objective). Consider objectives such as better or lower cost health care, greater access to health services, or better health outcomes or an improved ability to track and assess health outcomes; more accurate sensors or devices to detect or diagnose health conditions or to improve general wellness; improved education; environmental enhancements such as water conservation, energy cost reduction; infrastructure enhancements; economic improvement; more accessible/usable technology; increased job opportunities; protection of reasonable expectation of privacy, including anonymity; protection of rights or safeguards against discrimination.*		
3 - What are the potential risks to each stakeholder and which risk gating criteria(s) have been triggered? Considering all the factors relating to the AI application, what are the risks (real and/or perceived) to each identified stakeholder? *Explain the high-level potential impact and/or concern that the AI project and*	Section 4.1.2.2: • Assess the general criteria that qualify an AI project for a review by the IT Board/CIO IT Planners/Executives, Business Users	

A. **Purpose of AI Application and Accountability**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
application could create and what risks the project could create. If the project triggers any of the risk gating criteria, it will require initial approval before project commencement by the IT Board/CIO. (see above table for the risk gating criteria) *Consider the risks or increase in risks (real or perceived) to the identified stakeholder as a result of the application of the AI. Areas to consider include: perception of technology or data about them being used in an impactful way, an impact on the employee relationship, reduced status and/or well-being; damage to reputation or embarrassment; shock or surprise at the processing activity or the results of the processing; inappropriate discrimination, the possibility of inappropriate access to or misuse of information (e.g. insights or predicative data) by the organisation, including sensitive categories of data and directly identifiable data; manipulation of needs or desires/wants of the individual (i.e. creation of a need where one previously did not exist); a negative impact of the technology through a probability-based process, such as a score; Who will have access to information on the AI application and who won't? Will stakeholders who do not have access to this information or data or the insight suffer a setback compared to those who do? What does that setback look like? What new differences will there be between the "haves" and "have-nots" of this information? Would individual stakeholders be surprised by the activity related to them? Would the information use about individuals align with their perception of whether this data/information should be used this way? Determine whether there are other sensitivity issues with the potential use of insights and what aspect of use of potential insights might be considered unfair to the stakeholder. Are all stakeholders treated equally?*		

A. **Purpose of AI Application and Accountability**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
4 - Have all the governance processes and roles outlined in the AI Practice Guide been satisfied and who specifically will be fulfilling each requirement? *Roles and responsibilities are needed to operationalise AI and ensure accountability for ethics. Who does what? What is the nature or responsibilities of any governing bodies that will be established for this AI application? Are different teams responsible for the AI model validation and model development?*	Section 4.1.2.2: • Ensure oversight exist for key processes • Define roles and responsibilities IT Planners/Executives, Business Users	
5 - Who has ultimate decision-making authority for the AI Application?	Section 4.1.2.2: • Ensure oversight exist for key processes • Define roles and responsibilities IT Planners/Executives, Business Users	

B. **AI Governance Process – Project Planning**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
6 - Does the project plan account for each stage of the AI Lifecycle for the AI application including change management and any organisational changes? *Does the project map to the established Project Management requirements? Are functional and technical requirements fully accounted for and documented? Does a roadmap exist to account for possible future requirements, necessary AI application updates as required and roles and responsibilities?*	Section 4.1.2.2: • Consider all Ethical AI Principles throughout the AI Lifecycle • Ensure appropriate resourcing needs are met • Ensure oversight exist for key processes • Define roles and responsibilities IT Planners/Executives, Business Users	
7 - Do all project team members understand their roles and the project plan? Are staff and/or contract resources equipped with the skills and knowledge they need to take on the project responsibility? How are governance controls to be managed that will enable a consistent, robust, repeatable development or implementation process?	Section 4.1.2.2: • Ensure oversight exist for key processes • Define roles and responsibilities IT Planners/Executives, Business Users	

B. **AI Governance Process – Project Planning**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
Do all project stakeholders understand the probabilistic nature of AI algorithms, recognising that all outputs will not be one hundred percent certain and correct? If not, either AI upskilling programmes or trainings (potentially using external trainers), should be completed or new staff/consultants hired with the skills hired prior to AI project commencement. This is to ensure the Project Team have members that are equipped with the necessary skills and knowledge required to execute the project.		
8 - What is the defined and intended application/scope of the AI project? How did you choose the AI model's suitability for the task at hand? What are the defined business success criteria for an AI application? *(Tasks examples: Do you validate that the metrics used to select an AI model are appropriate? Do you evaluate AI models for reliability? Do you test for AI model sensitivity? Is there a process to identify AI model vulnerabilities?)*	Section 4.1.2.1: • Select and prioritise AI projects IT Planners/Executives, Business Users	
9 - (i) How has "fairness" been described? (ii) What steps are in place to measure and test for achieving this? *Given there is no single definition of fairness that will apply equally well to different AI applications, the goal is to detect and mitigate fairness-related harms as much as possible. AI applications can behave unfairly due to biases inherent in the data sets used to train them or biases that are explicit or implicitly reflected in decisions made by the development teams or can result in unfair behaviour when these applications interact with particular stakeholders after deployment. Types of harm and risk can include allocation, quality of service, stereotyping, denigration, over or underrepresentation. They can also be affected by trade-offs*	Section 4.1.2.2: • Consider all Ethical AI Principles throughout the AI Lifecycle Section 4.1.4.4: • Define and test for fairness and bias Section 4.1.4.5.4: • Perform regularisation and other forms of model selection IT Planners/Executives, Business Users, Data Scientists, Project Managers, System	

B. **AI Governance Process – Project Planning**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
between expected benefits and potential harms for different stakeholder groups. Mitigation measures include processes that scrutinise the system vision and what resulting potential fairness related harms to stakeholder groups. Consider defining and scrutinising the system architecture for example machine learning models, performance metrics user interfaces, data sets needed to develop and test the system. Scrutinise the production datasets against defined fairness criteria. Consider doing a ship review before launch and a code review. Build in regular product review meetings.	Analysts, System Architect	
10 - How will traceability be maintained across data, experiments, AI model versions and usage? How will you capture performance against success criteria? *Traceability and performance contribute to the explainability, transparency and reliability of the AI solution. The risk of not maintaining traceability is that the AI can perform unexpectedly and not be explainable to users.* *Mitigation measures could entail documenting methods used for designing and developing the algorithmic system. These could consist of:* *a) Rule-based AI applications: the method of programming or how the model was built;* *b) Learning-based AI applications: the method of training the algorithm, including which input data was gathered and selected, and how this occurred. Describe the methods used to test and validate the algorithmic system:* *a) Rule-based AI applications: the scenarios or cases used in order to test and validate;* *b) Learning-based model: information about the data used to test and validate. Describe the outcomes of the algorithmic system:* *The outcomes of or decisions taken by the algorithm, as well as potential other decisions that would result from different*	Section 4.1.4.3: • Complete documentation on data sources Section 4.1.5.4: • Enable traceability System Analysts, System Architects, Project Managers, Data Scientists	

B. **AI Governance Process – Project Planning**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
cases (for example, for other subgroups of users).		

B. **AI Governance Process – Project Ecosystem**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
11 - What specific types of and sources of data will be collected, tracked, transferred, used, stored or processed as part of AI model development or application? Is any of the data to be used personally identifiable? *Is the data identifiable to a person? Is the data anonymous (or de-identified) and what policy, processes and/or technical measures have been used to minimise the re-identification of the data to an individual? Does an up to date data dictionary exist for the data used by the project? Have you considered how combinations of characteristics may be used by the AI model to reveal a special category, which may result in processing that is unfair to the individuals represented by the data? Do you collect and use biometric data in your AI model? Source data could be from other internal sources or from externally obtained data. If there are personally identifiable data are used, a PIA should be conducted to identify privacy risks and perform mitigating steps to mitigate those risks identified as well as ensure compliance with the PD(P)O.*	Section 4.1.4.3: • Perform data validation • Obtain assurance for public or third-party data • Complete documentation on data sources System Analysts, System Architects	
12 - Do you log the data lineage to understand the source, path, license or other obligations and transformations of data being loaded into ML Models? Do you have procedures in place to validate the use of the data or that the AI model is compliant with any applicable licenses? Consider attaching a data map. *What processes are in place to ensure input data are fit for the purpose of AI application? Do you have procedures in place to validate the use of the data or that*	Section 4.1.4.3: • Obtain assurance for public or third-party data • Complete documentation on data sources System Analysts, System Architects, Project Managers, Business Users	

B. **AI Governance Process – Project Ecosystem**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
the AI model is compliant with any applicable licenses (if the data or technology is third-party provided)? *Risks for not logging the data lineage include technical risks where the data may not be understood or appropriate. This can impact Ethical Principles such as fairness, transparency and explainability and data privacy. Mitigation measures can include using a data map to map the data elements from their source to the models. This would provide an overview of all the data used and can be used as a checklist to identify whether each of these data sources has applicable licences.*		
13 - Is the data accurate enough for the purpose of the AI model and initiative activity? Is the dataset used credible and from a reliable source? What are your techniques for validating the reliability of source data? *AI Modelling data need to be appropriate in terms of the data sample, size and distributions to ensure the AI model makes meaningful and representative inferences. What steps are being taken to determine the accuracy of source data and if the source data will be accurate enough over time? Has consolidation/transformation impacted the data in such a way the accuracy is affected? Are there concerns about the quality of the final data set relative to the purpose of the activity? Ethical AI principles affected can include fairness, reliability robustness and security. Please refer to Section 4.1.4.3 "Data Extraction" for details on addressing data reliability.*	Section 4.1.4.3: • Perform data validation • Obtain assurance for public or third-party data System Analysts, System Architects	

B. **AI Governance Process – Third-party Questions**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
14 - Does your AI application or Big Data project contemplate using third-party technology or data as either a supplier or partner? (Yes/No)	Not applicable. This is a question to identify whether organisations have to complete this section.	

B. **AI Governance Process – Third-party Questions**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
Third parties are usually external vendors, providers or partners. If no, skip the third-party section.		
15 - If your organisation obtained AI models or datasets from a third-party, did your organisation assess and manage the risks of using these?	Section 4.1.4.3: • Obtain assurance for public or third-party data Section 4.1.3.2: • Define requirements of the AI application System Analysts, System Architects, Sourcing Team (i.e. Procurement), Project Manager, Business Users	
16 - What is/are the documentation requirements of third parties? Did you ask for and receive detailed documentation? Does the documentation satisfy the requirements established? *The applicability of the organisations data to the vendor AI model should be assessed. Back testing, model validation and outcomes analysis should be done on the organisation's intended portfolio of AI model use.* *Where documentation can't be provided, there can be a business decision for organisation to take on where the risk of not having this documentation which should be assessed against possible consequences. This should be considered on a case-by-case basis. Mitigation measures can include only using these models for lower risk AI applications (e.g. applications for internal use that do not trigger any of the initial triggers in the risk gating criteria within this assessment).*	Section 4.1.3.2: • Define requirements of the AI application Section 4.1.4.5.5: • Use trusted cloud/third-party services Sourcing Team (i.e. Procurement), Project Manager, Business Users	
17 - What are the processes for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report	Section 4.1.3.2: • Define requirements of the AI application	

B. **AI Governance Process – Third-party Questions**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
potential vulnerabilities, risks or biases in the AI application? *Organisation should request third parties to report any potential vulnerabilities, risks or biases in the AI application. This should be stated explicitly in the third-party engagement contract. Where this can't be performed organisations should assess the risk of potential vulnerabilities, risks or biases against possible consequences.*	Sourcing Team (i.e. Procurement), Project Manager, Business Users	

B. **AI Governance Process – Project Development**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
18 - What data types and sources will be used in the AI model development? Will the AI application use personal information as input data? What process have you put in place to ensure the quality and integrity of your data? Describe the data cleansing steps you will be using. Describe how you have determined that the data you process is adequate, relevant and limited to what is necessary to achieve the objectives of the model. *Complex data cleansing steps are prone to unintended user errors that are difficult to identify and may lead to erroneous modelling results. Will you need to conduct or amend a PIA? What type of data do users expect you to accurately share, measure or collect? Mitigation measures include implementing data cleansing steps with experienced system analysts/system architects and testing procedures.*	Section 4.1.4.3: • Perform data validation • Obtain assurance for public or third-party data Section 4.1.4.4: • Conduct Privacy Impact Assessment • Perform data anonymisation System Analysts, System Architects, Project Managers	
19 - How has the quality of training data been assessed? Were there enough total training samples? Were the samples well-representative of different social groups based on race, gender, colour, age, income, etc.? *Did you consider the diversity and representativeness of users in the data*	Section 4.1.4.4: • Compare training, validation and test data Project Managers, System Analysts, System Architects, Data Scientists	

B. **AI Governance Process – Project Development**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
being used? As a rule of thumb, the benchmark for training data: test data: validation data are commonly set at 60:20:20 or 70:15:15. If validation data are not used, the recommended sample size for training data to test data are 80:20. Refer to Section 4.1.4.3 "Data Extraction" for more details. Where quality and/or training data are not well representative, the potential impact of this should be captured in Section C of this assessment to assess the risk.		
20 - How did you test the performance of the AI model? Was the AI model well-trained and analysed through different metrics - Precision, Recall, F1Score, Accuracy, etc.? An example is using scikit metrics (https://scikit-learn.org/stable/modules/model_evaluation.html). What performance metrics did you consider, how did you perform them, and did you consider performance differences by subpopulations, e.g. protected groups? If the performance of the AI model is not tested, the impact of the AI model not performing well should be captured in Section C to assess the risk. Performance metrics should be used to mitigate the risk if needed.	Section 4.1.5.1: • Perform integration, system, decision and User Acceptance Testing Data Scientists, Project Managers	
21 - How have all technology or data security requirements been met? How are you verifying that your data sets have not been compromised or hacked? *Did your risk analysis include whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behaviour of the AI application? Did you put measures or systems in place to ensure the integrity and resilience of the AI application against potential attacks? What could bad actors do with this data if they had access to it? What is the worst thing someone could do with this data if it were stolen or leaked?*	Section 4.1.5.2: • Implement model/Application Programming Interface access security practices Data Scientists, Project Managers	

B. **AI Governance Process – Project Development**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
A Security Risk Assessment & Audit should be performed to identify, analyse and evaluate the security risks, and determine the mitigation measures to reduce the risks to an acceptable level.		
22 - Could the AI application have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use? How have these risks been identified and mitigated? *Did you consider different types of vulnerabilities and potential entry points for attacks such as:* • *Data poisoning (i.e. manipulation of training data); or* • *Model evasion (i.e. classifying the data according to the attacker's will);* • *Model inversion (i.e. infer the model parameters).* *Did you put measures in place to ensure the integrity, robustness and overall security of the AI application against potential attacks over its lifecycle? Did you perform penetration testing on the application? When conducting a security assessment of the AI application, consider involving an independent third party. Risk mitigation measures include adversarial training as referenced in Section 4.1.4.5.6.*	Section 4.1.4.5.5: • Use trusted cloud/third-party services Section 4.1.4.5.6: • Perform adversarial training Sourcing Team (i.e. Procurement), Data Scientists, Project Managers	
23 - Describe the validation processes that will be used. *Are there technical review processes? Is there any independent review? Integration, system, decision and UAT testing are all part of validation measures.*	Section 4.1.5.1: • Perform integration, system, decision and User Acceptance Testing Data Scientists, Project Managers	
24 - Will the AI application be replacing human decisions that require judgement or discretion? What is the automated decision the AI application will make? Will they	Section 4.1.4.2: • Determine the appropriate level of human intervention	

B. **AI Governance Process – Project Development**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
have a legal or similar impact on an individual? *Do you evaluate whether an AI model requires human intervention? Processing that aims at taking decisions on data subjects producing "legal effects concerning the individual" or which "similarly significantly affects the natural person". For example, the processing may lead to the exclusion or discrimination against individuals*	Project Managers, Business Users	
25 – (i) Can the AI model and the AI model output be explained in a simple way? (ii) What are the communications plans to achieve explainability to impacted stakeholders? (iii) Will the AI application be able to produce reasons for its decisions or recommendations when required? *Is the way your algorithms work transparently communicated to the people impacted by them? Is there any recourse for people who feel they have been incorrectly or unfairly assessed? (circumstances examples: to non-technical stakeholders; you consider the purpose and the context under which the explanation is needed; where technical explainability/explicit explanations may not be useful to the audience). When AI model and its output could not be explained in a simple way, consider drawing a workflow with all possible decision tree to enable end-users to visualise the model's logic. If an AI model can't be explained, the impact of that application and its decisions should be considered, and risk assessed.*	Section 4.1.4.2: • Define requirements for transparency and explainability of AI models Project Managers, Business Users	
26 - What are the other possible alternatives to the current AI applications that might be more manual and how do they perform relative to the AI application in terms of both accuracy-metrics and business, legal, economic, social risks, etc.? Is the manual method less risky than the AI	Section 4.1.4.1: • Define business requirements Project Managers, Business Users	

B. **AI Governance Process – Project Development**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
application? Is it costlier or requires higher investment?		

B. **AI Governance Process – System Deployment**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
27 – Describe your risk analysis that accounts for security or network problems such as cybersecurity and adversarial attacks. *(Risks examples: cause safety risks or damage due to unintentional behaviour of the AI system; the impact of data leakage; What could bad actors do with this data if they had access to it?)*	Section 4.1.5.1: • Perform integration, system, decision and User Acceptance Testing Data Scientists, Project Managers	
28 - Are all users treated equally? If not - and your AI algorithms and predictive technologies prioritise certain information or sets prices or access differently for different users - how would you handle consumer demands or government regulations that require all users to be treated equally, or at least transparently unequally?	Section 4.1.2.2: • Consider all Ethical AI Principles throughout the AI Lifecycle Data Scientists, Project Managers	
29 - What have you put in place to monitor and document the AI application's performance (e.g. accuracy) and acceptance criteria? What steps have been put in place to thoroughly test the AI application before deployment to see if the AI application breaks and if it behaves as intended? What is the rollback plan? *Could a low level of accuracy of the AI application result in critical, adversarial or damaging consequences? Did you put in place measures to ensure that the data (including training data) used to develop the AI application is up to date, of high quality, complete and representative of the environment in the AI application will be deployed?*	Section 4.1.6.1: • Set performance metrics • Incorporate quality assessments Data Scientists, Project Managers	

B. **AI Governance Process – System Deployment**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
30 – (i) Could the AI application cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility? (ii) Is there a well-defined process to monitor if the AI application is meeting the intended goals? (iii) Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility? *Did you test the AI model used on different demographic groups to mitigate systematic bias?*	Section 4.1.5.4: • Ensure repeatable and reproducibility end-to-end workflow Section 4.1.6.1: • Set performance metrics • Incorporate quality assessments Data Scientists, Project Managers	
31 - Did you have an adequate working definition of "fairness" that you apply in designing AI applications? Please describe. What is your strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI application, both regarding the use of input data as well as for the algorithm design? *What are your definitions of unfair bias in your use case? Describe the metrics used to evaluate each of them. Describe the existing best practices for detection, identification and mitigation of unfair biases. What is your risk analysis framework? Describe the risks of unfair bias identified for your use case and the groups described by end-user characteristics for which you evaluated bias.* *Refer to Question 9 above. This should be defined at Project Planning stage. If unfair bias is in the model, what impact could this have?*	Section 4.1.4.4: • Define and test for fairness and bias Project Managers, Business Users, Data Scientists	
32 - Describe the processes to test and monitor for potential negative discrimination (bias) during the	Section 4.1.6.1: • Set performance metrics • Incorporate quality assessments	

B. **AI Governance Process – System Deployment**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
development, deployment and use phases of the AI application? *Does the AI application potentially negatively discriminate against people? Describe the controls in place to mitigate any detected bias. Is the data used in your processing representative of the population you apply the AI application to?*	• Create anomaly detection techniques Data Scientists, Project Managers	
33 - Is the AI application designed to interact, guide or take decisions by human end-users that affect humans or society? Could the AI application generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision? Describe how end-users or other subjects are adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision. Note: This question is not applicable if the AI application is not related to individuals. *Could the AI application generate confusion for some or all end-users or subjects on whether they are interacting with a human or AI application?*	Section 4.1.5.2: • Provide information to users Data Scientists, Project Managers	
34 - How have you estimated the likely impact of your AI application when it provides inaccurate results? *Did you verify what harm would be caused if the AI application makes inaccurate predictions?*	Section 4.1.5.1: • Enable edge cases and exception handling Data Scientists, Project Managers	
35 - Describe the steps you have put in place to explain the decision(s) of the AI application to the users. Note: This question is not applicable if the AI application is not related to individuals. *In cases of interactive AI applications (e.g. chatbots, robot-lawyers), do you communicate to users that they are interacting with an AI application instead*	Section 4.1.5.2: • Provide information to users Data Scientists, Project Managers	

B. **AI Governance Process – System Deployment**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
of a human? Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI application? This relates to Ethical AI Principles such as Transparency and explainability so that users are aware of the AI decisions.		

B. **AI Governance Process – System Operation and Monitoring**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
36 - Do you have a plan in place to handle AI incident? Describe the detection and response mechanisms for undesirable adverse effects of the AI application for the end-user or subject. *Did you ensure a 'stop button' or procedure to safely abort an operation when needed? Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI application? Where there is no procedure to safely abort, is the impact acceptable based on potential for adverse effects of the AI application?*	Section 4.1.5.2: • Establish multiple layers of mitigation Section 4.1.5.3: • Track mistakes Section 4.1.6.1: • Incorporate quality assessments Section 4.1.6.2: • Create or leverage a communication plan for dealing with crises situations. Data Scientists, Project Managers	
37 - (i) Do you have a process in place to incorporate customer/user feedback? (ii) Do you have a process to identify AI application weaknesses? (iii) What is the escalation process to address significant issues that may be identified? *Do you have established processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI application? Upon AI application deployment, ongoing operational support must be established to ensure that the AI*	Section 4.1.5.3: • Track mistakes Section 4.1.6.1: • Incorporate quality assessments System Analysts, System Architects, Data Scientists, Project Managers	

B. **AI Governance Process – System Operation and Monitoring**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
application's performance remains consistent, reliable and robust.		
38 - How are you monitoring the ongoing performance of your AI model? What is the business continuity plan you have in place? What are your triggers for AI model maintenance and rollback? *Did you put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the AI application's reliability and reproducibility? Did you put in place measures that address the traceability of the AI application during its entire lifecycle? Did you put in place measures to continuously assess the quality of the input data to the AI application? Did you define tested failsafe fallback plans to address AI application errors of different origins and put governance procedures in place to trigger them?*	Section 4.1.5.4: • Ensure repeatable and reproducibility end-to-end workflow • Enable traceability Section 4.1.6.1: • Set performance metrics • Incorporate quality assessments Data Scientists, Project Managers	

B. **AI Governance Process – Compliance**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
39 - After deployment, what is the process to continually identify, review and mitigate risks of using the identified AI application? *Does your organisation perform active monitoring, review and regular AI model tuning when appropriate (e.g. changes to customer behaviour, commercial objectives, risks and corporate values)? This can mitigate risks related to the Ethical AI principles such as fairness, reliability, robustness and security as models are running under changing circumstances.*	Section 4.1.6.3: • Perform management/continuous review Project Managers, IT Planners/Executives	

B. **AI Governance Process – Compliance**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
40 - Have all key decision points of the AI application been mapped and do they meet all relevant legislation, internal policies or procedures? *This can be checked using this assessment and with reviews as outlined in section 4.1.6.3.*	Section 4.1.6.3: • Consult IT Board/CIO Project Managers, IT Planners/Executives	

C. **Impact – Beneficial Impact**	**Guideline**	**Assessment results/proposed mitigation plan**
41 - What are the benefits of this AI application to the organisation?	*Determine and describe what the benefits are that could be realised by the organisation. Consider factors such as increased revenue, lower costs, improved efficiency, enhanced employee satisfaction, engagement and productivity, enhanced workforce relationship, enhancement or maintenance of brand or reputation, assurance of compliance, fraud prevention, enhancement or maintenance of cyber or physical security, new or improved services, improved manner of marketing, improved ability to assess customer preferences, improvements to innovation or enabling greater, faster, more efficient innovation, improved research processes, improved ability to conduct research and find or enrol study subjects, or improved efficiency with studies, innovative ways to conduct research.*	

C. **Impact – Beneficial Impact**	**Guideline**	**Assessment results/proposed mitigation plan**
42 - What are the benefits to the defined impacted stakeholders? Could the AI application be used in a way that may result in a specific stakeholder or group of stakeholders being treated differently in a positive way from other groups of individuals?	*Determine and describe the positive impacts on the various stakeholders that are expected to come from the application of this technology/data activity. Are there identifiable expectations of individuals, groups of individuals for each beneficial use of the AI application? Determine what the potential positive goal of the difference in treatment is (if any).*	
43 - What are the benefits for society as a whole?	*Which social interest is served with the deployment of this AI application? How does the project/application contribute to or increase well-being? How will the project /application contribute to human values?*	
44 - What are the factors that may limit the realisation of any benefits to external stakeholders?	-	

C. **Impact – Negative Impact to Specific Stakeholders**	**Practice Guide Reference**	**Assessment results/proposed mitigation plan**
45 - Does the AI application potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?	-	
46 - Is it foreseeable that the potential application of data analytical insights or the data activity might seem surprising, inappropriate or discriminatory or might be	*Could the data or technology be used in a way that may result in a group of individuals being treated differently from*	

C. **Impact – Negative Impact to Specific Stakeholders**	**Practice Guide Reference**	**Assessment results/proposed mitigation plan**
considered offensive causing distress or humiliation?	*other groups of individuals?*	
47 - Are there potential negative impacts of the AI application on the environment? Could the AI application have a negative impact on society at large or democracy? Are there potential negative impacts of the AI application on users' mental and physical health, life, and property?	*Did you assess the societal impact of the AI application's use beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large?*	
48 - For data or technology activities that involve third parties (e.g. receiving or sourcing technology or data as part of this activity), what are the associated risks?	*Examples of third parties could include data brokers that sell blocks of information, data aggregators, providers of storage and computing tools, data trusts. Examples of risks could include data accuracy, data protection, downstream use monitoring and control, legitimate data collection (when done through third parties), data availability.*	
49 - Is there any likelihood the AI application could lead to any potential costs from the legal and business perspective?	*For example, lawsuits can potentially lead additional legal costs. Is it possible that the AI application might lead to such overheads?*	

C. **Impact – Controls**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
50 - What are the additional technical and/or procedural safeguards (mitigating controls) that are being implemented to prevent and mitigate risks should they occur?	Section 4.1.5.2: • Establish multiple layers of mitigation Data Scientists, Project Managers	

C. **Impact – Controls**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
Have appropriate governance and accountability measures and processes been established? Is the accuracy and/or quality of the data appropriate for the data activity? Does the relative accuracy of the data have an impact on individuals/groups?		
51 - Describe the mechanism used to externally explain how technology and data are used, and how benefits and risks to individuals that are associated with the processing are considered and/or addressed. *Determine what the transparency and individual accountability mechanisms are and whether they are appropriate for the information activity use. Does the application of the technology or information do anything your users do not know about, or would probably be surprised to find out about? What are the explainability mechanisms proposed that are to be used?*	Section 4.1.5.2: • Provide information to users Data Scientists, Project Managers	
52 - Describe the extent of human involvement in the processing. Have all considerations with respect to automated decision making been accounted for? *Have the humans (human-in-the-loop, human-out-of-the-loop, human-in-command) been given specific training on how to exercise oversight?*	Section 4.1.4.2: • Determine the appropriate level of human intervention IT Planners/Executives, Business Users, Project Managers	
53 - What is the mechanism to capture feedback by users of the AI application? Will there be a recourse process planned or established for users of the AI application that wish to challenge the decision? *Could the AI application benefit from additional review and input by an external party (e.g. Ethical AI Committee)*	Section 4.1.5.3: • Implement an easy-to-use feedback user interface Data Scientists, Project Managers	
54 - What is the plan and process(es) in place to assess the AI model performance over time, including model drift and changes in the model use environment to ensure that output stays statistically accurate?	Section 4.1.6.1: • Set performance metrics • Incorporate quality assessments • Re-train AI model	

C. **Impact – Controls**	**Practice Guide Reference/Responsibility**	**Assessment results/proposed mitigation plan**
	Data Scientists, Project Managers, IT Planners/Executives	

D. **Decision – Go/No-Go**	**Guideline**	**Assessment results/proposed mitigation plan**
55 - How effective are the mechanisms that facilitate the AI application auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI application's processes, outcomes, positive and negative impact)?	-	
56 - What are the additional regulatory requirements surrounding your use of AI? Are there other legal, cross-bordary requirements, policy, contractual, industry or other obligations linked to the collection, analysis and use(s) of data (or technology)? Have these all been addressed?	-	
57 - How effective are the overall controls and safeguards in reducing risk? Please use the Risk Assessment Tool as part of the Project Management Plan ("**PMP**") template to assess the risks from the impact recorded in Section C of this assessment.	-	
58 - Decision – Has an appropriate balance of benefits and mitigated risks supports the AI application processing activity and achieves alignment with the Ethical AI Principles. Are there any other factors that should be considered? Have the interests, expectations and rights of the stakeholders been effectively addressed, and what additional contextual based individual participation and choice factors should be considered?	*Describable, achievable net positive benefit outcomes (tangible benefits to people) have been demonstrated and that negative consequences have been mitigated to a satisfactory and demonstrable level. The proposed uses of technology and data meet all the Ethical AI Principles and values of diversity, inclusion, and*	

D. **Decision – Go/No-Go**	**Guideline**	**Assessment results/proposed mitigation plan**
Does the project or the AI application require escalation to a senior decision-making body (e.g. IT Board/CIO)?	*privacy as a fundamental human right.*	

APPENDIX D – AI STRATEGY TEMPLATE

The following template is for illustration and can be used for AI strategy/plan. Organisations can use other templates and strategy formats if preferred.

A. Executive Summary

Vision and Objectives	*State the vision and objectives of having the AI applications. What is the goal of having the AI applications/use cases and how can it benefit the organisation? This can be accomplished by explaining the intended outcomes of the AI projects and ways of achieving this.*
Alignment of AI application goals with organisation goals	*Map the AI application goal with the organisations goals. This must be aligned to ensure that the AI outcome conforms to the organisation's vision and mission. If there is a gap in the mapping between the AI application goal and organisation goals, organisation should then assess whether the AI application proposed is suitable for the organisation.*
Current Situation	*Highlight the current situation, including the major achievements over the past year(s), existing work in progress and any identified issues to be addressed that are related to AI application.*
Key Drivers and Targets	*Summarise the key driving forces for changes, such as the latest policy objectives, business strategy, operational requirements, and other improvement opportunities identified that are related to AI.*
Proposed AI Application/Use Case	*Identify all the AI use cases and rank the AI projects in order of importance. The importance of the projects should be based on how critical these projects to the organisation at this point. Also consider whether there is any short-term AI project that may help demonstrate value from AI in relatively fast, inexpensive and fewer effort methods.*
Implementation Plan	*Outline the list of proposed projects with individual project objectives to be achieved with a timeline to achieve the target outcomes and the associated resources requirements*

B. Current Situation (only applicable for organisations with existing AI projects)

Current Ethics and Legal Considerations	*Identify the current ethical considerations and any legal implications around the existing AI application. For example, data privacy is usually one of the key considerations in most AI applications.*
Current Technology and Infrastructure	*Identify the current technology and infrastructure used to develop and deploy the existing AI application. The current technology and infrastructure used should consist of the following data layers:* • *Data collection* • *Data storage* • *Data modelling* • *Data processing/analysing*
Current Skills and Talent	*Identify the current skills and talent acquired to develop and implement the existing AI application.*

Current Challenges	*Identify any current issues or challenges faced throughout the existing AI project Lifecycle.*
Current Change Management	*Identify changes brought about from the existing AI application such as staff impact, engagement and communication.*

C. Key Drivers and Targets

Target Ethics and Legal Considerations	*Identify the ethical considerations and any legal implications around the AI application. For example, data privacy is usually one of the key considerations in most AI applications. Organisation should consider all the 12 Ethical AI Principles to ensure that the to-be AI application is ethical.*
Target Technology and Infrastructure	*Identify the technology and infrastructure required to develop and deploy the AI application. A suitable technology and infrastructure required should consider the following data layers:* • *Data collection* • *Data storage* • *Data modelling* • *Data processing/analysing*
Target Skills and Talent	*Identify the skills and talent required to develop and implement the AI application. If there are skills gaps identified, organisation should determine whether training should be conducted to upskill the staff or if there is a need to hire new staff, or partner with an external AI provider.*
Identified Challenges	*Identify any foreseeable issues or challenges throughout the AI project Lifecycle. What actions can organisation take to ensure the successful delivery of the project?*
Target Change Management	*Plan for change management and identify changes brought about from the proposed AI application such as staff impact, engagement and communication. For example, the AI application may have an impact on job displacement, particularly if it involves automating certain tasks or processes.*

D. Proposed AI Applications

Requirement Specifications	*Specify the requirements of business, data, application, and technology (including AI model where applicable) as well as the overarching components (e.g. information security and data privacy) that are necessary and sufficient for subsequent development and implementation.*
Solution Options and Suitability	*Identify the strategic options for implementation based on aligned selection criteria (e.g. cost-benefit analysis).*
Recommended AI Applications	*Describe the recommended projects to the implemented for meeting the business objectives and achieving the target outcomes*

E. Implementation Plan

Implementation Strategy	*Define the strategic approach for implementation, including the inter-project dependencies, relative priorities of work and the strategic measures (based on the Ethical AI Principles and AI Lifecycle practices, etc.) to be adopted across all projects.*
High Level Roadmap	*Identify the major activities, milestones and expected deliverables alongside a timeline to achieve the target outcomes*
Resources Estimation	*Provide the estimated resource requirements (including staff and expenditure) for implementing each of the recommended projects*
Benefits and Impact	*Identify the intangible and tangible benefits, the anticipated impact and associated mitigation measures for implementation.*
Governance	*Define the governance structure and process for monitoring the progress and resolving any issues that may arise during the implementation stage.*

Suggested Quality Criteria

The quality criteria below are provided to assist the organisations to assess the quality of their AI Strategy/Plan. The criteria are illustrative and can be adapted by the organisations to their preferred strategy template.

A. **Executive Summary**

- Does the executive summary include necessary and sufficient information for assessment by organisations' senior management?
- Is the information accurate and consistent with the other parts of the report?

B. **Current Situation (only applicable for organisations with existing AI projects)**

- **Current Ethics and Legal Considerations**: Is the list of all current ethical considerations and any legal implications of AI applications complete and accurate?
- **Current Technology and Infrastructure:** Is the list of all current technologies used for supporting the current AI applications complete and accurate?
- **Current Skills and Talent**: Is the list of all skills and talent acquired through the existing AI projects complete and accurate?
- **Current Challenges**: Is the list of all issues or challenges identified throughout the AI project Lifecycle complete and accurate, with proper resolutions?
- **Current Change Management**: Is the list of all changes brought about from the AI applications complete and accurate, with proper change management plans?

C. **Key Drivers and Targets**

- **Target Ethics and Legal Considerations**: Are the ethical considerations and legal implications conform to the Ethical AI Principles and generally accepted by society?
- **Target Technology and Infrastructure:** Are the target technology and infrastructure conform to the Ethical AI Principles and capable to support the target AI application?
- **Target Skills and Talent**: Do the target talent and skills provide the necessary

skillsets with minimal redundancy to support the target AI application development?

- **Identified Challenges**: Do the issues or challenges identified include suggested resolution measures?
- **Target Change Management**: Do the changes identified include suggested change management plan?
- Have all the gaps between current and target ethics and legal, technology and infrastructure, skills and talent been identified?

D. **Proposed AI Applications**

- Have the requirement specifications fully defined the required capabilities for filling the gaps and migrating to the target state? Have the requirements on AI application and interoperability been identified and clearly specified?
- Are all the requirement specifications fully met by the identified solution options?
- Are the selection criteria among options aligned with the Ethical AI Principles and agreed with the key stakeholders?
- Are the selected solution options grouped logically into the recommended projects?
- Are the business values of all the recommended projects assigned properly with key stakeholders' consensus?
- Is the priority of the recommended projects agreed with the key stakeholders?

E. **Implementation Plan**

- Does the implementation plan include all the recommended projects?
- Does the high-level roadmap fully show the progression from current to target AI projects on a timeline with clear milestones and expected major deliverables in place?
- Are the identified risks properly mitigated and accepted by the key stakeholders?
- Are the management structure and governance mechanism for implementation clearly defined and agreed with the key stakeholders?