PCPD.org.hk

香港個人資料私隱專員公署
Office of the Privacy Commissioner for Personal Data, Hong Kong

Artificial Intelligence: Model Personal Data Protection Framework

Supporting Organisations:

中華人民共和國香港特別行政區政府
政府資訊科技總監辦公室
Office of the Government Chief Information Officer
The Government of the Hong Kong Special Administrative Region of the People's Republic of China

Hong Kong Applied Science and Technology Research Institute
香港應用科技研究院

Table of Contents

Foreword

With its wide range of applications, artificial intelligence (AI) opens up abundant business opportunities. Although many enterprises are actively embracing AI technology with a view to increasing revenue, reducing expenditure and boosting productivity, the risks associated with AI should not be overlooked. For example, an AI system trained on insufficient or poor quality data may generate inaccurate or biased results. Further, if the training dataset contains personal data, they may be inadvertently disclosed during the output process.

Inevitably, the new risks arising from the innovative applications of AI present regulatory challenges. In response to the rapid development of AI, regulators around the world have rolled out various laws and regulations, including the *Artificial Intelligence Act* adopted by the European Parliament in March 2024, which aims to regulate AI systems according to their risk level, and the *Interim Measures for the Management of Generative Artificial Intelligence Services* issued by our Motherland in July 2023 with a view to promoting the healthy development of generative AI and regulating its application.

I am pleased that the Office of the Privacy Commissioner for Personal Data has published the *Artificial Intelligence: Model Personal Data Protection Framework* and taken the initiative to provide guidance for Hong Kong enterprises, enabling them to reap the benefits of AI technology while brushing up on personal data privacy protection. This publication will significantly enhance the level of AI governance within enterprises and ensure the proper use of the technology.

Adopting a risk-based approach, the Framework provides a set of practical and detailed recommendations for local enterprises intending to procure, implement and use AI systems. It covers the entire business process and provides pragmatic recommendations for enterprises, whether they are procuring existing AI solutions or customising AI solutions based on their needs. To ensure the protection of personal data privacy and the safe, ethical and responsible use of innovative technology, I encourage enterprises to refer to the Framework and implement the measures suggested within it when procuring and using AI systems.

This is an opportune moment for the publication of the Framework, as our Motherland is currently focusing on the pursuit of new quality productive forces and has launched the "Artificial Intelligence +" initiative to foster industrial development through technological innovation. The Framework serves as a useful guidance for enterprises to utilise AI technology, thus promoting industrial innovation and upgrading. In the broader context, the Framework contributes to the development of Hong Kong's digital economy, strengthening the city's status as a global technology and innovation hub and proactively facilitating its integration with the development of our Motherland.

Prof Hon William WONG Kam-fai, MH

Member of the National Committee of the Chinese People's Political Consultative Conference

Legislative Council Member

Associate Dean (External Affairs), Faculty of Engineering, the Chinese University of Hong Kong

June 2024

Preface

The groundbreaking advancement of artificial intelligence (AI) is revolutionising our world in ways we never thought possible. My Office and I firmly believe that AI, although a double-edged sword, can be harnessed for the greater good provided that proper safeguards are in place, one of which is the implementation of a holistic personal data protection framework. As guardians of personal data privacy protection, we are committed to advocating for the sustainable use of AI in an ethical, responsible and privacy-friendly manner.

In August 2021, my Office took a significant step in this regard by publishing the *Guidance on the Ethical Development and Use of Artificial Intelligence*, which is one of the first leading guides in the Asia-Pacific region on the subject. Recognising that AI is a global challenge necessitating a global solution, we have striven to contribute at an international level by hosting international conferences on AI to facilitate meaningful dialogues among experts, and by co-sponsoring resolutions on responsible and trustworthy AI at the Global Privacy Assembly, a forum uniting over 130 data protection authorities. With the recent adoption of a historic resolution by the United Nations General Assembly promoting 'safe, secure, and trustworthy' AI and our Motherland's earlier release of the *Global AI Governance Initiative*, the momentum towards formulating a comprehensive personal data protection framework for AI is gaining strength on a global scale.

To support the *Global AI Governance Initiative* of the Motherland, my Office has developed the *Artificial Intelligence: Model Personal Data Protection Framework* ("Model Framework"), which targets organisations procuring, implementing and using AI systems that involve the use of personal data. The Model Framework aligns with general business processes and is structured to ensure the effective governance of AI systems that adheres to the three Data Stewardship Values and seven Ethical Principles advocated in our AI guidance of 2021. It provides internationally well-recognised, practical and step-by-step recommendations to assist organisations in harnessing the benefits of AI while safeguarding the personal data privacy of individuals.

The development of this Model Framework would not have been possible without the unwavering support of the two supporting organisations, the Office of the Government Chief Information Officer and the Hong Kong Applied Science and Technology Research Institute. I am truly indebted to our stakeholders, including members of my Office's Standing Committee on Technological Developments and industry experts, for their invaluable inputs and views. My heartfelt gratitude also goes to my team, particularly Ms Cecilia SIU Wing-sze, Ms Joyce LIU Nga-yan, Ms CHAN Gwen-long, and Mr Jackey CHEUNG Wai-yu, for their great dedication, meticulous research, and hard work in the drafting process, particularly in considering and consolidating the views of stakeholders and relevant best practices from other jurisdictions.

This guiding Model Framework, which focuses on the protection of personal data in the context of AI, is the first of its kind in the Asia-Pacific region. With AI security being one of the major fields of national security, I believe that this Model Framework will help nurture the healthy and safe development of AI in Hong Kong, facilitate Hong Kong's development into an innovation and technology hub, and propel the expansion of the digital economy not only in Hong Kong but also in the Greater Bay Area.

Ada CHUNG Lai-ling

Privacy Commissioner for Personal Data

June 2024

Introduction

1. Artificial intelligence ("AI") has no universal definition but generally refers to a family of technologies that mimic human intelligence and involve the use of computer programmes and machines to perform or automate tasks, including solving problems, providing recommendations and predictions, making decisions and generating contents by inferring from input data.

The 2021 AI Guidance

2. In August 2021, the Office of the Privacy Commissioner for Personal Data, Hong Kong ("PCPD") published the *Guidance on the Ethical Development and Use of Artificial Intelligence* ("2021 AI Guidance"), with recommendations that primarily target organisations that develop and use AI systems involving the use of personal data.

3. The 2021 AI Guidance recommends that organisations embrace three **Data Stewardship Values**, namely, (1) being respectful, (2) being beneficial, and (3) being fair. It encourages organisations to adopt the seven internationally recognised **Ethical Principles for AI**, namely (1) accountability, (2) human oversight, (3) transparency and interpretability, (4) data privacy, (5) fairness, (6) beneficial AI, and (7) reliability, robustness and security.

Figure 1: ***Data Stewardship Values and Ethical Principles for AI***

	Data Stewardship Values	Ethical Principles for AI
1	Being Respectful	• Accountability • Human Oversight • Transparency and Interpretability • Data Privacy
2	Being Beneficial	• Beneficial AI • Reliability, Robustness and Security
3	Being Fair	• Fairness

The Trend of AI Adoption

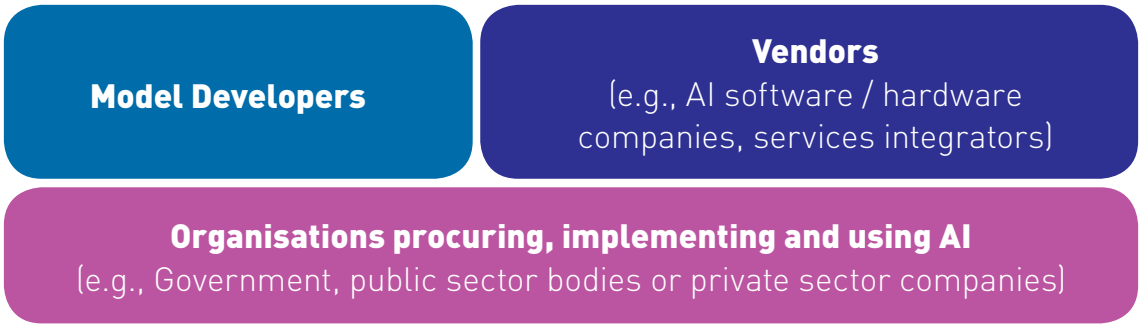
4. In recent years, AI has experienced seismic changes with the advent of foundation models[1]. Simply put, foundation models are AI models that have been trained on a vast amount of unstructured data, which allows them to be adapted for a wide range of tasks, operations and applications, and used for various purposes. Insofar as generative AI is concerned, there are numerous types of foundation models, such as language models, audio models, video models, and even multimodal models. Large language models ("LLMs"), for example, are foundation models trained on text data that can be adapted to facilitate tasks which require natural language processing[2], such as chatbots.

5. Notwithstanding the increasing supplies of small-scale language models, the development of large-scale foundation models can be costly and time-consuming for many organisations. As more organisations are adopting AI into their operations, there has been an increasing trend, especially among small and medium-sized enterprises, towards purchasing AI solutions from vendors and developers that are tailored to the purchasers' specific use cases, instead of developing AI systems from scratch. In this way, organisations can leverage customised AI systems or off-the-shelf solutions obtained from AI system developers and / or vendors, both of which augment their decision-making capability, automate processes, generate contents and extract insights from data. This practice splits the responsibilities for the ethical development and use of AI among different actors.

1 "Foundation model" generally refers to a machine learning model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of downstream distinctive tasks or applications, including simple task completion, natural language understanding, translation, and content generation.

2 According to the US National Institute of Standards and Technology, natural language processing (NLP) is a powerful computational approach that allows machines to meaningfully understand human spoken and written languages. Powering activities such as algorithmic searches, speech translation and even conversational text generation, NLP is able to help us communicate with computer systems to direct them to carry out a variety of tasks.

Figure 2: ***Landscape of AI Developers, Vendors and Organisations Procuring / Implementing / Using AI***



Focus of this Model Framework

6. **This Model Framework, which is based on general business processes, provides a set of recommendations on the best practices for any organisations procuring, implementing and using any type of AI systems that involve the use of personal data, which may include predictive AI and generative AI**. Apart from being supportive of the *Global AI Governance Initiative* promulgated by the Mainland in 2023, this Model Framework also reflects the prevailing norms and best practices of the international community. The adoption of this Model Framework can facilitate organisations in complying with well-established data protection principles, including data security, which is especially significant in the context of AI given the substantial volume of data typically involved.

7. In this Model Framework, the term "organisations" refers to organisations that procure AI solutions from third parties and engage in the handling of personal data in (a) customising an AI system to improve its performance for a specific domain or use case and / or (b) operating the AI system; and the term "AI supplier" refers to both AI developers and / or AI vendors (as the case may be) who provide AI solutions to the organisations. Organisations that develop in-house AI models are recommended to refer to the 2021 AI Guidance.

Compliance with the Personal Data (Privacy) Ordinance

8. **Organisations should ensure compliance with the requirements under the Personal Data (Privacy) Ordinance ("PDPO"), including the six Data Protection Principles ("DPPs") in Schedule 1 thereto, when handling personal data in the process of procuring, implementing and using AI solutions**. The six DPPs, which cover the entire life cycle of the handling of personal data from collection to destruction, represent the core requirements of the PDPO. See **Appendix A** for an overview of the DPPs.

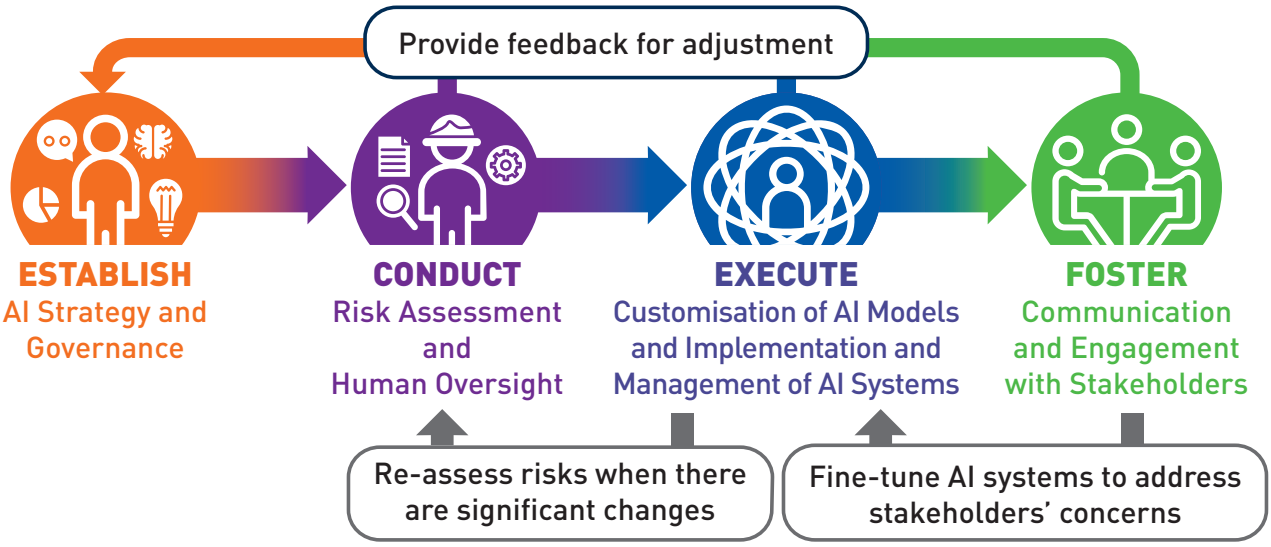
9. The recommendations in this Model Framework are by no means exhaustive. Organisations should adopt other measures as appropriate to comply with the PDPO and to adhere to the Data Stewardship Values and the Ethical Principles for AI when procuring, implementing and using AI solutions.

10. The PCPD advocates the adoption of a Personal Data Privacy Management Programme ("PMP") to ensure the responsible collection, holding, processing and use of personal data, thereby enhancing data governance. Good data governance goes hand in hand with governance for trustworthy AI. By incorporating the principles of AI governance and "privacy-by-design" into their existing PMP and / or data management practices, organisations can reinforce their commitment to personal data privacy protection and demonstrate their accountability.

Model Personal Data Protection Framework

11. To ensure that the Data Stewardship Values and the Ethical Principles for AI (see paragraph 3 above) are implemented, organisations should formulate appropriate policies, practices and procedures when they procure, implement and use AI solutions by taking into consideration the recommended measures in the following areas:
    - AI Strategy and Governance (Part I);
    - Risk Assessment and Human Oversight (Part II);
    - Customisation of AI Models and Implementation and Management of AI Systems (Part III); and
    - Communication and Engagement with Stakeholders (Part IV).

Figure 3: ***Model Personal Data Protection Framework***



12. **In general, organisations sourcing third-party AI solutions should adopt a risk-based approach to procuring, implementing and using AI systems, as part of a broader, holistic approach to AI governance in their organisations**. The recommendations in this Model Framework should be considered and adopted in proportion to the risks that an AI system may pose in context. In incorporating elements of this Model Framework into their existing workflows, organisations may leverage and adapt existing data governance, accountability, and third-party vendor management frameworks.

Part I AI Strategy and Governance

13. Buy-in from and active participation by top management (such as executive or board level) are essential ingredients of success in the ethical and responsible procurement, implementation and use of AI systems. **Organisations should have an internal AI governance strategy, which generally comprises an (i) AI strategy, (ii) governance considerations for procuring AI solutions, and (iii) an AI governance committee (or similar body) to steer the process**.

1.1 AI Strategy

Key principle: Accountability

14. Organisations should formulate an AI strategy to demonstrate the commitment of top management to the ethical and responsible procurement, implementation and use of AI. The AI strategy, which should provide directions on the purposes for which AI solutions may be procured, and how AI systems should be implemented and used, may include the following elements:

(i) Defining the functions that AI systems would serve in the technological ecosystem of the organisation;

(ii) Setting out ethical principles for the procurement, implementation and use of AI solutions that are specific and applicable to the organisation by referring to the Ethical Principles for AI;

(iii) Determining the unacceptable uses of AI systems in the organisation[3];

(iv) Establishing an AI inventory to facilitate the implementation of governance measures;

(v) Establishing specific internal policies and procedures regarding how to ethically procure, implement and use AI solutions, including an institutionalised decision-making process with criteria for internal escalation;

3 Organisations should identify use cases of AI where the potential risks are so high that they should not be allowed. The list of use cases should remain open to allow for the addition, removal or adjustment of use cases as AI technology evolves, as new risks come to light and / or as new risk-mitigating measures are adopted.

(vi) Ensuring that the appropriate technical infrastructure is in place to support lawful, responsible and quality AI implementation and use, ranging from data storage, management and processing tools, and computing resources and facilities, to machine learning operations for deployment and monitoring, etc;

(vii) Regularly communicating the AI strategy, policies and procedures to all relevant personnel, including internal staff at all levels and, where appropriate, external stakeholders such as business partners and customers;

(viii) Considering emerging laws and regulations that may be applicable to the procurement, implementation and use of AI, including data protection and intellectual property laws; and

(ix) Continuously reviewing and adjusting the AI strategy based on feedback from the implementation of Parts II, III and IV of the Model Framework.

1.2 Governance Considerations for Procuring AI Solutions

15. The procurement of AI solutions generally involves **engaging third parties to customise AI systems or buying / subscribing to off-the-shelf AI systems / services**. Such procurement practices typically include the following steps:

(i) Sourcing appropriate AI solutions and considering the expertise and reputation of AI suppliers;

(ii) Selecting the AI solution with AI models that are suitable for the organisation's purposes for using AI (factors to consider include the type(s) of machine learning algorithms (such as regression models, decision trees, random forests or neural networks), type(s) of learning models (such as the supervised learning model, the unsupervised learning model and the reinforcement learning model), and model size and complexity);

(iii) Collecting and preparing the organisation's data for customising the AI model (if necessary);

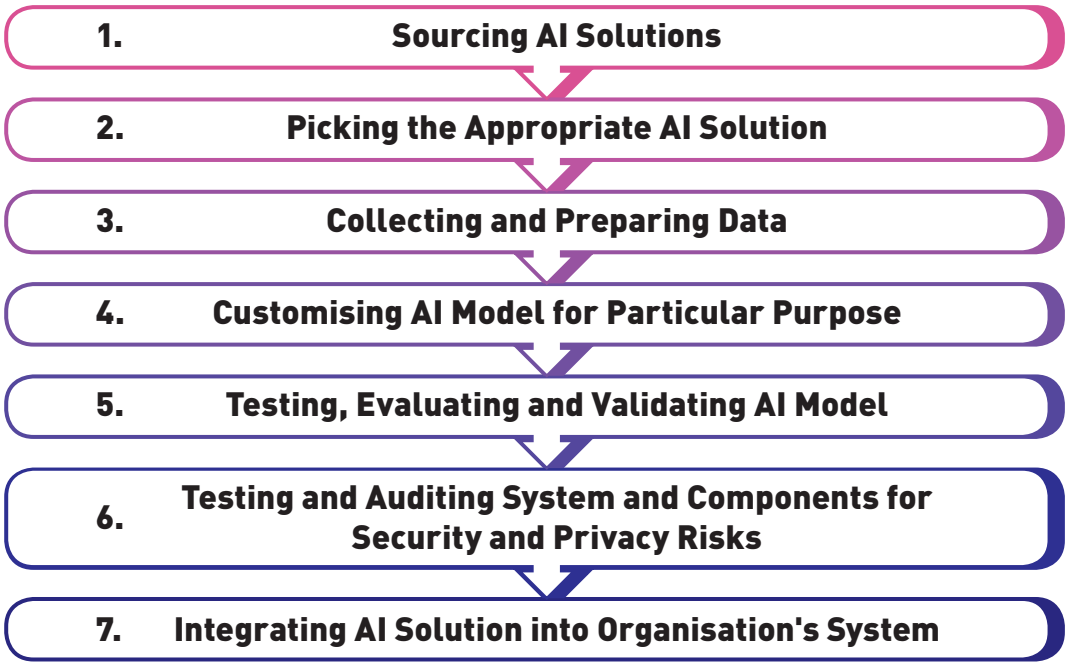
(iv) Customising the AI model for particular purpose (if necessary);

(v) Testing, evaluating and validating the AI model;

(vi) Testing and auditing the system and its components for security and privacy risks; and

(vii) Integrating the AI solution into the organisation's systems.

Figure 4: ***Process of Procurement and Implementation of AI Models***



16. An organisation intending to invest in AI solutions is recommended to consider the following governance issues:

(i) The purposes of using AI and the intended use cases for AI deployment;

(ii) The key privacy and security obligations and ethical requirements[4] to be conveyed to potential AI suppliers;

(iii) International technical and governance standards that potential AI suppliers should follow[5];

4 Among other things, these obligations and requirements should be aligned with the organisation's privacy policy (which should comply with the PDPO) and the Ethical Principles for AI. For example, depending on the use cases and circumstances, the obligations and requirements may address dataset fairness, the kinds of machine learning algorithms and types of learning suitable for addressing the organisation's purposes and how ethical expectations will be met (e.g., the transparency and explainability of different types of AI models; see section 2.3).

5 Organisations may refer to standards developed and published by professional associations such as the International Organization for Standardization (ISO) and Institute of Electrical and Electronics Engineers (IEEE). For example, ISO/IEC 27001:2022 and ISO/IEC 27002:2022 cover information security, ISO/IEC 27701:2019 covers personal data protection, ISO/IEC 23894:2023 covers risk management in AI and ISO/IEC 42001:2023 covers the establishment, implementation, maintenance and continual improvement of an AI management system within organisations.

(iv) The general criteria and procedures (e.g., by way of scoring) that will qualify an AI solution for review by the AI governance committee (or similar body) (e.g., the situations in which the AI use case is likely to result in high risk (see section 2.1));

(v) Any data processor agreements to be signed, if the procured AI solution involves the engagement of data processors (e.g., the AI solutions involve the development or customisation of AI models directly on a third-party platform and / or the AI solutions run directly on an "AI-as-a-service" cloud-based platform[6]);

(vi) The policy on handling the output generated by the AI system (e.g., where feasible, employing techniques to anonymise personal data contained in AI-generated content, label or watermark AI-generated content and filter out AI-generated content that may pose ethical concerns);

(vii) A plan to continuously analyse the business and technological landscapes to identify potential research or strategies that may help the organisation adopt "privacy-by-design" or "ethics-by-design" principles into their AI governance;

(viii) A plan to continuously monitor, manage and maintain the AI solution with assistance from the selected AI supplier, where appropriate (see section 3.3); and

(ix) Evaluation of the AI suppliers' competence during due diligence.

6 Organisations are encouraged to read the PCPD's Information Leaflet on Outsourcing the Processing of Personal Data to Data Processors for more information: https://www.pcpd.org.hk/english/publications/files/dataprocessors_e.pdf

Figure 5: ***Governance Considerations for Procuring AI Solutions***

17. During each of the stages of procurement and implementation of AI models (see figure 4), the degree of organisational involvement may differ, depending on the data provided and the instructions given for AI model development and / or customisation, etc. As an illustration of this point, the following scenarios depict different degrees of organisational involvement:

- A fully customised AI model developed by a third-party developer;
- A pre-trained AI model with minor customisation of the model to suit the organisation's needs;
- An off-the-shelf AI solution, including "AI-as-a-service" and cloud-based services (e.g., via application programme interface ("API")); or
- An AI model created by the organisation running its data through, or giving customised instructions via, automated machine learning services ("AutoML") on third-party platforms.

18. In each scenario, the organisation's engagement with third parties may raise data (including personal data) protection compliance issues, which should be clearly addressed in the service agreements signed between the parties.

Figure 6: ***Key Data (Including Personal Data) Protection Compliance Considerations (Non-exhaustive)***

Who the data user is

- The party who has control of the collection, holding, processing or use of the personal data is the data user (section 2 of the PDPO).
- For example, an organisation that determines the types of personal data to be used for customising, testing, validating and / or operating an AI system is likely to be considered a data user.

Who the data processor is

- The party who processes personal data on behalf of another person and does not process the data for its own purposes is a data processor (section 2 of the PDPO).
- For example, an AI supplier that does not decide on the input data and the output of an AI model in the processing of personal data for customisation and only provides a platform for the customisation of AI is likely to be a data processor.

Legality of cross-border transfer

- If the customisation and use of AI involve processing personal data on cloud platforms with data centres distributed across multiple jurisdictions, and organisations (as data users) transfer personal data to places outside Hong Kong, the data user:
  - Must comply with the relevant requirements of the PDPO, including the 6 DPPs; and
  - Should ascertain if there are any restrictions or regulations pertaining to cross-border or cross-boundary transfers of data back to the data user from the jurisdiction where the data are processed.

Data security considerations

- If an organisation as the data user transfers personal data to the data processor for processing in the customisation and / or use of AI, it must adopt contractual or other means to prevent unauthorised or accidental access, processing, erasure, loss or use of the personal data, in compliance with the requirements of DPP 4(2) of the PDPO.

19. The procurement team should work with the project team to select AI solutions, determine the degree of organisational involvement that is suitable for the purposes of the organisation[7], and **work with the legal and compliance teams to address any potential data protection compliance questions**.

7 For example, the desired levels of accuracy and interpretability of the output of the AI system, as well as barriers to the implementation of the system in the organisation's IT infrastructure, may be considered.

1.3 Governance Structure

Key principles: Accountability / Human Oversight

20. Expertise in different fields, such as computer engineering, data science, cybersecurity, user experience design, law and compliance, and public relations is recommended for the procurement, implementation and use of AI systems. **An internal governance structure with sufficient resources, expertise and authority should be established to steer the implementation of the AI strategy and oversee the procurement, implementation and use of AI system**. An AI governance structure may include the following elements:

    (i) An AI governance committee (or similar body), which reports to the board and oversees the whole life cycle of all AI solutions from procurement, implementation, and use to termination. The AI governance committee should have oversight across the business and not be constrained by division (i.e., risk and compliance, finance or sales, etc.);

AI Governance Committee

Participation by senior management and interdisciplinary collaboration should be the most significant attributes of an AI governance committee. A cross-functional team with a mix of skills and perspectives should be established, including business and operational personnel, procurement teams, system analysts, system architects, data scientists, cybersecurity professionals, legal and compliance professionals (including data protection officer(s)), internal audit personnel, human resources personnel and customer service personnel.

A C-level executive (such as a chief executive officer, chief information officer / chief technology officer, chief privacy officer or similar senior management position) should be designated to lead the cross-functional team.

(Optional) Independent AI and ethics advice may be sought from external experts. An additional ethical AI committee may be established to conduct an independent review when a project is sufficiently large, with a considerable impact and / or a high profile, and its ethical value may be challenged.

(ii) Clear roles and responsibilities for different divisions or personnel;

Examples of roles and responsibilities:

- Procurement teams should obtain AI solutions in accordance with the internal policies and procedures set out in the organisational AI strategy;
- System analysts, system architects and data scientists should focus on the customisation, implementation, monitoring and maintenance of AI solutions, and on the organisation's internal data governance processes;
- Legal and compliance professionals should focus on ensuring compliance with relevant laws and regulations (including data protection laws) as well as internal policies regarding the procurement, implementation and use of AI systems;
- Human reviewers should focus on reviewing the decisions and output of AI systems;
- Business and operational personnel should use AI in accordance with the policies and procedures of the organisations; and
- Customer service and public relations personnel should communicate with stakeholders, including customers, regulators and the general public, and address their concerns.

(iii) Adequate resources in terms of both finance and manpower; and

Cases where adequate resources (e.g. experts with relevant technical skills, experience and expertise) are required include:

- Conducting risk assessments when necessary to identify and mitigate risks, including privacy, security and ethical risks, arising from the use of AI, and adopting risk-mitigating measures accordingly;
- Establishing internal data governance processes and information systems that allow the monitoring, documentation and review of the implemented AI solutions; and
- Providing adequate training to relevant personnel (see section 1.4 below).

(iv) Effective internal reporting mechanisms for reporting any system failure or raising any data protection or ethical concerns to facilitate proper monitoring by the AI governance committee.

Organisations should establish an AI strategy and an AI governance committee (or similar body) to steer the procurement, implementation and use of AI systems.

1.4 Training and Awareness Raising

Key principle: Accountability

21. To ensure that AI-related policies are properly applied, adequate training should be provided to all relevant personnel to ensure that they have the appropriate knowledge, skills and awareness to work in an environment using AI systems.

Figure 7: ***Examples of Training***

Recommended Personnel	Training Topics
System analysts / architects / data scientists	• Compliance with data protection laws, regulations and internal policies; cybersecurity risks
AI system users (including business and operational personnel)	• Compliance with data protection laws, regulations and internal policies; cybersecurity risks; general AI technology
Legal and compliance professionals	• General AI technology and governance
Procurement staff	• General AI technology and governance
Human reviewers	• Detection and rectification of any unjust bias, unlawful discrimination and errors / inaccuracies in the decisions made by AI systems or presented in the content
All staff performing work relating to AI system	• Benefits, risks, functions and limitations of the AI system(s) used by the organisation

Roles of human reviewers

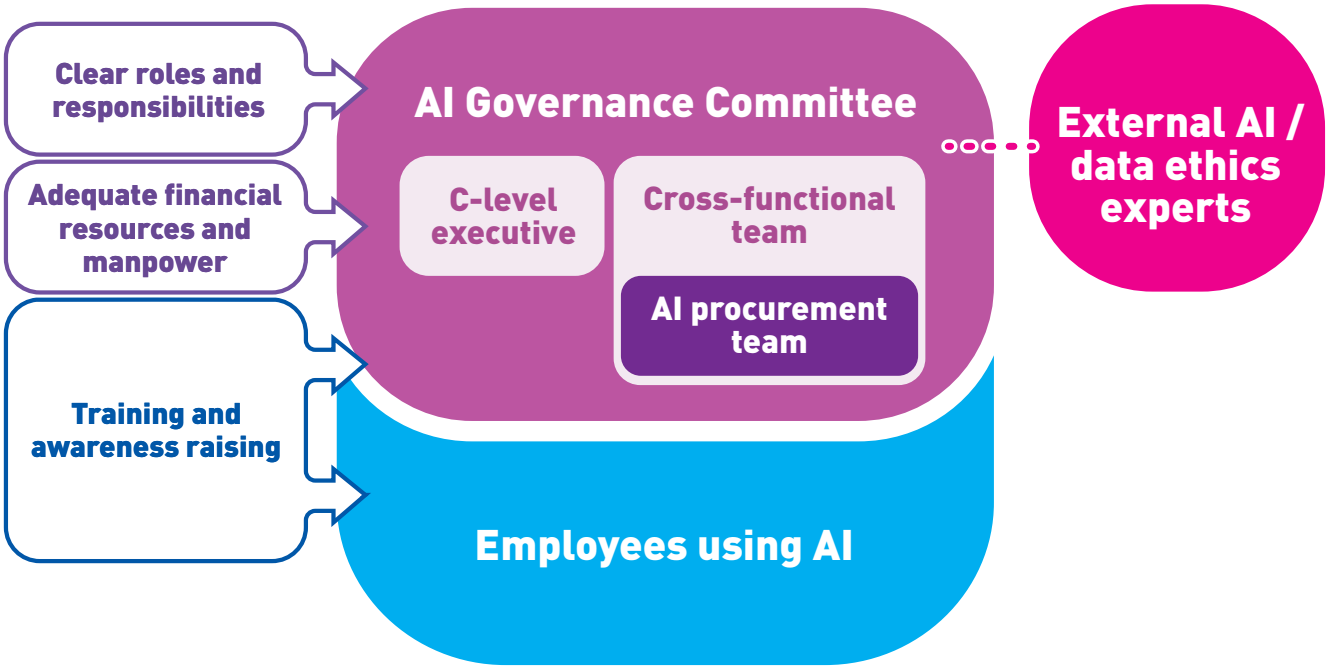
To ensure that human reviewers perform their duties conscientiously and that human oversight is not merely a gesture, relevant personnel should be able to assess and interpret the recommendations made by AI, and / or review the contents generated by AI. Reviewers should be able to properly exercise their discretion and authority to veto the recommendations made by AI or flag problematic decisions and / or contents and should alert the AI supplier when necessary.

Organisations may consider requesting, where appropriate, the AI supplier to provide information on and explanations of AI output to enable effective exercise of human oversight.

22. As part of the PMP, any personal data privacy protection training covering the requirements of the PDPO and the organisation's privacy policies should also cover the collection and use of personal data in the procurement, implementation and use of AI systems.

23. In addition, the importance of ethical AI and applicable principles should be conveyed to all relevant personnel through staff meetings or other internal communications to cultivate and promote an ethical and privacy-protecting culture.

Figure 8: ***Governance Structure***



Part II Risk Assessment and Human Oversight

24. Procured AI solutions containing AI models that were originally developed for general use or for multiple purposes may be implemented in an organisation's operations for specific use. The risk levels of different AI systems thus depend on how the organisation uses the systems and the specific purposes for which they are used. For example:

- An AI system which assesses the credit worthiness of individuals tends to carry a higher risk than a system used to present individuals with personalised advertisements because the latter is unlikely to have a significant impact on individuals, while the former may deny them access to credit facilities.
- A generative AI tool used for internal translation is less likely to have a significant impact on individuals than a generative AI chatbot generating direct responses to customer enquiries.
- An AI system with full autonomous decision-making capabilities may be riskier than a system that involves some degree of human operation (e.g., one that provides recommendations to human actors), especially if the decisions significantly affect individuals.

25. **A risk-based approach should be adopted in the procurement, use and management of AI systems. Comprehensive risk assessment is necessary for organisations to systematically identify, analyse and evaluate the risks, including privacy risks, involved in the process. A risk management system should be formulated, implemented, documented and maintained throughout the entire life cycle of an AI system**[8]. For AI use cases with risks that have been determined as unacceptable in the organisation's AI strategy (see section 1.1), they should be disallowed.

Comprehensive risk assessment is necessary for organisations to systematically identify, analyse and evaluate the risks, including privacy risks, involved in the procurement, use and management of AI systems.

8 The AI governance committee may consult frameworks such as the ISO/IEC 23894:2023 (Information technology - Artificial intelligence - Guidance on risk management) and the US National Institute of Standards and Technology's AI Risk Management Framework in integrating risk management into the life cycle of AI systems.

26. **Risk assessments should be conducted by a cross-functional team during the procurement process or when significant updates are made to an existing AI system**. The cross-functional team should include privacy compliance personnel to identify privacy-related risks (e.g., via a Privacy Impact Assessment). Depending on the circumstances, individuals from different social, cultural and religious backgrounds and of different genders and races (or experts with relevant knowledge) may need to be consulted to identify potential unjust bias; unlawful discrimination; adverse impact on individuals' rights, freedom and interests; and wider societal impact in the use of AI. **All risk assessments should be properly documented, and the results should be reviewed in line with the organisation's AI policies as endorsed by the AI governance committee**.

Figure 9: ***Process of Risk Assessment***

2.1 Risk Factors

Key principles: Beneficial AI / Data Privacy / Fairness

27. **As the use of AI often involves the use of personal data, it is essential to address data privacy risks**. To protect personal data privacy, organisations should consider the following factors in a risk assessment:

(i) The allowable uses of the data for customising procured AI solutions and / or to be fed into AI systems to make decisions, having regard to DPP 3 of the PDPO[9];

(ii) The volume of personal data (having regard to DPP 1 of the PDPO[10]):

- Required for customising AI models;
- Collected by the AI system during operation (e.g., surveillance, systematic evaluation and monitoring may involve the large-scale collection of personal data); and
- Required to develop and train the AI solution by the AI supplier, and whether anonymisation techniques have been applied, as far as possible, to adhere to the data minimisation principle;

(iii) The sensitivity[11] of the data involved, having regard to DPP 4 of the PDPO[12];

9 DPP 3 stipulates that personal data must not be used for new purposes without the prescribed consent of the data subjects.

10 DPP 1 stipulates that the amount of personal data to be collected shall be adequate but not excessive in relation to the purpose of collection.

11 Personal data that are generally considered to be more sensitive include biometric data, health data, financial data, location data, personal data about protected characteristics (e.g., gender, ethnicity, sexual orientation, religious beliefs, political affiliations), and the personal data of vulnerable groups, such as children.

12 DPP 4(1)(a) stipulates that all practicable steps shall be taken to ensure that any personal data (including data in a form in which access to or processing of the data is not practicable) held by a data user is protected against unauthorized or accidental access, processing, erasure, loss or use having particular regard to the kind of data and the harm that could result if any of those things should occur.

(iv) The quality of the data involved, taking into account the source, reliability, integrity, accuracy (having regard to DPP 2 of the PDPO), consistency, completeness, relevance and usability of the data[13];

(v) The security[14] of personal data used in an AI system, taking into account how personal data may be transferred in and out of the AI systems across the organisation's technological ecosystem[15], and whether guardrails on AI-generated output are in place to mitigate the risk of personal data leakage, having regard to DPP 4 of the PDPO[16]; and

(vi) The probability that privacy risks (e.g., the excessive collection, misuse or leakage of personal data) will materialise and the potential severity of the harm that might result.

28. From a wider ethical perspective, and insofar as the use of AI systems may have an impact on the rights, freedom or interests of stakeholders, especially individuals, the risk assessment should also take into account:

(i) The potential impacts (including benefits and harms) of the AI system on the affected individuals, the organisation and the wider community;

(ii) The probability that the impacts of the AI system on individuals will occur, as well as the severity and duration of the impacts[17]; and

(iii) The adequacy of mitigation measures (both technical and non-technical) to minimise the risk of harm (see section 2.2 and Part III).

13 DPP 2 requires a data user to take all practicable steps to ensure that personal data is accurate having regard to the purpose for which the personal data is used.

14 Using third-party-built or maintained AI solutions requires cautious assessment of the security risks, as the AI solution may rely simultaneously on numerous forms of software and hardware developed in-house and / or based on open-source codes and frameworks (see section 3.2).

15 DPP 4 requires a data user to take all practicable steps to safeguard the security of personal data held by the data user.

16 DPP 4(1)(e) stipulates that all practicable steps shall be taken to ensure that any personal data (including data in a form in which access to or processing of the data is not practicable) held by a data user is protected against unauthorized or accidental access, processing, erasure, loss or use having particular regard to any measures taken for ensuring the secure transmission of the data.

17 For example, taking into account the AI's degree of autonomy, its capability of interacting with the environment directly, the complexity of that environment and the complexity of the decisions to be made by the AI should be considered.

29. Potential impacts on individuals as a result of the use of AI systems may affect their legal rights, human rights (including privacy rights), employment or educational prospects, as well as their access and eligibility to services and so on. **An AI system likely to produce an output that may have such significant impacts on individuals would generally be considered high risk**.

In assessing whether the potential impacts on individuals may be significant, organisations may begin by considering the potential types of harm[18] to individuals that might result from the use of the AI system.

An AI system used to produce an output (such as the assessment of job applicants) that, in certain instances, has a high likelihood of causing severe and long-lasting harms to individuals, and the risks of which cannot be adequately mitigated, should be considered high risk.

Figure 10: Factors to Consider in Risk Assessment of AI Systems (Non-exhaustive)

Requirements under the PDPO	Volume, sensitivity and quality of data
Security of data	Potential impact on individuals, the organisation and community
Probability, severity and duration of impact	Mitigation measures

18 For example, financial harm, bodily harm, discrimination, loss of control of personal data, lack of autonomy, psychological harm, and other adverse effects on rights and freedoms should be considered.

2.2 Determining the Level of Human Oversight

Key principle: Human Oversight

30. The primary objective of a risk assessment is to identify the potential risks and adopt corresponding risk mitigation and management measures. In adopting a risk-based approach, the types and extent of risk mitigation measures should correspond with and be proportionate to the levels of the identified risks. Residual risks that cannot be eliminated should be communicated to the end-users of the AI system and / or any individuals impacted by the system. In any event, residual risks should be reduced to an acceptable level. Residual risks are considered acceptable when they are minimised to the greatest extent reasonably practicable and when the potential benefits to stakeholders significantly outweigh the risks presented by the AI system.

In adopting a risk-based approach, the types and extent of risk mitigation measures should correspond with and be proportionate to the levels of the identified risks.

31. **Human oversight is a key measure for mitigating the risks of using AI**. The risk assessment would indicate the appropriate level of human oversight required in the use of the AI system. Ultimately, human actors should be held accountable for the decisions and output made by AI.

32. In general, an AI system with a higher risk profile, i.e., one likely to have a significant impact on individuals, requires a higher level of human oversight than an AI system with a lower risk profile. Therefore:

    (i) A high-risk AI system should take a "human-in-the-loop" approach, where human actors retain control of the decision-making process to prevent and / or mitigate errors or improper output and / or decisions made by AI.

(ii) An AI system with minimal or low risks may take a "human-out-of-the-loop" approach, whereby the AI system is given the capability to adopt output and / or make decisions without human intervention to achieve full automation / fully automated decision-making.

(iii) If neither approach is suitable, such as when the risks are non-negligible or if the "human-in-the-loop" approach is not cost-effective or practicable, organisations may consider a "human-in-command" approach, whereby human actors make use of the output of the AI system and oversee the operation of the AI system and intervene whenever necessary.

Figure 11: Risk-based Approach to Human Oversight

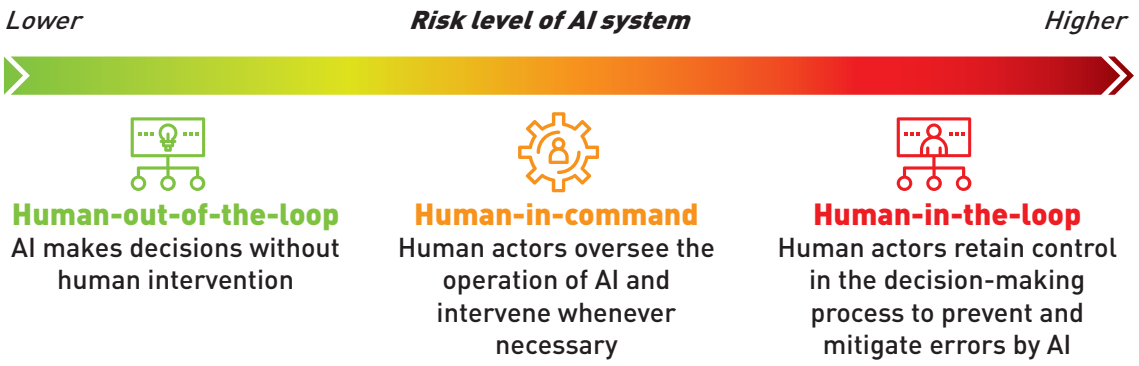


Figure 12: Examples of AI Use Cases that May Incur Higher Risk

- Real-time identification of individuals using biometric data
- Evaluation of individuals' eligibility for social welfare or public services
- Assessment of job applicants, evaluation of job performance or termination of employment contracts
- Evaluation of the creditworthiness of individuals for making automated financial decisions
- AI-assisted medical imaging analytics or therapies

33. **Depending on the level of risk, an appropriate level of human oversight should be incorporated in the organisation's use of AI systems**. Organisations are recommended to understand from the AI supplier whether and how human reviewers have been involved in the training and development of the AI models to reduce the risk of significant adverse impacts on individuals materialising during deployment. Organisations may also need to request the AI supplier to provide information and explanation about AI output to enable effective performance of human oversight in their use of the AI system.

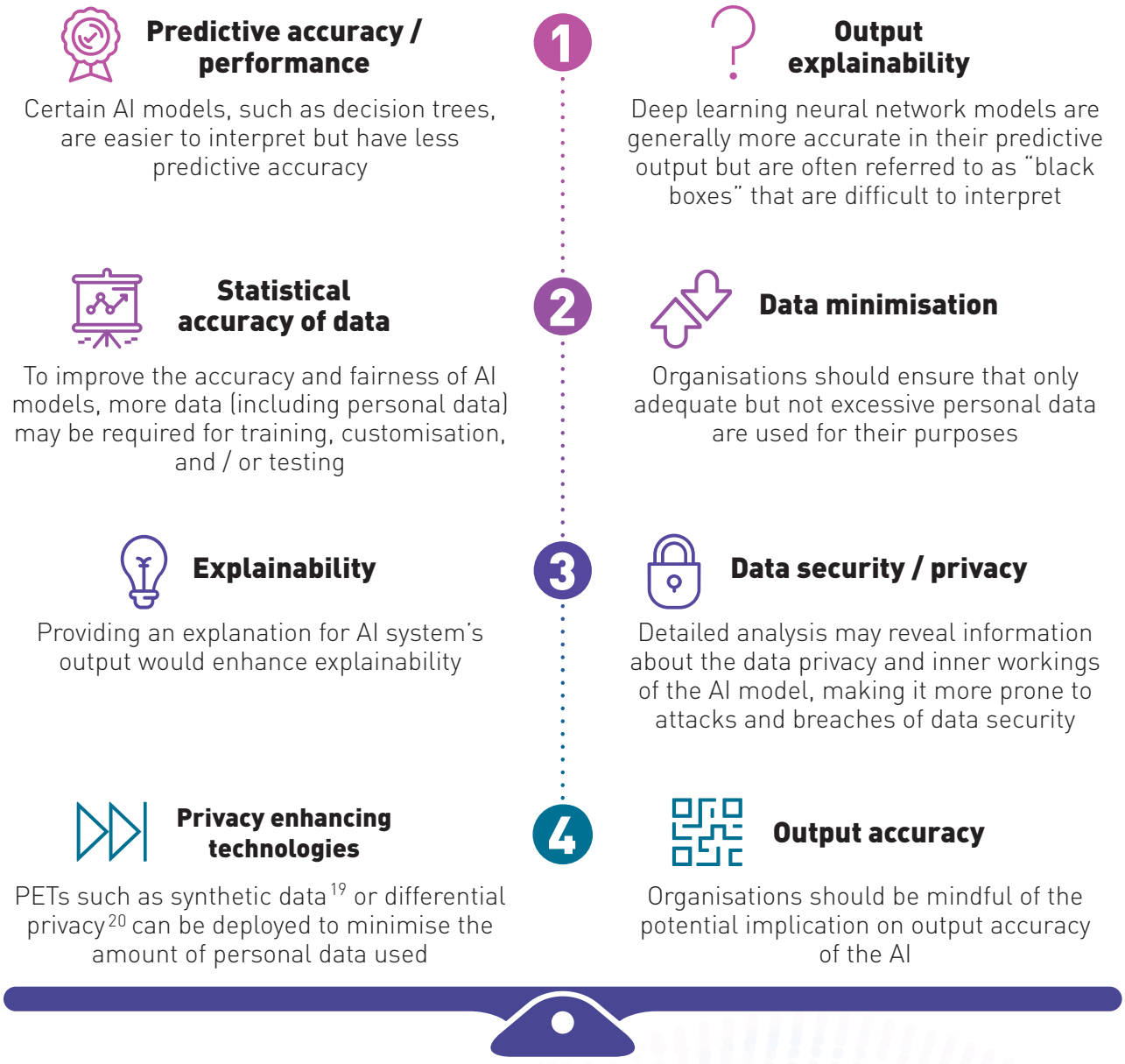
2.3 Risk Mitigation Trade-offs

34. When seeking to mitigate AI risks to comply with the Ethical Principles for AI, organisations may need to strike a balance when conflicting criteria emerge (see Figure 13) and make trade-offs between the criteria.

35. **Organisations may need to consider the context in which they are deploying the AI to make decisions or generate contents and thus decide how to justifiably address the trade-offs that arise**. For example, explainability may be valued over output accuracy in a context where a decision affects a customer's access to services, and a human reviewer would need to explain the AI system's decision to the customer. **Organisations are recommended to document their assessments of such trade-offs, including the rationale for the final decisions**.

36. In any event, organisations are reminded that any applicable legal requirements, including the requirements of the PDPO, must be complied with.

Figure 13: ***Examples of Risk Mitigation Trade-offs***



19 Synthetic data refers to a dataset that has been generated artificially and is not related to real people.

20 Differential privacy is an approach to privacy protection in the release of datasets, usually by adding noises (i.e., making minor alterations) to the datasets before release. Unlike de-identification, differential privacy is not a specific process, but a quality or condition of datasets that a process can achieve. A released dataset achieves differential privacy if it is uncertain whether a particular individual's data is included in it. Differential privacy is generally considered to have stronger protection of privacy than de-identification.

Part III Customisation of AI Models and Implementation and Management of AI Systems

37. Apart from procuring the right AI solution to achieve an organisation's purposes, the quantity and quality of data involved in the customisation of AI models and the use of AI systems will have a significant impact on the usability, accuracy and reliability of an AI system. The primary goal of customisation is to use the data to improve the AI solution's performance by providing more domain / context-specific information. In this Model Framework, "customisation" refers to the process of adjusting or adapting a pre-trained AI model, including the fine-tuning[21] and / or grounding[22] of the AI models, to meet the purposes of an organisation in its use of AI.

38. AI models may continue to learn and evolve and the environment in which an AI system operates may also change. Therefore, continuous monitoring, review and user support are required after the adoption of an AI model to ensure that the AI systems remain effective, relevant and reliable.

Figure 14: ***Major Customisation and Management Processes***

21 Fine-tuning is the process of taking AI models trained on large and general datasets and updating / adapting them for using other specific data for a specific purpose or need.

22 Grounding is the process of linking AI models to verifiable real-world knowledge and examples from external sources. One of the most popular methods of grounding for generative AI models is Retrieval-Augmented Generation, which augments the capabilities of an LLM by adding an information retrieval system that provides grounding data, to improve the performance of the LLM in specific use cases or domains.

3.1 Data Preparation for Customisation and Use of AI

Key principles: Data Privacy / Fairness

39. Internal proprietary data, often involving personal data, may be used in both the customisation and decision-making or output stages. **Good data governance in the customisation and operation of AI not only protects individuals' personal data privacy but also ensures data quality, which is critical to the robustness and fairness of AI systems**. Poorly managed data may result in the "garbage in, garbage out" problem and may have an adverse effect on the results that an AI system produces (e.g., unfair output of predictive AI and "hallucinations" by generative AI[23]).

Good data governance is critical to the robustness and fairness of AI systems.

40. Organisations should give due consideration to the data governance practices of their upstream AI supplier and of the source(s) of the training data. The legality and quality of the training data used could affect the quality, robustness, and fairness of an AI solution, and its compliance with applicable legal requirements.

41. Organisations should take the following steps in the preparation of datasets for the customisation and use of AI:

(i) **Measures must be adopted to ensure compliance with the requirements under the PDPO**, including:

- Collecting an adequate but not excessive amount of personal data by lawful and fair means - see DPP 1;
- Refraining from using personal data for any purpose that is not compatible with the original purpose of collection, unless the express and voluntary consent of the data subjects has been obtained, or the personal data have been anonymised - see DPP 3;

23 Poor data governance may not be the sole cause of "hallucinations" by generative AI. "Hallucinations" tend to be inherent in generative AI models which use the transformer architecture, but can be minimised effectively through mechanisms such as grounding and prompt-engineering.

- Taking all practicable steps to ensure the accuracy of personal data before use - see DPP 2(1);
- Taking all practicable steps to ensure the security of personal data - see DPP 4;
- Erasing or anonymising personal data when the original purpose of collection has been achieved - see DPP 2(2) and section 26 of the PDPO;
- Upon or before collecting personal data, taking all practicable steps to ensure that the data subjects are informed of the required information, such as the classes of persons to whom the data may be transferred, especially where AI suppliers are involved - see DPP 1;
- Taking all practicable steps to ensure that the required information regarding the organisation's policies and practices in relation to personal data is made available, for example, via a privacy policy - see DPP 5; and
- Implementing systems that would help the organisation to respond to requests from data subjects - see DPP 6.

(ii) **Minimising the amount of personal data involved in the customisation and use of AI models reduces privacy risks (DPP 1), taking into account that adequate data may need to be collected to ensure accurate and unbiased results**. Organisations should adopt the following practices and techniques, where appropriate:

- Collecting and using only the personal data that are necessary to customise and / or operate the AI for the particular purposes and discarding the data containing characteristics of individuals that are not necessary for such purposes, while access to broader datasets may allow training models to produce more accurate and fairer output. Data scientists and subject matter / domain experts may be consulted in advance with the AI suppliers to identify the necessary and adequate amount of data required for customisation;

Example 1: A fashion retail platform is purchasing a third-party developed AI chatbot that it will customise to provide fashion recommendations to its customers. The company may find it necessary to use the past purchases and browsing histories of different segments of its customer groups to fine-tune the chatbot. However, the use of personal data, such as customers' names, contact details and certain demographic characteristics, would not be necessary.

- Considering the appropriate size and complexity of the AI model. If the intended purpose does not require a sophisticated or customised model, select a simpler and smaller model that requires less data for customisation, or an off-the-shelf model that would not require data for customisation at all;
- Using anonymised[24], pseudonymised[25] or synthetic data to customise and feed into AI models, where appropriate[26];
- Applying PETs, such as "differential privacy" techniques, to datasets before releasing them for use when customising AI models;
- Erasing personal data from the AI system when the data are no longer required for the customisation and use of AI[27]; and
- Revisiting the need to use personal data if using an AI model that is an expert system[28] and does not need large quantities of data for customisation would be sufficient to achieve the same purposes.

24 Anonymised data refers to a dataset that has been processed in such a manner that no individual can be identified from it. As anonymised data cannot be used to identify individuals, they are not personal data.

25 Pseudonymised data refers to a dataset that has had all personally identifiable information removed from it and replaced with other values, preventing the direct identification of individuals without additional information. Pseudonymised data are personal data because individuals can still be identified indirectly with the aid of additional information.

26 These three data minimisation techniques may not apply to certain types of non-text data, such as images.

27 For example, if personal data are loaded into a generative AI system during the grounding process in response to an individual's queries, the personal data should be discarded after fulfilling the request.

28 An expert system is "a form of AI that draws inferences from a knowledge base to replicate the decision-making abilities of a human expert within a specific field." (Source: IAPP AI Glossary) Expert systems may be built by creating a set of rules according to expert knowledge of the field, without relying on data and machine learning.

(iii) **The quality of the data used to customise and use an AI model should be managed (DPP 2)**, especially for high-risk AI models. The data should be accurate, reliable, complete, relevant, lawfully obtained[29]and representative of the target population, and the data should not be discriminatory or contain unjust bias in relation to the purposes for which customisation is being conducted. In this regard, organisations should consider the following:

- Understanding the source, accuracy, reliability, integrity, consistency, completeness, relevance and usability of the data used for model customisation;
- Conducting relevant data preparation processes, such as annotation, labelling, cleaning, enrichment and aggregation;
- Identifying outliers and anomalies in datasets and removing or replacing these values as necessary while maintaining a record of such actions;
- Testing the customisation data for fairness before using it to customise AI models;

Example 2: Unjust bias may inherently exist in datasets if certain groups of individuals are under or over-represented. To address this issue, sampling techniques may be used to rebalance the class distribution, such as random over-sampling (i.e., duplicating samples from the minority class) and random under-sampling (i.e., deleting samples from the majority class)[30].

- Setting aside a portion of the dataset to be used as "holdout" data / test data[31] for validating and / or testing the AI model after customisation; and
- Designating personnel to regularly review the need to further customise the AI models with more data to ensure its effectiveness.

29 For example, under the PDPO, personal data should be collected in a lawful manner. Other applicable laws including intellectual property laws should also be considered.

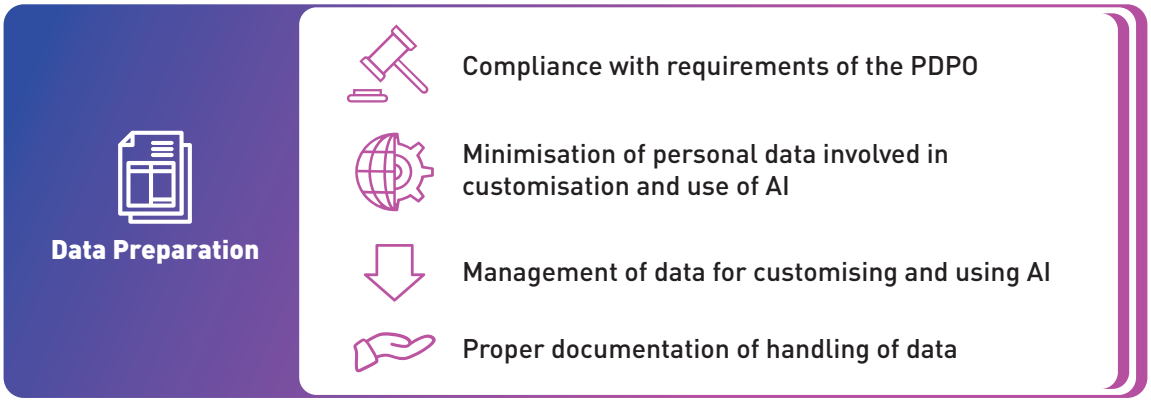
30 Other techniques to mitigate possible bias include re-weighing the input data / features fed into a neural network, and removing the influence of characteristics (e.g., race) on which bias may be based and their proxies.

31 In supervised learning, a holdout dataset reliably represents the training / customisation dataset, but because it has not been seen by the AI model, it can be used to test whether the model would still perform effectively when deployed outside the initial training dataset.

(iv) **The handling of data for the customisation and use of AI should be properly documented** to ensure that the quality and security of data are maintained over time, and to ensure compliance with the requirements of the PDPO. The necessary documentation should cover:

- The sources of the data;
- The allowable uses of the data;
- How the data used were selected from the pool of available data;
- How the data were collected, curated and transferred within the organisation and to the AI supplier (if applicable);
- Where the data is stored; and
- How data quality is maintained over time.

Figure 15: ***4 Aspects of Data Preparation***



3.2 Customisation and Implementation of AI Solutions

Key principles: Transparency and Interpretability / Reliability, Robustness and Security

Customising, testing and validating

42. If customisation is necessary, organisations need to apply the prepared data to customise the procured AI model to suit the specific needs and purposes of AI use[32].

32 For example, to process internal documents and data, to assist with the drafting of documents of a particular domain of expertise or to generate content in a particular corporate style.

43. **In proportion to the level of risks involved, there should be rigorous testing and validation of the AI models to ensure that they perform as intended, and their reliability, robustness and fairness should be evaluated before deployment, especially where the models have been customised**. Recommended measures include:

(i) Validating the AI system with respect to privacy obligations and ethical requirements including fairness, transparency and interpretability;

(ii) Testing the AI model for errors[33] to ensure its reliability, robustness and fairness. System analysts, system architects and data scientists may be consulted, for example, to:

- Compare the AI decisions with decisions made by human beings or traditional non-AI models and compare AI-generated content with real-world data;
- Test the AI model for fairness and accuracy using fairness metrics[34] and accuracy metrics[35] that are appropriate for the context;
- Test the customised AI model with "holdout" data / test data to ensure that it does not overfit its training / customisation dataset and performs effectively[36];
- Use edge cases and potential malicious input to test the AI models; and
- Conduct repeatability and reproducibility[37] tests of the AI system;

(iii) For AI-generated content, implementing mechanisms to:

- Ensure that any disclosure of personal data is compliant with the PDPO, where applicable;

33 For example, regression testing (i.e., testing performed to confirm the recent code / programme changes that does not affect the existing AI application's performance negatively).

34 Fairness can be defined mathematically by different metrics (demographic parity, equality of opportunity, etc.) in a classification model. Certain metrics of fairness are mutually incompatible and cannot be satisfied simultaneously. The organisation should select the suitable metrics to use in a given context.

35 Accuracy can be defined mathematically by different metrics (e.g., accuracy, precision, recall, F1 score, specificity) which test different types of errors in a classification model. Certain metrics of accuracy are mutually incompatible and cannot be satisfied simultaneously. The organisation should select the suitable accuracy metric(s) to use in the given context to know what to optimise for.

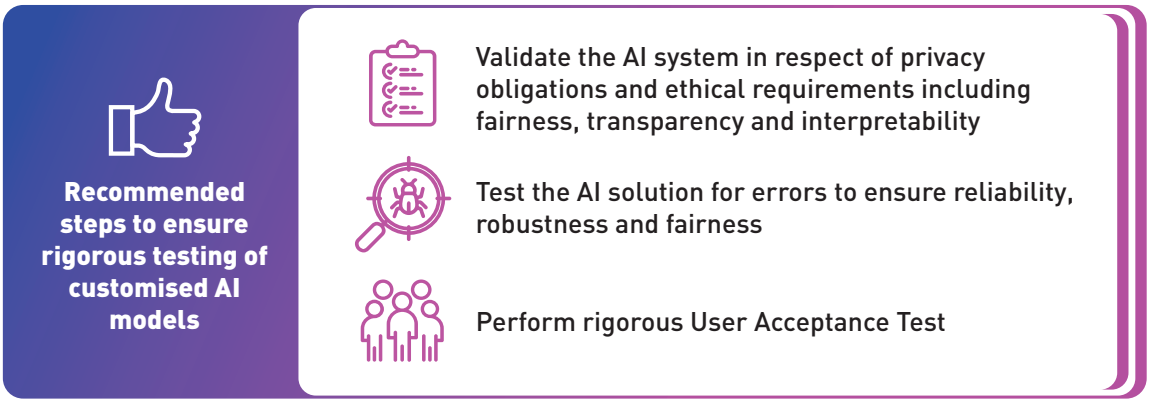
36 Overfitting is where "an [AI] model becomes too specific to the training data and cannot generalise to unseen data, which means it can fail to make accurate predictions on new datasets" (IAPP AI Glossary). Overfitting generally makes an AI system more prone to attacks which may compromise personal data contained in the training / customisation dataset.

37 Reproducibility refers to whether an AI system produces the same results when the same datasets or methods of prediction are used. Reproducibility is important in assessing the reliability of an AI system.

- Identify the content's generated nature (e.g., by labelling and watermarking), where feasible and appropriate; and
- Filter out content which may raise ethical concerns (e.g., biased output, harmful content), where feasible; and

(iv) Performing rigorous User Acceptance Test before integrating the AI solution into the organisation's systems.

Figure 16: Customising, Testing and Validating AI solutions



Integration and hosting

44. Organisations may need to take into account other considerations for compliance with the PDPO, depending on how the AI solution is to be integrated, i.e., whether it will be hosted on an on-premises server or on a cloud server provided by a third party. Hosting an AI system within the organisation's own premises naturally gives the organisation more control over data security than hosting on a third-party cloud. However, the organisation should determine whether it has the expertise to securely run and protect the on-premises system. If the organisation deploys the AI solution on a third-party cloud[38], and personal data are processed in its use, the organisation should, by way of contractual agreement, address issues including:

(i) Compliance with the PDPO (and any other applicable laws) in cross-border data transfers (where applicable);

(ii) Each party's roles and responsibilities as a data user or data processor (as the case may be) as defined under the PDPO; and

38 Organisations are encouraged to read the PCPD's Information Leaflet on Cloud Computing for more information: https://www.pcpd.org.hk/english/resources_centre/publications/files/IL_cloud_e.pdf.

(iii) The data security requirements of each party, including physical and technical controls.

45. In general, organisations should take a holistic approach to the security testing of all components of their AI systems. Implementing third-party-developed AI solutions generally requires adjustments to the organisation's own tech stack, which may encompass security risks. In particular, open-source frameworks, which emphasise transparency by making the source code publicly available, are common in machine learning. Although the answer to whether open-source software is more secure than closed-source software is far from definitive, research has suggested that dependence on open-source machine learning frameworks with externally developed and maintained code may lead to additional security risks. **In any event, organisations implementing AI solutions with open-source components should observe industry-best security practices in maintaining code and managing security risks**[39]**, and pay due attention to security advisories and alerts**. Similarly, organisations which use APIs to programmatically connect the AI solutions to internal applications and systems should carefully review the security of the APIs and follow industry best practices[40].

Ensuring system security and data security

46. Organisations should, in proportion to the level of risks involved, consider adopting the following measures (and involve the AI supplier, where appropriate) to ensure that an AI system is robust, reliable and secure:

(i) Implementing measures (e.g., red teaming) to minimise the risk of attacks against machine learning models, such as malicious input / prompts or training data being fed into the AI system (data poisoning attacks), or the deliberate generation of incorrect or unsafe output (adversarial attacks)[41];

39 Organisations may refer to the information provided by InfoSec: https://www.infosec.gov.hk/en/best-practices/business/open-source-security.

40 For example, by limiting calls that can be made via APIs to the organisation's AI system, and carrying out penetration tests.

41 Other attacks targeting AI models which may affect personal data privacy include model inversion attacks and membership inference attacks, which may seek to uncover personal data contained in training / customisation datasets.

(ii) Implementing internal guidelines for staff on the acceptable input to be fed into and the permitted / prohibited prompts to be entered into AI systems;

Example 3: A law firm is customising a third-party developed AI chatbot to assist its employees in drafting legal documents and performing clerical tasks. Taking into consideration whether the AI chatbot is hosted on-premises or on cloud, the firm is advised to caution its employees against inputting personal data and / or confidential information of its clients when using the AI chatbot.

(iii) Establishing multiple layers of mitigation to prevent system errors or failures at different levels or in different modules of the AI system;

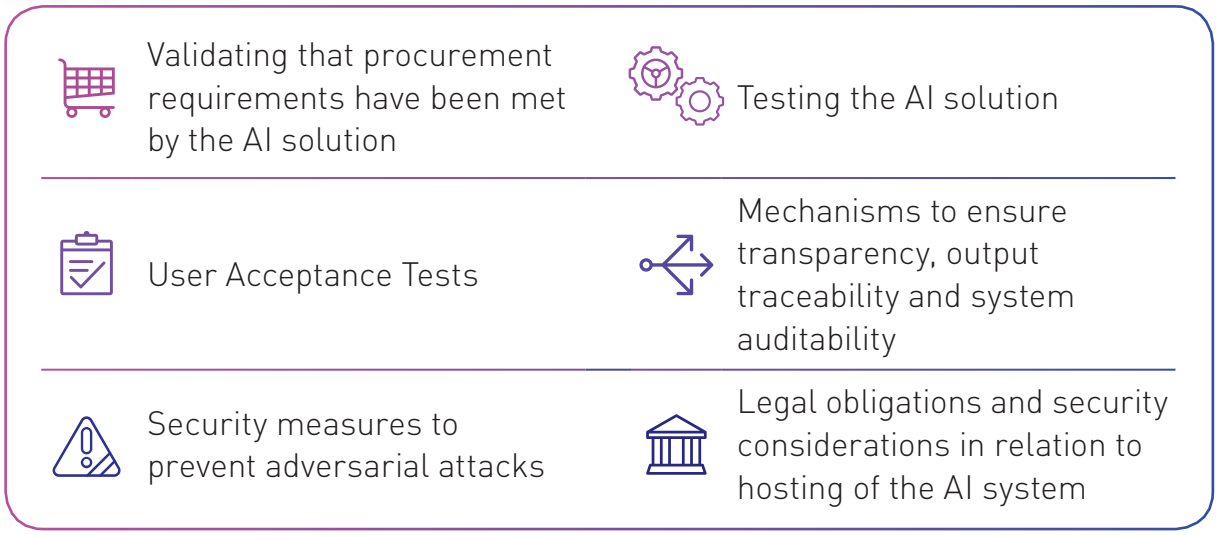
(iv) Establishing contingency plans (e.g., an AI Incident Response Plan - see section 3.3) to promptly suspend the AI system and trigger fallback solutions if necessary;

(v) Establishing mechanisms to ensure that the operations of the AI system is sufficiently transparent to enable end-users to interpret its output; and

(vi) Establishing mechanisms to enable the traceability[42] and auditability of the AI system's output by, for example, where appropriate and in accordance with the data minimisation principle, automatically recording events (i.e., logs) while the AI system is operating.

42 Traceability refers to the ability to keep track, typically by means of documentation, of the development and use of an AI system, including the training and decision-making processes and the data used. Ensuring traceability can help enable auditability.

Figure 17: ***Examples of AI Solution Implementation Considerations***



3.3 Management and Continuous Monitoring of AI Systems

Key principles: Reliability, Robustness and Security / Human Oversight

47. **AI systems should be monitored and reviewed continuously because the risk factors related to their use may change over time**. An AI model itself may also evolve as it learns over time, which affects the reliability, robustness, integrity and security of the AI systems.

48. High-risk AI systems would necessitate more frequent and stringent monitoring and review than low-risk systems. Organisations should consider incorporating the following review mechanisms:

    (i) Maintaining proper documentation on the procurement, risk assessment, risk mitigation measures taken, data sources, data preparation, customisation, testing and validation, implementation and use of the AI system, and considering whether the AI supplier has similar documentation practices;

(ii) Monitoring and logging input to the AI systems (e.g., prompts, queries and requests) to facilitate the prevention of misuse, performance of audits and investigation of any data breach incidents, where appropriate and in accordance with the data minimisation principle[43];

(iii) Conducting re-assessments of the AI system to identify and address new risks, especially when there is a significant change to the functionality or operation of the AI system or to the regulatory or technological environments[44];

(iv) Conducting, and considering requesting the AI supplier to conduct (where necessary and appropriate), a periodic review of the AI models to ensure that they are operating and performing as intended;

(v) Monitoring AI models for any "model drift" or "model decay"[45], and correcting it and involving the AI supplier where necessary and appropriate, to ensure the accuracy of the AI system's output despite changes in the real-world environment, for example, by regularly fine-tuning and re-training the AI model with new data;

(vi) Establishing ongoing feedback and operational support channels with the AI supplier to continuously manage the AI system, which could include feedback from both internal users of the AI system and individuals impacted by the AI system (see Part IV);

(vii) Ensuring that an appropriate level of human oversight of the AI system is in place, taking into account the risk profile of the AI system;

43 For example, organisations are recommended to handle, anonymise and appropriately erase these logs in accordance with a robust data management process.

44 Simple security patches and bug-fixing usually do not trigger the need for re-assessing the risks of an AI system.

45 "Model drift" or "model decay" is where the accuracy or performance of a model degrades over time due to either changes in the environment or target variable on which the AI model produces output ("concept drift") or changes in the input data that the AI model is using to produce output ("data drift").

Human oversight should aim to prevent and minimise the risks posed by AI to individuals. Personnel who exercise human oversight should:

- Understand the capacities and limitations of the AI system, to the extent possible;
- Remain aware of the tendency to over-rely on the output produced by AI (i.e., "automation bias");
- Correctly interpret and assess the output produced by AI;
- Flag and, where appropriate, disregard, override or reverse the output produced by AI if it is abnormal; and
- Intervene and interrupt the operation of the AI system where appropriate, with the assistance of information on the AI system's output from the AI supplier.

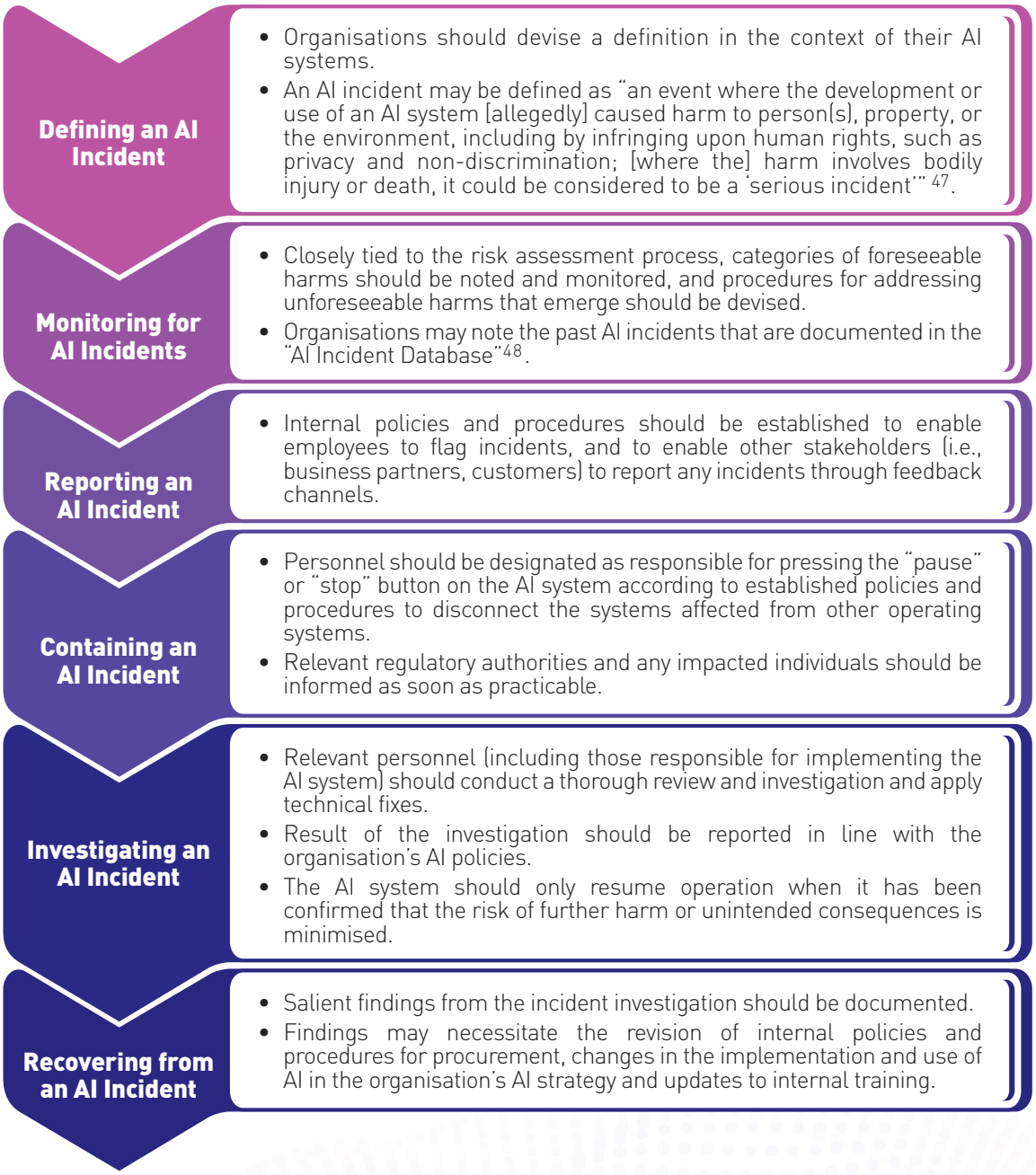
(viii) Maintaining robust security measures throughout the AI system's life cycle, from customisation, implementation, use and monitoring to termination; and

(ix) Regularly evaluating the wider technological landscape to identify gaps in the existing technological ecosystem of the organisation and making adjustments to the AI strategy and governance structure as necessary.

49. Organisations are recommended to consider establishing an **AI Incident Response Plan** to monitor and address incidents that may inadvertently occur[46]. The plan may encompass elements such as:

46 If a data breach incident occurs as part of an AI incident, the organisation should simultaneously engage its data breach response plan.

Figure 18: *AI Incident Response Plan*

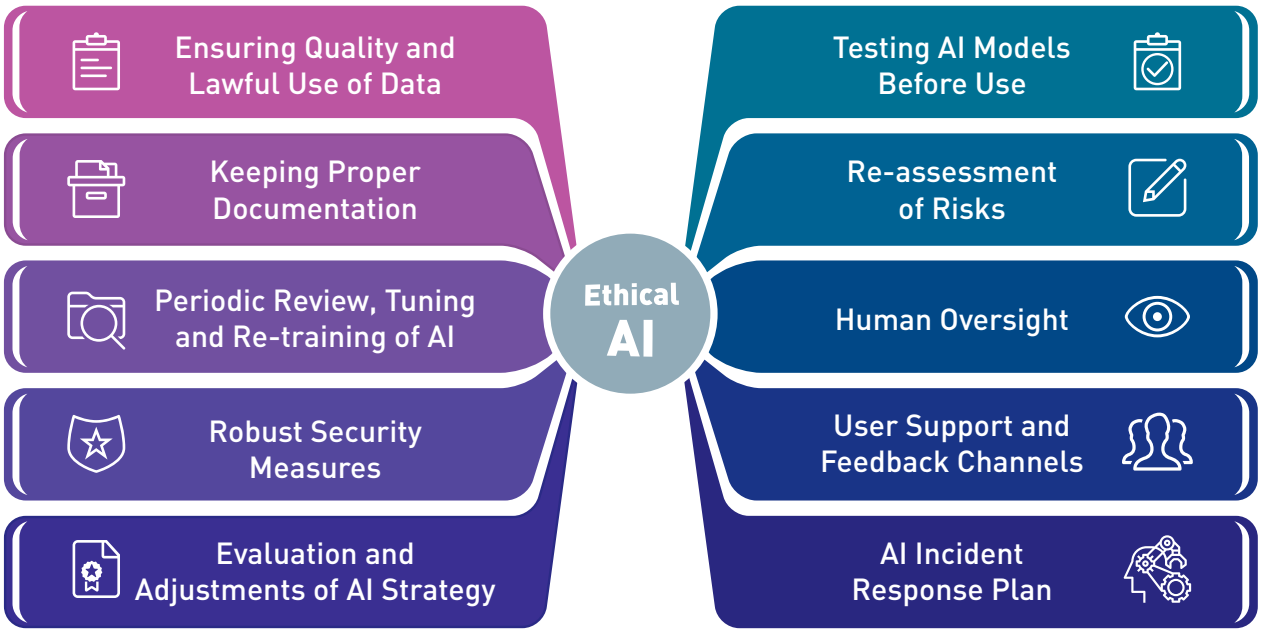


47 OECD (2023), "Stocktaking for the development of an AI incident definition", *OECD Artificial Intelligence Papers*, No. 4, OECD Publishing, Paris, https://doi.org/10.1787/c323ac71-en.; https://oecd.ai/en/wonk/incidents-monitor-aim

48 https://incidentdatabase.ai/

50. **Internal audits (and independent assessments, where necessary) should be conducted periodically to ensure that the use of AI continues to comply with the relevant policies of the organisation and align with its AI strategy**. The results should be reported to the board, top management and governance bodies, such as the audit committee.

Figure 19: ***Management of AI Systems***



Part IV Communication and Engagement with Stakeholders

Key principle: Transparency and Interpretability

4.1 Information Provision

51. An organisation's use of AI should be transparent to stakeholders to demonstrate the organisation's adherence to the "Transparency and Interpretability" principle. **Organisations should communicate and engage effectively and regularly with stakeholders, in particular internal staff, AI suppliers, individual customers and regulators. The level of transparency will vary depending on the stakeholder. Effective communication is essential to building trust**.

52. Where personal data are involved in the customisation and use of AI, organisations must communicate the required information to the data subjects concerned in accordance with DPP 1(3) and DPP 5 of the PDPO, including, but not limited to:

    (i) The purpose for which the personal data are used, e.g., for AI training and / or customisation, or facilitating automated decision-making and so on;

    (ii) The classes of persons to whom the data may be transferred, e.g., the AI supplier; and

    (iii) The organisation's policies and practices in relation to personal data in the context of customisation and use of AI.

53. In addition, to enhance transparency and openness, organisations should consider the following when communicating with stakeholders, especially staff, individual customers and regulators:

    (i) Clearly and prominently disclosing the use of AI systems unless the use is obvious in the circumstances and context;

(ii) Providing adequate information[49] on the purposes, benefits, limitations and effects of using AI systems in their products or services[50]; and

(iii) Disclosing the results of risk assessment of their AI systems[51].

54. In cases where the AI supplier may be better placed than the organisation to provide the above information, especially information about the technical aspects of an AI system, the organisation is recommended to coordinate closely with the AI supplier throughout procurement and beyond and, where necessary, leverage their expertise to address any concerns raised by stakeholders.

4.2 Data Subject Rights and Feedback

55. Where an organisation using AI processes personal data, it should take note that data subjects have the right to submit data access requests and data correction requests respectively under sections 18 and 22 of the PDPO. Organisations may engage the AI supplier to fulfil these requests where necessary.

56. For an AI system that produces decisions / output that may have a significant impact on individuals, organisations should, to the extent possible, provide channels for individuals to provide feedback, seek explanation, and / or request human intervention. Organisations should also carefully consider whether to provide individuals with the option to opt out from using the AI system.

57. More broadly, organisations are recommended to establish user feedback channel for both internal staff and / or customers, encourage the communication of feedback to adjust the relevant AI systems and / or convey the feedback to the AI supplier, where appropriate.

49 Organisations may consider disclosing relevant information about AI systems using AI model cards, which are "short documents provided with machine learning models that explain the context in which the models are intended to be used, details of the performance evaluation procedures and other relevant information" (https://iapp.org/news/a/5-things-to-know-about-ai-model-cards/).

50 Subject to whether the disclosure would compromise commercially sensitive or proprietary information.

51 Subject to whether the disclosure would compromise commercially sensitive or proprietary information.

4.3 Explainable AI

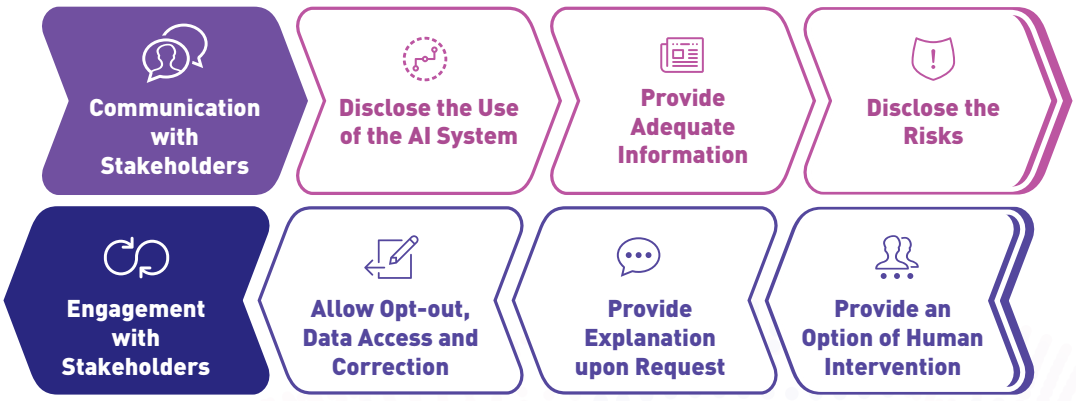
58. Making the decisions and output of AI explainable is the key to building trust with stakeholders. Explanations, where feasible, may include the following information especially when the use of the AI system may have a significant impact on individuals[52]:

(i) How and to what extent AI has been involved in the decision-making process, including a high-level overview of the key tasks for which the AI system is deployed and the involvement of human actors (if any);

(ii) How personal data has been used in the automated or AI-assisted decision-making or content generation processes and why those data are considered relevant and necessary; and

(iii) The major factors leading to the automated decisions / output by the AI system (global explainability), and the major factors leading to the individual decisions / output (local explainability). If it is not feasible to provide an explanation, then that should be made explicit.

Figure 20: ***Communication and Engagement with Stakeholders***



52 Organisations may consider referencing the guidance on *Explaining Decisions Made with AI* published by the Information Commissioner's Office, UK and The Alan Turing Institute in 2020 for more advice on how automated decisions made by AI may be meaningfully explained.

59. Organisations may consider engaging the AI supplier, where appropriate, who may be better placed to explain the decisions and output of the AI systems. In deciding on the types of information to be disclosed and the level of details, organisations should consider, among others, the stakeholders' comprehension of the information, their needs and whether the disclosure would adversely impact the security and legitimate purposes of the AI system. For example, in the case of an AI system used to detect customer fraud or other crimes, the organisation may not need to disclose the relevant indicators used by the AI system, lest the customers learn how to bypass the system. However, for AI systems designed to assist employees' internal use and that are customised with internal data, the organisation may consider providing an option to trace the source of the information utilised by the AI system to produce the output / decision to ensure its accuracy, where feasible and appropriate[53].

4.4 Language and Manner

60. Communication with stakeholders, particularly consumers, should be in plain language that is clear and understandable to lay persons, and such communication should be drawn to the attention of stakeholders. Communication may also be included in an organisation's privacy policies.

Communication with stakeholders should be in plain language that is clear and is understandable to lay persons, and be drawn to the attention of stakeholders.

53 For example, where Retrieval-Augmented Generation was involved in the customisation process.

Acknowledgement

The Office of the Privacy Commissioner for Personal Data (PCPD) would like to thank the following individuals and organisations, as well as major AI suppliers, for giving us invaluable feedback during our consultation, by alphabetical order:

Supporting Organisations

Hong Kong Applied Science and Technology Research Institute

Office of the Government Chief Information Officer

Members of the Standing Committee on Technological Developments of the PCPD

Ir Alex CHAN, General Manager, Digital Transformation Division, Hong Kong Productivity Council

Mr Alan CHEUNG, Chief Director, Artificial Intelligence and Trust Technologies, Hong Kong Applied Science and Technology Research Institute

Adj. Professor Jason LAU, Director, ISACA International Board of Directors

Dr Gregg LI, Founding Director and President, Orion Astropreneur Space Academy

Professor the Hon William WONG Kam-fai, MH, Associate Dean (External Affairs), Faculty of Engineering, The Chinese University of Hong Kong

Professor S M YIU, Professor and Deputy Head, Department of Computer Science, The University of Hong Kong

Organisations

AI & Humanity Lab, The University of Hong Kong

Asia Securities Industry & Financial Markets (ASIFMA)

Centre for Information Policy Leadership

Deloitte

Ernst & Young

Hong Kong Association of Banks

Hong Kong Computer Society

Hong Kong Monetary Authority

Hong Kong Productivity Council

Hong Kong Science and Technology Parks Corporation

Research Centre for Sustainable Hong Kong, City University of Hong Kong

Appendix A - Data Protection Principles under the Personal Data (Privacy) Ordinance

The Personal Data (Privacy) Ordinance (Cap. 486) ("PDPO") governs the collection, holding, processing and use of personal data by both private and public sectors. The PDPO is technology-neutral and principle-based. The Data Protection Principles ("DPP") in Schedule 1 to the PDPO represent the core requirements of the PDPO and cover the entire life cycle of the handling of personal data from collection to destruction.

DPP 1 - PURPOSE AND MANNER OF COLLECTION

DPP 1 provides that personal data shall only be collected for a lawful purpose directly related to a function or activity of the data user. The means of collection shall be lawful and fair. The data collected shall be necessary and adequate but not excessive for such purpose.

Data users shall also be transparent as regards the purpose of collection and the potential classes of persons to whom the personal data may be transferred, and the data subjects' right and means to request access to and correction of their personal data. Usually, the information is presented in a Personal Information Collection Statement.

DPP 2 - ACCURACY AND DURATION OF RETENTION

DPP 2 requires data users to take all practicable steps to ensure that personal data is accurate and is not kept longer than is necessary for the fulfillment of the purpose for which the data is used. Section 26 of the PDPO contains similar requirements for the erasure of personal data that is no longer required.

If a data user engages a data processor for handling personal data, the data user must then adopt contractual or other means to prevent the personal data from being kept longer than is necessary by the data processor.

DPP 3 - USE OF DATA

DPP 3 prohibits the use of personal data for any new purpose which is different from and unrelated to the original purpose of collection, unless express and voluntary consent has been obtained from the data subjects.

DPP 4 - DATA SECURITY

DPP 4 requires data users to take all practicable steps to protect the personal data they hold against unauthorized or accidental access, processing, erasure, loss or use.

If a data user engages a data processor in processing the personal data held, the data user must adopt contractual or other means to ensure that the data processor complies with the aforesaid data security requirement.

DPP 5 - OPENNESS AND TRANSPARENCY

DPP 5 obliges data users to take all practicable steps to ensure certain information, including their policies and practices in relation to personal data, the kind of personal data held and the main purposes for which the personal data is held, is generally available to the public.

DPP 6 - ACCESS AND CORRECTION

DPP 6 provides data subjects with the right to request access to and correction of their own personal data.

DPP 6 is supplemented by the detailed provisions in Part 5 of the PDPO which covers the manner and timeframe for compliance with data access requests and data correction requests, the circumstances in which a data user may refuse such requests, etc.

Appendix B - Main Publication Reference List

- Infocomm Media Development Authority and Aicadium, Singapore, *Generative AI: Implications for Trust and Governance* (2024)[54]
- Infocomm Media Development Authority, Singapore and AI Verify Foundation, *Proposed Model AI Governance Framework For Generative AI Fostering a Trusted Ecosystem* (2024)[55]
- ISACA, *Artificial Intelligence: A Primer on Machine Learning, Deep Learning and Neural Networks* (2024)[56]
- Organisation for Economic Cooperation and Development, France, *AI principles* (2024)[57]
- Personal Data Protection Authority, Singapore, *Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems* (2024)[58]
- Commission Nationale de l'Informatique et des Libertés, France, *AI how-to sheets* (2023)[59]
- Competition and Markets Authority, UK, *AI Foundation Models: Initial Report* (2023)[60]
- Cyberspace Administration of China, the People's Republic of China, *Global AI Governance Initiative* (2023)[61]
- Information Commissioner's Office, UK, *Guidance on AI and data protection* (2023)[62]
- ISACA, *The Promise and Peril of the AI Revolution: Managing Risk* (2023)[63]
- International Organization for Standardization, *ISO/IEC 23894:2023 Information technology - Artificial intelligence - Guidance on risk management* (2023)[64]
- International Organization for Standardization, *ISO/IEC 42001:2023 Information technology - Artificial intelligence management system* (2023)[65]

54 https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf
55 https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf
56 https://store.isaca.org/s/store#/store/browse/detail/a2S4w000008Kn59EAC
57 https://oecd.ai/en/ai-principles
58 https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/advisory-guidelines-on-the-use-of-personal-data-in-ai-recommendation-and-decision-systems.pdf
59 https://www.cnil.fr/en/ai-how-sheets
60 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1185508/Full_report_.pdf
61 https://www.mfa.gov.cn/eng/wjdt_665385/2649_665393/202310/t20231020_11164834.html
62 https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection-2-0.pdf
63 https://www.isaca.org/resources/white-papers/2023/the-promise-and-peril-of-the-ai-revolution
64 https://www.iso.org/obp/ui/en/#iso:std:iso-iec:23894:ed-1:v1:en
65 https://www.iso.org/standard/81230.html

- Meta, *Llama 2 - Responsible Use Guide* (2023)[66]
- National Cyber Security Centre, UK, and Cybersecurity and Infrastructure Security Agency, US, *Guidelines for secure AI system development* (2023)[67]
- National Institute of Standards and Technology, US Department of Commerce, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (2023)[68]
- National Technical Committee 260 on Cybersecurity of Standardization Administration, the People's Republic of China, *Practical Guidance of Cybersecurity Standards - Labelling Methods for Content Generated by Generative Artificial Intelligence Services* (2023)[69]
- Organisation for Economic Cooperation and Development, *Advancing Accountability in AI: Governing and Managing Risks throughout the Lifecycle for Trustworthy AI* (2023)[70]
- Office of the Government Chief Information Officer, Hong Kong SAR, China, *Ethical Artificial Intelligence Framework (Customised version for general reference by public)* (2023 revised edition)[71]
- Office of the Privacy Commissioner, Canada, *Principles for responsible, trustworthy and privacy-protective generative AI technologies* (2023)[72]
- United Nations AI Advisory Body, *Interim Report: Governing AI for Humanity* (2023)[73]
- United Nations Educational, Scientific and Cultural Organization, *Recommendation on the Ethics of Artificial Intelligence* (2023)[74]
- World Economic Forum, *Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector* (2023)[75]
- Information Commissioner's Office, UK, *AI and data protection risk toolkit* (2022)[76]

66 https://llama.meta.com/responsible-use-guide/
67 https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf
68 https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
69 https://www.tc260.org.cn/upload/2023-08-25/1692961404507050376.pdf
70 https://www.oecd.org/sti/advancing-accountability-in-ai-2448f04b-en.htm
71 https://www.ogcio.gov.hk/en/our_work/infrastructure/methodology/ethical_ai_framework/doc/Ethical_AI_Framework.pdf
72 https://www.priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/gd_principles_ai/
73 https://www.un.org/en/ai-advisory-body.
74 https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence
75 https://www.weforum.org/publications/adopting-ai-responsibly-guidelines-for-procurement-of-ai-solutions-by-the-private-sector/
76 https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/

Appendix B - Main Publication Reference List

- ISACA, *Developing an Artificial Intelligence Governance Framework* (2022)[77]
- Microsoft, *Microsoft Responsible AI Standard, v2 - General Requirements* (2022)[78]
- Wiley: *Trustworthy AI - A business guide for navigating trust and ethics in AI* (2022)[79]
- National Governance Committee for the New Generation Artificial Intelligence, the People's Republic of China, *Guidance on the Ethics of the New Generation AI* (2021)[80]
- Office for Artificial Intelligence, UK, *Guidelines for AI procurement* (2021)[81]
- Office of the Privacy Commissioner for Personal Data, Hong Kong SAR, China, *Guidance on the Ethical Development and Use of Artificial Intelligence* (2021)[82]
- Infocomm Media Development Authority and Personal Data Protection Commission, Singapore, *Model Artificial Intelligence Governance Framework* (2020 second edition)[83]
- Information Commissioner's Office, UK, and The Alan Turing Institute, *Explaining Decisions Made with AI* (2020)[84]
- Google, *Responsible AI Practices*[85]
- International Association of Privacy Professionals, *Key Terms for AI Governance*[86]

77 https://www.isaca.org/resources/news-and-trends/newsletters/atisaca/2022/volume-38/developing-an-artificial-intelligence-governance-framework
78 https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf
79 https://www.wiley.com/en-us/Trustworthy+AI%3A+A+Business+Guide+for+Navigating+Trust+and+Ethics+in+AI-p-9781119867951
80 https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html
81 https://www.gov.uk/government/publications/guidelines-for-ai-procurement/guidelines-for-ai-procurement
82 https://www.pcpd.org.hk/english/resources_centre/publications/files/guidance_ethical_e.pdf
83 https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf
84 https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/
85 https://ai.google/responsibility/responsible-ai-practices/
86 https://iapp.org/resources/article/key-terms-for-ai-governance/

PCPD.org.hk

香港個人資料私隱專員公署
Office of the Privacy Commissioner for Personal Data, Hong Kong

Tel : 2827 2827
Fax : 2877 7026
Address : Unit 1303, 13/F., Dah Sing Financial Centre, 248 Queen's Road East, Wanchai, Hong Kong
E-mail : communications@pcpd.org.hk

PCPD Website: pcpd.org.hk

Download this Publication

Disclaimer

The information and suggestions provided in this publication are for general reference only. They do not serve as an exhaustive guide to the application of the law and do not constitute legal or other professional advice. The Privacy Commissioner makes no express or implied warranties of accuracy or fitness for a particular purpose or use with respect to the information and suggestions set out in this publication. The information and suggestions provided will not affect the functions and powers conferred upon the Privacy Commissioner under the Personal Data (Privacy) Ordinance.

June 2024