ASEAN Guide on

AI Governance and Ethics

Contents

Executive Summary

What the Guide is about

This document serves as a practical guide for organisations in the region that wish to design, develop, and deploy traditional AI technologies in commercial and non-military or dual-use applications. This Guide focuses on encouraging alignment within ASEAN and fostering the interoperability of AI frameworks across jurisdictions. It also includes recommendations on national-level and regional-level initiatives that governments in the region can consider implementing to design, develop, and deploy AI systems responsibly.

Guiding Principles for the Framework

Transparency and Explainability

Fairness and Equity

Security and Safety

Robustness and Reliability

Human-centricity

Privacy and Data Governance

Accountability and Integrity

4 Key Components

Internal governance structures and measures

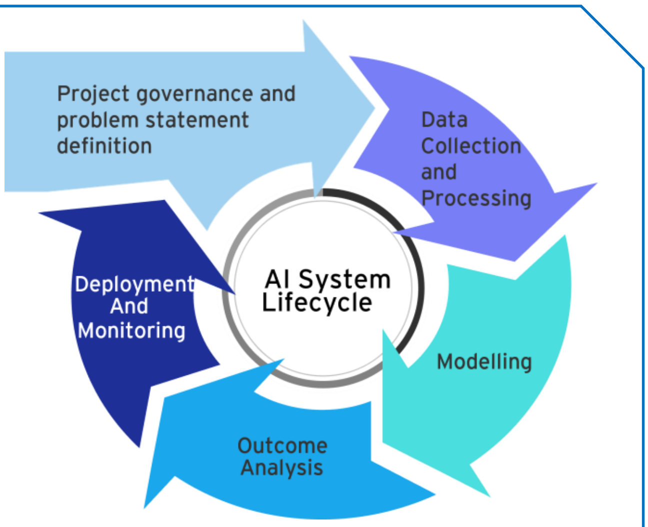
- Multi-disciplinary, central governing body, such as an AI Ethics Advisory Board, to oversee AI governance efforts
- Develop standards, guidelines, tools, and templates to help organisations design, develop, and deploy AI responsibly
- Clearly lay out the roles and responsibilities of personnel involved in the responsible design, development and/or deployment of AI

Determining the level of human involvement in AI-augmented decision-making

- Conduct relevant risk impact assessments to determine level of risk
- Three broad categories of human involvement based on level of risk – human-in-the-loop, human-over-the-loop, human-out-of-the-loop
- Mitigating risks helps build trust towards the acceptance and greater use of AI technologies in the region

Operations management

- The AI System Lifecycle consists of various stages and is often an iterative process
- Conduct risk-based assessments before starting any data collection and processing or modelling
- Mitigate risks of unjust bias due to insufficiently representative training, testing and validation datasets


Project governance and problem statement definition
Data Collection and Processing
Modelling
Outcome Analysis
Deployment And Monitoring
AI System Lifecycle

Stakeholder interaction and communication

- Develop trust with stakeholders throughout the design, development, and deployment of AI
- Provide general disclosure of when AI is used in products and/or service offerings
- Put in place measures to help employees adapt to an AI-augmented work environment

National-level Recommendations

Nurturing AI talent and upskilling workforce

Work closely with public and private sectors to ensure that a country's workforce can adapt to the new ways of working and possesses enough digital skills to interact effectively with AI systems.

Supporting AI innovation ecosystem and promoting investment in AI start-ups

Work closely with public and private sectors to create a supportive environment for AI development, where companies are able to access and leverage data, digital technologies, and infrastructure.

Investing in AI research and development

Keep abreast of the latest developments in AI and encourage research related to the cybersecurity of AI, AI governance, and AI ethics to ensure that the safety and resiliency of AI systems and tools also advance in parallel with new use cases.

Promoting adoption of useful tools by businesses to implement the ASEAN Guide on AI Governance and Ethics

Deploy tools to enable the implementation of AI governance in operations and ensure that documentation and validation processes are more efficient.

Raising awareness among citizens on the effects of AI in society

Raise awareness of the potential risks and benefits of AI so citizens can make informed decisions about the appropriate use of AI and take appropriate actions to protect themselves from harmful uses of AI systems.

Regional-level recommendations

Setting up an ASEAN Working Group on AI Governance to drive and oversee AI governance initiatives in the region

The Working Group can consist of representatives from each of the ASEAN member states who can work together to roll out the recommendations laid out in this Guide, as well as provide guidance for ASEAN countries who wish to adopt components of this Guide, and where appropriate, include consultation with other industry partners for their views and input.

Adaptation of this Guide to address governance of generative AI

Risks include:

- Mistakes and anthropomorphism
- Factually inaccurate responses and disinformation
- Deepfakes, impersonation, fraudulent and malicious activities
- Infringement of intellectual property rights
- Privacy and confidentiality
- Propagation of embedded biases

Governance should include:

- Adaptation of existing frameworks and tools
- Guidance on developing a shared responsibility framework
- Guidance on increasing the capacity to manage risks of generative AI
- Guidance on how to distinguish AI-generated content versus authentically generated ones

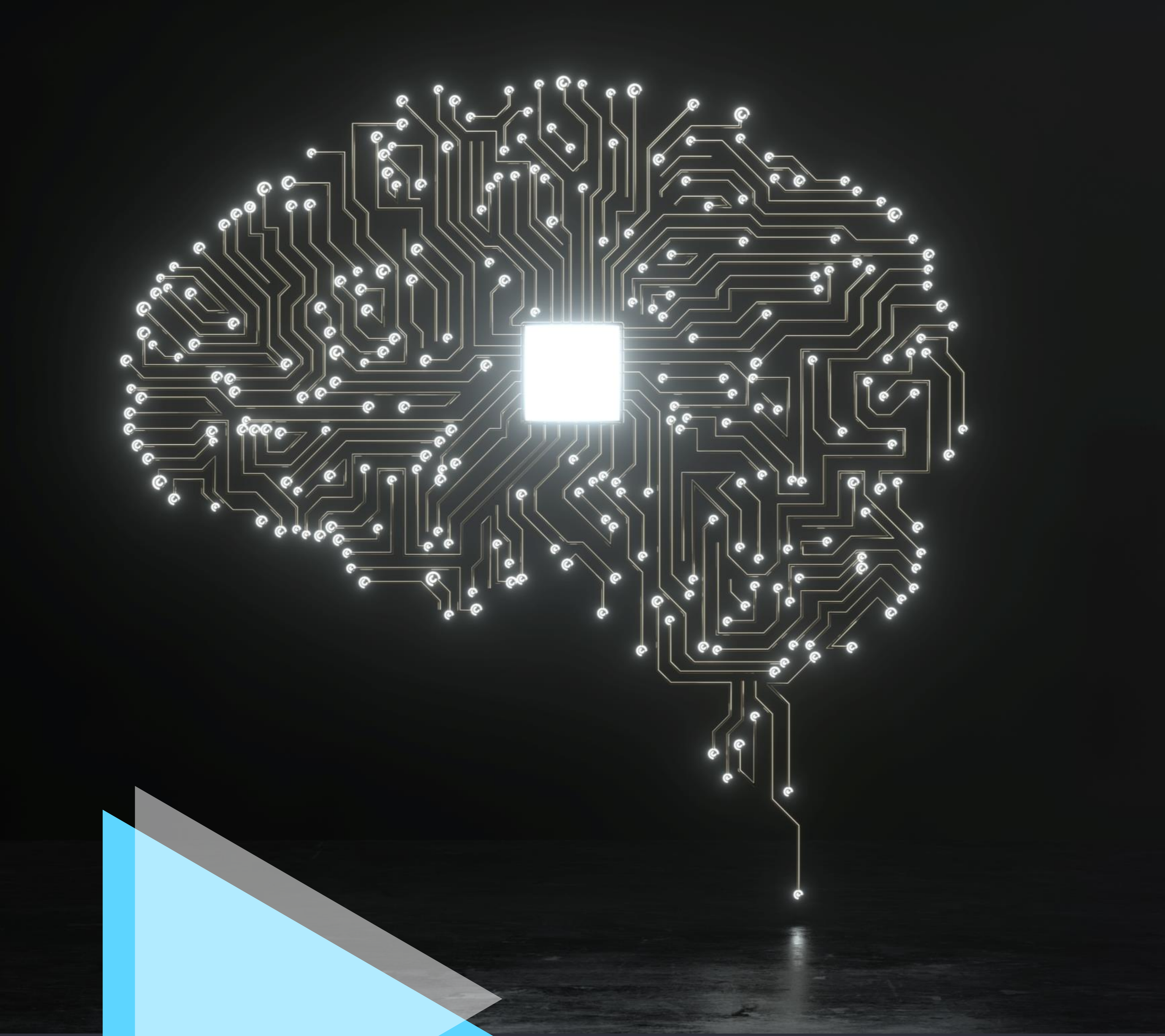
Compiling a compendium of use cases demonstrating practical implementation of the Guide by organisations operating in ASEAN

A compendium of use cases showcases the commitment of these organisations to AI governance and helps them promote themselves as responsible AI practitioners.

Use Cases

Illustration of components of the ASEAN Guide on AI Governance and Ethics through use cases of organisations operating in ASEAN that have implemented AI governance measures in AI design, development, and deployment.

- Gojek
- UCARE.AI
- Aboitiz Group
- EY
- Smart Nation Group (SNG), Singapore
- Ministry of Education, Singapore



Introduction

A. Introduction

Artificial Intelligence (AI) is the discipline of making analytical machines intelligent, enabling an organisation to function appropriately and with foresight. Unlike other technologies, some forms of AI adapt on its own, learning through use, so the decisions it makes today may be different from those it makes tomorrow. AI and automation have been hot topics, both for their transformative potential and for their capacity to introduce new opportunities by disrupting old models. Southeast Asia is no exception. AI systems should be treated differently from other software systems because of its unique characteristics and risks. Capabilities of AI systems fuelled by techniques evolution and breakthroughs are quickly outpacing the monitoring and validation tools. The development of AI is also decentralised due to low barriers to entry and the proliferation of open-source technologies. Given the profound impact that AI potentially brings to organisations and individuals in ASEAN, it is important that the decisions made by AI are aligned with national and corporate values, as well as broader ethical and social norms.

Also, the ASEAN Digital Masterplan 2025 (ADM2025) developed by ASEAN Member States ("AMS") envisions ASEAN as a leading digital community and economic bloc, powered by secure and transformative digital services, technologies, and ecosystem. In that context, the ADM2025 has identified Enabling Action (EA) 2.7 that suggests the development and adoption of a regional policy to deliver best practice guidance on AI governance and ethics. In recent years, governments and international organisations have begun issuing principles, frameworks and recommendations on AI ethics and governance. Examples include Singapore's Model AI Governance Framework[1] and OECD's Recommendation of the Council on AI[2]. However, there has not yet been an intergovernmental common standard for AI that defines the principles of AI governance and provides guidance for policymakers in the region to utilise AI systems in a responsible and ethical manner. In the process of drafting this Guide, existing AI governance frameworks and guidelines such as UNESCO's Recommendation on the Ethics of Artificial Intelligence[3] and EU's Ethics Guidelines for Trustworthy AI were referenced[4].

The ASEAN Guide on AI Governance and Ethics aims to empower organisations and governments in the region to design, develop, and deploy traditional AI systems responsibly and increase users' trust in AI.

[1] Infocomm Media Development Authority, "Model Artificial Intelligence Governance Framework Second Edition" (21 January 2020) < https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf >

[2] Organisation for Economic Co-operation and Development, "Recommendation of the Council on Artificial Intelligence" (22 May 2019) < https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 >

[3] United Nations Educational, Scientific and Cultural Organisation, "Recommendation on the Ethics of Artificial Intelligence" (23 November 2021) < https://unesdoc.unesco.org/ark:/48223/pf0000381137 >

[4] European Commission "Ethics Guidelines for Trustworthy AI" < https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai > (8 April 2019)

1. Objectives

This document serves as a practical guide for organisations in the region that wish to design, develop, and deploy traditional AI technologies in commercial and non-military or dual-use applications. This Guide focuses on encouraging alignment within ASEAN and fostering the interoperability of AI frameworks across jurisdictions. It also includes recommendations on national-level and regional-level initiatives that governments in the region can consider implementing to design, develop, and deploy AI systems responsibly.

The ASEAN Guide on AI Governance and Ethics encompasses a broad range of considerations that needs to be tailored to the organisations designing, developing, and deploying AI depending on the nature of the industry, complexity of the technology, and associated risks of the AI systems. Local government authorities are encouraged to refer to this Guide when formulating their policies and approaches.

While adoption of the framework laid out herein is voluntary, this Guide can help organisations build trust among stakeholders and the public as well as align their AI practices with international standards and best practices, among others. Organisations are encouraged to refer to the guidelines in this document to understand how to assess the risks associated with AI and take measures to design, develop, and deploy AI responsibly in the context of their organisations.

This Guide is meant to be a living document that should be periodically reviewed and assessed by relevant ASEAN sectorial bodies, in consultation with industry partners, to ensure that it is up to date with the latest regulations and advancements in the AI space. Updates to this Guide may be published subsequently to keep up with evolutions and growth in governance and standards.

2. Assumptions

AI systems need to be managed holistically, including its ecosystem and all components – human operator, Internet of Things (IoT), robotics, traditional technology, vendors, etc. In addition to the recommendations set out in this Guide pertaining to AI governance, organisations are also encouraged to follow and refer to international standards and best practices in related fields like information security management systems (ISMS), data management and governance, software development and testing, cybersecurity, IoT, etc.

Developers and deployers need to adhere to applicable national laws and regulations, including sector-specific laws and constitutions when designing, developing, and deploying AI technologies. The ASEAN Guide on AI Governance and Ethics does not replace or supersede any existing or upcoming laws and only serves as a guide for responsible design, development, and deployment of AI in the region. Before deploying AI, it is also important for developers and deployers to consider the relevant legal and regulatory requirements of the respective countries where the AI systems will be deployed, as well as the use context for legal, policy, and regulatory concerns.

Given the fast-evolving space of AI, developers and deployers of AI systems should be mindful of the latest state of the art vis-à-vis governance tools and technologies and conduct the relevant assessments to evaluate the feasibility and usefulness of these tools in the implementation of AI governance practices in the design, development, and deployment of AI systems.

3. Target Audience

The ASEAN Guide on AI Governance and Ethics is an ASEAN-endorsed framework for organisations in the region to refer to when designing, developing, and deploying traditional AI technologies. It provides guidelines and recommendations for a diverse range of individuals and organisations along the entire value chain. These include AI developers and deployers, academic professionals, and everyone that is interested in utilising or scaling up AI systems. The Guide also includes sections on national-level and regional-level recommendations that are more targeted towards policymakers in ASEAN.

4. Definitions

The definitions of several key terms used in this Guide are set out below.

ASEAN: refers to the Association of Southeast Asian Nations. It is a political and economic union of 10 member states in Southeast Asia, which promotes intergovernmental cooperation and facilitates economic, political, security, military, educational, and sociocultural integration between its members and countries in the Asia-Pacific.

Artificial Intelligence (AI): is an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments.

AI system: is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy.

Deep Learning: is a subset of machine learning. It is very loosely based on the information-processing architecture of the brain – albeit far from matching its ability – enabling systems to find ways of re-representing input data that facilitate making highly accurate predictions.

Deployer: is an entity that uses or implements an AI system, which could either be developed by their in-house team or via a third-party developer.

Developer: is an entity that designs, codes, or produces an AI system.

Machine Learning: is a subfield in AI where algorithms learn by identifying patterns and correlations within data using statistical techniques to enhance performance, all without being explicitly programmed.

User: is an entity or person (internal or external) that interacts with an AI system or an AI-enabled service and can be affected by its decisions.

Guiding Principles for the Framework

B. Guiding Principles for the Framework

The seven guiding principles below help to ensure trust in AI and the design, development, and deployment of ethical AI systems. They also provide guidance on how AI systems should be designed, developed, and deployed in ways which consider the broader societal impact.

1. Transparency and Explainability

Transparency refers to providing disclosure on when an AI system is being used and the involvement of an AI system in decision-making, what kind of data it uses, and its purpose. By disclosing to individuals that AI is used in the system, individuals will become aware and can make an informed choice of whether to use the AI-enabled system.

Explainability is the ability to communicate the reasoning behind an AI system's decision in a way that is understandable to a range of people, as it is not always clear how an AI system has arrived at a conclusion. This allows individuals to know the factors contributing to the AI system's recommendation.

In order to build public trust in AI, it is important to ensure that users are aware of the use of AI technology and understand how information from their interaction is used and how the AI system makes its decisions using the information provided.

In line with the principle of transparency, deployers have a responsibility to clearly disclose the implementation of an AI system to stakeholders and foster general awareness of the AI system being used. With the increasing use of AI in many businesses and industries, the public is becoming more aware and interested in knowing when they are interacting with AI systems. Knowing when and how AI systems interact with users is also important in helping users discern the potential harm of interacting with an AI system that is not behaving as intended. In the past, AI algorithms have been found to discriminate against female job applicants and have failed to accurately recognise the faces of dark-skinned women. It is important for users to be aware of the expected behaviour of the AI systems so they can make more informed decisions about the potential harm of interacting with AI systems. An example of transparency in an AI-enabled ecommerce platform is informing users that their purchase history is used by the platform's recommendation algorithm to identify similar products and display them on the users' feeds.

In line with the principle of explainability, developers and deployers designing, developing, and deploying AI systems should also strive to foster general understanding among users of how such systems work with simple and easy to understand explanations on how the AI system makes decisions. Understanding how AI systems work will help humans know when to trust its decisions. Explanations can have varying degrees of complexity, ranging from a simple text explanation of which factors more significantly affected the decision-making process to displaying a heatmap over the relevant text or on the area of an image that led to the system's decision. For example, when an AI system is used to predict the likelihood of cardiac arrest in patients, explainability can be implemented by informing medical professionals of the most significant factors (e.g., age, blood pressure, etc.) that influenced the AI system's decision so that they can subsequently make informed decisions on their own.

Where "black box" models are deployed, rendering it difficult, if not impossible to provide explanations as to the workings of the AI system, outcome-based explanations, with a focus on explaining the impact of decision-making or results flowing from the AI system may be relied on.

Alternatively, deployers may also consider focusing on aspects relating to the quality of the AI system or preparing information that could build user confidence in the outcomes of an AI system's processing behaviour. Some of these measures are:

- Documenting the repeatability of results produced by the AI system. Some practices to demonstrate repeatability include conducting repeatability assessments to ensure deployments in live environments are repeatable and performing counterfactual fairness testing to ensure that the AI system's decisions are the same in both the real world and in the counterfactual world. Repeatability refers to the ability of the system to consistently obtain the same results, given the same scenario. Repeatability often applies within the same environment, with the same data and the same computational conditions.
- Ensuring traceability by building an audit trail to document the AI system development and decision-making process, implementing a black box recorder that captures all input data streams, or storing data appropriately to avoid degradation and alteration.
- Facilitating auditability by keeping a comprehensive record of data provenance, procurement, pre-processing, lineage, storage, and security. Such information can also be centralised digitally in a process log to increase capacity to cater the presentation of results to different tiers of stakeholders with different interests and levels of expertise. Deployers should, however, note that auditability does not necessarily entail making certain confidential information about business models or intellectual property related to the AI system publicly available. A risk-based approach can be taken towards identifying the subset of AI-enabled features in the AI system for which implemented auditability is necessary to align with regulatory requirements or industry practices.
- Using AI Model Cards, which are short documents accompanying trained machine learning models that disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information.

In cases where AI systems are procured directly from developers, deployers will have to work together with these developers to achieve transparency. More on this will be covered in later sections of the Guide.

2. Fairness and Equity

Deployers should have safeguards in place to ensure that algorithmic decisions do not further exacerbate or amplify existing discriminatory or unjust impacts across different demographics and the design, development, and deployment of AI systems should not result in unfair biasness or discrimination. An example of such safeguards would include human interventions and checks on the algorithms and its outputs. Deployers of AI systems should conduct regular testing of such systems to confirm if there is bias and where bias is confirmed, make the necessary adjustments to rectify imbalances to ensure equity.

With the rapid developments in the AI space, AI systems are increasingly used to aid decision-making. For example, AI systems are currently used to screen resumes in job application processes, predict the credit worthiness of consumers and provide agronomic advice to farmers. If not properly managed, an AI system's outputs used to make decisions with significant impact on individuals could perpetuate existing discriminatory or unjust impacts to specific demographics. To mitigate discrimination, it is important that the design, development, and deployment of AI systems align with fairness and equity principles. In addition, the datasets used to train the AI systems should be diverse and representative. Appropriate measures should be taken to mitigate potential biases during data collection and pre-processing, training, and inference. For example, the training and test dataset for an AI system used in the education sector should be adequately representative of the student population by including students of different genders and ethnicities.

3. Security and Safety

AI systems should be safe and sufficiently secure against malicious attacks.

Safety refers to ensuring the safety of developers, deployers, and users of AI systems by conducting impact or risk assessments and ensuring that known risks have been identified and mitigated. A risk prevention approach should be adopted, and precautions should be put in place so that humans can intervene to prevent harm, or the system can safely disengage itself in the event an AI system makes unsafe decisions - autonomous vehicles that cause injury to pedestrians are an illustration of this. Ensuring that AI systems are safe is essential to fostering public trust in AI. Safety of the public and the users of AI systems should be of utmost priority in the decision-making process of AI systems and risks should be assessed and mitigated to the best extent possible. Before deploying AI systems, deployers should conduct risk assessments and relevant testing or certification and implement the appropriate level of human intervention to prevent harm when unsafe decisions take place. The risks, limitations, and safeguards of the use of AI should be made known to the user. For example, in AI-enabled autonomous vehicles, developers and deployers should put in place mechanisms for the human driver to easily resume manual driving whenever they wish.

Security refers to ensuring the cybersecurity of AI systems, which includes mechanisms against malicious attacks specific to AI such as data poisoning, model inversion, the tampering of datasets, byzantine attacks in federated learning[5], as well as other attacks designed to reverse engineer personal data used to train the AI. Deployers of AI systems should work with developers to put in place technical security measures like robust authentication mechanisms and encryption. Just like any other software, deployers should also implement safeguards to protect AI systems against cyberattacks, data security attacks, and other digital security risks. These may include ensuring regular software updates to AI systems and proper access management for critical or sensitive systems. Deployers should also develop incident response plans to safeguard AI systems from the above attacks.

It is also important for deployers to make a minimum list of security testing (e.g. vulnerability assessment and penetration testing) and other applicable security testing tools. Some other important considerations also include:

a. Business continuity plan

b. Disaster recovery plan

c. Zero-day attacks

d. IoT devices

[5] A byzantine attack in federated learning is a type of malicious act where one or more of the devices or servers involved in the federated learning process behaves erratically or provides misleading updates to the central model, with the intent to corrupt or manipulate the learning process or outcomes.

4. Human-centricity

AI systems should respect human-centred values and pursue benefits for human society, including human beings' well-being, nutrition, happiness, etc.

It is key to ensure that people benefit from AI design, development, and deployment while being protected from potential harms. AI systems should be used to promote human well-being and ensure benefit for all. Especially in instances where AI systems are used to make decisions about humans or aid them, it is imperative that these systems are designed with human benefit in mind and do not take advantage of vulnerable individuals.

Human-centricity should be incorporated throughout the AI system lifecycle, starting from the design to development and deployment. Actions must be taken to understand the way users interact with the AI system, how it is perceived, and if there are any negative outcomes arising from its outputs. One example of how deployers can do this is to test the AI system with a small group of internal users from varied backgrounds and demographics and incorporate their feedback in the AI system.

AI systems should not be used for malicious purposes or to sway or deceive users into making decisions that are not beneficial to them or society. In this regard, developers and deployers (if developing or designing in-house) should also ensure that dark patterns are avoided. Dark patterns refer to the use of certain design techniques to manipulate users and trick them into making decisions that they would otherwise not have made. An example of a dark pattern is employing the use of default options that do not consider the end user's interests, such as for data sharing and tracking of the user's other online activities.

As an extension of human-centricity as a principle, it is also important to ensure that the adoption of AI systems and their deployment at scale do not unduly disrupt labour and job prospects without proper assessment. Deployers are encouraged to take up impact assessments to ensure a systematic and stakeholder-based review and consider how jobs can be redesigned to incorporate use of AI. Personal Data Protection Commission of Singapore's (PDPC) Guide on Job Redesign in the Age of AI[6] provides useful guidance to assist organisations in considering the impact of AI on its employees, and how work tasks can be redesigned to help employees embrace AI and move towards higher-value tasks.

5. Privacy and Data Governance

AI systems should have proper mechanisms in place to ensure data privacy and protection and maintain and protect the quality and integrity of data throughout their entire lifecycle. Data protocols need to be set up to govern who can access data and when data can be accessed.

Data privacy and protection should be respected and upheld during the design, development, and deployment of AI systems. The way data is collected, stored, generated, and deleted throughout the AI system lifecycle must comply with applicable data protection laws, data governance legislation, and ethical principles. Some data protection and privacy laws in ASEAN include Malaysia's Personal Data Protection Act 2010, the Philippines' Data Privacy Act of 2012, Singapore's Personal Data Protection Act 2012, Thailand's Personal

[6] Personal Data Protection Commission Singapore, "A Guide to Job Redesign in the Age of AI" (2020) <https://file.go.gov.sg/ai-guide-to-jobredesign.pdf >

Data Protection Act 2019, Indonesia's Personal Data Protection Law 2022, and Vietnam's Personal Data Protection Decree 2023.

Organisations should be transparent about their data collection practices, including the types of data collected, how it is used, and who has access to it. Organisations should ensure that necessary consent is obtained from individuals before collecting, using, or disclosing personal data for AI development and deployment, or otherwise have appropriate legal basis to collect, use or disclose personal data without consent. Unnecessary or irrelevant data should not be gathered to prevent potential misuse.

Data protection and governance frameworks should be set up and adhered to by developers and deployers of AI systems. These frameworks should also be periodically reviewed and updated in accordance with applicable privacy and data protection laws. For example, data protection impact assessments (DPIA) help organisations determine how data processing systems, procedures, or technologies affect individuals' privacy and eliminate risks that might violate compliance[7]. However, it is important to note that DPIAs are much narrower in scope than an overall impact assessment for use of AI systems and are not sufficient as an AI risk assessment. Other components will need to be considered for a full assessment of risks associated with AI systems.

Developers and deployers of AI systems should also incorporate a privacy-by-design principle when developing and deploying AI systems. Privacy-by-design is an approach that embeds privacy in every stage of the system development lifecycle. Data privacy is essential in gaining the public's trust in technological advances. Another consideration is investing in privacy enhancing technologies to preserve privacy while allowing personal data to be used for innovation. Privacy enhancing technologies include, but are not limited to, differential privacy, where small changes are made to raw data to securely de-identify inputs without having a significant impact on the results of the AI system, and zero-knowledge proofs (ZKP), where ZKP hide the underlying data and answer simple questions about whether something is true or false without revealing additional information[8].

6. Accountability and Integrity

There needs to be human accountability and control in the design, development, and deployment of AI systems. Deployers should be accountable for decisions made by AI systems and for the compliance with applicable laws and respect for AI ethics and principles. AI actors[9] should act with integrity throughout the AI system lifecycle when designing, developing, and deploying AI systems.

Deployers of AI systems should ensure the proper functioning of AI systems and its compliance with applicable laws, internal AI governance policies and ethical principles. In the event of a malfunction or misuse of the AI system that results in negative outcomes, responsible individuals should act with integrity and implement mitigating actions to prevent similar incidents from happening in the future.

[7] TechTarget, "data protection impact assessment (DPIA) (July 2023) < https://www.techtarget.com/searchcio/definition/data-protection-impact-assessment-DPIA >

[8] Organisation for Economic Co-operation and Development, "Emerging Privacy Enhancing Technologies: Current Regulatory & Policy Approaches" (2023) < https://www.oecd-ilibrary.org/docserver/bf121be4-en.pdf?expires=1693746418&id=id&accname=guest&checksum=E8DFEFFD75D87E3C3DBB30DE8B6773C9 >

[9] AI actors can be defined as any actor involved in at least one stage of the AI system life cycle, and can refer to both natural and legal persons, such as researchers, programmers, engineers, data scientists, end-users, business enterprises, universities and public and private entities, among others.

To facilitate the allocation of responsibilities, organisations should adopt clear reporting structures for internal governance, setting out clearly the different kinds of roles and responsibilities for those involved in the AI system lifecycle. AI systems should also be designed, developed, and deployed with integrity – any errors or unethical outcomes should at minimum be documented and corrected to prevent harm to users upon deployment.

7. Robustness and Reliability

AI systems should be sufficiently robust to cope with errors during execution and unexpected or erroneous input, or cope with stressful environmental conditions. It should also perform consistently. AI systems should, where possible, work reliably and have consistent results for a range of inputs and situations.

AI systems may have to operate in real-world, dynamic conditions where input signals and conditions change quickly. To prevent harm, AI systems need to be resilient to unexpected data inputs, not exhibit dangerous behaviour, and continue to perform according to the intended purpose. Notably, AI systems are not infallible and deployers should ensure proper access control and protection of critical or sensitive systems and take actions to prevent or mitigate negative outcomes that occur due to unreliable performances.

Deployers should conduct rigorous testing before deployment to ensure robustness and consistent results across a range of situations and environments. Measures such as proper documentation of data sources, tracking of data processing steps, and data lineage can help with troubleshooting AI systems.

AI Governance Framework

C. AI Governance Framework

This section of the Guide includes guidance on measures promoting the responsible use of AI that organisations should adopt in the following key areas:

Internal governance structures and measures

Adapting existing or setting up internal governance structure and measures to incorporate values, risks, and responsibilities relating to algorithmic decision-making.

Determining the level of human involvement in AI-augmented decision-making

A methodology to aid organisations in setting its risk appetite for use of AI, i.e., determining acceptable risks and identifying an appropriate level of human involvement in AI-augmented decision-making.

Operations management

Issues to be considered when developing, selecting, and maintaining AI models, including data management.

Stakeholder interaction and communication

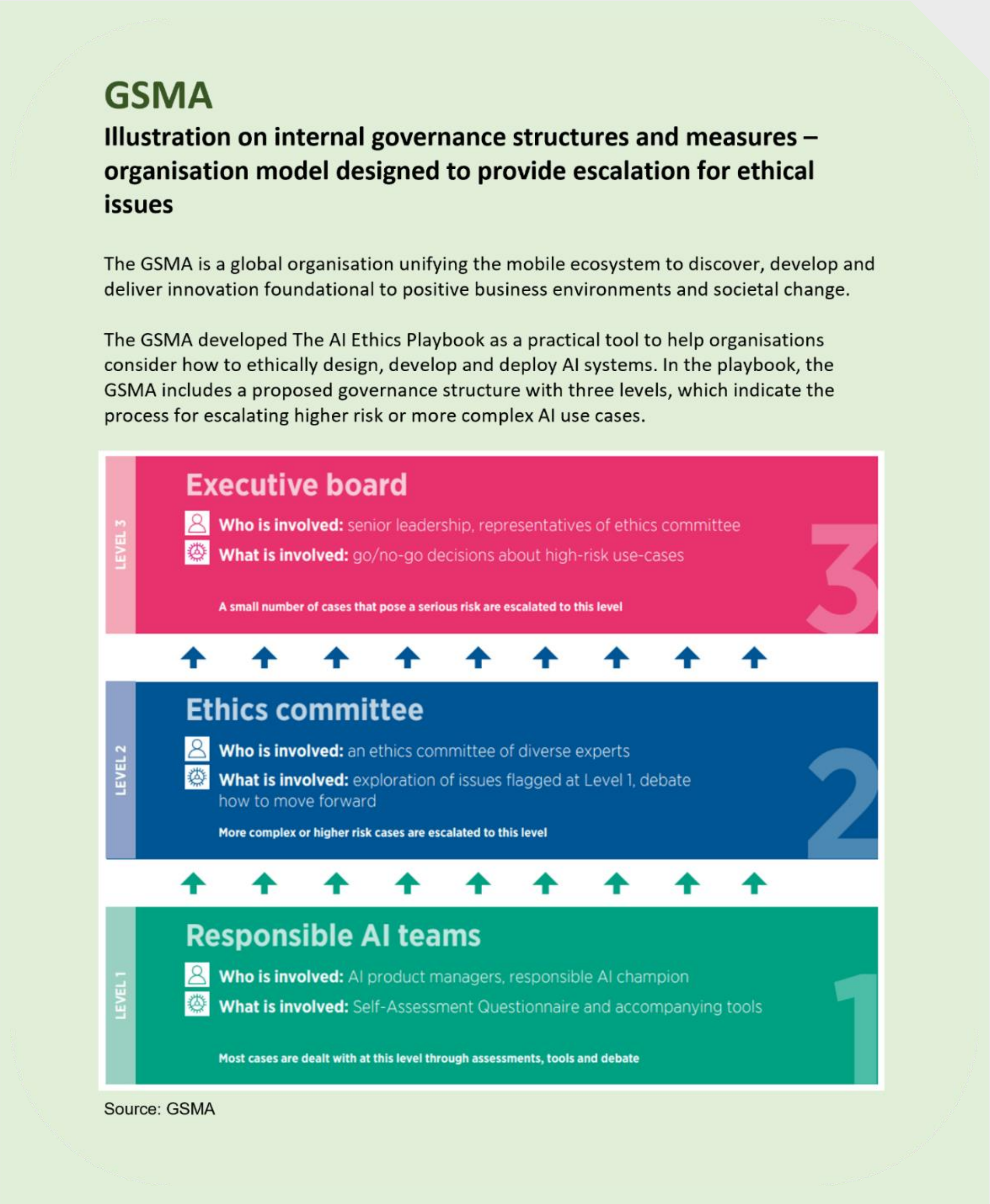
Strategies for communicating with an organisation's stakeholders, and the management of relationships with them.

1. Internal governance structures and measures

Internal governance structures need to be put in place for companies to have oversight of how AI systems are designed, developed, and deployed across the organisation. Deployers may choose to set up a new governance structure or adapt existing ones. For example, existing risk management structures can be used to evaluate and manage the risks of AI systems as well. Deployers can also consider setting up a multi-disciplinary, central governing body, such as an AI Ethics Advisory Board or Ethics Committee, to oversee AI governance efforts, provide independent advice, and develop standards, guidelines, tools, and templates to help other teams design, develop, and deploy AI responsibly. The AI Ethics Board or Committee should strive to be multi-disciplinary, and advisors can be drawn from ethics, law, philosophy, technology, privacy, regulations, science, and other relevant domains. To adequately reflect the diversity of society, there is also value in considering a governing body that is sufficiently representative of stakeholders and a range of voices. The ethical implications of AI can be complex to navigate and setting up an Ethics Committee of experts from different disciplines and geographies ensures that the issue is approached in a holistic manner.

Notably, the degree of centralisation or decentralisation of the governance structure needs to be suitable for the organisational structure and culture. This entails identifying the appropriate balance between flexibility and rigidity to ensure optimal business and process execution. In the case where the business needs to be nimble and responsive to changes in operational requirements, it might be more effective to go with a more decentralised approach, where AI governance considerations and decisions are made on a more frequent basis at the operational level.

Internal governance structures can also be designed for escalation of ethical issues, where AI systems and use cases that are of higher risk are escalated to a governing body with higher authority for review and decision-making. One possible example of an internal governance structure with three levels of escalation from GSM Association’s (GSMA) AI Ethics Playbook[10] is shown below:



GSMA

Illustration on internal governance structures and measures – organisation model designed to provide escalation for ethical issues

The GSMA is a global organisation unifying the mobile ecosystem to discover, develop and deliver innovation foundational to positive business environments and societal change.

The GSMA developed The AI Ethics Playbook as a practical tool to help organisations consider how to ethically design, develop and deploy AI systems. In the playbook, the GSMA includes a proposed governance structure with three levels, which indicate the process for escalating higher risk or more complex AI use cases.

LEVEL 3 — **Executive board** — 3

Who is involved: senior leadership, representatives of ethics committee

What is involved: go/no-go decisions about high-risk use-cases

A small number of cases that pose a serious risk are escalated to this level

LEVEL 2 — **Ethics committee** — 2

Who is involved: an ethics committee of diverse experts

What is involved: exploration of issues flagged at Level 1, debate how to move forward

More complex or higher risk cases are escalated to this level

LEVEL 1 — **Responsible AI teams** — 1

Who is involved: AI product managers, responsible AI champion

What is involved: Self-Assessment Questionnaire and accompanying tools

Most cases are dealt with at this level through assessments, tools and debate

Source: GSMA

[10] GSM Association, “The AI Ethics Playbook” (2022) < https://www.gsma.com/betterfuture/wp-content/uploads/2022/01/The-Mobile-Industry-Ethics-Playbook_Feb-2022.pdf >

Policies and standards for the ethical development of AI can consist of multiple components. An example would be a corporate code of conduct that provides guidance on the ethical use of data and AI. In addition, AI design principles are established to define and shape AI governance and accountability measures. These measures are intended to protect users, uphold societal norms, and ensure compliance with applicable laws and regulations.

Along with the development of AI governance policies and standards, oversight mechanisms also need to be put in place to ensure that AI governance guidelines are followed within the deployer's organisation.

As the introduction of new internal governance frameworks and structures by the deployer may lead to a major shift in the way departments and teams operate, deployers need to have a development plan that considers organisational readiness, technical skills, technology, and the need to raise overall awareness of AI ethics and governance across the organisation.

Clarity of roles and responsibilities for personnel involved in the responsible design, development and/or deployment of AI is important to ensure that the relevant individuals are aware of their duties. Roles that such individuals play include, but are not limited to:

- Establishing roles and terms of reference for oversight bodies who undertake governance and review.
- Ensuring that the composition of any committee or oversight team is representative of an appropriate range of functions such as legal, finance, safety, product, or service functions, etc.
- Conducting risk assessments and managing the potential risks of deploying AI systems.
- Ensuring that the model training, testing, and selection processes adhere to AI governance criteria set out by the central governing body.
- Ensuring that AI governance documentation and/or communication artefacts are adequate, maintained and updated.

To effectively implement AI governance, deployers need to ensure that proper guidance and training resources are provided to the individuals involved in the governance process and that broader awareness is raised across the organisation. Relevant personnel should understand the potential legal and ethical considerations for the development and deployment of AI, and their responsibility to safeguard interests of users impacted by the AI system's decisions. They should also be aware of the benefits, risks, and limitations of using the AI system and how to interpret the system outputs. This enables them to detect potential harm and assess when mitigating actions need to be taken, in line with internal AI governance standards and processes.

As AI technologies are ever-evolving, internal governance structures also need to be periodically reviewed and assessed to make sure that they align to the culture and organisational structures of companies and satisfy the code of conduct and ethical policies of the companies, while ensuring proper knowledge transfer in the event of any changes. Deployers should conduct periodic assessments of the sufficiency and effectiveness of the governance model, as well as controls across the AI system lifecycle, from problem identification and design to system development and deployment. To ensure that internal governance structures keep pace with the developments in the AI space, deployers should also train or improve the capacity of personnel involved in the design of such internal governance controls and policies. This could include sending personnel for additional training by relevant professional bodies. For example, the Singapore Computer Society offers a Certificate in AI Ethics and Governance to help professionals who wish to acquire more knowledge in applying ethical AI practices in organisations.

In considering the above recommendations for internal governance structures, developers and deployers should also take heed of factors such as the organisation's size and capabilities. For example, setting up a multi-disciplinary, central governing body, may be too onerous for smaller companies without the resources for such dedicated use of manpower. Organisations can instead take a more risk-based approach and focus on managing the risks that the governance structure seeks to address.

ABOITIZ GROUP

Illustration on establishing internal governance structures and measures

Aboitiz Group recognises that AI and ML algorithms are integral assets of the group, therefore, it is imperative to have a strategic AI governance framework to ensure that the algorithms and programs are properly managed, to support the day-to-day operations of different strategic business units within the group.

The Group has established the following internal governance structures and measures for oversight of AI:

- Ethical considerations related to the use of AI aligned with corporate values
- Clear and defined roles and responsibilities for the ethical use of AI technology
- All AI-related processes and decisions must be vetted by the management committee
- A multi-stakeholder approach with Model Governance Management Committee which is composed of the representatives of the strategic business units — the Chief Information Security Officer (CISO), Data Protection Officer (DPO), Chief Operations Officer, Chief Data Officer, Chief Risk Officer, Chief Technology and Operations Officer, Audit, Risk, and Compliance AI, AI and Innovation Center of Excellence, Chief Marketing Officer, and Senior Managers (Stakeholders of AI project)

SMART NATION GROUP (SNG), SINGAPORE

Illustration on internal governance structures and measures - approval gates at different stages of LLM product development

Singapore's National AI Office (NAIO) established guidelines for product teams in the government building custom LLM products, as well as an AI workgroup (with stakeholders from across government) to oversee the rigorous testing and safe deployment of products.

To encourage experimentation while also ensuring ample review of LIM products, product teams need only seek approval from the central AI workgroup from the beta testing phase onwards.

Embarking on internal experiments Proof-of-concepts or experiments within a development team	**No approval required**
Embarking on beta testing With test users outside of the product development team	**Approval gate 1** Seek approval from the central AI workgroup prior to embarking on beta testing. To maintain an agile process, teams provide key product details via a form, and AI workgroup members comment/endorse concurrently
Deploying the product To all intended users	**Approval gate 2** Seek central approval from the central AI workgroup prior to full deployment. Teams outline key feedback and improvements via a form, and AI workgroup members comment/endorse concurrently

2. Determining the level of human involvement in AI-augmented decision-making

How severe could the potential negative effects be?

How many people are or could be affected by the AI system?

How likely is it for the AI system to cause a negative impact?

AI systems are different from legacy technologies and may pose unfamiliar risks. For example, AI systems' processing speeds and decision-making capabilities are quickly outpacing monitoring and validation tools. AI systems are also increasingly used in applications where their decisions significantly impact the life of humans or business performance. Some other risks related to AI systems include discrimination due to bias, security weaknesses, potential for AI system malfunction or unexpected behaviour.

As a result of the nature of AI systems, they can pose significant risks when things go awry, especially when they are used to make significant decisions that can potentially cause harm. A robust risk management approach should be taken at every stage of the AI system lifecycle, assessing, and mitigating the risks of AI at every stage. Such risks include financial, reputational, ethical, and legal risks among others. This helps build trust towards the acceptance and greater use of AI technologies in the region.

This section is intended to help deployers determine the appropriate extent of human oversight in AI-augmented decision-making based on the risks assessed. Risk assessments identify the risks for different stakeholders and determine how to address these risks throughout the AI system lifecycle and across the entire value chain. To facilitate periodic risk assessments, regardless of the level of human intervention, system and user behaviour should be recorded for auditing purposes.

Having clarity on the objective of using AI is a key first step in determining the extent of human oversight. During the design phase of AI systems, deployers should first establish the intended commercial objectives of the AI system, ensure that it is compatible with the principles above, and assess it against the potential risks of using the AI system for operations. Some objectives can include ensuring consistency in decision-making, improving operational efficiency and reducing costs, or introducing new product features to increase consumer choice. The above objectives can then be weighed against the risks of deploying the AI system in the organisation. This assessment should be guided by corporate values which may reflect the societal norms or expectations of the operating region.

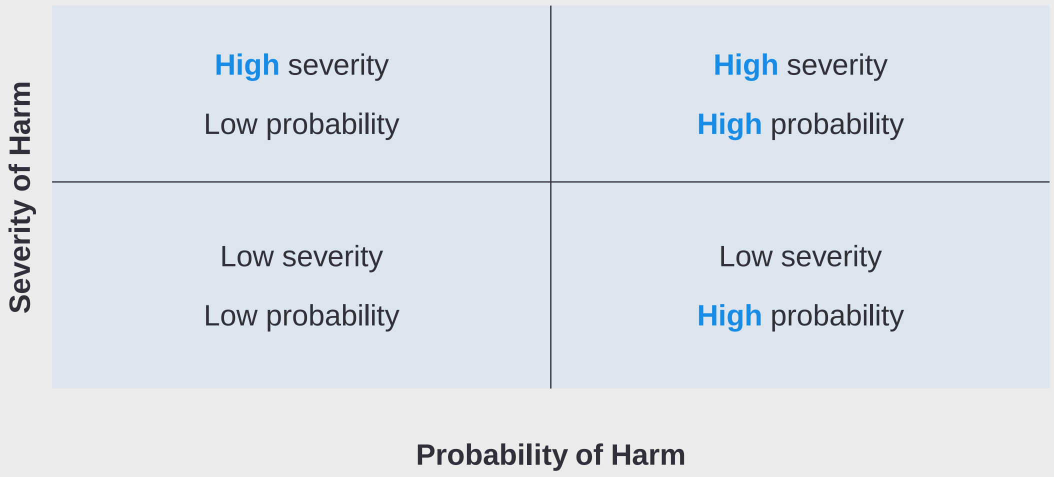
Considering how interconnected the world is now, ASEAN-operated organisations may offer AI services to customers and stakeholders who reside and operate outside the region or in culturally distinct areas within the region. In diverse regions like Southeast Asia, it is especially important for deployers to consider the unique local norms and values in different countries when assessing risks of using AI systems. For example, some products or topics may be insensitive and unacceptable in some countries but not others. Companies should also take into account the differing levels of digital maturity across ASEAN. When designing, developing, and deploying AI, deployers should take into account these considerations and take steps to ensure that the views of different cultures are respected.

In some cases, risks may only be presented when a sufficiently large group of people interact with AI. For example, if a substantial number of people use the same AI-enabled stock recommendation technology, market volatility could be increased due to herding behaviour.

Core values of the organisation can also be used as a guide to assess the risks and objectives of AI systems. Corporate values signify what the company stands for and any AI system that goes against these values is likely to warrant a review, with clearly documented rationale for any deviations.

As new AI systems are introduced and existing models are iterated, the identification of commercial objectives, risks and appropriate level of human involvement is a process that needs to be continually reviewed and improved. Deployers should continue to identify risks, and review risk management and mitigation plans to ensure that there are updated response plans for new risks. Deployers should also keep proper documentation of risk impact assessments (refer to Annex A for an example of a risk impact assessment template) done, for audit purposes and to instil trust and reassurance for users of AI. Such documentation will also be useful when faced with challenges by individuals, organisations, and other stakeholders. Other resources like PDPC's Implementation and Self-Assessment Guide for Organisations[11] or the US National Institute of Standards and Technology (NIST) AI Risk Management Framework Playbook[12] are useful references to understand how to mitigate the risks associated with the use of AI systems.

To determine the level of risk and the category of human involvement required in AI-augmented decision-making, developers and deployers can evaluate AI solutions along two axes – the probability and the severity of harm to users and individuals involved in the AI system lifecycle. The definition of "harm" and the computation of probability and severity will depend on the context, varying from sector to sector.



Severity of Harm		
	High severity Low probability	**High** severity **High** probability
	Low severity Low probability	Low severity **High** probability

Probability of Harm

[11] Personal Data Protection Commission Singapore, "Implementation and Self-Assessment Guide for Organisations" (January 2020) < https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgisago.pdf >

[12] National Institute of Standards and Technology, "NIST AI RMF Playbook" < https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook >

In computing severity or probability of harm, factors such as how poor system outcomes could impact the lives and livelihoods of people, whether the integrity of markets could be compromised, the potential to violate the privacy of an individual or the protection of his or her personal data, as well as the durability and reversibility of potential harm could be considered.

Minimal Risk: AI systems that have low severity and probability of harm are permitted to function autonomously with minimal human intervention. However, the AI system should still be subject to information and transparency obligations and used responsibly.

Medium Risk: Depending on the nature of the AI system and industry it is used in; AI systems need to have an appropriate level of human control (full control or supervisory control) to ensure that there is oversight on AI-augmented decisions.

High Risk: AI systems that have high severity of harm and/or high probability of harm should be carefully evaluated and highly controlled by humans to ensure that AI systems are not able to independently make decisions with unintended or dangerous outcomes.

In addition, other factors that deployers in various contexts may consider relevant, could include: (a) the nature of harm (i.e., whether the harm is physical or intangible in nature); (b) the reversibility of harm, and as a corollary to this, the ability for individuals to obtain recourse; and (c) whether it is operationally feasible or meaningful for a human to be involved in a decision-making process (e.g., having a human-in-the-loop would be unfeasible in high speed financial trading, and be impractical in the case of driverless vehicles).

Harm also needs to be assessed according to the objective of AI and the nature of the industry in which it is used. For example, an AI system that is used in a healthcare setting to aid diagnosis or provide care for patients is likely to have a lower threshold for severity as compared to a recommendation system that is used to recommend products on an ecommerce website. ASEAN governments should also ensure that their understanding of harm is updated and in line with technological advancements.

In general, AI systems that have high severity and probability of harm should adopt a human-in-the-loop approach where humans can assume full control of the system and decide when it is safe to execute decisions. These assessments should be made for all user types and deployers are encouraged to give special consideration to impact on vulnerable and/or marginalised populations.

Based on the risk assessment of AI systems, there can be three broad categories of human involvement in AI-augmented decision-making – human-in-the-loop, human-over-the-loop, and human-out-of-the-loop.

- **Human-in-the-loop:** AI system only provides recommendations that humans use as an input to make decisions. Humans have full control over decision-making and AI can only provide supporting information.

For example, in the case where AI is used to predict medical conditions in patients, doctors or medical professionals are ultimately the ones that perform the diagnosis and dispense the appropriate treatments. With human-in-the-loop, humans need to have enough understanding of the factors influencing the AI system's decision and how it makes its decision in order to determine if the recommendation, prediction, or decision is accurate, fair, and/or safe. The human should take the time and effort to make such an assessment rather than simply accepting the AI system's response for efficiency's sake. Deployers should also be cautious of the risk of automation bias (aka "rubber stamping risk") where the human gets used to approving the AI system's outputs because of its high accuracy and misses the occasional AI error due to "muscle memory" where he/she becomes used to clicking on "approve".

- Another factor to consider is the significance of the AI system's outputs to the human making the decision i.e., whether the AI system's outputs is the sole input to the decision (extremely significant) or one of a dozen inputs (less significant).
- **Human-over-the-loop:** Humans play a supervisory and monitoring role and can intervene in the decisions of the AI system when it does not behave as intended, encounters unexpected events, or presents potential harm to humans. For example, even in the "full self-driving" mode of some autonomous vehicles, humans still need to have their hands on the wheel and eyes on the road so they can take over immediately if needed[13]. As supervisors of AI, humans can also alter the parameters during the operation of AI.
- **Human-out-of-the-loop:** AI system has complete control over the execution of decisions and does not need to rely on human intervention. The AI system has full control without the option of human override. For example, recommendation algorithms are able to autonomously push products and/or services to users based on their usage patterns and behavioural profiles, all without a human screening and approving its decisions.

EY

Illustration on determining the level of human involvement in AI-augmented decision-making

EY's purpose is building a better working world, and as a leader in Artificial Intelligence, this means that EY is committed to developing and deploying trusted AI solutions both internally and for its clients.

EY adopts an AI Model Risk Tiering approach to assess and classify the models as High, Medium, or Low risk. The key areas of risks associated with AI, such as use case design, ethics, data, privacy, algorithmic, performance, compliance, technology, and business risks are evaluated to assign a risk tier for every AI model. Based on the risk tier, appropriate monitoring and human oversight is put in place for the AI models.

[13] NPR, "Cars are getting better at driving themselves, but you still can't sit back and nap" (22 December 2021) < https://www.npr.org/2021/12/22/1064598337/cars-are-getting-better-at-driving-themselves-but-you-still-cant-sit-back-and-na >

SMART NATION GROUP (SNG), SINGAPORE

Illustration on determining the level of human involvement in AI-augmented decision-making – determining the level of risk and corresponding mitigating measures of LLM use cases

Singapore's National AI Office (NAIO) established guidelines for product teams in the government building custom LLM products, as well as an AI workgroup (with stakeholders from across government) to oversee the rigorous testing and safe deployment of products.

NAIO takes a risk-based approach when advising product teams on required mitigating measures. The level of risk varies, depending on AI products':

I. Task (productivity and language tools VS factual information retrieval) [lower risk VS higher risk]

II. Audience (internal-facing VS external-facing) [lower risk VS higher risk]

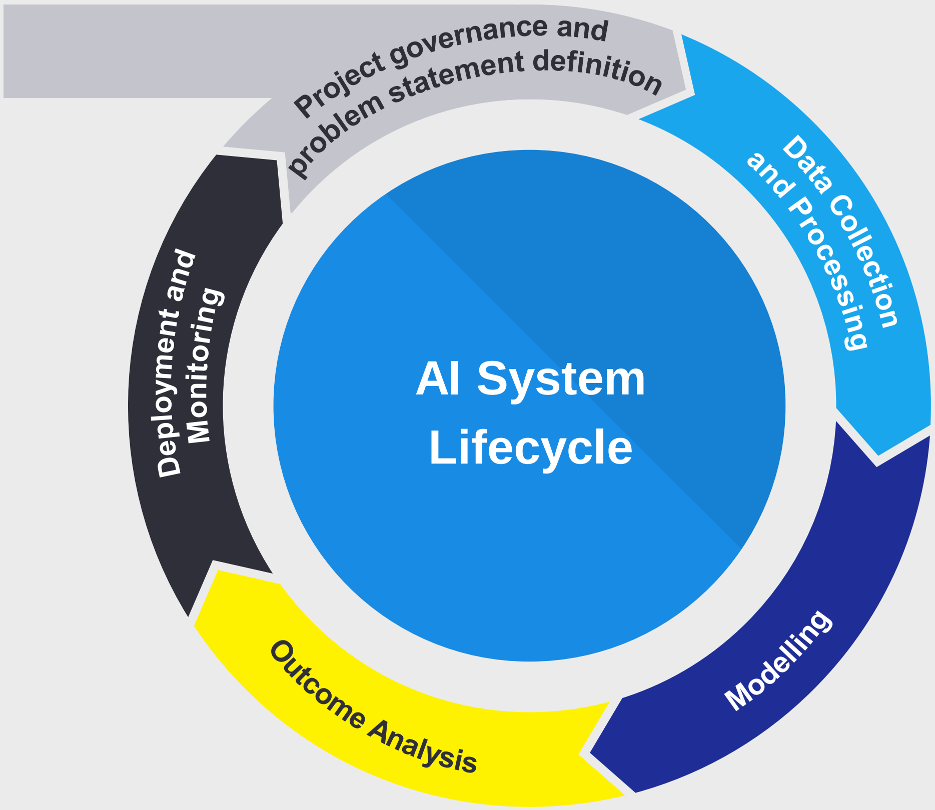
For example, for public-facing AI products, the product should be made robust to adversarial attack. The corresponding mitigating measures include the product teams engaging in efforts to increase the robustness of their products, such as via robustness tests to improve performance against adversarial prompts, red teaming or bug-bounty programmes, and rate-limiting queries so users are deterred from brute-force attacks.

3. Operations management

The AI System Lifecycle consists of the following stages:

- **Project governance and problem statement definition**
- **Data Collection and Processing**
- **Modelling**
- **Outcome analysis**
- **Deployment and Monitoring**

Developers and deployers should bear in mind that the system lifecycle is not always unidirectional and can be a continuous iterative process of fine-tuning AI systems that have already been developed and deployed.


AI System Lifecycle
Project governance and problem statement definition
Data Collection and Processing
Modelling
Outcome Analysis
Deployment and Monitoring

The subsequent sections cover the key areas that need to be considered in each stage of the AI System Lifecycle. There are also examples of questions that teams should ask as they go through each stage included in speech bubbles (💬) throughout the section.

Project governance and problem statement definition

What is the business driver(s) for the AI system?

What are the actions that will be autonomously performed by the AI agent?

Does the use of the AI agent align with the organisation's code of conduct and ethical policies?

Will AI actually improve the problem, or cause new ones?

AI governance should be built into AI systems by design. In this phase of the AI system lifecycle, deployers should ensure that business purpose, governance and key stakeholders are properly identified and aligned. AI systems should be designed, developed, and deployed in response to organisational needs and aligned with business strategies and goals. There should be conversations about the purpose, range, and portfolio for which the AI system is designed, developed, deployed, as well as how it might impact users, considering the degree of human oversight. When designing and developing AI systems, deployers and developers also need to bear in mind the principles of human-centricity, fairness and equity, transparency and explainability, safety and security, robustness and reliability, accountability and integrity, and privacy and data governance.

Deployers should conduct risk-based assessments of the AI systems before starting any data collection and processing or modelling. Risk assessments will help deployers identify potential safety risks of foreseeable uses of the AI system, including the potential for accidental or malicious misuse, and create a plan to assess and mitigate these risks. Measures and safeguards for risk mitigation should be properly documented because it allows organisations to keep a record of the responsible design behind the development and deployment of the AI system as well as the justifiability of outcomes.

According to the risks assessed, deployers should put in place mitigation measures to manage the risks relating to the AI system. Some measures include adopting an appropriate level of human involvement to oversee the AI system’s decision-making process, regular monitoring and maintenance, developing procedures to respond to previously unknown risks, and implementing mechanisms for the AI system to safely disengage itself in the event of potential harm.

Attention should also be paid to the potential environmental impact of the use of the AI system. In terms of measuring environmental impact, one way to do so is by estimating energy consumption throughout the design, development, and deployment process. Developers and deployers should ensure that energy consumption levels are within the appropriate range and take actions to reduce energy consumption where needed.

If the AI system or part of it is designed, developed, or deployed using a third-party developer or vendor, deployers should take actions to resolve or mitigate any non-compliance.

Before progressing further, roles and responsibilities for the design, development, and deployment of the AI system in a trusted and responsible manner need to be clearly defined and developed with the accompanying accountability mechanisms. Stakeholders who are involved in the development, review and/or approval of the problem definition and requirements of the AI system should be aware of the roles they play and be adequately equipped with the skills, knowledge and tools needed to carry out their duties.

Data collection and processing

What is the source of training and validation data used to train the AI system?

Is there a risk that the training data may not be representative of the population for one or more meaningful attributes?

Is personally identifiable or sensitive data used to make decisions or generate output?

Are there the necessary permission/approvals/ consent or lawful basis to use the data, especially for personal data, and even more critically for third-party sourced data?

Is a data protection impact assessment required? If yes, has it been done?

What is the amount of data being used and the duration that it is being held?

An AI system is only as good as the data used to develop it. Accordingly, data used for model training, testing, and validation should be sufficiently representative to mitigate risks of unjust bias. This can be done through constant monitoring of datasets used and variable performance of the model across different target population sub-groups. If developers are supplying AI systems to deployers, they should provide appropriate disclosure or supporting documentation that summarises the types of data used to train the AI system, and how they have managed potential bias. Types of bias that developers and deployers should look out for include:

- Representation bias – where training datasets poorly represent the real-world population the AI system intends to serve.
- Societal bias – where human biases and inherent cognitive biases are reflected in the behaviour of the AI systems.
- Labelling bias – where the process of labelling the training dataset can involve subjective decisions that can be a vector for introducing human biases into the AI system.
- Measurement bias – where features and labels are proxies for desired quantities, potentially leaving out important factors that affect the AI system's performance.[14]
- Activity bias – where AI systems get their training data only from their most active users and exclude those less active.[15]
- Proxy bias – where data directly relating to protected characteristics (e.g., race) are not used in the dataset, but other present data features with a correlation to such protected characteristics result in biases in the dataset.

[14] National Institute of Standards and Technology, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" (March 2022) < https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf >

[15] National Institute of Standards and Technology, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" (March 2022) < https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf >

AI systems trained on biased datasets could be less effective in performing their role and could even make unfair decisions or recommendations that may harm or prejudice users or individuals impacted by the decision. This could in turn lead to reputational harm or even legal liabilities for the organisation as well as a broader loss of societal trust in AI.

To mitigate the risks involved with the presence of existing bias in the datasets, developers and deployers should adopt the use of heterogenous and/or representative datasets, where data is gathered from various reliable sources. Developers and deployers must reasonably avoid the use of protected characteristics like gender, ethnicity, etc., in AI systems to drive decisions, but are encouraged to use them to assess AI system outcomes for unjust bias. Developers and deployers should also ensure that AI systems are trained on enough data relevant to their language(s), region, and industries. During data processing, developers and deployers should also try to avoid premature removal of data attributes that could help identify inherent bias and consider assigned roles for assessing and/or detecting bias in the processing pipeline for accountability. To mitigate the potential of labelling bias, the personnel responsible for labelling the data should be provided with clear guidelines to establish an objective and repeatable process for individual labelling decisions. In certain domains where the risk of labelling bias is high, labellers should also have adequate subject matter expertise and be provided training to recognise potential unconscious biases. A quality assurance mechanism can also be set up to monitor label quality.

Using different datasets for model training, testing and validation can also help developers check for systematic bias in AI systems. After the AI system is trained, it can be tested using data from different demographic groups. Deployers should regularly evaluate AI systems for bias and take steps to minimise cases where particular groups are systematically disadvantaged or advantaged by the AI system's output and make corrections when such cases are observed. One example of fairness testing is provided in AI Verify's Testing Framework[16] where fairness assessment includes technical tests to check whether an AI system produces different outcomes across different demographic groups or sensitive attributes such as gender or race. The AI system can also be validated with a validation dataset for further checks of bias.

Even after the AI system has been developed and deployed, it is important that deployers continue to review the system, datasets, and model metrics (e.g., data drift, precision, recall, bias, fairness) periodically and make reasonable effort to ensure the accuracy, relevance, and reliability of the data and outcomes. The data used for system development and testing should be similar to the data the AI system faces in the live production environment. Over time, data drift may occur, where the distribution of input data during the live production environment changes, leading to system performance degradation. Where the dataset is found to be outdated, new data obtained from the system during production can be used as new input data for model iteration. When the AI system has a continuous learning loop where system outputs during production is fed back to the model(s) as training and testing data, reinforcement bias might be introduced since the training data has gone through the system once and will carry any inherent biases the system has. Deployers who employ AI systems that utilise continuous learning loops should carefully evaluate the appropriateness of using system outputs as new training and testing data and take measures to mitigate any bias identified.

Other good data governance and management practices that will help the deployers ensure quality data for model training include maintaining a data provenance record and putting in place data quality measures. A data provenance record documents the end-to-end data lineage – where the data originally came from and how it was changed throughout its lifecycle. Knowing where the data originally came from, how it was

[16] Organisation for Economic Co-operation and Development, "Singapore's A.I. Verify builds trust through transparency" (16 August 2022) < https://oecd.ai/en/wonk/singapore-ai-verify >

collected and how it was processed will help developers and deployers understand how data accuracy and integrity is maintained over its lifecycle. It is also worthwhile noting that providing such information would also likely be material for fulfilling the consent and notice requirements under various data protection and privacy laws in ASEAN. Deployers may look at the data from its end-use and backdate it to its source or start from the data's source and follow its journey through to its end-use. For a more complete view, deployers may also choose to look at the end-to-end data lineage – starting from both the data's source and its end use. When data is obtained from a third-party provider, it can be difficult to establish its true origin. In this case, it is advisable for deployers to carefully assess the feasibility and risks of using that data before proceeding further. Where AI systems are not developed in-house, deployers will need to work with developers for such assessments.

Developers and deployers may reference the relevant ISO standards for data robustness, quality, and other data governance practices. One example is the ISO 2700 Series, which is a series of standards that provides guidelines for the protection of personal data in the cloud.

Developers should also pay attention to data quality throughout the system lifecycle. There are a few factors that may affect data quality and some examples include:

Accuracy: How well the values in the datasets match true characteristics of the entities described by the dataset.

Completeness: Whether the dataset can represent the entirety of the entities described by the dataset, both in terms of attributes and items, and whether it is relevant to the problem statement or scope of coverage of the AI system.

Credibility: Whether the data originated from a reliable source and whether its veracity can be ascertained.

Relevance and representativeness: Whether the data fits the context of the data collection and objective of the AI system and is accurately reflective of real-world demographics that the AI system will be exposed to.

Human interactions: Tracking how the data has been edited or amended by humans, and whether such amendments or edits are material to the usefulness and/or accuracy of the data.

Good data governance practices are a key factor in ensuring data quality. These include:

Documentation: Ensuring that documentation of data sources, data transformations, and data processing steps are kept up to date.

Data Lineage: Ensure that data can be followed across its lifecycle, from its source to its current target. This involves tracking every transformation step and link between data points to provide a clear understanding of how data evolves. One way of doing this is by keeping a data provenance record, which allows an organisation to ascertain the quality of the data based on its origin and subsequent transformation, trace potential sources of errors, update data, and attribute data to their sources.

Data storage: Define storage standards (data formats, structures, locations, etc.) to ensure consistency of data storage. Data retention policies also need to be defined to ensure that data is stored for the appropriate length of time and disposed according to the appropriate procedures.

Data security and privacy: Implement measures (e.g., encryption, access controls, anonymisation, etc.) to secure and protect data from unauthorised access and comply with privacy and data protection regulations.

When sourcing and processing data for AI development, developers should assess if a data protection impact assessment is needed and ensure that the data sourced for analytics is not used in a manner incompatible with its intended use or that data has not been improperly disclosed. Personal data should only be collected in accordance with applicable privacy and data protection policies of the organisation and applicable legal requirements of the country of operation. In the data processing phase, it is important that model development datasets have been reviewed to minimise use of potentially sensitive or personal data. Anonymised data should be used where possible and if it does not compromise on model quality. In the event that developers and deployers obtain personal data from third party sources, there should be appropriate due diligence to check and ensure that the third-party source is authorised to collect and disclose personal data on behalf of the individual, or that the source had obtained the necessary consent for disclosure of the personal data.

To minimise impact on the environment, developers and deployers should also make sure that the amount of data being held and the duration it is being held for is kept to a minimum to reduce data centres' energy requirements.

UCARE.AI

Illustration on operations management - documenting data lineage, ensuring data quality and mitigating bias

UCARE.AI is a Singapore-based deep-tech start-up, with a proprietary award-winning online ML and AI platform built on a cloud-based microservices architecture that provides real-time predictive insights, which can be applied to the healthcare sector and beyond.

UCARE.AI logged data consistently across multiple components and collected data in a secure and centralised log storage. In ensuring data quality, the company was also careful to transform its data into a usable format so that the properly formatted data could be used to build AI models. The company also prioritised creating AI models that were unique to clients, obtaining reliable datasets from the client to build models instead of using third-party datasets. Such a practice provided distinctions between patients' profiles and the features selected for each AI model differed for each hospital, contributing to greater accuracy in the bill estimations for patients. Another pertinent part of AI model development was minimising the risk of bias. For this, the objective and consistent machine predictions gave patients customised, data-driven predictions of their hospital bills instead of those subjected to human biases in algorithm development.

Modelling

Does a regulator require that the organisation can explain how the AI system arrived at its outcomes?

Does the AI system produce repeatable and reproducible results?

Is there proper documentation to track the AI model training and selection process?

During the system development process, developers should assess the approach and evaluate if AI systems are explainable, repeatable, reproducible, and robust. With the complexity of AI systems, it may not always be feasible to achieve all the above. Developers should adopt a risk-based approach to identify which model attributes are more relevant and necessary. There are two areas to consider in the risk-based approach – which features or functions of the model(s) in the AI system have the greatest impact on the users and which measures are likely to help establish more trust with users.

Explainability

Explainability is about explaining how AI systems function and how they arrive at certain decisions. To build trust in AI, it is important that humans understand how AI systems make predictions. However, in some cases such as "black box" models, it can be difficult to explain a model. Nonetheless, developers and deployers of AI systems containing such models can still work towards explainability by explaining how the predictions of AI system are used in the whole decision-making process. Explainability practices that developers and deployers can adopt when developing and deploying AI systems include:

- Auditability refers to the readiness of an AI system to undergo an assessment of its algorithms, data, and design processes. Ensuring proper documentation of the training and selection processes of the AI system, reasons for certain decisions made and measures taken to mitigate any risks found during the risk assessments will help developers keep track and remain accountable for the decisions made during the AI deployment process. Some developers may employ the use of automated machine learning which automates some or all the steps in the AI system lifecycle, such as feature engineering and hyperparameter tuning. Even for the steps that are automatically performed by automated machine learning, developers are encouraged to still consider how to incorporate transparency and explainability in such automated workflows. For example, documenting the data sources, hyperparameters, algorithms, and optimisation techniques initially selected by developers can help developers and deployers better understand how changes to these input parameters can potentially alter the machine learning outcomes.
- Where "black box" models are deployed, rendering it difficult, if not impossible to provide explanations as to the workings of the AI systems, outcome-based explanations, with a focus on explaining the impact of decision-making or results flowing from the AI systems may be relied on. Another method that can be considered is AI Model Cards, which are short documents accompanying trained machine learning models that disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information[17].
- Deployers may also consider focusing on alternative aspects of the quality of the AI system or preparing information that could help demonstrate and justify the outcomes of an AI system's processing behaviour (e.g., documenting the repeatability of results produced by the AI system). If need be, they can work with developers to develop such information. Some practices to demonstrate repeatability include conducting repeatability assessments to ensure deployments in live environments are repeatable and performing counterfactual fairness testing to ensure that the AI system's decisions are the same in both the real world and in the counterfactual world.
- For AI-enabled products and devices, it is advisable for developers and deployers to include descriptions of the expected behaviour of the AI into accompanying technical specifications or product documentation. Such descriptions can include the rationale behind why certain features or models were selected during the AI system development process. Provision of such

[17] OECD.AI Policy Observatory, "Catalogue of Tools & Metrics for Trustworthy AI" (15 September 2022) <https://oecd.ai/en/catalogue/tools/model-cards >

information helps deployers remain accountable to individuals and/or regulators and helps AI systems become more explainable and transparent. As far as possible, deployers that do not develop in-house AI systems and procure them from developers should appropriately govern their relationships with these developers through contracts that allocate liability in a manner agreed between parties. Deployers can consider including an obligation for developers to help support them in meeting transparency and explainability needs in the contractual terms. This could help deployers obtain the necessary support from third-party developers on retrieving or understanding the relevant information about how the AI system functions and its expected behaviour. An example of information that deployers can obtain from third-party developers would be AI Model Cards, which provide details on how the AI model functions, model development and testing process, and limitations of the model. Some examples of AI Model Cards can be referenced from Google's Face Detection Model Card[18] and Salesforce's Einstein Optical Character Recognition (OCR) Model Card[19].

- Deployers can also collaborate with developers to conduct joint audits and assessments for transparency and explainability.
- Deployers can also use explainability tools to evaluate the quality of their explanations. Some examples of such tools include AI Verify's Toolkit[20], OmniXAI[21], AIX360, Shapley Additive Explanations (SHAP)[22], and Local Interpretable Model-agnostic Explanations (LIME)[23]. It should be noted that most of the tools mentioned provide technical (explicit) explanation and are meant to be used by technical users (i.e., data scientists).

Even with the use of tools to explain AI systems, it may be difficult for the layman, and even experts, to understand exactly how AI systems work. In these cases, deployers can consider using implicit explanations of AI instead. For example, deployers could use comparisons such as "users with similar profiles as you were recommended these similar products" to help users understand to some degree how the AI system uses other users' data and their purchase history to recommend products to them. Counterfactuals may also be a useful means to explain how changes in variables or data affect outcomes of AI systems. For example, informing users that "you would have been approved if your average debt was 15% lower".

Although explainability is generally encouraged to promote transparency and build trust in AI systems, there may be situations where it does not make sense to disclose important information about how AI systems function. For example, the workings of fraud detection AI systems need to remain confidential to the organisation so that bad actors are not able to circumvent them. There also needs to be an appropriate balance between transparency and ensuring that companies' Intellectual Property (IP) is protected. Certain algorithms may be essential to business operations and bottom line, such as trading algorithms for robo-advisors, and disclosing such information will put confidential or proprietary business information at stake.

[18] Google, "Face Detection Model Card v0" < https://modelcards.withgoogle.com/face-detection >

[19] Salesforce, "Salesforce Einstein Model Cards" < https://resources.docs.salesforce.com/latest/latest/en-us/sfdc/pdf/salesforce_ai_model_cards.pdf>

[20] Infocomm Media Development Authority, "Fact Sheet – Open-sourcing of AI Verify and set up of AI Verify Foundation" (2023) < https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2023/06/7-jun---ai-annoucements---annex-a.pdf >

[21] Salesforce, "Welcome to OmniXAI's documentation!" (2022) < https://opensource.salesforce.com/OmniXAI/latest/index.html >

[22] SHAP, "Welcome to the SHAP documentation" (2018) < https://shap.readthedocs.io/en/latest/ >

[23] C3.ai, "What is Local Interpretable Model-Agnostic Explanations (LIME)" < https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/#:~:text=What%20is%20Local%20Interpretable%20Model,to%20explain%20each%20individual%20prediction. >

UCARE.AI

Illustration on explainability in the model lifecycle

UCARE.AI deployed its AI-Powered Cost Predictor (AlgoExpect^{TM}) in Parkway's four Singapore hospitals to provide dynamic real-time predictions of bill size at pre-admission at 82% accuracy.

UCARE.AI has incorporated explainability directly into the AI Cost Predictor model. Along with providing prediction, it is also able to tell on demand what are the important features that contribute to each prediction result. Client applications consuming the model's prediction service are also able to ask the model to "explain" each prediction result without going through UCARE.AI's support team. This helps users to understand the model predictions and provides greater transparency for the model's performance and instils greater trust.

The shared professional trust and respect between UCARE.AI and its clients in turn helped to build the recognition of the company as a reliable and trusted partner in data management and developer of AI models.

Robustness

Robustness helps establish trust in the performance of AI systems, especially in unpredictable circumstances. Robustness refers to the ability of the AI system to still function as intended or in a safe manner in the face of errors during execution or erroneous input. It can be measured by the extent to which the AI system can function correctly with invalid input or environmental stress. Robustness is an especially important attribute for building trust with users as it shows that the AI system can withstand a range of input environments.

AI systems are only as good as the data used to train and test them. It is difficult for the AI system to be trained on every possible precondition and scenario that it will face, especially when it is deployed in the real-world with dynamic human interactions. When faced with an unfamiliar input, the system might produce insensible or unintended outputs.

Developers and/or deployers can assess robustness by testing the AI system on various foreseeable erroneous input and scenarios. This can be done via adversarial testing, which is a series of tests to expose the system to a broad range of unexpected inputs and mitigate any unintended behaviour before the deployment of the AI system in live environments.

Even with the use of adversarial testing, AI systems are not immune to changes in inputs and operation environments that occur over time. To address this, some deployers or developers may choose to adopt continuous learning practices, where the learned parameters of the AI system are not fixed and can continue to change as the AI system is deployed in the live environment and learns from data it receives. However, it is still important for deployers to closely monitor the AI system and ensure that it does not learn unintended behaviour in the process.

Outcome analysis

How does the organisation determine if the AI system developed is fit-for-purpose?

How does the performance of the AI system compare to industry standards?

After the design and development process, deployers need to confirm that the AI system's outcomes are fit for purpose, achieve the desired level of precision and consistency, and are aligned with ethical, lawful, and fair design criteria.

During the project governance and problem statement definition phase, deployers should have clearly documented the intended purpose of the AI system and assessed that the risks associated with the AI system will be mitigated with appropriate measures. Risk identification and analysis are also necessary to address the root cause of the risks encountered.

In the outcome analysis phase, business stakeholders should be involved to observe the performance of the AI system and validate that it fulfils the purposes laid out at the start. Organisations can also consider conducting acceptance tests, which may cover functional and non-functional aspects, including security and performance evaluations. In some cases, developers and deployers may also choose to compare the actual Return on Investment (ROI) of the AI system against the planned ROI that was estimated during the planning phase.

To facilitate the outcome analysis process, it is important to establish a clear communication channel between the technical team and other stakeholders to ensure mutual understanding of AI system performance and potential improvements.

Fairness testing should also be conducted at this stage to ensure that the AI system does not make decisions that can result in unintended discrimination of certain demographics of users.

UCARE.AI

Illustration on operations management – robust model testing before deployment

UCARE.AI deployed its AI-Powered Cost Predictor (AlgoExpect^{TM}) in Parkway's four Singapore hospitals to provide dynamic real-time predictions of bill size at pre-admission at 82% accuracy.

UCARE.AI worked with its clients to create a validation framework to strengthen the AI model's accuracy, making sure to obtain patients' feedback on the framework for further fine-tuning. The Cost Predictor's AI model then underwent User Acceptance Testing, where the end business users from each hospital were invited to test the solution and provide feedback on various predictions.

GOJEK

Illustration on operations management - conducting outcome analysis for AI models

Gojek is an Indonesian on-demand multi-service platform and digital payment group based in Jakarta. In order to provide a pleasant and efficient experience to all actors of the platform, Gojek leverages AI in various sectors, including, but not limited to, driver-order matching, cartography and fraud detection.

Before deployment of machine learning models, Gojek tests the models' performance metrics against a set of predefined offline benchmarks. Such benchmarks allow evaluation of the AI model's performance in a controlled environment with a fixed set of data, which helps Gojek measure the variation in model outputs for successive iterations under the same operating conditions. Conducting offline benchmarking also allows Gojek to test model performance without any real-world risk or harm to end users.

Deployment and Monitoring

How critical is it if the AI system’s malfunction is not monitored or reported?

Would the failure of the AI system (unexpected or corrupted results) raise concern over any areas?

Deployers need to ensure that AI systems are scalable and deployable with the right technical infrastructure. Mechanisms should be established to measure the processing power use of all AI systems in deployment to ensure that there is enough processing power to run all the AI systems deployed by the organisation and that energy consumption is sustainable and within expectations. Measures should also be put in place to ensure that the underlying architecture is secure and resilient, so that AI operations would not be interrupted by adverse events.

To ensure ease of maintenance of AI systems, deployers can also maintain a system and technology inventory that keeps a record of programming languages, software packages, third-party vendor programs and hardware that have been reviewed and approved for use.

Deployers should set up mechanisms to monitor the AI system’s performance, for example, by establishing monitoring and reporting systems as well as processes to ensure that the appropriate level of management is aware of the performance and other issues relating to the deployed AI system. The scope, frequency, and timeliness of monitoring activities, as well as the relevant follow-up actions, and definition of metrics and thresholds required by the supporting teams to flag any errors should be defined.

Where appropriate, monitoring of deployed AI systems can include autonomous monitoring to effectively scale human oversight. AI systems can be designed to report on the confidence level of their predictions, and explainability features can focus on why the AI system had a certain level of confidence.

As part of monitoring, it is also good practice to revalidate the AI system for early detection of underperformance. Technical interventions like continuous system monitoring for performance drift issues can also be implemented. To ensure safety, testing or revalidation may also need to assess the degree to which an AI system generalises well and fails gracefully.

In cases where the performance of the AI system is found to have dropped over time, regular tuning of the models in the AI system is a measure that allows developers and deployers to keep AI systems relevant and up to date. Developers and deployers should have internal policies and processes to mandate regular model tuning, allowing them to iterate models based on updated training and testing datasets. These new datasets should incorporate new data from the production environment, to more accurately reflect the deployment environment that the AI system operates in. Regular tuning may also be necessary in cases where commercial objectives and risks change. For example, the function and scope of an AI-powered chatbot may change according to consumers' preferences and needs.

After tuning the models in the AI system, it is also important that developers and deployers continue to test the AI system in environments that best reflect the dynamism of the live environment. This ensures that the AI system does not just learn from regularities in the environment but is able to perform as intended when faced with dynamic inputs.

UCARE.AI

Illustration on operations management – continuous monitoring of deployed AI models

UCARE.AI deployed its AI-Powered Cost Predictor (AlgoExpect^{TM}) in Parkway's four Singapore hospitals to provide dynamic real-time predictions of bill size at pre-admission at 82% accuracy.

After the deployment of the Cost Predictor, UCARE.AI continuously monitored and iterated the algorithm, improving the data and simplifying the process for better accuracy. This continual training of the AI models ensured that the algorithms remained up-to-date and functioned with more precision after each data input.

EY

Illustration on operations management – continuous monitoring of deployed AI models

EY's purpose is building a better working world, and as a leader in Artificial Intelligence, this means that EY is committed to developing and deploying trusted AI solutions both internally and for its clients.

EY bears in mind that it can be a continuous process of fine-tuning models, even after they have been deployed. Leveraging EY Fabric capabilities, deployed models can be monitored to ascertain the performance and need for corrective action over its full life cycle. Each model's input data, output data and reference data are monitored to determine if any corrective measures to re-train or re-develop the model are needed.

4. Stakeholder interaction and communication

It is important that appropriate steps are taken to develop trust with stakeholders throughout the design, development, and deployment of AI.

In line with the principle of transparency, deployers should consider providing general disclosure of when AI is used in their product and/or service offerings, information such as the type of AI system used, the intended purpose of the AI system and how the AI system affects the decision-making process in relation to users. One example is the use of chatbots on websites – deployers should inform users that the answers are provided by an AI-powered algorithm, and they are not interacting directly with human customer service agents.

Where applicable, deployers may also consider putting a disclaimer that any inputs made to a chatbot or any AI-powered application will be used as additional data to improve or train the AI system in order to give data awareness to the user. This, however, needs to be coupled with a risk-based approach as bad actors may use this opportunity to skew the learning of the AI system if they know that such data might be used for continuous learning of the AI system.

The deployment of AI systems might cause a major shift in the roles and responsibilities of certain individuals in the company. For example, some employees may find that some aspects of their jobs are now being augmented or made redundant by AI-enabled technologies. Deployers need to be sensitive about the change management aspect of AI system deployment and make sure that there is adequate communication with employees to help them understand the AI system deployed and how it changes previous workflows.

To help employees adapt to an AI-augmented work environment, deployers should also look into wider awareness raising, providing training and education opportunities for employees to understand how to work with AI systems and how much benefit it brings. They should also try to redesign existing jobs so that AI augments existing skillsets of workers rather than renders them redundant[24].

As communication needs differ for different types of stakeholders, deployers can consider developing a standardised policy that dictates what level of information, who to provide the information and how to provide the information to stakeholders. This ensures that there is consistency and common understanding across the organisation on stakeholder communication. As deployers set out to develop such policies, they can first start by identifying the target audience and the purpose and context of interaction – for example, if they are internal or external stakeholders and what role they play in the AI system lifecycle. In the case where the target audience are users of AI systems, deployers can consider providing information related to the needs of the user as they navigate the interaction with the system, such as being informed whether AI is used and understanding how the AI system is expected to behave in normal circumstances. In the case where decisions made by the AI system will affect users, developers and deployers should provide more specific and detailed information relating to how the AI system arrives at a decision and how users will be affected by the decisions that the AI system makes.

Deployers should also put in place feedback mechanisms for users and other stakeholders to give feedback on the performance and output of the AI system. This will allow deployers to make iterative adjustments to the AI system to ensure it continues to perform as intended. Feedback channels and mechanisms for managing communications with aggrieved individuals should also be implemented and adapted to assist individuals who have queries or concerns about the impact of decisions or outcomes made by AI systems.

[24] For further guidance, organisations may wish to refer to Singapore's Personal Data Protection Commission, "A Guide to Job Redesign in the Age of AI" (2020) < https://file.go.gov.sg/ai-guide-to-jobredesign.pdf >

Some examples of feedback mechanisms include:

Feedback channels

Channels that could be used for users to raise feedback or queries. Where users find inaccuracies in their personal data which has been used for AI-augmented decisions affecting them, such channels can also allow them to correct their data. Such channels can be managed by the Data Protection Officer or Quality Service Manager where appropriate.

Decision review channels

Avenues for individuals to request a review of AI system's decisions that have materially affected them. Where the effect of a fully autonomous decision on a user may have significant impact, it would be reasonable to provide an opportunity for the decision to be reviewed by a human.

Where feasible and practical, deployers can also consider if users should be given the choice to opt out of the AI-enabled service and if yes, how this option can be exercised by the users. In the case where users are not given a choice to opt out, deployers should implement processes to engage with and obtain feedback from users or individuals impacted by the output of the AI system. Some factors that deployers can consider in determining whether users should be given the choice to opt out include:

- Degree of risk/harm to the individuals
- Availability of alternative decision-making mechanisms
- Technical feasibility
- Reversibility of the decision made
- Complexity and inefficiency of maintaining parallel systems

Sometimes, the AI system is required to learn from real-life input data (i.e., training the AI system based on user inputs). Before making use of users' data for AI system development, deployers are encouraged to inform users that their data will be used for subsequent model training so that users can understand and, where required, provide informed consent to how their data will be used. Since data from users are used to train AI systems in such cases, developers should adhere to applicable data privacy and protection laws and ensure that proper measures are put in place to safeguard the security of user data. Deployers should also look into setting up acceptable user policies that outline the boundaries for interacting with the AI systems. Some boundaries include prohibiting the use of malicious and offensive language or images and abuse of the AI system.

Deployers that do not develop AI systems in-house may contract a developer to help them in the design and development of AI systems. To facilitate the obtaining of information needed by companies to facilitate stakeholder management, deployers may wish to institute obligations for developers to provide the relevant information as part of the contract between the parties. Depending on the nature of the AI system and the risks involved, some categories of information that deployers can request from AI developers could include:

Data – sources, types of data, etc.

Model development process – feature engineering, model selection process, etc.

Safeguards within the model – measures put in place to mitigate negative effect of AI on humans.

Deployers may also use off-the-shelf or open-source solutions in AI systems. In these cases, data used or AI system design and development considerations may not be available to them. To ensure alignment with the principles and recommendations in this Guide, developers and deployers can consider the following actions:

- Gain as thorough an understanding as possible of the open-source or off-the-shelf solution that it intends to use, including its capabilities, other use cases, etc. This will help in assessing the suitability of the solution for the AI use case and identify any governance considerations that need to be addressed.
- Evaluate the significance of information that is not available to them and assess the corresponding risk of the absence of such information to determine the feasibility of using the open-source or off-the-shelf solution. For example, a developer that develops facial recognition systems needs to know the type of demographic of the dataset that the open-source AI system was trained on to assess if it is adequately representative of the population. Absence of such information may result in them being unable to evaluate whether the AI system will be fit for purpose in their specific local context.

Developers can play a key role in supporting deployers in implementing the AI governance framework laid out in this Guide. Some areas include:

- Incorporating features that promote transparency and explainability into the AI system, so that deployers can better understand the decision-making process of the system. For example, providing descriptions (e.g. data format, data recency, pre-processing steps, etc.) of the datasets used in the development process and providing visibility on the more significant features that affect the AI system’s decisions.
- Sharing the limitations and potential risks associated with the AI system for deployers to determine whether to conduct their own impact assessments and/or if it is acceptable for their use case.
- For deployers that are less mature in AI governance practices, third-party developers can potentially offer guidance and expertise on AI governance best practices and can assist deployers in establishing policies and controls for deployment of AI systems.

GOJEK

Illustration on stakeholder interaction and communication

Gojek is an Indonesian on-demand multi-service platform and digital payment group based in Jakarta. AI is leveraged for user-base growth and maintaining engagement of consumers through automated allocation of promotions, under budget constraints. Automated allocation of promotions identifies users with high incremental engagement potential given incentives and prioritises promotion allocation accordingly while estimating cost of campaign.

Consumers interact with promotional campaigns deployed by Campaign Managers. As such, they provide implicit feedback on the relevance of campaigns, which is captured in model online metrics. Thanks to this mechanism, the Data Science team and Campaign Managers can take informed decisions in model version management.

SMART NATION GROUP (SNG), SINGAPORE

Illustration on stakeholder interaction and communication

Singapore’s National AI Office (NAIO) established guidelines for product teams in the government building custom LLM products, as well as an AI workgroup (with stakeholders from across government) to oversee the rigorous testing and safe deployment of products.

To mitigate the risks of misuse of LLM products, NAIO recommends that all products should include visual UX cues to educate users on proper use. Such visual cues include reminders to always double check and adapt generated output for use.

Product teams are also encouraged to consider other education efforts such as workshops, guidebooks, and EDMs to raise awareness and literacy of LLM products.

MINISTRY OF EDUCATION, SINGAPORE

Illustration on stakeholder interaction and communication

Singapore’s Ministry of Education (MOE) is developing an AI-enabled Adaptive Learning System (ALS) for deployment within the Student Learning Space (SLS), Singapore’s national online learning portal. The ALS is one of MOE’s three educational AI use cases announced under the National AI Strategy launched in November 2019.

Various stakeholders were engaged for their views during the design and development of the ALS. Ideas and feedback from policymakers, curriculum and technical experts, as well as users (teachers and students) were sought and incorporated at the planning, building, and piloting phases.

To continually improve the performance of the ALS, teachers are also able to give suggestions for improvements directly to MOE. Based on their feedback on wanting more control and visibility about the content and assessment items being presented to students, the next iteration of the ALS will enable teachers to add their own resources into the resource pool for recommendation to their students, enhancing the quality of resources and improving the ALS’ recommendation algorithm.

National-level Recommendations

D. National-level Recommendations

This section is intended for policy makers.

This section covers national-level recommendations to ensure responsible design, development, and deployment of AI systems. It is also important to note that cross-border collaboration is essential for promoting responsible AI governance and ensuring that best practices and safeguards are shared within ASEAN and across different regions. Such collaboration can help facilitate the exchange of knowledge, expertise, and resources among different countries and regions.

Nurturing AI talent and upskilling workforce

With the increasing use of AI in a variety of industries, it is important to ensure that a country's workforce can adapt to the new ways of working and possesses enough digital skills to interact effectively with AI systems.

There are a few levels of AI capabilities that can be developed:

- Baseline level to operate AI applications;
- Intermediate level to maintain and rectify issues with AI applications; and
- Higher level to develop AI capabilities

Depending on the nature of jobs and industries, governments and policymakers should consider implementing measures to upskill the workforce according to the AI capabilities required.

Governments can collaborate closely with both public and private sectors to establish platforms and resources for employees to upskill themselves in their respective fields. Companies in private sectors bring with them industry-specific expertise and knowledge that can be used to curate training materials for AI-enabled technologies in different sectors. Public and private agencies can collaborate to organise industry-specific AI courses and forums to help personnel learn about the types of AI systems in the market and how they will increasingly play the role of operator of AI systems instead of manually performing tasks on their own. In Singapore for example, the Singapore Computer Society collaborated with IMDA and Nanyang Technological University to offer a professional certification program on AI Ethics and Governance. Industry players can also offer industry-based AI programmes where workers and graduating students can participate in short stints to learn more about the industry and its relevant AI technologies.

Individuals should also be able to acquire deeper insights into AI through schools and educational institutes. The curriculum of STEM disciplines needs to be periodically reviewed to keep up to date with the newest technological advancements in the data analytics and AI space. For other disciplines, it is also important that the concept of AI is introduced so students are aware of what AI can do and how it might automate part of their jobs in the future. Educational institutes can also invite AI experts to share knowledge and cross-pollinate ideas on use cases for AI and their corresponding AI governance considerations.

Beyond the technical knowledge of AI, there is also a need to increase individuals' understanding of ethical principles and how they are used to identify, evaluate, and mitigate the ethical implications of AI systems. For example, the principle of fairness and equity guides the identification and analysis of bias in AI systems, as well as the development of mitigation strategies. A robust curriculum for AI ethics should be developed and incorporated into STEM courses or training for AI practitioners to ensure that such individuals have a strong ethical foundation to guide future design, development, and deployment of AI technologies.

Certain groups of workers may experience significant changes in their job roles due to the introduction of AI. For example, automation of production and testing steps in manufacturing facilities will mean that certain job roles are made obsolete, since AI-augmented machines can work effectively and efficiently round-the-clock. Field agronomists may find that their on-farm services are less needed as AI-driven forecasting and agronomic advisory applications give farmers the guidance to optimally produce throughout the season. In these cases, it is important for policymakers to provide adequate support to affected employees as they transition to new job roles, such as offering training and upskilling programmes and access to other employment opportunities for those who wish to make a mid-career switch.

Governments should also collaborate with the private sector to determine the pool of future AI-trained graduates who will be needed to help their countries achieve longer-term digital economy objectives. To best develop this pipeline of talent and ensure educational institutions are equipped with curricula that align most closely with AI industry trends and their advances, governments (including respective Ministries of Education) should work with the private sector to:

- Co-develop curricula that are relevant and up to date with industry trends.
- Facilitate AI private sector and research professionals giving lectures to students.
- Encourage and help foster professional internship and co-opt programmes among educational institutions and the private sector.
- Encourage professors and lecturers at educational institutions to be seconded to private sector AI firms. This will ensure the latest AI technologies and trends are reflected in curricula, which will help to mitigate training time needed for fresh graduates once they go into the workforce.

Supporting AI innovation ecosystem and promoting investment in AI start-ups

Governments should take actions to foster the growth of the AI innovation ecosystem. A supportive environment for AI development should be created, where companies are able to access and leverage data, digital technologies, and infrastructure. There should be support for businesses to build the foundational digital infrastructure they need to implement AI, such as organising information sessions for senior leaders to learn about the types of infrastructure they can invest in and the leading industry solutions that can help them scale up AI design, development, and deployment efforts. A short-term grant could also be offered to companies who wish to kickstart the adoption of AI in operations.

AI development is a data-intensive process that requires the ingestion and processing of vast amounts of data. Some businesses may face difficulties in trying to obtain the required data and subsequently storing it in an appropriate manner. Governments can bridge these data gaps by making selected government data freely available for businesses to use and develop AI systems that could benefit national interests. In addition, encouraging data sharing within and between public and private sectors will help to create a rich data environment for AI to thrive.

Along with the provision of freely accessible data, support on setting up effective data governance measures and structure can also be provided to ensure that organisations are able to access and use the data in a responsible and ethical manner. For example, developers and deployers can be introduced to data anonymisation processes for data privacy and protection or data lineage tools to help them understand the end-to-end lifecycle of the data that is used during AI design, development, and deployment. Governments should encourage the adoption of privacy enhancing technologies (PETs). PETs refer to digital solutions that allow information to be collected, processed, analysed, and shared while protecting data confidentiality and privacy[25]. The use of PETs would be especially useful in making selected government data freely available to organisations for AI development and deployment.

To support the development of data pools that can help the development of AI, facilitating cross-border data flow will also play an important role. To enable this, governments should work with other governments to establish data sharing and data transfer mechanisms, to help facilitate companies' access to various data sources that can help develop and train AI systems. It is also useful for governments to work towards consistency and interoperability between national data protection legal frameworks and AI governance efforts. This should include highlighting key areas of interaction between data protection and AI, including lawful grounds for processing personal data as training data for AI systems, data protection impact assessments (where required), transparency obligations, rights of data subjects that should be respected (such as access, erasure, and correction), and parental consent and children's rights.

Beyond cross-border data flows, it is also worthwhile for governments to explore collaboration across the wider ecosystem, which includes regulatory bodies, non-profit organisations, AI industry leaders, and other stakeholders who are involved in the design, development, and deployment of AI systems. There needs to be discussions around different collaboration models where required stakeholders within the ecosystem contribute their own expertise at different levels of the AI journey. Notably, private and public sector collaboration ensures that both public and private entities develop and deploy AI in a secure and ethical manner.

Government agencies can also provide funding and grants to AI start-ups to help their employees gain the skills and knowledge to design, develop, and deploy better AI systems. This can be done through subsidising course fees for AI and AI ethics courses or providing start-ups with incentives when they adopt AI governance tools. Equipping start-ups with the skilled manpower that they need to develop and deploy AI responsibly is also crucial for building trust in AI and helping AI adoption to scale across the region.

Voluntary and mutually agreed knowledge sharing and exchange is also key to enriching the AI innovation ecosystem in countries. To facilitate the sharing of knowledge and best practices for AI, government bodies can collaborate with industry partners to organise forums where successful AI use cases are presented and experts on AI in the respective sectors share about their experience and the new developments in the AI space.

To facilitate the investment process in AI start-ups, governments can work with relevant parties to develop an online platform or marketplace where AI start-ups can post about their offerings and how they are adhering to AI governance principles, and investors can easily view all the AI start-ups and their needs in one place. Providing such a platform will not only increase the visibility of AI start-ups, but also encourage more AI companies to adopt AI governance practices as it is seen as a value proposition to potential investors.

Promoting investments in AI start-ups is important to ensure that they can scale sufficiently and subsequently develop into bigger players in the AI space. Governments should formulate policies to attract investment in AI start-ups, such as offering incentives to deployers that employ their services and encouraging AI adoption in both public and private sectors.

[25] Organisation for Economic Co-operation and Development, "Emerging privacy-enhancing technologies" (8 March 2023) < https://www.oecd-ilibrary.org/docserver/bf121be4-en.pdf?expires=1693742862&id=id&accname=guest&checksum=FD0CEDF75CDC287045BE68749B811CBE >

Investing in AI research and development

Investing in AI research and development ensures that countries are kept abreast of the latest developments in AI and can implement cutting-edge AI solutions for national problems faced by end users. Encouraging research related to the cybersecurity of AI, AI governance, and AI ethics is also key to ensuring that the safety and resiliency of AI systems and tools also advance in parallel with new use cases. Governments can fund AI research initiatives by providing grants for the hiring of required AI and STEM-related talent or developing a development sandbox for researchers to conduct pilots of their products. Setting up such AI sandboxes not only enables researchers to iterate upon their solutions in preparation for a formal launch into the market, but also provides a regulated environment where relevant authorities can ensure compliance of AI systems to regulations.

To boost AI research capabilities, governments can build a talent pool of AI experts who will specifically contribute to the design, development, and deployment of AI. Scholarships for post-graduate AI research or STEM-related undergraduate courses will help attract more locals and foreign talents to study and eventually work in the AI space. Private companies can also be incentivised to upskill their employees who are already working in STEM-related fields and build up their own internal AI research and development capabilities.

Ensuring that there is proper digital infrastructure is also another enabler for AI research and development. Governments can work with partners in the private sector to set up and invest in digital infrastructure such as creating a use case laboratory with appropriate data storage and computing power to improve research efficiency. Governments can also make AI research and development more accessible to organisations of all sizes by developing open-source AI platforms which are available for free and provide the necessary tools and computing resources to train and test AI systems more efficiently.

Governments can also work to establish triple helix partnerships between the research community, private sector, and government agencies to facilitate cross-pollination of ideas and healthy discourse on pioneering ideas and technologies.

Via open data and data sharing, governments can ensure that there is a thriving and rich data environment for AI systems to be trained. The quality and diversity of the large pool of data will ensure accurate and effective AI systems, accelerate the pace of innovation while also reducing the cost and barriers of entry into AI development. While more data is being made available to organisations looking to develop or deploy AI systems, it is also important that organisations work towards ensuring that AI systems have been trained on enough data relevant to their language(s), region, and industries.

Promoting adoption of useful tools by businesses to implement the ASEAN Guide on AI Governance and Ethics

AI governance processes can involve many rounds of checks and reviews by different stakeholders and may result in high costs when too many of such processes are done manually. For example, manual data validation steps can require employees to spend a significant amount of time inspecting a large number of datasets. As developers and deployers scale up AI systems, it is unsustainable to have employees manually inspect data and the model training process. Developers and deployers can make use of tools to enable the implementation of AI governance in operations and ensure that documentation and validation processes are more efficient.

One example of such tools is model provenance tools that keep track of and record every step of the system lifecycle. Model provenance tools can document details like where the training and testing data came from, data processing steps performed, feature extraction process and model selection process. Without the use of such tools, employees will need to manually record every detail and human error may result in inaccurate model provenance records. In order to scale up the use of AI in businesses, model provenance tools are essential to provide visibility across the lifecycle of many AI systems.

Another example of such tools is fairness tools to assess and mitigate fairness issues with AI systems. Such tools enable developers and deployers to test AI systems against a series of bias and fairness metrics and implement mitigation algorithms to correct unfair biasness. Computing fairness metrics can be computationally intensive and difficult for developers to perform manually. As more AI systems are integrated into operations, it is important that users' trust in AI are not negatively affected by inherent or other biases. Employing the use of AI fairness tools will enable developers and deployers to efficiently assess fairness metrics and gain the trust of users by designing, developing, and deploying fair AI systems.

The inability to understand how AI systems function and make decisions is one of the key reasons for mistrust in AI. Developers and deployers can use explainable AI tools to understand and interpret predictions made by machine learning systems. A What-If Tool is an example of an explainable AI tool that provides a user-friendly interface for developers to better understand the AI system. Through the What-If Tool, developers can change system inputs and re-run the system to observe how the system output changes in response to specific changes in input. This allows humans to have a better understanding of which features are more dominant in the system's decision-making process and validate some aspects of the AI system's behaviour.

Organisations can also adopt tools that have a range of testing capabilities for different principles. For example, AI Verify can be used to conduct technical tests and process checks on AI systems against principles of explainability, fairness, as well as robustness. There are also process checks for transparency, security, accountability, data governance and human-centricity[26].

In order to fully reap the benefits of such useful tools, employees will also need to know how they work and how to use them to design, develop, and deploy AI responsibly. Government agencies can provide subsidies or grants for employees who wish to go for training on how to operate and leverage such tools for AI design, development, and deployment.

Given that AI is continually evolving, it is also important to engage the wider developer community in co-developing tools. For example, in Singapore, IMDA has set up the AI Verify Foundation to harness the collective power and contributions of the global open-source community to develop the AI Verify testing tool for the responsible use of AI. The Foundation will boost AI testing capabilities and assurance to meet the needs of companies and regulators globally.

Apart from selecting such tools to deploy in the AI system lifecycle, it is important that developers and deployers also adopt best practices in data governance, software development and cybersecurity to ensure that such tools do not compromise the well-being of users.

[26] AI Verify Foundation, "What is AI Verify?" (2023) < https://aiverifyfoundation.sg/what-is-ai-verify/ >

Raising awareness among citizens on the effects of AI in society

By raising awareness of the potential risks and benefits of AI, citizens can make informed decisions about the appropriate use of AI and take appropriate actions to protect themselves from harmful uses of AI systems.

This can be done by engaging with the public through roadshows, social media platforms and other public forums to demonstrate AI technologies and their potential impact on society. Such public engagements can also be done in collaboration with private companies which provide AI-enabled services that are widely used in citizens' daily lives. Seeing how AI is used in these platforms that are commonly used by citizens will help them relate better to AI and be more aware of the risks of using AI-enabled technologies.

Policymakers and the relevant government agencies can also look at how to incorporate AI into general education to help citizens understand how AI works and the ethical considerations for responsible use of AI.

Government agencies can also consider implementing AI-enabled technologies for citizen-facing applications or platforms. For example, employing the use of chatbots and/or virtual assistants to provide 24/7 customer service to citizens and answer frequently asked questions. Along with the deployment of such AI technologies, a general disclosure should also be provided to inform citizens that they will be interacting with an AI system and what the AI system is expected to do when under normal behaviour. There should also be an avenue for citizens to provide feedback on the performance of the AI system, especially in cases where it malfunctions, so that the appropriate mitigating actions can be taken to prevent further misinformation or harm.

As AI systems increasingly become part of operations, it is important that vulnerable citizens are protected from malicious actors using AI with ill intentions, such as to generate or disseminate misinformation and online falsehoods. Governments should increase public education efforts on AI systems and the potential pitfalls so that citizens are empowered to be discerning consumers of information. One way to do this is through media literacy education to educate the public on how to identify and verify credible sources of information and combat misinformation from AI systems.

Regional-level Recommendations

E. Regional-level Recommendations

This section is intended for policy makers.

Setting up an ASEAN Working Group on AI Governance to drive and oversee AI governance initiatives in the region

The ASEAN Working Group on AI Governance can lead the technical and operational implementation of AI governance action plans in the region. The Working Group can consist of representatives from each of the ASEAN member states who can work together to roll out the recommendations laid out in this Guide, as well as provide guidance for ASEAN countries who wish to adopt components of this Guide, and where appropriate, include consultation with other industry partners for their views and input. The ASEAN Working Group can also lead efforts to further international cooperation on AI governance approaches, including through engaging ASEAN's various dialogue partners on AI governance issues.

The responsibilities of the Working Group can include (non-exhaustive):

1. Development and implementation of regional tools for AI governance
2. Provision of support and recommendations for policies to nurture AI talent and upskill workforce across ASEAN
3. Provision of support and recommendations for initiatives to support the AI ecosystem in ASEAN and promote investment in AI start-ups
4. Provision of platforms to encourage cross-pollination of ideas between the AI research and development community and ASEAN
5. Keeping abreast of new developments in the AI space (e.g., generative AI) and related emerging technology and making recommendations on updates to be made to the AI governance frameworks in the region

Adaptation of this Guide to address governance of generative AI

Generative AI, refers to a branch of artificial intelligence where new content, such as text, imagery, and audio, is created by algorithms in response to prompts. Unlike traditional AI systems that are primarily used for tasks such as recommending, filtering, and making predictions as a primary output, generative AI systems learn the underlying distribution of the training data and generate new and original data that resembles the input they were trained on. Generative AI systems have recently been able to produce outputs that are remarkably realistic and sometimes indistinguishable from human-created content. While generative AI shows promising transformative impacts in numerous fields including art, entertainment, scientific research and healthcare, the responsible development and deployment of generative AI also raises complex ethical, legal, and societal issues that require careful attention.

In particular, generative AI brings with it unique risks that may require new approaches to governing it. Some of these risks include:

- **Mistakes and anthropomorphism:** Generative AI models are not fool-proof and can sometimes make mistakes. However, these mistakes often appear to be highly coherent and human-like, and take on anthropomorphisation, commonly known as “hallucinations”. For example, some language models have created convincing but erroneous responses to medical questions, or generated software code that is susceptible to vulnerabilities.
- **Factually inaccurate responses and disinformation:** Generative AI systems can be used to generate and propagate false or misleading information to a larger scale, shaping public perception and eroding trust in reliable sources of information. One example of this is the dissemination of fake news, which may become increasingly difficult to identify if malicious actors use generative AI to generate convincing text, images, and videos with the intention to mislead others.
- **Deepfakes, impersonation, fraudulent and malicious activities:** Generative AI carries certain risks of being used for impersonation due to its ability to generate realistic content, such as text, images, and voices. One such risk is the use of generative AI to create deepfakes, which are manipulated videos or images that convincingly depict individuals saying or doing things they never actually did. Deepfakes can be utilised to deceive others into divulging confidential or personal information by posing as real individuals in online interactions or social networks. Another example is in the use of generative AI to develop phishing emails. In the past, phishing emails are often characterised by misspellings and grammatical errors. However, using generative AI to generate these emails omits this common “red flag”, thus making it more difficult for a user to distinguish between legitimate emails and compromised ones.
- **Infringement of intellectual property rights:** The development of generative AI systems requires huge amounts of data for model training, validation, and testing. This raises concerns about the use of copyrighted materials as some of these data collected and used may be copyrighted and generative AI developers may face legal repercussions if found using them without permission and in the absence of relevant fair use exceptions under local copyright laws.

  Generative AI systems can also learn from copyrighted material, such as images and music, without proper authorisation from the copyright holder. If the generated content closely resembles the style and/or form of the original copyrighted work and are insufficiently transformative, it may infringe upon the rights of the copyright holder.
- **Privacy and confidentiality:** Unlike traditional AI systems, generative AI systems have a tendency to memorise – which refers to the ability of a machine learning model to remember and reproduce specific examples from its training data[27]. Malicious actors may be able to reconstruct training data by querying the generative AI system, and this can compromise the privacy of individuals represented by the dataset, especially in cases where sensitive datasets (e.g., medical datasets) are used. In a corporate setting, employees may unconsciously disclose confidential information during their interaction with generative AI systems.

[27] Berkeley Artificial Intelligence Research, “Evaluating and Testing Unintended Memorisation in Neural Networks” (13 August 2019) < https://bair.berkeley.edu/blog/2019/08/13/memorization/ >

Propagation of embedded biases: Generative AI systems have the ability to capture and reflect the biases present in the training dataset. If not properly addressed, these biases can be inherited and result in biased or toxic output that reinforces biased or discriminatory stereotypes. For example, image generation systems prompted with “African worker” may generate images of individuals in tattered clothing and rudimentary tools, while simultaneously generating images of wealthy individuals when prompted with “European worker”[28]. This highlights the risk of propagation of biases from foundation models to downstream models trained from them, which perpetuates such biases and stereotypes.

Generative AI brings with it unique risks and the principles and components of AI governance defined in this Guide may need to be adapted to ensure responsible design, development, and deployment of generative AI. This is something that has already caught the attention of regulatory authorities across the world. In the region, Singapore’s IMDA and Aicadium have jointly published a discussion paper - “Generative AI: Implications for Trust and Governance”[29], which shares some of Singapore’s policy thinking on how to develop an ecosystem for trustworthy and responsible adoption of generative AI while building on existing frameworks and tools.

One area worth exploring is the allocation of responsibility and liability. Here, there is space for policymakers to facilitate and co-create with developers a shared responsibility framework. Such a framework aims to clarify the responsibilities of all parties in the AI system life cycle, as well as the safeguards and measures they need to respectively undertake.

Noting the increased attention on the role and responsibility of developers in generative AI, guidelines for ensuring developer accountability and integrity in the design and development of generative AI systems could also be developed.

This is also something that can be built off existing governance frameworks, which will need to be adapted to manage the unique risks of generative AI that have been detailed above. For example, given that generative AI systems are primarily used for content generation, it would be useful to provide specific guidance on how governance processes can be adapted to provide users with the ability to distinguish AI-generated content versus authentically generated ones through transparency. One way is to create a digital “watermark” that can be tagged to content generated by generative AI systems so users know when they are interacting with synthetically generated content and can make more informed decisions.

With the introduction and rapid uptake of generative AI systems, it is imperative that ASEAN builds on this current framework to develop governance guidelines for generative AI. This is also a good initial project for the proposed ASEAN Working Group to take up.

[28] Infocomm Media Development Authority and Aicadium, “Generative AI: Implications for Trust and Governance” (2023) < https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf >

[29] Infocomm Media Development Authority and Aicadium, “Generative AI: Implications for Trust and Governance” (2023) < https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf >

Compiling a compendium of use cases demonstrating practical implementation of the Guide by organisations operating in ASEAN

A compendium of AI governance practices could be useful to practically illustrate how various organisations and/or government agencies in ASEAN have implemented or aligned their AI governance practices with the ASEAN Guide on AI Governance and Ethics. It also shows how organisations have tailored the Guide to their industry needs and nature of business and benefited from the responsible use of AI.

Building a compendium of such AI governance examples in the region not only helps featured organisations stand out from others, but also builds trust with users and other stakeholders that use AI-enabled solutions. Showcasing the commitment of these organisations to AI governance also helps them promote themselves as responsible AI practitioners to other users outside ASEAN.

As more companies and government agencies are featured on the compendium, it is likely that organisations which have not implemented AI governance solutions will be inspired to do so as well, establishing a virtuous cycle of trust in AI.

Conclusion

Conclusion

One of the aims and purposes of ASEAN is to promote active collaboration and mutual assistance on matters of common interest in the economic, social, technical, and scientific fields. This Guide is collaboratively developed by all ASEAN member states and will serve as a regional best practice guidance on AI governance and ethics.

This Guide also serves as a practical guide for organisations and government agencies in the region that wish to design, develop, and deploy AI technologies. It is not meant to be exhaustive and static but rather, a living document that is periodically reviewed and enhanced as needed.

Member states, developers and deployers operating in their jurisdictions are recommended to apply, on voluntary basis, the provisions in this Guide. Nothing in this Guide may be interpreted as replacing or changing any party's legal obligations or rights under any member state's laws.

Annex A: AI Risk Impact Assessment Template

Annex A: AI Risk Impact Assessment Template

The Risk Impact Assessment below is adapted from Singapore's PDPC's Implementation and Self-Assessment Guide for Organisations.[30]

Purpose of AI Risk Impact Assessment

The purpose of the assessment below is to identify potential risks and vulnerabilities associated with the AI system and ensure that the design, development, deployment, and monitoring of the AI system complies with the components set out in the Guide above. The assessment also ensures consistent and systematic documentation of information that might be useful to the risk assessment of AI systems.

Who should use this assessment?

Developers and deployers of AI systems as well as AI governance/AI ethics committees within organisations.

How should this assessment be used?

The assessment below sets out a list of questions, based on and organised according to the four key areas described in Guide, for developers and deployers to consider in a systematic manner. The list of questions in the assessment below are a mix of open-ended and yes/no questions - developers or deployers filling in the assessment should fill in the 'Response' column respectively. Developers and deployers should refer to the Guide for definitions of terms and explanation of concepts used in the assessment below. In addition to the points called out by the assessment below, developers and deployers are free to implement other measures that best fit the purpose and context of their AI deployment, as appropriate.

When using the assessment, developers and deployers should consider whether the questions and practices are relevant to their unique business context and industry.

[30] Personal Data Protection Commission Singapore, "Companion to the Model AI Governance Framework– Implementation and Self-Assessment Guide for Organisations" (January 2020) < https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgisago.pdf >

<table>
<tr><th>#</th><th>Guiding Question</th><th>Useful industry examples, practices and guides for consideration</th><th>Response</th></tr>
<tr><td colspan="4"><b>1. Objectives of deploying AI</b>
This section guides developers and deployers on how to include ethical considerations in the development of the AI system’s problem statement and business objectives</td></tr>
<tr><td>1.1</td><td>Has a clear purpose in using the identified AI system been defined (e.g., operational efficiency and cost reduction)?</td><td>• Consider whether the AI system is able to address the identified problem or issue</td><td>[ ]</td></tr>
<tr><td>1.2</td><td>Has an assessment been done to verify that the expected benefits of implementing the AI system in a responsible manner (as described in the Guide) outweighs the expected costs?</td><td>• Consider whether to conduct a cost-benefit analysis
• Consider whether it is useful to leverage benchmarks and case studies for similar AI systems (e.g., PDPC’s Compendium of Use Cases vols 1 and 2[31])</td><td>[ ]</td></tr>
<tr><td>1.3</td><td>Is the use of the AI system consistent with the organisation’s core values and/or societal expectations?</td><td>• Consider developing a Code of Ethics for the use of AI that is in line with or can be incorporated into the organisation’s mission statement</td><td>[ ]</td></tr>
<tr><td>1.4</td><td>Does the AI system align with one or more of the organisation’s strategic priorities?</td><td>• Consider developing a list of strategic priorities for the organisation</td><td>[ ]</td></tr>
<tr><td colspan="4"><b>2. Internal governance structures and measures</b>
This section guides developers and deployers to develop appropriate internal governance structures to ensure that the relevant principles and recommendations set out in the Guide are adhered to</td></tr>
<tr><td>2.1</td><td>Is there an existing governance structure that can be leveraged to oversee the use of the AI system? If not, is there a governance structure in place to oversee the use of the AI system?</td><td>• Consider whether there is a need to adapt existing governance, risk, and compliance (GRC) structures to incorporate AI governance processes for the AI system
• Consider whether it is necessary to implement a process where each department’s head is accountable for the controls and policies that pertain to the respective areas, overseen by subject matter experts such as chief security officer or data protection officer</td><td>[ ]</td></tr>
</table>

[31] Infocomm Media Development Authority, “Compendium of Use Cases: Practical Illustrations of the Model AI Governance Framework” (2020) < https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgaigovusecases.pdf >

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
2. Internal governance structures and measures This section guides developers and deployers to develop appropriate internal governance structures to ensure that the relevant principles and recommendations set out in the Guide are adhered to			
		• Consider a sandbox type of governance to testbed and deploy AI systems, before fully-fledged governance structures are put in place	[ ]
2.2	Does the organisation’s board and/or senior management support and participate in AI governance?	• Consider whether it is useful to form a committee/board that is chaired by the senior management and include senior leaders from the various departments • Consider having senior management set clear expectations/directions for AI governance within the organisation	[ ]
2.3	Are the roles and responsibilities of the personnel involved in the various AI governance processes clearly defined?	• Consider defining separate roles and responsibilities for business and technical staff • Business staff responsible for defining business goals and business rules, and checking that an AI system behaves consistently with those goals and rules • Technical staff responsible for data practices, security, stability, error handling	[ ]
2.4	Are the personnel involved in various AI governance processes properly trained and equipped with the necessary resources and guidance to perform duties?	• Consider the importance and relevance of hiring talent with the right skillsets • Consider educating key internal stakeholders to increase awareness of AI governance guidelines • Consider educating employees, particularly those using the AI system or with customer-facing roles, to identify and report potential ethical concerns relating to AI design, development, or deployment	[ ]

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
3. Determining the level of human involvement in AI-augmented decision-making This section guides developers and deployers in determining the appropriate extent of human oversight in their AI-augmented decision-making process			
3.1	What is the probability and severity of harm on users or developers and deployers of the AI system? Please elaborate on the potential harm.	• Consider whether it is necessary to list all internal and external stakeholders, and the impact on them accordingly • Consider whether it is necessary to assess risks from a technical perspective and from a personal data protection perspective (e.g., PDPC's Guide to Data Protection Impact Assessments[32])	[ ]
3.2	What is the appropriate level of human involvement in AI-augmented decision-making?	• Human-in-the-loop, human-over-the-loop, human-out-of-the-loop • Consider severity and probability of risks and other relevant factors	[ ]
3.3	What are some mechanisms that must be put in place to continually identify, review, and mitigate risks after deployment of the AI system?	• Consider whether it is useful to determine and implement an appropriate review period for retraining the AI model(s) • Consider defining key performance indicators for the AI system's performance and putting in place measure to alert relevant employees when AI system performance deteriorates • Consider developing scenario-based response plans in the event that risk management efforts fail	[ ]

[32] Personal Data Protection Commission Singapore, "Guide to Data Protection Impact Assessments" (2021) <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/DPIA/Guide-to-Data-Protection-Impact-Assessments-14-Sep-2021.pdf >

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
3. Determining the level of human involvement in AI-augmented decision-making This section guides developers and deployers in determining the appropriate extent of human oversight in their AI-augmented decision-making process			
3.4	For safety-critical systems, how will relevant personnel be able to assume control where necessary? Or will the AI system be able to safely disengage itself where necessary?	• Consider whether it is necessary and feasible to put in place controls to allow the graceful shutdown of the AI system and/or bring it back to safe state, in the event of system failure • When an AI system is making a decision for which it is significantly unsure of the answer/prediction, consider designing the AI system to be able to flag these cases and triage them for a human to review	[ ]
4. Operations management This section guides developers and deployers in adopting responsible measures (as set out in the principles and Guide above) in the operations aspect of the design, development, and deployment of AI systems			
Ensuring personal data protection			
4.1	Are there accountability-based practices in data management and protection?	• Consider adopting industry best practices and engineering standards to ensure compliance with relevant data protection laws, such as PDPA[33] • Consider referring to OECD Privacy Principles[34] • Consider whether the AI system can be trained on pseudonymised or de-identified data	[ ]
Understanding the lineage of data			
4.2	Are there measures in place to trace the lineage of data (i.e. backward data lineage, forward data lineage and end-to-end data lineage)?	• Consider developing and maintaining a data provenance record • Consider whether it is useful to create a data inventory, data dictionaries, data change processes and document control mechanisms	[ ]

[33] Infocomm Media Development Authority, "Trusted Data Sharing Framework" (2019) < https://www.imda.gov.sg/-/media/imda/files/programme/data-collaborative-programme/trusted-data-sharing-framework.pdf >

[34] Organisation for Economic Co-operation and Development, "OECD Privacy Principles" (2010) < http://oecdprivacy.org/ >

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
4. Operations management This section guides developers and deployers in adopting responsible measures (as set out in the principles and Guide above) in the operations aspect of the design, development, and deployment of AI systems			
		- Consider whether it is useful to establish a data policy team to manage tracking of data lineage with proper controls	
Understanding risks of using datasets from a third party			
4.3	What are the risks of using datasets from third party?	- Consider obtaining datasets only from trusted third-party sources that are certified with proper data protection practices - Consider adopting the practices within IMDA's Trusted Data Sharing Framework[35] when establishing data partnerships (e.g., create a common "data-sharing language")	[ ]
Ensuring data quality			
4.4	Can the accuracy of the dataset in terms of how well the values in the dataset match the true characteristics of the entity described by the dataset be verified?	- Consider reviewing data in detail against its metadata - Consider whether it is useful to develop a taxonomy of data annotation to standardise the process of data labelling	[ ]
4.5	Is the dataset used complete in terms of attributes and items?	- Consider whether it is useful to conduct validation schema checks (i.e., testing whether the data schema accurately represents the data from the source to ensure that there are no errors in formatting and content)	[ ]
4.6	Is the dataset used up to date?	- N.A.	[ ]
4.7	Is the dataset used relevant?	- N.A.	[ ]

[35] Infocomm Media Development Authority, "Trusted Data Sharing Framework" (2019) < https://www.imda.gov.sg/-/media/imda/files/programme/data-collaborative-programme/trusted-data-sharing-framework.pdf >

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
4. Operations management This section guides developers and deployers in adopting responsible measures (as set out in the principles and Guide above) in the operations aspect of the design, development, and deployment of AI systems			
4.8	Is the dataset used well-structured and in machine-understandable form?	- Consider setting up an extraction, transformation, and loading (ETL) process	[ ]
4.9	If any human has filtered, applied labels, or edited the data, are there measures to ensure the quality of data?	- Consider whether it is necessary to assign roles to the entire data pipeline to enforce accountability. This would allow an organisation to trace who manipulated data and by which rule	[ ]
Minimising inherent bias			
4.10	Are there measures in place to mitigate unintended biases in the dataset used by the AI system?	- Consider taking steps to mitigate inherent bias in datasets, especially where social or demographic data is being processed for an AI system whose output directly impacts individuals	[ ]
4.11	Are there any data attributes that will be prematurely removed?	- Consider whether removed data is significant to detecting bias	[ ]
4.12	Will there be systemic bias that may result from data collection devices?	- N.A.	[ ]
4.13	Is the dataset used to produce the AI system fully representative of the actual data or environment the AI system may operate in?	- Consider benchmarking data distributions against population statistics to identify and quantify how representative the data is - Consider whether it is necessary to use a heterogeneous dataset (i.e., data collected from different demographic groups or from a variety of reliable sources)	[ ]

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
4. Operations management This section guides developers and deployers in adopting responsible measures (as set out in the principles and Guide above) in the operations aspect of the design, development, and deployment of AI systems			
Different datasets for training, testing and validation			
4.14	Are different datasets used for training, testing and validation of the AI model(s)?	• After training of the AI model, consider validating the AI model using a separate validation dataset	[ ]
4.15	Will the AI system be tested on different demographic groups to mitigate systematic bias?	• Consider whether it is necessary to test the results of different AI systems to identify potential biases produced by a certain system	[ ]
Periodic review and updating of datasets			
4.16	Are there measures in place to periodically review and update datasets to ensure its accuracy, quality, currency, relevance and reliability?	• Consider whether it would be useful to schedule regular reviews of datasets • Consider whether it would be necessary to update the dataset periodically with new data that was obtained from the actual use of the AI system deployed in production or from external sources • Consider exploring if there are tools available that can automatically notify your organisation when new data becomes available	[ ]
Explainability in Algorithm and System			
4.17	Can how the AI system functions and arrives at a particular prediction be explained?	• Consider implementing supplementary explanation strategies to explain AI systems, especially for systems that are less interpretable. Examples of these strategies include the use of surrogate models, partial dependence plots, global variable importance/interaction, sensitivity analysis, counterfactual explanations, or self-explaining and attention-based systems.	[ ]

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
4. Operations management This section guides developers and deployers in adopting responsible measures (as set out in the principles and Guide above) in the operations aspect of the design, development, and deployment of AI systems			
		• Consider putting in place a factsheet outlining the details on how the AI system operates, including how the model(s) was/were trained and tested (with what types of data), its performance metrics, fairness and robustness checks, intended uses and maintenance • Consider whether it is relevant to request assistance from the AI solution provider to explain how the identified AI solution functions	[ ]
4.18	Where explainability cannot be practically achieved, can lesser alternatives be considered?	• Consider conducting repeatability tests in a production environment • Consider performing counterfactual fairness testing	[ ]
Robustness in Algorithm and System			
4.19	Are there measures in place to ensure that the AI system is sufficiently robust?	• Consider whether it is relevant to conduct adversarial testing on the AI system to ensure that it is able to handle a broader range of unexpected input variables (e.g., unexpected changes or anomalies) • Consider whether it is necessary to put in place back-up systems, protocols, or procedures in the event the AI system produces unacceptable/inaccurate results, or fails	[ ]

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
4. Operations management This section guides developers and deployers in adopting responsible measures (as set out in the principles and Guide above) in the operations aspect of the design, development, and deployment of AI systems			
Active monitoring, review and tuning in Algorithm and System			
4.20	Are there measures in place for active monitoring, review and regular model tuning when appropriate (e.g., changes to customer behaviour, commercial objectives, risks, and corporate values)?	• Consider updating the AI system with new data points – set up an automated pipeline to update the system with newer data points via the extraction, transformation, and loading (ETL) process, and retrain the model(s) periodically when new data points are added • Consider whether it is useful to gather feedback from AI system users via multiple channels (e.g., email distribution lists, in-app feedback, and periodic user discussion forums)	[ ]
Traceability in Algorithm and System			
4.21	Will relevant information such as datasets and processes that yield the AI system's decisions be documented?	• Consider whether it is useful to track the AI system's decision-making process and performance using standard documentation (e.g., project objectives, scope, data and input values, error logs, etc.) • Consider whether it is useful to ensure that all data relevant to traceability is stored appropriately and retained for durations relevant to the industry	[ ]
Reproducibility in Algorithm and System			
4.22	Will an independent team be engaged to check if they can produce the same or very similar results using the same AI method based on the documentation relating to the model made by your organisation?	• Consider whether it is useful to make available replication files (i.e., files that replicate each step of the AI system's developmental process) to facilitate the process of testing and reproducing behaviour • Consider whether it is relevant to check with the original developer on whether the system's results are reproducible	[ ]

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
4. Operations management This section guides developers and deployers in adopting responsible measures (as set out in the principles and Guide above) in the operations aspect of the design, development, and deployment of AI systems			
Auditability in Algorithm and System			
4.23	Are there measures in place to ensure relevant documentation, procedure and processes that facilitate internal and external assessments of the AI system?	• Consider whether it is useful to keep a comprehensive record of data provenance, procurement, pre-processing, how the data has been processed, lineage of the data, storage, and security • Consider whether it is useful to centralise information digitally in a process log	[ ]
5. Stakeholder Interaction and Communication This section guides developers and deployers in implementing good communication practices to promote transparency and inspire trust among their stakeholders when designing, developing, and deploying AI systems			
5.1	Who are the various internal and external stakeholders that will be involved and/or impacted by the development and deployment of the AI system?	• Consider customising the communication message for the different stakeholders who are impacted by the AI system • Consider providing different levels of explanation for data, model, human involvement, etc.	[ ]
5.2	Are there measures in place to inform the relevant stakeholders that AI is used in products and/or services?	• Consider whether it is necessary to provide information on the role and extent that AI played in the decision-making process (e.g., statistical results and inferences) in plain language and in a way that is meaningful to the individuals impacted by the AI system (e.g. infographics, summary tables and simple videos) • Consider publishing a privacy policy on your organisation's website to share information about AI governance practices (e.g., data practices, and decision-making processes), which can include disclosure of third-party engagement, depiction of data flow and identification of standards that the organisation is compliant with	[ ]

#	Guiding Question	Useful industry examples, practices and guides for consideration	Response
5. Stakeholder Interaction and Communication This section guides developers and deployers in implementing good communication practices to promote transparency and inspire trust among their stakeholders when designing, developing, and deploying AI systems			
		• Consider informing users how the AI-enabled features are expected to behave during normal use	[ ]
5.3	Will users be given the option to opt out of the AI system by default or only on request?	• Consider informing users of the consequences of choosing to opt out, if such an option is available	[ ]
5.4	Are there communication channels in place for users to provide feedback or make enquires?	• Consider providing a hotline or email contact of relevant personnel such as a data protection officer or quality service manager on the organisation's website • Consider providing an avenue for individuals to submit updated data about themselves	[ ]

Annex B: Use Cases

Annex B: Use Cases

Aboitiz Data Innovation

GT2025

Together we are building the PH's first **techglomerate**

Aboitiz Group is over a century old conglomerate established in the Philippines with presence all over Asia such as Singapore, Thailand, Sri Lanka, Vietnam, Malaysia, Myanmar, Indonesia, and China. Its investments are in power, banking, and financial services, food, land, construction, shipbuilding, infrastructure, and data science and artificial intelligence. With its vision to build the first Philippine techglomerate by 2025, the group is the forefront of innovation and technology in the country.

Illustration on establishing internal governance structures and measures

Aboitiz recognises that AI and ML algorithms are integral assets of the group, therefore, it is imperative to have a strategic AI governance framework to ensure that the algorithms and programs are properly managed, to support the day-to-day operations of different strategic business units within the group. Consistent monitoring and evaluation of programs will aid to achieve the goal of implementing reliable and ethical AI-driven decisions. Effective AI governance helps the group improve existing AI-driven processes and decisions and mitigate the risks and challenges posed.

Core objectives of the policy is to support effective governance of the use of AI and ML programs within the group. This includes:

- Establishing appropriate responsibilities for all AI-related programs and decisions.
- Establishing clear risk assessment protocols of AI-driven decisions and appropriate measures to address the different levels of risk; and
- Regular review of existing organizational values and incorporation of ethical principles into the use of AI.

In addition to the Model Governance Policy, the group has also established the following **Internal Governance Structures and Measures for oversight of AI:**

- Ethical considerations related to the use of AI aligned with corporate values.
- Clear and defined roles and responsibilities for the ethical use of AI technology.
- All AI-related processes and decisions must be vetted by the management committee; and
- A multi-stakeholder approach with Model Governance -Management Committee which is composed of the representatives of the strategic business units - the Chief Information Security Officer (CISO), Data Protection Officer (DPO), Chief Operations Officer, Chief Data Officer, Chief Risk Officer, Chief Technology and Operations Officer, Audit, Risk, and Compliance AI, AI and Innovation Centre of Excellence, Chief Marketing Officer, and Senior Managers (Stakeholders of AI project).

Illustration on determining the level of human involvement in AI-augmented decision-making

To assess the **Risk Appetite for AI** usage the group has established a risk assessment procedure (i.e. pre-development AI risk assessment and pre-deployment risk assessment) to measure the different levels of risk. Its focus is on the impact of AI decisions on individuals which aims to help the group determine the appropriate levels of human involvement in AI-driven decisions and other remedial measures if necessary.

Illustration on operations management

Aboitiz implements AI governance measures throughout the AI project/system lifecycle from the design, development, deployment, and monitoring phase.

Illustration on stakeholder interaction and communication

Aboitiz adopts appropriate strategies to facilitate effective communication with stakeholders to harness collaboration and innovation.

EY

Building a better
working world

EY is one of the largest professional services firms in the world, providing a wide range of professional services to clients through four integrated service lines - Assurance, Consulting, Strategy and Transactions, and Tax. EY's purpose is building a better working world, and as a leader in Artificial Intelligence, this means that EY is committed to developing and deploying trusted AI solutions both internally and for its clients. Thus, EY has developed a unique perspective and toolset to help guide their employees and clients when developing AI systems. By leveraging these tools and framework, many organisations have benefited from the ability to rapidly grow AI adoption through trust. Its close adherence to the AI governance principles also provides EY and its clients with a significant level of confidence on the reliability of the developed AI implementations.

Illustration on developing guiding principles for internal governance structures and measures

EY is committed to the responsible implementation of AI systems and as such has defined a set of principles to adhere to when implementing and deploying AI within EY. The principles, **Accountability | Data Rights | Reliability, Safety, & Security | Transparency & Explainability | Fairness & Inclusiveness | Professional Responsibility** are aligned to the principles in the ASEAN Guide on AI Governance and Ethics. They outline key areas that guide oversight while implementing AI solutions. EY has already built foundational offerings for operationalisation of Trusted AI, which leverages its **Trusted AI Framework** to assess governance across the full AI program. The Trusted AI framework applies a continuous innovation process throughout the AI model lifecycle, which includes Agile Governance, Purposeful Design & Vigilant Supervision. This framework helps to determine the level of oversight required and the corresponding considerations for oversight management.

Illustration on operations management

EY has developed a language-agnostic XOps (DataOps, DevOps, MLOps, ModelOps etc.) platform called EY Fabric, a global technology platform, that leverages open-source technologies & frameworks to **standardize all aspects of model lifecycle**. It helps to reduce the development time with Standardised Model Development, Automated Model Documentation, Shortened Model Validation and Automated Model Deployment & Monitoring, while reducing risks by ensuring reproducibility and repeatability. EY's application of the Trusted AI Framework with EY Fabric provides confidence while automating the repetitive processes in model development. This helps to achieve efficiency and reduce development time.

Additionally, EY bears in mind that it can be a continuous process of fine-tuning models, even after they have been deployed. Leveraging EY Fabric capabilities, deployed models can be monitored to ascertain the performance and need for corrective action over its full life cycle. Each model's input data, output data and reference data are monitored to determine if any corrective measures to re-train or re-develop the model are needed.

Illustration on determining the level of human involvement in AI-augmented decision-making

EY adopts an AI Model Risk Tiering approach to assess and classify the models as High, Medium, or Low risk. The key areas of risk associated with AI, such as use case design, ethics, data, privacy, algorithmic, performance, compliance, technology, and business risks are evaluated to assign a risk tier for every AI model. Based on the risk tier, appropriate monitoring and human oversight are put in place for the AI models.

Gojek is an Indonesian on-demand multi-service platform and digital payment group based in Jakarta. Today, it is a super-app offering more than 20 services across several countries in Southeast Asia. As part of GoTo group, it contributed to 2.2% of Indonesia's GDP in 2022 creating employment opportunities for 1.1 to 1.7 million people (source, 29 March 2023), and connecting merchants (restaurants, grocery stores) with consumers. In order to provide a pleasant and efficient experience to all actors of the platform, Gojek leverages artificial intelligence (AI) technologies in various sectors, including, but not limited to, driver-order matching, cartography and fraud detection.

AI is also leveraged for user-base growth and maintaining engagement of consumers through automated allocation of promotions, under budget constraints.

This automated allocation of promotions identifies users with high incremental engagement potential given incentives and prioritises promotion allocation accordingly while estimating cost of campaign.

Illustration on operations management - conducting outcome analysis for AI models

Before deployment of machine learning models, Gojek tests the models' performance metrics against a set of predefined offline benchmarks. Such benchmarking allows evaluation of the AI model's performance in a controlled environment with a fixed set of data, which helps Gojek measure the variation in model outputs for successive iterations under the same operating conditions. To ensure repeatability, Gojek monitors and ensures that the variation is within an appropriate threshold before the model can be deployed. Conducting offline benchmarking also allows Gojek to test model performance without any real-world risk or harm to end users.

Illustration on operations management - monitoring deployed AI models

For the deployed AI models, Gojek also implements continuous monitoring to measure their online performance. The model performance is tracked over time in the production environment and improvement areas are identified. For example, if model accuracy starts to decline over time, it could be an indication of model drift and mitigating actions will be taken to ensure the model continues to perform optimally over time.

Illustration on internal governance structures and measures - defining roles and responsibilities for AI

Gojek has clearly defined roles and responsibilities for the development of AI for automated allocation of promotions.

a) **The Data Science team** is responsible for modelling the coupon-user interactions, maintaining a weekly model training pipeline and providing rollout/experiment allocation strategies week-on-week

b) **The Campaign Managers** are responsible for defining discount constructs, campaign budgets and traffic between rollout and experiment. The usage of AI models is transparent to the managers, and they feedback to the Data Science team if any changes in the model performance is noticed or if any changes are required in coupon allocation campaign strategies.

Illustration on stakeholder interaction and communication

Consumers interact with promotional campaigns deployed by Campaign Managers. As such, they provide implicit feedback on the relevance of campaigns, which is captured in model online metrics. Thanks to this mechanism, the Data Science team and Campaign Managers can take informed decisions in model version management.

UCARE.AI

unlock the impossible

[Use case 1] Data Security, Privacy & Transparency

UCARE.AI (https://www.ucare.ai/home/news/) is a Singapore-based deep-tech start-up, with a proprietary award-winning online ML and AI platform built on a cloud-based microservices architecture that provides real-time predictive insights, which can be applied to the healthcare sector and beyond. UCARE.AI's solutions have been used by customers such as Parkway Pantai, Singapore's Ministry of Health, Grab and Great Eastern Life Assurance to manage risk, contain cost and maximise efficiency.

Among their various solutions, UCARE.AI's AI-powered Cost Predictor (AlgoExpect^{TM}) works with hospitals to deliver accurate estimations of hospital bills to patients.

Illustration on operations management - good data accountability and adhering to privacy and data protection policies

UCARE.AI invested its efforts in good data accountability practices and treated the use of AI with openness and transparency. This provided tremendous benefits to patients in terms of seamless experiences in hospitals, greater certainty over their medical expenses and less re-financial counselling.

As a first step, when handling personal data for AI model development, UCARE.AI adhered to the requirements of various personal data protection laws and draft bills in its operating regions. Singapore's PDPA (2012) is one such law UCARE.AI kept in mind. Besides obtaining consent prior to any collection and use of personal data, UCARE.AI also made efforts to securely encrypt sensitive data. Its connectors were also designed to automatically detect such sensitive data and where possible, the algorithm was trained to minimise the use of this data in developing the AI model.

To further boost efforts in data protection, UCARE.AI anonymised client data at source before using it for development, thereby minimising the risk of inappropriate access to personal data. This also ensured that in the unlikely event of a breach, personal information could not be easily used to trace back to an individual.

The company also actively reinforced its commitment to data protection, cataloguing and evaluating every use of data that could be accessed by clients.

Illustration on operations management - documenting data lineage, ensuring data quality and mitigating bias

Understanding the lineage of data was also central in the accountable use of AI. Knowing this, UCARE.AI **logged data consistently** across multiple components and collected data in a secure and centralised log storage. In **ensuring data quality**, the company was also careful to transform its data into a usable format so that the properly formatted data could be used to build AI models. The company also prioritised creating AI models that were unique to clients, obtaining reliable datasets from the client to build models instead of using third-party datasets. Such a practice provided distinctions between patients' profiles and the features selected for each AI model differed for each hospital, contributing to greater accuracy in the bill estimations for patients.

Another pertinent part of AI model development was minimising the risk of bias. For this, the objective and consistent machine predictions gave patients customised, data-driven predictions of their hospital bills instead of those subjected to human biases in algorithm development.

Illustration on Transparency in the Use of AI and Data

To build greater confidence and trust in the use of AI, UCARE.AI was mindful to be transparent in its use of AI with various stakeholders. UCARE.AI not only disclosed the exact parameters used in developing the AI model to its clients, but also provided detailed explanations on all algorithms that had any foreseeable impact on operations, revenue, or customer base. UCARE.AI made a conscious decision to declare the use of AI in its analysis and prediction of bill amounts to Parkway's data managers and its patients.

Illustration on stakeholder interaction and communication

Clients with concerns about bill predictions were also encouraged to highlight them through UCARE.AI's communication channels. This gave stakeholders the necessary assurance on UCARE.AI's policies and processes for responsible AI use.

[Use case 2] Oversight, Validation & Monitoring

UCARE.AI, a deep-tech start-up with a proprietary award-winning online ML and AI platform built on a cloud-based microservices architecture that provides real-time predictive insights to help insurers, hospitals, pharmacies and governments manage risk, reduce medical cost, and maximise efficiency, with the end goal of making healthcare affordable to all.

UCARE.AI deployed its AI-Powered Cost Predictor (AlgoExpect^{TM}) in Parkway's four Singapore hospitals to provide dynamic real-time predictions of bill size at pre-admission at 82% accuracy. Based on this success, Parkway launched the Price Guarantee Programme for six hospital procedures. In its commitment to help patients with accurate cost estimations, UCARE.AI understood that trust was essential in driving adoption of its AI solutions. To achieve this, the company turned to the Model AI Governance Framework, aligning its practices in AI governance to those in the Framework to ensure reliability in its AI solutions.

Illustration on internal governance structures and measures - defining clear roles for internal oversight of AI

As a critical part of AI governance is oversight, UCARE.AI put in place certain internal governance measures, which includes **assigning clear roles and responsibilities for the ethical development and deployment of AI**. UCARE.AI's projects all include a primary and secondary data science lead to concurrently develop AI models for the same problem statement. Once completed, the data science leads would present their results to UCARE.AI's internal team, which consists of the CEO, CTO, CSO, project managers and the client services team for validation. During the project, UCARE.AI also conducted weekly check-ins with its clients to ensure quicker and more reliable iterations of its AI models. A final step before submission of the models to the client was to have UCARE.AI's appointed medical advisors assess the models' outputs for accuracy. After the models and its results have been submitted to the client for blind testing and approval, UCARE.AI's QA team reviewed and ensured that the model was production-ready before deployment.

Illustration on operations management - robust model testing and continuous monitoring of deployed AI models

UCARE.AI also conducted rigorous feasibility studies before developing the Cost Predictor. These studies helped address potential risks such as reduced accuracy in forecasted healthcare costs. With the studies, UCARE.AI then worked with its clients to create a validation framework to strengthen the AI model's accuracy, making sure to obtain patients' feedback on the framework for further fine-tuning. The Cost Predictor's AI model then underwent User Acceptance Testing, where the end business users from each hospital were invited to test the solution and provide feedback on various predictions.

After the deployment of the Cost Predictor, UCARE.AI **continuously monitored and iterated the algorithm**, improving the data and simplifying the process for better accuracy. This continual training of the AI models ensured that the algorithms remained up-to-date and functioned with more precision after each data input. More importantly, the methodology of continuous validation of the AI models with client inputs helped to boost confidence in the accuracy of the AlgoExpect^{TM}'s predictive insights.

Illustration on Explainability in the model lifecycle

UCARE.AI has incorporated explainability directly into the AI Cost Predictor model. Along with providing prediction, it is also able to tell on demand what are the important features that contribute to each prediction result. Client applications consuming the model's prediction service are also able to ask the model to "explain" each prediction result without going through UCARE.AI's support team. This helps users to understand the model predictions and provides greater transparency for the model's performance and instils greater trust.

Smart Nation
SINGAPORE

Singapore’s National AI Office (NAIO) was established under the Smart National Digital Government Office (SNDGO) to set the national agenda for AI and catalyse efforts across research, industry, and Government stakeholders to work on national AI priorities.

As part of its efforts to govern the use of Large Language Models (LLMs) in the public sector, NAIO has established guidelines for product teams in the government building custom LLM products, as well as an AI workgroup (with stakeholders from across government) to oversee the rigorous testing and safe deployment of products.

Illustration on internal governance structures and measures – approval gates at different stages of LLM product development

To encourage experimentation while also ensuring ample review of LLM products, product teams need only seek approval from the central AI workgroup from the beta testing phase onwards.

Embarking on internal experiments Proof-of-concepts or experiments within a development team	**No approval required**
Embarking on beta testing With test users outside of the product development team	**Approval gate 1** Seek approval from the central AI workgroup prior to embarking on beta testing. To maintain an agile process, teams provide key product details via a form, and AI workgroup members comment/endorse concurrently
Deploying the product To all intended users	**Approval gate 2** Seek central approval from the central AI workgroup prior to full deployment. Teams outline key feedback and improvements via a form, and AI workgroup members comment/endorse concurrently

Illustration on determining the level of human involvement in AI-augmented decision-making – determining the level of risk and corresponding mitigating measures of LLM use cases

NAIO takes a risk-based approach when advising product teams on required mitigating measures. The level of risk varies, depending on AI products':

I. Task (productivity and language tools VS factual information retrieval) [lower risk VS higher risk]

II. Audience (internal-facing VS external-facing) [lower risk VS higher risk]

For example, for public-facing AI products, the product should be made robust to adversarial attack. The corresponding mitigating measures include the product teams engaging in efforts to increase the robustness of their products, such as via robustness tests to improve performance against adversarial prompts, red teaming or bug-bounty programmes, and rate-limiting queries so users are deterred from brute-force attacks.

Illustration on stakeholder interaction and communication

To mitigate the risks of misuse of LLM products, NAIO recommends that all products should include visual UX cues to educate users on proper use. Such visual cues include reminders to always double check and adapt generated output for use.

Product teams are also encouraged to consider other education efforts such as workshops, guidebooks, and EDMs to raise awareness and literacy of LLM products.

Ministry of Education
SINGAPORE

Singapore’s Ministry of Education (MOE) formulates and implements education policies on education structure, curriculum, pedagogy, and assessment. It oversees the management and development of Government-funded schools, and the Institute of Technical Education, polytechnics, and universities.

MOE is developing an AI-enabled Adaptive Learning System (ALS) for deployment within the Student Learning Space (SLS), Singapore’s national online learning portal. The ALS is one of MOE’s three educational AI use cases announces under the National AI Strategy launched in November 2019.

The ALS provides a personalised learning pathway for each student, recommends learning resources, practice questions and customised feedback based on their level of readiness. These recommendations are based on the student’s mastery of pre-requisite concepts, their preferences and learning needs, responses to learning content, and are enhanced over time.

Illustration on determining the level of human involvement in AI-augmented decision-making

Students can use the ALS for self-directed learning at their own pace while teachers can monitor students’ performance and progress through the Learning Progress Dashboard and provide timely interventions where necessary. For greater learning support, students can also share their ALS learning artefacts with teachers for feedback and guidance.

ALS also allows teachers to recommend specific subtopics and concepts for each student and supplement ALS’ recommendations with their professional insights.

Illustration on operations management – continuous monitoring of deployed AI models

The team overseeing the ALS receives monthly reports on the models’ performance and accuracy. This enables the team to monitor the accuracy of ALS in recommending appropriate learning resources and ensure its performance remains above a pre-determined threshold.

Illustration on stakeholder interaction and communication

Various stakeholders were engaged for their views during the design and development of the ALS. Ideas and feedback from policymakers, curriculum and technical experts, as well as users (teachers and students) were sought and incorporated at the planning, building, and piloting phases.

To continually improve the performance of the ALS, teachers are also able to give suggestions for improvements directly to MOE. Based on their feedback on wanting more control and visibility about the content and assessment items being presented to students, the next iteration of the ALS will enable teachers to add their own resources into the resource pool for recommendation to their students, enhancing the quality of resources and improving the ALS’ recommendation algorithm.

Illustration on human-centricity

Students' learning readiness is a key determinant in how the ALS provides personalised and effective support for each learner. For example, questions are matched to the readiness of the students, so that all students are more likely to experience success and be motivated to learn.

To enable this, the design team engaged in user research during a pilot phase with a small number of schools, and a subsequent deep dive phrase. The data collected was then used to inform the subsequent refinements in the design and development of the ALS.

During development, the team also considered the choice of curriculum and pedagogical models, subject matter, assessment items and follow-up recommendations to enhance human centricity in the system.