

Secure AI@Work Enablement Campaign

Safe and Responsible Use of AI at Work

A practical guide on AI usage rules for all staff

Prepared by the Digital Policy Office and the Hong Kong Internet Registration Corporation Limited, with drafting support from NyxLab Limited

Version: v1.1, English, August 2026

Contents

Contents	2
Overview.....	4
Part 1: Introduction	6
1.1 Purpose of This Training	6
1.2 Overview of This Training	6
Part 2: Overview of the AI Use Lifecycle	7
2.1 Before Using AI: Establish a Secure Working Environment.....	9
2.2 While Using AI: Follow the Four Golden Rules	10
2.3 After Using AI: Complete the Required Review and Reporting Procedures.....	11
Part 3: Understanding AI and How It Works	12
3.1 Basic Concepts of Generative AI.....	12
3.2 How AI Tools Handle Input Data	14
3.3 Why AI Outputs May Be Inaccurate: Understanding AI Hallucinations	15
3.4 Differences Between Approved AI Tools and Personal AI Subscribed Accounts	18
3.5 Writing Safe and Effective Prompts	20
3.6 Three Core Principles for Staff Using AI.....	21
Part 4: Four Golden Rules for Secure AI Use	23
4.1 Rule 1: Do Not Input Sensitive or Personal Data	23
4.2 Rule 2: Verify AI Output Before Use	25
4.3 Rule 3: Use Only Approved AI Tools	27
4.4 Rule 4: Report Anomalies and Incidents to IT Department Promptly	28
Part 5: Workplace Applications of AI	30
5.1 Using AI to Draft Emails.....	31
5.2 Using AI for Translation.....	31
5.3 Using AI for Data Analysis	32
5.4 Recent Significant Incidents Relating to AI Use	33
Part 6: Other AI-Related Risks.....	34
6.1 Copyright and Intellectual Property Considerations.....	34
6.2 AI Image Generation and Deepfake Risks	35
6.3 Prompt Injection Attacks	36
6.4 Risk of Bias in AI Output.....	37
6.5 AI Agents: From Content Generation to Autonomous Task Execution.....	38
Part 7: Practical Guidance for Secure AI Use	39
7.1 Recommended Practices and Practices to Avoid	39
7.2 Secure AI Use Checklist.....	41

Part 8: Follow-Up Actions and Support 43

8.1 Understanding the Organisation’s AI Governance Requirements 43

8.2 Support Resources and Enquiry Channels.....43

Appendix I: Glossary 44

Appendix II: Further Examples of Workplace AI Use 46

Appendix III: References and Methodology 49

Overview

Audience	Staff across all types of organisations in Hong Kong, including administrative, frontline, and back-office staff in small and medium enterprises, schools, and non-governmental organisations.
Duration	Approximately 60 minutes of self-study or reading time.
Objectives	To strengthen staff awareness and capability in the secure use of AI in the workplace through practical workplace scenarios and case studies, and to mitigate operational AI risks such as accidental data leakage, over-reliance on AI output, and improper use of AI-generated content.

This booklet is prepared with reference to relevant Hong Kong laws and workplace practices, as well as international best practices for responsible AI usage. It aims to help staff use AI securely and in compliance with applicable standards. Detailed references and the methodology are provided in Appendix III.

Case Study: An Incident Caused by Input Error

In 2023, an employee of a multinational technology company inadvertently entered internal company data into an AI tool, leading to a high-profile data leakage. Similar incidents have occurred frequently in recent years. As AI tools become increasingly integrated into workplace processes, seemingly routine actions such as inputting data, drafting prompts, or uploading documents can result in data leakage, non-compliance or other business risks.

This booklet is not intended to prohibit the use of AI. Through clear principles, practical workplace scenarios, and common case studies, it is designed to help readers understand the essentials of secure AI use in approximately 60 minutes. It explains how AI can improve efficiency while mitigating risk, highlights high-risk practices to avoid, and details incident handling and reporting protocols. Upon completion, readers will be able to use AI more securely, prudently, and confidently at work, while improving efficiency and managing the associated risks.

Common Workplace AI Risk Scenarios

Common Problems	Related Training Topic
Failure to recognise that entering customer data into a free or personal subscribed AI tool may cause data leakage or even create compliance risks under the Personal Data (Privacy) Ordinance.	Section 4.1; Section 5.1; Section 5.2; Section 5.3; Appendix II.
Over-reliance on AI output, including submitting reports or sending emails generated by AI without verification, resulting in errors or false information.	Section 3.3; Section 4.2; Section 5.

Common Problems	Related Training Topic
Insufficient understanding of how AI tools handle data, including the risk that inputted data may be accessed, retained, or further processed.	Section 3.2; Section 3.4.
Use of free or personal subscribed AI accounts for work without adequate understanding of the associated data protection and confidentiality risks.	Section 2.1; Section 3.4; Section 4.3.
Use of AI-generated translations, summaries or analyses without human review, resulting in errors in deliverables.	Section 5.1; Section 5.2; Section 5.3.
Uncertainty about the standard protocols for handling and reporting an AI malfunction or incident.	Section 2.3; Section 4.4.

Part 1: Introduction

1.1 Purpose of This Training

This booklet applies to all staff within the organisation, not only the information technology (IT) department or management. Staff whose duties involve data processing, content creation, analysis and research, customer service, administrative support, or any other use of AI tools should understand and comply with the basic principles and practical requirements set out in this booklet. As AI is becoming an increasingly important feature of the modern workplace, with applications across departments and business processes, every staff member, regardless of seniority or function, is responsible for using AI securely, responsibly, and in accordance with the organisation's policies and procedures to safeguard the organisation's information security, data privacy, and reputation.

AI tools are rapidly being integrated into workplace processes, and staff can leverage them to draft emails, summarise documents, translate content, and handle enquiries, significantly improving efficiency. However, using AI without adequately understanding how the tool collects, processes, and uses input data may expose staff, customers, and the organisation to data security, privacy, compliance, and operational risks.

No technical or legal background is required to read this booklet. It is written in a clear, accessible, and practical manner to explain the basic principles and best practices for using AI at work.

1.2 Overview of This Training

This booklet is structured around three stages: before using AI, while using AI, and after using AI. It identifies the principal risks, precautions, and practical requirements at each stage of the AI use lifecycle.

- Part 2 introduces the three-stage framework and the issues to be aware of before, during, and after using AI.
- Part 3 explains the fundamental concepts of AI and its operation.
- Part 4 outlines the four core rules to follow whenever using AI tools.
- Part 5 demonstrates secure and responsible AI use through practical workplace scenarios.
- Part 6 covers common risks, including copyright disputes, deepfakes, prompt injection attacks, and algorithmic bias.
- Part 7 provides a printable quick-reference summary for workplace use.
- Part 8 lists internal and external support resources, responsible functions, and enquiry channels to help staff obtain appropriate assistance with AI-related questions.
- Appendix I: Glossary
- Appendix II: Further Examples of Workplace AI Use
- Appendix III: References and Methodology
- Knowledge Test (online version only)

Part 2: Overview of the AI Use Lifecycle

Secure use of AI requires effective management throughout the AI use lifecycle. A failure in tool selection, data input, use of output, post-use review, or follow-up may create information security, data privacy, and business risks for the organisation.

Accordingly, staff should apply the appropriate controls at each of the three stages: Before using AI, While using AI, and After using AI. This helps establish consistent and sustainable best practices.

Staff should apply the three-stage process below whenever using AI for work:

Stage	Action Staff Should Take	Reason for Risk Control
1. Before Using AI	• Confirm that the organisation has approved the AI tool for use. • Verify that the tool is obtained from an authorised source, such as the supplier's official website, an official application store, or a link provided by the IT department. • Sign in using the work account designated by the organisation. Do not use a personal account for work-related matters. • Confirm that the proposed task and purpose fall within the organisation's permitted scope of AI use.	Many AI-related incidents arise before use, including through unapproved tools, software obtained from unverified sources, or personal accounts used for work data. Completing these checks helps identify risks at an early stage and reduces the likelihood of an incident.
2. While Using AI	Apply the four golden rules whenever using AI (Refer to Part 4): • Do not input sensitive data or personal data. • Verify AI output before use. • Use only AI tools approved by the organisation. • Report any anomaly or suspected incident promptly to the IT department, your supervisor, or the designated responsible officer.	Compliance with these requirements safeguards the organisation's data and business operations. Failure to comply may increase the risk of data leakage, dissemination of incorrect information, and operational disruption. The four golden rules provide practical controls for preventing common risks associated with workplace AI use.
3. After Using AI	• Review AI output carefully and verify all facts, data, and cited references.	Post-use review and follow-up are required to confirm that AI output is accurate, used appropriately, and consistent with the organisation's governance and compliance

Stage	Action Staff Should Take	Reason for Risk Control
	- Where required by organisational policy or workplace requirements, disclose the use of AI. - Report unusual situations immediately, including anomalous content, suspected data leakage, or unusual tool behaviour to the IT department or your supervisor.	requirements. These controls also support the early identification and mitigation of potential issues.

2.1 Before Using AI: Establish a Secure Working Environment

Before using any AI tool, staff should confirm that the organisation has approved the tool, that the tool is obtained from a trusted source, and that the account complies with organisational requirements. An AI tool may be used for work only after all three conditions have been satisfied.

Where the organisation has not published an approved AI tools list, staff should apply the following precautionary controls:

- When handling work data, particularly personal, student, case-related, financial, contractual, or other non-public data, use only securely configured work-account tools provided by the organisation.
- If a tool's approval status cannot be confirmed, do not use it and obtain confirmation from your supervisor, the IT department, or the designated responsible function.
- Do not enter real work data into a free, trial, or personal subscribed AI tool, even for testing, evaluation, or trial use.

Pre-Use Checklist: Before using an AI tool, confirm the following:

- The AI tool appears on the organisation's approved AI tools list or has been expressly approved for work by the IT department.
- The AI tool was obtained from the supplier's official website, an organisation-designated distribution platform, or another trusted source, rather than a third-party download platform, cracked software source, sideloaded application, or unofficial application store.
- The AI tool is accessed through a work account provided by the organisation (an enterprise subscription), and not a free or personal subscribed account.
- The proposed task complies with the organisation's AI use requirements and falls within the permitted scope for the relevant role.
- The proposed input contains no personal, customer, confidential, or other sensitive information; if there is business need to process the relevant information, de-identification or other safeguards have been applied in accordance with the organisation's requirements.

Risk Warning

Unapproved AI tools, applications from unverified sources, and personal accounts used for work data are common sources of AI risk. Pre-use checks help identify and reduce these risks before they result in data leakage, non-compliance, or other information security incidents.

2.2 While Using AI: Follow the Four Golden Rules

After completing the pre-use assessment and confirming that the tool, account, and proposed data-use arrangements comply with organisational requirements, staff should observe the following four core principles throughout their use of the AI tool to ensure compliance with the organisation's data protection, information security, and compliance requirements.

- **Do not input personal data or other sensitive information.**

Personal data, customer data, confidential information, or other sensitive information shall not be entered into an AI tool unless the organisation has approved the use and all applicable handling requirements are satisfied.

- **Verify all AI-generated content.**

AI-generated content may be incorrect, incomplete, outdated, or misleading, and should not be adopted, published, or relied upon for decision-making without appropriate verification.

- **Use only AI tools and accounts approved by the organisation.**

Staff should, in accordance with the organisation's requirements, use approved AI tools and designated accounts, and shall not independently adopt unassessed or unapproved alternatives to process work data.

- **Identify and report anomalies promptly.**

Report any risk of data leakage, anomalous output, unauthorised access, system anomalies, or other suspected incident to the IT department or the designated responsible officer promptly in accordance with the organisation's procedures.

Many AI-related incidents arise from inadequate controls while using AI, rather than technical failure. Following these rules reduces the risks of data leakage, dissemination of incorrect information, non-compliance, and information security incidents, ensuring the use of AI complies with the organisation's governance requirements and risk management standards.

Part 4 explains how these rules apply in different workplace scenarios and outlines the relevant Hong Kong laws, common risks, and potential consequences of non-compliance.

2.3 After Using AI: Complete the Required Review and Reporting Procedures

Work processes do not end when AI generates content. The user shall independently determine whether the final output is accurate, complete, suitable for its intended purpose, and compliant with organisational policies, applicable laws and regulations, and relevant professional requirements. Effective post-use review, verification, and reporting are essential to secure, reliable, and accountable AI use.

Post-Use Checklist

- Conduct a comprehensive review of all AI-generated content, including text, data, analytical results, quotations, references, and recommendations.
- Verify and cross-check all relevant factual statements, figures, quotations, and references against reliable and authoritative original sources.
- Confirm that the final output meets the business needs and task requirements, and accept responsibility for any content adopted or published. The AI tool is merely an assistant and is not the party responsible for the deliverable.
- Where organisational policy, regulatory requirements, customer requirements, or professional standards require disclosure of AI involvement, provide an appropriate disclosure or labelling regarding AI usage.
- If any anomaly is identified during use, including abnormal tool operation, suspected data leakage, dissemination of incorrect information, unauthorised access, or any other potential risk, report it to the IT department, your supervisor, or the designated responsible officer in accordance with the established procedures.

Importance of Post-Use Review

While AI tools can improve efficiency and support decision-making, they cannot replace professional judgement, management oversight, or accountability mechanisms. Unverified content, including incorrect information, inaccurate analysis, misquotations, or outdated data, may be incorporated into work products and consequently affect the organisation's operations, reputation, legal liability, and compliance position.

By establishing rigorous post-use review and reporting protocols, the organisation can strengthen quality control over AI output, identify and address potential risks at an early stage, and ensure that its use of AI continues to comply with information security, data protection, risk management, and corporate governance requirements.

Part 3: Understanding AI and How It Works

3.1 Basic Concepts of Generative AI

AI tools commonly used in the workplace are mostly forms of generative AI. Using AI securely and effectively requires an understanding that generative AI processes information and responses differently from the way humans understand, analyse, and assess facts.

How Generative AI Works

In general, generative AI does not retrieve answers by consulting a verified database. Instead, it generates responses by predicting the content most likely to follow based on patterns, correlations, and regularities learned from large volumes of data. Therefore, AI-generated content may not be factual and should not be treated as a verified source of information.

Model Training and Response Generation

When a user enters a question or instruction, the AI analyses the input and uses learned patterns to predict the words, sentences, and passages most likely to form a response.

AI-generated content is therefore the product of calculation and prediction, not item-by-item retrieval from an authoritative source. For example, AI may complete the statement “The capital of France is” with “Paris” because that correlation is strongly represented in its training data, not because it has verified the answer against an official source.

While AI responses may appear reasoned or analytical, the underlying process remains based principally on pattern recognition and probabilistic prediction.

Capabilities and Limitations of Modern AI Tools

Some of the latest generation of AI tools can retrieve information from the internet or an organisation's internal knowledge base before generating a response. This is generally known as retrieval-augmented generation (RAG). When the latest information is needed, retrieval may improve the timeliness and relevance of a response, but it does not make AI a reliable fact-checking mechanism.

An AI tool may cite a recently retrieved web page, document, or data source without necessarily being able to assess whether the material is accurate, the content is outdated, the source is authoritative, the information is biased or erroneous, or the content was generated by another AI system.

Real-time retrieval may improve convenience and efficiency, but it does not necessarily improve the reliability of the information obtained.

Understanding the Limitations of AI

Unless relevant context is provided in the prompt or made available through an authorised connection, an AI tool does not understand the organisation's operations, customer circumstances, staff roles, internal documents, or organisational knowledge.

While AI is trained on large volumes of data, this does not enable it to determine independently whether a source is reliable. AI cannot reliably identify authoritative sources, or distinguish information that may be inaccurate, outdated, or misleading.

Even an AI tool with real-time web-search capability can only read the information it retrieves. It may not be able to assess independently the accuracy, completeness, or reliability of that content.

Moreover, AI does not necessarily know whether its own response is correct. Whether the content is factual, inferential, erroneous, or fabricated, it may be presented in the same fluent and confident tone. This is known as an AI hallucination and is one of the inherent risks of generative AI.

Key Points for Staff

The risk of AI lies not merely in occasional error, but in its capacity to present incorrect statements in a manner that appears reasonable, credible, or even professional. Appropriate review, verification, and assessment of AI output therefore remain an important safeguard for ensuring work quality.

Because generative AI generates responses through probabilistic prediction rather than retrieving a single fixed answer from a database, the same question may produce different responses at different times or in different contexts. This does not necessarily indicate a malfunction, but is a normal operational characteristic of generative AI.

Staff should understand that AI output is not ordinarily a sole or final answer, but a suggestion or draft for reference. Responsibility for the final judgement and decision remains with the user.

Core Principle: AI-generated content is reference information only. Apparent plausibility does not equate to accuracy. Before any AI output is incorporated into a work product, business decision or external communication, it should be subjected to appropriate verification, professional judgement, and the necessary human oversight.

3.2 How AI Tools Handle Input Data

Data entered into a free or personal subscribed AI tool may be used, pursuant to the supplier's terms and policies, to improve product features, optimise model performance, or support the operation of related services. In practice, staff should therefore treat free or publicly available AI tools as external platforms and refrain from entering any actual work data, particularly personal data, student data, case data, customer data, financial data, contract content or other internal confidential information.

International guidance on generative AI consistently emphasises that users are responsible for understanding how input data will be collected, used, stored, and retained. Every staff member must understand these basic principles before using an AI tool, as they are essential to safeguarding data security and privacy.

Where the organisation licenses an enterprise-version AI tool, it will ordinarily enter into appropriate service agreements and data processing arrangements with the supplier. The enterprise terms will govern matters such as data use, access controls, and safeguards. In general, content entered through an organisational work account will not be used to train the supplier's publicly available models.

However, use of an enterprise-version tool does not permit unrestricted data input. Staff shall continue to comply with the organisation's data classification requirements and the principle of data minimisation by entering only the minimum information necessary to complete the task and avoiding disclosure of information beyond what is genuinely required.

Importantly, these enterprise protections generally apply only where the relevant tool is accessed through a work account configured and managed by the organisation. They do not ordinarily extend to free or personal subscribed AI accounts registered by staff. Accordingly, even where the same AI tool is used, the applicable data protection arrangements may differ materially depending on the account type.

Core Principle: AI tools are not internal organisational databases. Before entering any data, assess its sensitivity, the business need, and the applicable compliance requirements. Ensure only the minimum information necessary to complete the task is disclosed.

3.3 Why AI Outputs May Be Inaccurate: Understanding AI Hallucinations

An AI hallucination occurs when generative AI produces content in a confident and persuasive manner despite lacking accurate information or sufficient understanding. This may include stating false facts, citing non-existent sources, misinterpreting source material, or providing flawed analysis.

AI hallucination is not merely an isolated system failure or occasional technical error, but an inherent risk arising from the way generative AI uses existing data and probabilistic prediction to generate content. A response may therefore be fluent and logically structured while still being inaccurate, incomplete, or unsupported by reliable evidence.

In a corporate or workplace setting, adopting AI output without verification may introduce inaccurate information into reports, decision papers, external communications, customer responses, or compliance documents. This may adversely affect the organisation's operations, reputation, compliance position, and risk management.

AI hallucinations are more likely to arise in the following circumstances:

- **Specific facts, figures, or statistics are involved**
AI may provide inaccurate or outdated information about dates, amounts, market data, operational indicators, statutory time limits, historical events, or other specific facts and figures.
- **Legislation, policies, cases or academic references are involved**
AI may misstate statutory provisions, policy documents, research reports, case names, or other sources, and may even generate references that do not exist.
- **Recent developments or current events are involved**
Because an AI model's knowledge may be limited to information available up to a particular point in time, it may not reflect the latest regulatory requirements, market developments, news events, or technological advances.
- **Professional, specialist or highly complex matters are involved**
Where relevant training data is limited, AI may infer from similar scenarios and produce content that does not reflect the actual circumstances or applicable professional requirements.

Key Points for Risk Management

AI-generated content should be treated as reference material that supports work and informs thinking, rather than as verified evidence. Whether the content is used for analysis, decision-making, report writing, customer communications, or external publication, the user shall exercise professional judgement and verify it against reliable and authoritative sources.

The organisation should establish appropriate review and verification controls to ensure that significant decisions, key data, legal, or compliance content, and information concerning the organisation's external position are adequately confirmed before use. Effective human oversight and proportionate quality controls can reduce the operational, legal, and reputational risks arising from AI hallucinations.

Scenario: A Fabricated Citation

A staff member preparing a briefing on AI adoption by Hong Kong enterprises asks a free AI chat tool: “What is the current rate of AI adoption among Hong Kong small and medium enterprises, and is there an official report supporting that figure?”

Roughly what is the AI adoption rate among Hong Kong SMEs at the moment? Is there any official report to support the figures?

Copilot

Here is a concise but careful answer that cites only official or semi-official sources and responds directly to the question.

Conclusion (in one sentence)

The AI adoption rate among Hong Kong SMEs is currently around 30% to 50%. In the wording of official policy and statistical reports, this is usually summarised as 'about half'.

Is there any 'official report' to support this? Yes, and more than one.

The two types of official source most often cited in government and policy documents are:

1 'Digital Economy Pulse Report' (government policy level)

Issuing / using body:

- Relevant HKSAR Government policy and research bodies (used to assess the overall digital economy and the level of SME digitalisation).

Concise answer (may be quoted directly)

According to the 'Digital Economy Pulse Report', the AI adoption rate among Hong Kong SMEs is currently 'about half', that is, roughly 45% to 55%.

Why is the phrase 'about half' used?

The 'Digital Economy Pulse Report' itself rarely uses a single, hard 'AI adoption percentage'. Instead, it:

- places AI, automation and data analysis within a basket of indicators for 'digital economy participation' and 'intelligent application'; and
- then estimates, based on whether SMEs have used intelligent tools (for example customer-service chatbots, data analysis, process automation) or are at a trial or early-adoption stage, to reach an overall estimate.

In the official summary and policy interpretation of the report, the core description of SMEs is:

Among SMEs, about half already engage with AI or intelligent applications, but most are still at an early or limited-use stage and are some way from deep integration.

The tool produces a fluent and apparently complete analysis containing specific adoption figures, purported research findings, and a citation to a report presented as an official or authoritative source. Its format resembles a conventional market-research summary and may appear credible.

Does the 'Digital Economy Pulse Report' really exist?

 Copilot

To answer briefly and directly:

Conclusion (very important)

The 'Digital Economy Pulse Report' is not a genuine, officially named Hong Kong Government report.

- ✓ The Hong Kong Government has never published a document formally called the 'Digital Economy Pulse Report'.
- ✓ This name does not appear on news.gov.hk, policyaddress.gov.hk, itib.gov.hk, digitalpolicy.gov.hk or Legislative Council documents.
- ✓ No publication titled 'Digital Economy Pulse' can be found on Chinese or English government websites, in Legislative Council documents or in Policy Addresses.

However, if a user incorporates such content into a management briefing, board paper or external communication without verification and circulates it to business partners, they may be disseminating false information, potentially affecting organisational decision-making and credibility. In fact, the report cited by the AI may not exist at all, and its title, statistics or quotations may not be verifiable through any authoritative source. This often occurs because the AI predicts how a research report would ordinarily be written based on patterns learned from large volumes of existing text, and then generates content that appears plausible but has no genuine factual basis.

Core Principle: AI cannot reliably determine whether its own output is correct. Accurate and inaccurate information may be presented with equal confidence. Responsibility for verification remains with the user and the organisation.

Key Points for Risk Management

Do not rely directly on statistics, research findings, quotations, or third-party views supplied by AI. Verify all figures, quotations, and references independently against official websites, regulatory materials, original research, or other authoritative sources.

Management decisions, compliance matters, public statements, and external communications particularly require appropriate review and verification. Information whose source or accuracy cannot be confirmed should not be presented as fact or used as a basis for decision-making.

3.4 Differences Between Approved AI Tools and Personal AI Subscribed Accounts

Many organisations have established, or are in the process of developing, an approved AI tools list that governs which AI tools staff may use for work purposes. An approved AI tool is one that has been assessed by the IT department or the designated responsible function, taking into account data security, privacy protection, compliance requirements, and technical risks, and has been formally made available to staff through appropriate procurement, licensing, or contractual arrangements.

By contrast, a personal subscribed AI account is a public AI service that a staff member registers for and accesses using a personal email address or personal account. Even where the user pays for the service personally, such an account remains fundamentally different from an enterprise account provided by the organisation. Personal accounts are generally not covered by data protection agreements between the organisation and the AI supplier, and they fall outside the organisation's monitoring, audit, and risk-management arrangements. As a result, they may not meet the organisation's information security and compliance requirements.

How to confirm which AI tools are approved

Staff who need to confirm which AI tools are approved for work may take the following steps:

- Ask the IT department, the HR department, or the designated responsible function for the organisation's approved AI tools list.
- Consult the organisation's internal policies, information security guidance, and relevant work instructions.
- Confirm that the AI tool is formally provided by the organisation and is accessed through an authorised work account.

Where no approved AI tools list has been published

- If the organisation has not yet published an approved AI tools list, staff should adopt a precautionary approach until one is available.
- When processing work data, including customer data, personal data, business information, contract content, or internal documents, staff should use only enterprise-version tools provided and managed by the organisation.
- Do not enter work data into a personal account or an unapproved AI tool.
- Before using a new AI tool for work, submit an application through the organisation's established process to the IT department or the designated responsible function. The tool may be used only after the relevant function has completed its risk assessment and granted formal approval.

Where does the data go? Enterprise versus personal AI tools

Enterprise Version / Approved



Data Protection

A data processing agreement (DPA) has been signed, clearly governing data use, access rights and the division of responsibilities.



Information Security

Encryption, access control and audit mechanisms help support corporate security requirements and internal oversight.



Contractual Protection

The supplier's responsibilities are set out in the contract, including service scope, data processing and incident response.



Account Governance

Company accounts are managed centrally by the enterprise, making it easier to manage permissions, track usage and hand over accounts when staff leave.

Input is protected by contract and corporate policy, and the use of data is clearer and more controllable.

Personal Version / Unapproved



Open Input

The purpose of the input may be unclear. In some cases the content may be used to improve the service or to train models.



Public Cloud

How data is stored, retained and processed may not be transparent, and retention periods or cross-border transfers may be unknown.



Supplier Model

Third-party services and models may change beyond the enterprise's control, and data handling may change as the supplier updates its policies.



Lack of Contractual Protection

There may be no enterprise-level commitment or definition of responsibility, making accountability harder if problems arise.

The subsequent use of the input may not be fully visible, and the enterprise may have limited control over it.

3.5 Writing Safe and Effective Prompts

Before using an AI tool for work, staff should understand the basic principles of prompt drafting. A prompt that is clear, complete, and specific is more likely to produce output that meets practical requirements, while reducing risks relating to data security and content accuracy.

The following five principles help staff improve efficiency while maintaining information security and output quality:

1. **Prioritise context and purpose over raw data.**

When requesting assistance from an AI tool, begin by describing the work context, objective, and type of support required, rather than pasting source text, a complete email, or an entire spreadsheet. In most cases, AI can assist with drafting, structuring, or providing suggestions based on a description alone, without receiving the underlying data.

2. **Specify the required output format.**

State the intended format and length clearly, for example, “list three key points”, “prepare a reply of approximately 150 words”, or “organise the content under four headings”. The more specific the instructions, the easier it is to apply the output directly in a workplace setting.

3. **Request verifiable sources.**

Where facts, figures, research findings, or policy information are involved, ask the tool to identify its sources, for example: “Please list your sources so that I can verify them independently.” For tools with web-search capability, this practice improves transparency and reduces the risk of AI hallucinations.

4. **Request disclosure of assumptions and uncertainties.**

Ask the tool to state the assumptions underlying its response and any matters that remain unconfirmed, for example: “What has not yet been verified?”, or “What assumptions does this recommendation rely on?” This enables the user to understand the limitations of the output and to avoid overreliance on AI-generated content.

5. **Apply the principle of data minimisation.**

Even when using an approved enterprise-version AI tool, provide only the minimum information necessary to complete the task. Before use, delete, de-identify, or replace with generic descriptions any unnecessary personal data, confidential information, or sensitive content, thereby reducing the risk of data leakage and improper use.

3.6 Three Core Principles for Staff Using AI

Sections 3.1 to 3.4 establish three core principles that apply whenever staff use AI.

Core Principle	Description
1. Reliability of AI Output	AI generates responses by analysing language patterns in large volumes of data rather than by retrieving and verifying facts from an authoritative database, even where it appears fluent and well-structured. Its output may therefore be affected by limitations in the timeliness, coverage, and balance of the training data, resulting in content that is inaccurate, incomplete, outdated, or fabricated. Staff shall be particularly aware that AI typically presents its output in a confident and assertive tone regardless of whether the content is correct.
2. Protection of Input Data	Once data is entered into an AI tool, the content is generally transmitted to a third-party supplier's systems for processing. At that point, the data may no longer fall entirely within the organisation's established information security and data protection controls. In the case of free or personal subscribed AI tools, input content may be used for model training or product improvement under the applicable terms of service. Staff must therefore assess the sensitivity of the data and the suitability of the tool before providing any input.
3. User Judgement and Accountability	The primary function of AI is to support efficiency, assist with analysis and facilitate decision-making, rather than to replace professional knowledge, business judgement, or management responsibility. Although AI-generated content is typically produced rapidly, expressed fluently and may appear persuasive, this does not mean the output is accurate, complete, or suitable for a particular context. In practice, users may lower their level of scrutiny because AI-generated content appears reasonable, adopting it without adequate review and verification. This tendency is known as automation bias — the over-reliance on system output that weakens the independent judgement and oversight that should be applied. Staff must therefore exercise professional judgement at all times when using AI, and must review, verify, and assess all output appropriately. Regardless of whether content was generated with AI assistance, the person who sends, submits, approves, publishes, or signs the final content remains responsible for its accuracy, suitability, and compliance.

These three core principles address the reliability of AI output, protection of input data, and user judgement and accountability. Understanding these concepts helps staff develop a more comprehensive awareness of the risks and responsibilities associated with AI use, and supports the adoption of appropriate practices. Part 4 translates these principles into specific, actionable workplace requirements through the four golden rules.

The four golden rules are mandatory requirements for workplace AI use. They are grounded in practical workplace applications, intended to be clear, specific, and capable of consistent application, and reflect relevant Hong Kong laws, regulatory requirements and international best practices.

Together, the four rules provide a framework for secure and responsible AI use. Staff should apply them consistently when using AI to support work, thereby managing risks to data security, privacy, content accuracy and compliance.

Part 4: Four Golden Rules for Secure AI Use

4.1 Rule 1: Do Not Input Sensitive or Personal Data

Golden Rule 1: Do not input sensitive or personal data into an AI tool

Before entering any content into an AI tool, confirm under the organisation's policy that the data may be used with that tool.

Do not use an unapproved free or personal subscribed AI tool. Even when using an approved enterprise-version AI tool, comply with the organisation's acceptable use policy and data classification requirements. However, note that using an enterprise-version tool does not permit unrestricted input of any data.

If there is any uncertainty about whether data may be entered into an AI tool, consult your supervisor or the IT department before use.

Work involving personal data, customer data, or confidential business information may be undertaken only with an approved enterprise-version AI tool subject to appropriate contractual or data processing safeguards. Unapproved free or publicly available AI services should not be used.

This rule is particularly important because information entered into a third-party AI platform outside the organisation's control may be used for model training, stored in systems or jurisdictions that the organisation cannot manage, or disclosed as a result of a data leakage incident affecting the supplier.

Entering customer or employee personal data into a third-party AI tool for a purpose beyond the original purpose of collection or the data subject's reasonable expectations may create compliance risk under Data Protection Principle 3 of the Personal Data (Privacy) Ordinance. It may also breach confidentiality obligations arising from customer commitments, commercial contracts, or confidentiality agreements.

Once information is leaked, it may be difficult or impossible to remove or recover. Even if the disclosure is not discovered immediately, the information may remain in a third-party system and create an ongoing risk for the organisation.

Common Categories of Sensitive or Personal Data Include:

- Customer names, Hong Kong identity card numbers, telephone numbers, residential addresses, and email addresses.
- Financial records, banking details, credit card numbers, and salary information.
- Any person's medical or health information.
- HR data, performance appraisal results, and remuneration arrangements.
- Unpublished business plans, contract content, tender documents, and other trade secrets.
- Passwords, access tokens, or application programming interface (API) keys of any kind.

Particular care is required when using free or publicly available AI tools because, depending on the applicable terms, information entered by a user may be used by the supplier for model training or product improvement.

Data Minimisation: The Less Information Entered, the Lower the Risk

Even when using an approved enterprise-version AI tool, staff should follow the principle of data minimisation and enter only the minimum information required to complete the task.

Where practicable, de-identify information, omit unnecessary data, and use fictitious examples instead of real data.

Examples of De-identification:

- Customer name → “Customer A” or “the customer”.
- Organisation name → “the organisation” or “our organisation”.
- Contract → replace the original with fictitious text of a similar structure.
- Email → summarise the relevant points in your own words without identifying information.
- Figures → use sample figures or placeholders, such as “X”, instead of actual amounts.
- Hong Kong identity card number, telephone number, or address → should not be entered in principle. If a format must be demonstrated, use fictitious data, such as “A123456(7)” or “9XXX-XXXX”.

Practical Advice

- For any work involving customer data, Hong Kong identity card numbers, contact details, financial data, medical information, or confidential business information, use only enterprise-version AI tools that have been approved by the organisation and are subject to appropriate data protection arrangements.
- If approval for using an AI tool could not be confirmed, treat it as unapproved and consult the IT department before use.
- For tasks involving large volumes of sensitive data, highly confidential information, or material used in significant decision-making, consider not only whether an approved tool is available, but also whether AI use is necessary and whether manual processing would better meet risk management and compliance requirements.
- In all cases, apply the principle of data minimisation and do not provide an AI tool with more information than is necessary for the task.

4.2 Rule 2: Verify AI Output Before Use

Golden Rule 2: The user, not the AI tool, is responsible for the final content

AI may assist with drafting, analysis, and organisation, but responsibility for adopting, publishing, submitting, or externally providing such content remains with the user. All AI-generated content should therefore be reviewed and verified before it is incorporated into a work product.

Before adopting AI output, complete the following steps:

1. Review the entire output.
2. Verify each material fact and figure.
3. Check all quotations, citations, sources, and references.
4. Use the output only after confirming that it is accurate, suitable, and responsive to the task.

Specific information generated by AI, including statutory provisions, regulatory requirements, figures, dates, names, reference numbers, quotations, and policy requirements, shall be checked against an original or authoritative source before use. Users should always bear in mind that an AI output's professional tone or authoritative language does not establish that its content is accurate. Presentation and accuracy must be assessed separately.

Using unverified AI-generated information in customer advice, government filings, contracts, management reports, or public communications may cause incorrect decisions, non-compliance, financial loss, reputational damage, legal liability, or customer claims.

In professional fields such as law, accountancy, healthcare, finance, and consultancy, providing customers or clients with incorrect or misleading information may amount to professional negligence and may give rise to professional indemnity claims, disciplinary action, or other regulatory consequences.

AI-generated content is often presented in a fluent, comprehensive, and persuasive manner. The more detailed and professional the content appears, the easier it is to assume that it has already been verified and to overlook the necessary verification procedures. Staff must therefore apply appropriate scrutiny and exercise sound judgement when assessing AI output.

Before using AI output, confirm the following:

- **Are the facts accurate?**
Have all figures, dates, individuals' names, organisation names, and factual statements been checked against reliable and authoritative sources?
- **Are the sources genuine and verifiable?**
Where AI cites statutory provisions, research reports, statistics, or other references, have they been checked against the original documents or other authoritative sources?
- **Does the output meet the task requirements?**
AI may depart from the main issue, confuse concepts, or provide information that is relevant but does not address the actual requirement. Confirm that the output is suitable for its intended purpose and responds directly to the task.
- **Are you prepared to accept responsibility for the content?**

Before submitting, publishing, adopting, or relying on the output, confirm that you have fully understood, reviewed, and approved the content, and are prepared to accept responsibility for it on behalf of yourself or the organisation, as applicable.

4.3 Rule 3: Use Only Approved AI Tools

Golden Rule 3: Use only AI tools approved for work

All work-related AI use should be undertaken through an approved enterprise-version AI tool configured or licensed in the organisation's name and accessed through an authorised work account.

A tool not included on the approved AI tools list should not be used for work data. Where a new tool is required, apply to the IT department through the organisation's established process and do not use it until the relevant risk assessment and approval have been completed.

Using a public AI service through a personal account, a personal subscribed AI tool, or any tool not approved by the organisation for work-related matters is a form of shadow AI.

Shadow AI is the use of an AI tool for work without formal organisational approval or outside the IT department's management and monitoring arrangements. The relevant question is whether the tool appears on the approved AI tools list, not its brand, place of origin, or version.

Shadow AI is a common source of organisational AI risk because the organisation may be unable to maintain effective oversight of how data is transferred and processed, or whether data has left its control. If data leakage, system misuse, or another incident occurs, the organisation may also be unable to intervene promptly. Where third-party personal data is involved, the practice may further impair the organisation's ability to fulfil its obligations under the Personal Data (Privacy) Ordinance and other applicable laws and regulations.

Maintaining an approved AI tools list is a core AI governance control. The IT department will generally assess:

- Whether the supplier may use input data to train or improve its models.
- Where data will be stored, how it will be transmitted and what encryption measures will apply.
- Whether appropriate data processing and protection agreements are in place between the organisation and the supplier.
- The arrangements for audit logs, access controls, and monitoring; and whether the tool complies with the organisation's information security, data privacy, and regulatory requirements.

Workplace Requirements

- Do not use a personal subscribed AI tool for work data, regardless of its functionality. This includes customer data, internal documents, commercially sensitive information, contract content, and drafts of external communications.
- Use only AI tools provided and approved by the organisation, accessed through an authorised work account. Consult the IT department if the permitted scope is unclear.
- If the organisation's existing approved tools do not meet a business need, raise the requirement through formal channels with management or the IT department. The responsible function can then assess whether another enterprise-version tool should be procured, configured, or approved. Do not bypass the organisation's procedures by obtaining a personal subscription.
- Use work devices, email, and accounts only for authorised work purposes. Personal accounts should not be used for organisational work.

4.4 Rule 4: Report Anomalies and Incidents to IT Department Promptly

Golden Rule 4: Report any AI-related anomaly or suspected incident promptly

Report the matter to the IT department or your supervisor as soon as possible. Prompt reporting helps contain risk, reduce the impact, and preserve the organisation's options for remediation.

Although errors may not always be avoidable, delaying or concealing a report can allow the problem to escalate and may lead to more serious consequences.

Report anomalies and incidents to the IT department or your supervisor promptly. Examples include inadvertently entering a customer's personal data into a public AI tool, sending a customer AI-generated content that is subsequently found to be inaccurate, relying on an unauthorised commitment generated or suggested by AI, or observing suspected abnormal behaviour by an AI tool.

Prompt incident reporting enables the IT department and management to take appropriate action, including recalling communications, asking the AI service provider to delete relevant data, assessing under the Personal Data (Privacy) Ordinance whether notification to the Office of the Privacy Commissioner for Personal Data or affected persons is required, and implementing other containment measures. By contrast, delayed reporting may significantly reduce the availability, effectiveness, and practicality of remedial action, allowing an otherwise manageable event to develop into a more serious incident.

Identifying AI-related Incidents

Effective implementation of this rule begins with understanding what may constitute an AI-related incident. Common examples include:

- Inadvertently entering personal data, customer data, or the organisation's confidential information into a public AI tool.
- Discovering that AI-generated content used in an external document, customer deliverable, or official communication contains inaccurate information.
- Becoming aware that a colleague is using an AI tool that has not been approved by the organisation for work.
- Suspecting that sensitive information may have been disclosed or leaked through the use of an AI tool.

Workplace Requirements

- Report a suspected AI-related incident immediately. Staff are not required to determine its severity before reporting.
- Management and the relevant specialist function are responsible for risk assessment and classification. Staff shall provide accurate and complete information at the earliest opportunity.

Report the Matter to the Appropriate Function

Report incidents involving personal data or data privacy to your supervisor and the designated responsible function. Report system, account, or technical anomalies to the IT department. Report issues involving external communications, commitments, or service delivery to your supervisor and follow the organisation's established procedures.

Provide Complete and Accurate Information

When reporting an incident, provide the date and time of the incident, the AI tool used, the data or content involved, any action already taken, and the known or potential scope of the impact. Do not omit information or present facts selectively because of concerns about personal responsibility.

Cooperate with the Investigation and Response

After reporting an incident, follow the organisation's established incident-handling procedures and cooperate with any investigation or follow-up action undertaken by the relevant functions. Unless authorised, do not contact affected persons, make external statements, or attempt to resolve the incident independently. Doing so may increase the risk or interfere with the organisation's response.

Importance of Prompt Reporting and Incident Containment

The purpose of prompt reporting is not to assign blame. It is to enable personnel with the appropriate expertise and authority to intervene quickly, contain the risks and minimise potential loss. Timely reporting allows the organisation to implement effective remedial measures and limit the impact of an incident. By contrast, delaying or concealing a report may increase the risks, costs, and complexity of the response.

Part 5: Workplace Applications of AI

This Part explains how the guidance applies to common workplace uses of AI. It addresses three principal scenarios. Additional examples are set out in Appendix II.

Each use case identifies the types of information that may and may not be entered into an AI tool. The examples reflect common workplace scenarios in small and medium enterprises, schools, and non-governmental organisations, including communications with customers, parents, donors, and other stakeholders, and the management of student and service-user cases. They are intended to help staff use AI safely and responsibly in different workplace settings.

Six common AI use scenarios in daily work

For each task, a green tick means safe use and a red cross means unsafe use.

 Email drafting ✓ Using an enterprise-approved tool to help refine tone and draft content. ✗ Entering client emails, account numbers or other identifiable data.	 Document summarisation ✓ Generating summaries of public documents or materials approved for sharing. ✗ Summarising confidential contracts, internal documents or non-public information.	 Content translation ✓ Translating marketing materials or general public content. ✗ Translating named contracts, HR records or other sensitive documents.
 Data analysis ✓ Analysing de-identified, aggregated or public data. ✗ Entering client lists, payroll data or other sensitive data.	 Information research ✓ Helping to organise research directions and cross-checking the original sources. ✗ Treating AI-generated content directly as a reliable source of information.	 Client communication ✓ Drafting replies for staff to review and approve before sending. ✗ Automatically sending unreviewed messages to clients, or entering real client data.

The three principal scenarios are set out below. For additional examples, refer to Appendix II, “Further examples of workplace AI use”.

5.1 Using AI to Draft Emails

AI tools may assist staff in refining wording, adjusting tone and drafting responses to routine enquiries, thereby improving the clarity and professionalism of emails. However, email correspondence often contains personal data, business information, or other identifying details. Entering original correspondence into an AI tool may create data leakage and raise compliance risks.

The following practices apply where a staff member uses AI to draft a response to a major donor's complaint about delayed receipt confirmation:

Recommended Practices	Practices to Avoid
• Describe the situation in general terms. For example: "Draft a professional and courteous response to a customer complaint." • Remove all names, contact details, and other identifying information before submitting content to the AI tool. • Ask AI to suggest improvements to tone, wording, structure, and clarity. • Verify all facts, arrangements, and commitments before adopting the draft. • Disclose AI assistance in preparing the content where required by organisational policy or other applicable requirements.	• Do not enter the entire message along with the sender's name, organisation, email address, or other contact details. • Do not enter customer account numbers, reference numbers, case numbers, or other identifiers. • Do not provide confidential business information, pricing arrangements, contract terms, or unpublished commitments. • Do not send an AI-generated response without first reviewing and verifying it. • Do not use a personal AI account or an unapproved AI tool to process work emails.

5.2 Using AI for Translation

AI tools may assist with translating public information, internal notices, and routine administrative documents. However, AI-generated translations of documents with legal, contractual, regulatory, or other material effect should not be relied upon without appropriate review. Before use, the translation should be reviewed and confirmed by a suitably qualified person who is responsible for its accuracy.

In addition to translation quality, staff should consider data protection and confidentiality requirements. Information that is unsuitable for processing by the relevant AI tool should not be entered.

The following practices apply when using AI to translate a twelve-page Chinese legal opinion into English for an overseas client:

Recommended Practices	Practices to Avoid
• Use AI primarily to translate public materials, internal notices, general correspondence, and other low-risk content.	• Do not use a free, personal subscribed, or unapproved AI tool to translate contracts or documents containing personal data, identifying information, confidential

Recommended Practices	Practices to Avoid
• If a document contains personal data, third-party information, or internal information, use only an approved enterprise-version AI tool and comply with the organisation's data classification and data minimisation requirements. • Ensure that an appropriately qualified person reviews and confirms any translation intended for external publication, involving legal, contractual, regulatory, or other material matters. • Use the organisation's approved terminology, glossary, and standard translations consistently. • Verify all proper nouns, legal terms, figures, dates, and other material details before using the translation.	business information, or confidentiality provisions. • Do not enter human resources records, medical information, Hong Kong identity card numbers, or other personal data into an unsuitable AI tool. • Do not assume that an AI-generated translation is accurate without review. Do not release externally any translation affecting legal liability, regulatory requirements or contractual rights, and obligations without appropriate verification. • Do not enter an entire confidential document into an AI tool without first assessing the data protection and confidentiality risks.

5.3 Using AI for Data Analysis

AI tools may assist staff in organising data, preparing summaries, identifying trends, and generating preliminary charts. When using AI for data analysis, staff shall address two principal risks:

First, the input data may contain personal data, customer data, or other sensitive information.

Second, AI-generated calculations, classifications, analytical conclusions, or charts may be inaccurate and must be verified before use.

The following practices apply when using AI to analyse customer purchasing behaviour from the previous year to support planning for the next quarter:

Recommended Practices	Practices to Avoid
• Provide AI only with data that has been de-identified or aggregated. • Independently verify all material figures, classifications, and conclusions. • Connect data sources only through tools and accounts approved by the organisation. • Retain an unaltered copy of the source data to support subsequent review, re-performance, and investigation of discrepancies. • Obtain appropriate further review and approval where the results will support management reporting, external communications, or significant decisions.	• Do not upload customer lists, account information, payroll data, or other identifying information to an unapproved AI tool. • Do not submit AI-generated calculations, analyses, or recommendations without independent verification. • Do not use an AI-generated chart without checking its axes, calculations, classifications, legend, and data sources. • Do not use AI for formal filings, financial analysis, or other regulated calculations without completing the required review.

Recommended Practices	Practices to Avoid
	Do not treat AI-generated analysis as a final conclusion without validating it against the business context, source data, and professional judgement.

5.4 Recent Significant Incidents Relating to AI Use

The incidents below illustrate how risks relating to data security, content accuracy, and accountability have arisen in workplace settings. Although AI may improve efficiency, organisations, and staff shall maintain appropriate controls and oversight.

Source-code data leakage at a technology company (2023)

Engineers in the semiconductor division of a technology company reportedly used ChatGPT to assist with source-code-related tasks and entered confidential information into the external AI tool. Following the incident, the company imposed further restrictions on certain employees' use of external generative AI services. The incident demonstrates that, even where AI is used to improve workplace efficiency, entering confidential data, commercially sensitive information, or unpublished content into a tool that has not been approved by the organisation may create significant data leakage and confidentiality risks.

United States lawyer citing AI-fabricated cases (2023)

A New York lawyer used ChatGPT to assist with legal research while preparing documents for civil litigation and incorporated some of the results into documents submitted to the court. The court subsequently found that several of the cited cases did not exist and had been fabricated by AI. The incident serves as a reminder that AI-generated information, citations, cases, research findings, and references may be inaccurate or entirely fabricated. All factual information and references supplied by AI shall be verified independently against reliable and authoritative sources before use.

Canadian airline chatbot incident (2024)

An airline used an AI chatbot on its official website to respond to passenger enquiries about its bereavement fare policy. A passenger subsequently relied on information provided by the chatbot, which did not reflect the airline's actual policy. The dispute resulted in a court holding the airline responsible and requiring it to refund the fare difference. The incident demonstrates that organisations may remain responsible for inaccurate information communicated externally through AI tools. AI applications used to communicate with customers, the public, or other external stakeholders should therefore be subject to appropriate human oversight and review.

Data exposure at a Chinese Mainland AI company (2025)

In early 2025, security researchers reportedly identified an internet-accessible ClickHouse database operated by a Chinese Mainland AI company. The database contained records including user conversation logs, application programming interface (API) keys, and back-end system information. The incident illustrates that a supplier's information security and data protection arrangements can affect the security of user data even where the AI service itself appears to operate normally. When selecting an AI service, organisations should assess not only its functionality and performance, but also the supplier's information security standards, data protection arrangements, compliance capabilities, and incident-response mechanisms.

Part 6: Other AI-Related Risks

AI-related risks may arise not only from staff use, but also from the characteristics of the technology, supplier arrangements, external cyber threats, and applicable legal, compliance, and ethical requirements. Staff should understand these risks and apply appropriate preventive measures to protect organisational data, maintain business operations, and reduce potential exposure.

6.1 Copyright and Intellectual Property Considerations

AI-generated content may raise copyright, intellectual property, and permitted-use issues. Because ownership and liability may be uncertain, AI output should not be published or delivered externally as a final product without appropriate review. Staff should use AI to support ideation and drafting, not as the sole source of final content.

Practical Implications for the Workplace:

- Treat AI-generated content as a draft, reference, or starting point, not as material ready for publication or delivery.
- For external publication or delivery, use licensed materials, the organisation's own content, or human-authored and approved works wherever practicable.
- Where AI has made a material contribution, ensure that an appropriate person reviews and approves the final content before publication or delivery.
- Disclose AI assistance where required by organisational policy, customer requirements, contractual terms, or professional standards.

6.2 AI Image Generation and Deepfake Risks

AI can generate highly realistic images, audio, and video. Such content may be mistaken for genuine material and used in phishing, impersonation, or social engineering attacks.

Examples include impersonating management to request urgent payment, impersonating human resources to obtain personal data, or posing as a supplier to change payment instructions.

Staff should exercise heightened caution when receiving urgent, confidential, or exceptional requests by voice, video, or image. Verify each request through an independent channel, particularly where it concerns funds, sensitive data, contracts, or other significant business matters.

Do not bypass established verification procedures because of time pressure or asserted urgency.

If deepfake content or another suspicious communication is suspected, cease the interaction and report it promptly under Golden Rule 4. Unless organisational procedures require otherwise, do not challenge the sender or disclose information about the investigation.

Practical Implications for the Workplace:

- For images published or delivered externally, use licensed stock images, the organisation's own materials, or professionally commissioned photographs wherever practicable.
- Before using AI to generate an image, voice recording, or video of a named individual, obtain the necessary authorisation and approval in accordance with the organisation's procedures.
- If a voice message, video, or video call has an unusual tone, content, or deadline, particularly where it requires immediate action, involves funds, or concerns confidential matters, verify it through an independent channel before acting.
- Treat voice or video content requiring an immediate response as only one source of information. For any significant decision, obtain final confirmation through a telephone call, email, or other formal channel initiated independently.

6.3 Prompt Injection Attacks

A prompt injection attack occurs when hidden instructions are embedded in a document, webpage, email, image, or other content to influence an AI tool's behaviour or output.

The AI tool may follow the hidden instructions instead of the user's request, produce misleading output or perform unauthorised actions. The risk is greater where the tool can access other systems, documents, or data sources.

Staff should not assume that external content is safe or trustworthy, and should remain alert even where it is received from known customers or partners or obtained from public websites.

Practical Implications for the Workplace:

- Exercise caution when providing external content to AI for processing, including customer documents, third-party website content, supplier data, and uploaded images. Do not assume that such content is safe or trustworthy.
- If an AI tool departs from the original instructions, produces abnormal results, or requests sensitive information, such as login credentials or access rights, stop using it immediately and notify IT.
- Use only enterprise-version AI tools and accounts approved by the organisation.
- Do not submit confidential data or sensitive business information to an AI tool along with external content of unknown or suspicious origin.

6.4 Risk of Bias in AI Output

AI outputs may reflect patterns and perspectives present in the data on which the AI was trained. Where the training data contains bias, imbalance, or historical limitations, AI-generated content may perpetuate or amplify those biases. Extra care is therefore required when AI is used in contexts involving individual rights and interests, access to opportunities, or decisions affecting how individuals are treated.

Relevant contexts include recruitment, promotion, performance management, and decisions affecting customers, students, beneficiaries, or staff. AI may support these activities, but authorised personnel should retain professional judgement, management responsibility, and final decision-making authority.

Practical Implications for the Workplace:

- Do not make a final decision about an individual solely on the basis of AI output.
- AI may assist with drafting job descriptions, summarising key points from résumés, generating interview questions, or providing analytical support. Authorised staff should make all substantive decisions, including candidate shortlisting, ranking, appointment, and promotion, in accordance with established procedures.
- Review AI output intended for customers, students, or beneficiaries for biased, unfair or potentially misleading language, assumptions, examples, or recommendations.
- For any decision affecting an individual's rights and interests, treatment, or development opportunities, retain a record of the basis for the final human judgement to ensure appropriate transparency and accountability.

6.5 AI Agents: From Content Generation to Autonomous Task Execution

An AI agent may be authorised to perform tasks on a user's behalf, including sending emails, retrieving system data, updating records, organising documents, and interacting with internal platforms or business systems.

Unlike a general AI assistant, which mainly provides suggestions or responses, an AI agent can perform actual operations. Therefore, before deploying or using an AI agent, staff should understand its functions, data access, permissions, and authorised use cases.

Incorrectly executed or poorly configured AI agent operations may result in misdirected emails, inappropriate changes to document permissions, improper modification of data, erroneous payments, or other adverse business consequences. High-risk operations involving funds, external communications, data deletion, approvals, or access management require human review and appropriate record-keeping.

In addition, some AI agents may be authorised to access email, calendars, cloud-based documents, collaboration platforms, or other internal systems. If access permissions are configured too broadly, an AI agent may access data beyond what is necessary for its assigned task, thereby increasing data protection and information security risks. Before enabling an AI agent, the organisation should therefore verify that its access permissions comply with the principle of least privilege and grant only the minimum permissions necessary to perform the designated task.

Practical Implications for the Workplace:

- Before enabling an automated execution function, confirm that it has been approved by the organisation and identify the data and systems that it can access.
- Require appropriate human review and confirmation before any high-risk operation involving funds, external communications, approval processes, or data deletion is carried out.
- Configure each AI agent's access permissions in accordance with the principle of least privilege. Do not grant access to email, documents, calendars, or systems beyond what the authorised task requires.
- If abnormal behaviour, unauthorised operations, or potential security issues are identified, disable the relevant function immediately and notify the IT department or the designated responsible officer in accordance with established procedures.

Part 7: Practical Guidance for Secure AI Use

7.1 Recommended Practices and Practices to Avoid

This Part may be printed as a quick-reference guide or shared through team platforms and internal communication channels, enabling staff to consult it readily at work. The table summarises the principal requirements for workplace use of AI and supports prompt and consistent decision-making.

Task	Recommended Practices	Practices to Avoid
Drafting Emails	Describe the workplace situation in general terms. Use AI to improve tone, structure, or wording, and use only enterprise-version AI tools approved by the organisation.	Do not enter an incoming message into an AI tool along with the sender's name, contact details, account data, or other identifying information. Do not use a personal AI account to process work emails.
Summarising Documents	Summarise only documents that are public or contain no sensitive data. Read and check the source document before adopting the summary.	Do not upload confidential documents, contracts, or documents containing personal data. Do not rely solely on an AI-generated summary without consulting the source document.
Translation	Use AI to assist with translating general and low-risk content. Ensure that an appropriate person reviews any translation intended for external publication, involving legal, contractual, or regulated content.	Do not use a free, personal subscribed, or unapproved AI tool to translate documents containing personal data, identifying information, contractual content, human resources records, or other sensitive information.
Data Analysis	Provide AI only with data that has been de-identified or aggregated. Verify all key figures, charts, and conclusions before adopting the analytical results.	Do not upload customer lists, payroll data, or other identifying information. Do not adopt AI-generated charts or analytical conclusions without verification.
Research	Use AI to understand background information, organise research topics, and develop a preliminary framework. Consult and verify original sources before citing any information.	Do not treat AI as an authoritative or original source. Do not adopt statistics, dates, statutory provisions, regulatory requirements, or other information without verification.

Task	Recommended Practices	Practices to Avoid
Communicating with Customers	Use AI to assist with preparing an initial draft. A staff member should review and confirm every external message before it is sent.	Do not allow AI to respond directly to customers without human review. Do not enter genuine customer messages, account data, or unpublished commitments into an AI tool.

7.2 Secure AI Use Checklist

Complete the following checklist before, during, and after each use of an AI tool. If the answer to any question is “No”, suspend use of the tool or the relevant output and seek guidance from your supervisor, IT, or the designated responsible officer. Proceed only after the issue has been resolved.

#	Question	Answer
Before Using AI		
A1	Is the user accessing an approved enterprise-version AI tool through a work account?	Yes / No
A2	Is the proposed use within the organisation’s permitted scope for AI?	Yes / No
A3	Does the user understand how the AI tool and its supplier will process the input data?	Yes / No
A4	If the task involves legislation, contractual terms, professional standards, or regulatory requirements, has the use of AI been authorised?	Yes / No
While Using AI		
B1	Has the task been described in general terms without entering personal data, customer data, confidential information, or other sensitive information?	Yes / No
B2	Has the proposed input been checked to confirm that it contains no unnecessary, sensitive, or protected information?	Yes / No
B3	Is AI being used solely as a supporting tool, without replacing professional judgement, decision-making, or approval procedures?	Yes / No
B4	If required by organisational policy, have the prompts and outputs been retained for audit or future reference?	Yes / No
B5	Has the need to verify the content against current, reliable, and authoritative sources been assessed?	Yes / No
After Using AI		

#	Question	Answer
C1	Has all AI-generated content been read and reviewed in full?	Yes / No
C2	Have all facts, figures, dates, quotations, and references been verified against original or authoritative sources?	Yes / No
C3	Does the output meet the original purpose and practical requirements of the task, without irrelevant content, misinterpretations, or material that was not requested?	Yes / No
C4	Is the responsible user prepared to confirm the final output and accept responsibility for it?	Yes / No
C5	If required by organisational policy, contract, or other applicable requirements, has the use of AI been disclosed appropriately?	Yes / No

Interpreting the Results and Follow-Up

If the answer to any question in A1 to A4 is “No”, suspend use of the AI tool until the applicable requirements and risks have been clarified. Consult your supervisor, IT, or the designated responsible officer where necessary.

If the answer to any question in B1 to C5 is “No”, make any required revisions or additions and complete the necessary verification and review before adopting, submitting, sending, or publishing the output.

Part 8: Follow-Up Actions and Support

8.1 Understanding the Organisation's AI Governance Requirements

Staff can consult their supervisor, the IT department, or the designated responsible officer for guidance on AI use at the workplace regarding the following matters:

- The organisation's approved AI tools list.
- The organisation's AI policy and acceptable use rules.
- Incident reporting procedures and contact channels for AI-related incidents, data leakage, and anomalies.
- Additional guidance for particular roles or functions, including customer service, human resources, compliance, finance, and other regulated activities.

8.2 Support Services and Reference Resources

Organisations seeking support with AI governance, AI strategy, or staff capability development may refer to the following resources:

- **Secure AI@Work Enablement Campaign**
Provides AI training and practical resources for organisations.
<https://cyberhub.hk/en/ai/campaign>
- **Digital Policy Office (DPO)**
A policy office under the Government of the Hong Kong Special Administrative Region, responsible for promoting the development of digital government and the digital economy.
<https://www.digitalpolicy.gov.hk/en/>
- **Office of the Privacy Commissioner for Personal Data (PCPD)**
Provides guidance and resources on personal data protection, data governance, and AI privacy management.
<https://www.pcpd.org.hk/index.html>

Thank you for completing the all-staff training: Safe and Responsible Use of AI at Work.

AI is rapidly becoming an essential tool in the modern workplace. Its full potential can be realised only when it is used securely, responsibly, and in compliance with organisational requirements, safeguarding data, protecting privacy, and upholding the organisation's reputation.

Promoting the secure and responsible use of AI is not solely a matter of technology or policy. It depends on the careful judgement and professional practice of every staff member in their day-to-day work.

Secure AI use starts with you. Use AI wisely, exercise sound judgement, and play your part in protecting the organisation's information security and the trust placed in it.

Appendix I: Glossary

This Appendix defines the principal technical and governance terms used in the booklet. English technical terms are included to support cross-reference with policies, contracts, and other guidance.

Term	Explanation
AI agent	An AI tool that can automatically execute specific workflows according to preset objectives, rules or authorisation, such as sending emails, submitting applications, querying system data or handling payments. Unlike general AI tools that only provide recommendations or respond to questions, an AI agent can perform actual operations on behalf of the user within the scope of authorisation.
AI hallucination	The generation by AI of inaccurate, unverifiable or entirely fabricated content presented in a manner that appears reasonable and authoritative. AI hallucination is not necessarily a system error, it is an inherent risk associated with the operating mechanisms of generative AI.
Approved AI tools list	A list of AI tools that staff may use for work after the organisation has completed a risk assessment and granted formal approval. The list is ordinarily maintained and updated by a designated responsible officer, the IT department or the relevant governance function.
De-identification	The removal or replacement, before an AI tool is used, of information capable of identifying an individual, such as names, Hong Kong identity card numbers, email addresses, telephone numbers, account numbers or case numbers. De-identification reduces data leakage and privacy risks.
Deepfake	Audio, video or images generated, modified or synthesised using AI technology with sufficient realism to cause a person to believe that the content is genuine. Deepfakes may be used for fraud, impersonation or the dissemination of false information.
Enterprise-version AI tools	Versions of AI tools procured, licensed or configured by the organisation under commercial contracts. They ordinarily provide more comprehensive data protection, permission management, audit functions, and technical support than free or personal subscribed AI services.
Generative AI	Artificial intelligence technology that generates text, images, audio, video, code, or other content in response to user instructions. It operates principally by identifying patterns in training data and predicting likely output.
Non-governmental organisation (NGO)	An organisation outside government that performs social service, educational, charitable, or other non-profit functions. In Hong Kong, schools, social-service organisations, and charities may fall

	within this category and remain subject to applicable laws and internal AI use requirements.
Personal Data (Privacy) Ordinance (PDPO)	Hong Kong's principal personal data protection legislation, namely Chapter 486 of the Laws of Hong Kong, which regulates how organisations collect, use, retain, disclose, and protect personal data, and prescribes the Data Protection Principles and the related rights of data subjects with which data users should comply.
Personal data	Data relating directly or indirectly to an identifiable living individual, as defined in the Personal Data (Privacy) Ordinance. Examples include names, contact details, Hong Kong identity card numbers, and financial information.
Privacy by design	The principle of incorporating personal data protection requirements into the system or service design, from the project planning stage and throughout the entire lifecycle of the AI system or application, so as to reduce privacy risks and enhance compliance.
Prompt	An instruction, question, or explanation entered into an AI tool to specify the context, purpose, requirements, constraints, and intended output.
Prompt injection	If an attacker embeds hidden instructions in documents, websites, images or other AI-readable content to induce AI to deviate from its original task and perform unauthorised operations or disclose data. This type of risk is listed as an important threat in the OWASP risk classification for large language models (LLM01).
Retrieval-augmented generation (RAG)	A technique combining information retrieval with generative AI. Before generating a response, the system retrieves information from specified sources, such as an organisational knowledge base, document system, or the internet, and uses that information to improve the relevance and accuracy of the response.
Shadow AI	The use by staff of AI tools for work without the organisation's formal approval, oversight or assessment. Shadow AI may result in uncontrolled data flows and create information security, privacy, copyright and compliance risks.

Appendix II: Further Examples of Workplace AI Use

Using AI to Draft and Summarise Documents

AI is particularly suitable for assisting with summarising documents, identifying key points, preparing template text and improving structure. However, care is required where a document contains personal data, confidential information, or other sensitive material. AI-generated summaries may omit important terms, limitations, or risks and should not replace review of the source document.

Suppose a staff member receives a 40-page logistics services agreement submitted by a new supplier and needs to prepare a summary of the key terms for management to assist in evaluating whether to enter into the agreement.

The following practices may be observed when using AI to assist with document processing:

Recommended Practices	Practices to Avoid
• Use AI to draft public-facing material such as marketing copy, frequently asked questions, and general guidance. • Summarise only documents that are public, authorised for use, or free from personal and sensitive information. • Review the source document before making any decision or taking action based on the summary. • Use AI to assist with structure, outlines, and headings, not to replace professional judgement or decision-making.	• Do not upload confidential contracts, customer deliverables, or regulatory reports to a free or unapproved AI tool. • Do not upload employee performance records, identifying financial information, or human resources documents. • Do not use an AI summary instead of reviewing the source document. • Do not rely on an AI summary of legal or compliance material without reviewing the material terms and risks.

Using AI for Information Gathering and Research

When approaching an unfamiliar subject, AI can help staff quickly develop a foundational understanding, for example, learning about the regulatory landscape of a new industry, gaining an overview of a prospective client's business, or grasping the basics of an emerging technology. AI improves the efficiency of information gathering and broadens the scope of initial inquiry, but it should not be treated as the sole or definitive source of information. Before making any professional judgement, recommendation, or decision, staff should verify the accuracy and completeness of AI-generated content against original materials and other reliable sources.

Suppose a staff member is preparing for a meeting with the management of a large retail group and needs to present the organisation's digital transformation advisory services.

The following practices may be observed when using AI to gather background information:

Recommended Practices	Practices to Avoid
• Use AI to identify background concepts, trends, and issues for further research. • Verify facts, data, and views against the relevant original sources. • Confirm information against reliable, authoritative, or official sources before use. • Pay attention to the temporal limitations and possible knowledge cut-off of the AI training data; verify recent developments, current policies, and newly released data independently.	• Do not cite an AI chatbot as an authoritative or original source. • Do not use AI-generated statistics, dates, cases, or market information without checking the original source. • Do not use AI as the sole research source for legal, regulatory, medical, or other professional matters. • Do not use AI-generated content as a substitute for professional judgement, management decisions, or expert advice.

Using AI to Support Customer Communications

AI can assist with drafting standard responses, frequently asked questions, and routine external communications, thereby improving efficiency and reducing preparation time. However, AI may not understand the organisation's business policies, customer commitments, scope of services, or compliance requirements. It may also introduce commitments that have not been authorised, use inappropriate language, or depart from the organisation's established communication style. All AI-generated external communications shall therefore be reviewed and approved by appropriately authorised staff before they are sent.

Suppose a staff member receives an enquiry from a long-standing corporate customer asking whether a discount will be available when its contract is renewed the following year.

When using AI to help draft the response, the staff member may follow the practices below:

Recommended Practices	Practices to Avoid
• Use AI to prepare an initial draft for routine enquiries and standard communications. • Ensure that authorised staff review and approve every customer communication before it is sent. • Check that the draft complies with the organisation's external communication policies and procedures. • Apply enhanced review where the communication concerns a regulated or professional sector, including financial services, healthcare, law, or education.	• Do not allow AI to send a customer response without human review and approval. • Do not enter customer names, contact details, account information, or other actual personal data into a free or unapproved AI tool. • Do not assume that the AI tool understands internal policies, service terms, authority limits, or workflows. • Do not use output without verification where it may contain unauthorised commitments, discounts, product features, or commercial terms.

Appendix III: References and Methodology

This booklet is prepared with reference to the authoritative documents listed below. The guidance reflects relevant Hong Kong legal and workplace requirements and international best practices concerning the responsible use of AI. The concepts have been incorporated into each part of the booklet.

Source	Document	Main contribution to this booklet
Digital Policy Office	Ethical Artificial Intelligence Framework Hong Kong Generative AI Technology and Application Guideline	These documents provide AI ethical principles, governance frameworks, AI lifecycle management and impact assessment guidance applicable in Hong Kong. They form a principal reference for organisational AI governance and risk management. This booklet does not reproduce them verbatim, but consolidates relevant laws, standards, and practical guidance into a framework for Hong Kong enterprises, small and medium enterprises, schools, and non-governmental organisations, supporting management in implementing the responsible use of AI.
Office of the Privacy Commissioner for Personal Data, Hong Kong	Artificial Intelligence: Model Personal Data Protection Framework Checklist on Guidelines for the Use of Generative AI by Employees	These documents explain organisations' responsibilities and best practices when processing personal data in AI application scenarios, including data classification, data processing agreements, human oversight mechanisms, staff use standards, stakeholder communication, and incident management arrangements.
Faculty of Law, The University of Hong Kong	Guidelines on the Responsible Use of Generative Artificial Intelligence for Research	This emphasises that users shall bear the ultimate responsibility for AI-generated content, including verifying the output, disclosing the use of AI, identifying hallucination risks, and properly keeping records of use.
International Organisation for Standardisation (ISO)	ISO/IEC 42001:2023, Artificial Intelligence Management System	This provides the international standard requirements for an AI management system, covering the accountability structure, the division of roles, objective setting, risk management, monitoring, and continual improvement, which can serve as an important reference for the organisation to assess and improve the maturity of its AI governance.

Open Web Application Security Project (OWASP)	OWASP Top 10 for LLM Applications, v2.0	This provides a classification of the main security risks of large language model applications, including prompt injection, sensitive data leakage, excessive authorisation, and insecure plug-in integration, helping management understand the key security controls for generative AI tools and AI assistants.
---	--	---