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

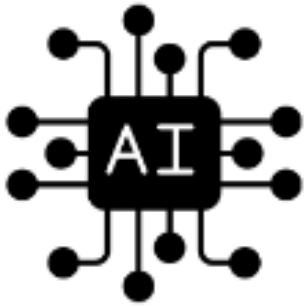
「人工智能與私隱保障：發展與安全並重」研討會

個人資料私隱專員
鍾麗玲女士

2024年7月30日

人工智能(AI)
簡介

定義



- 沒有通用的定義
- 泛指一系列**模仿人類智能**及**以電腦程式和機器**透過所輸入的數據**執行解難、提供建議和預測、作出決策**及**生成內容**等工作（或將其自動化）的**科技**

特定用例

- 自動化流程
- 分析數據
- 加強其決策能力
- 生成內容

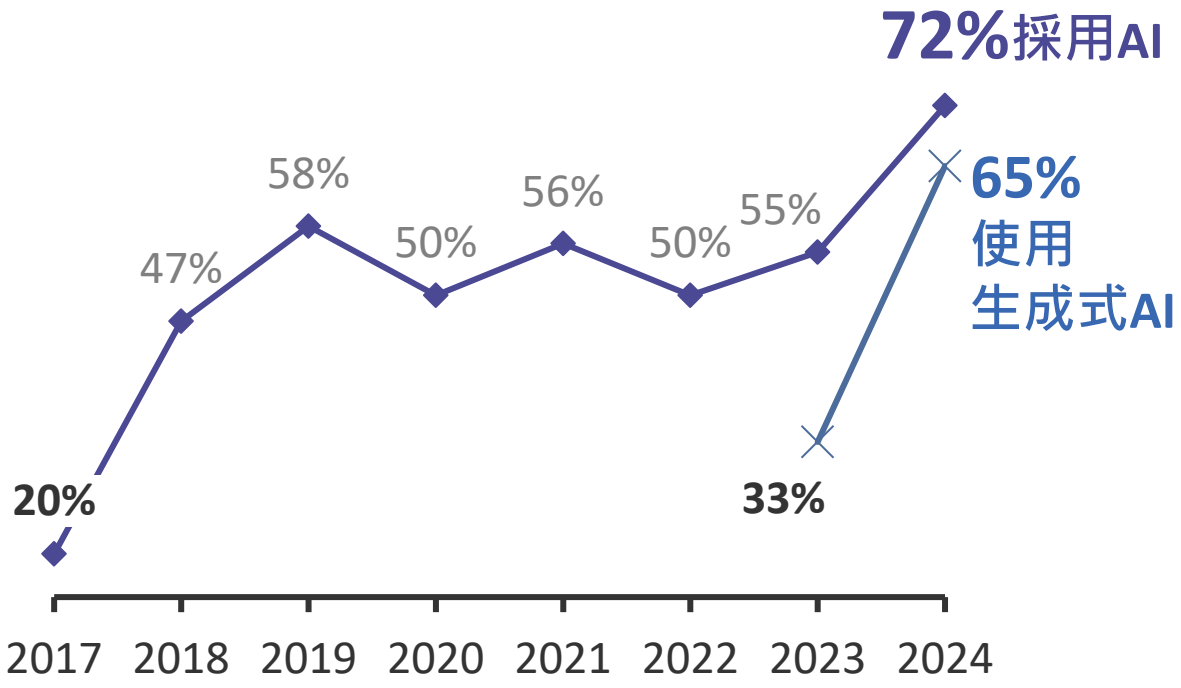
例子

- 處理保險索償文件
- 社交媒體內容個人化建議
- 對機構或個人進行信用評分
- AI聊天機械人

趨勢
機構正積極採用 AI

全球機構AI(包括生成式AI)採用率於2024年大幅上升

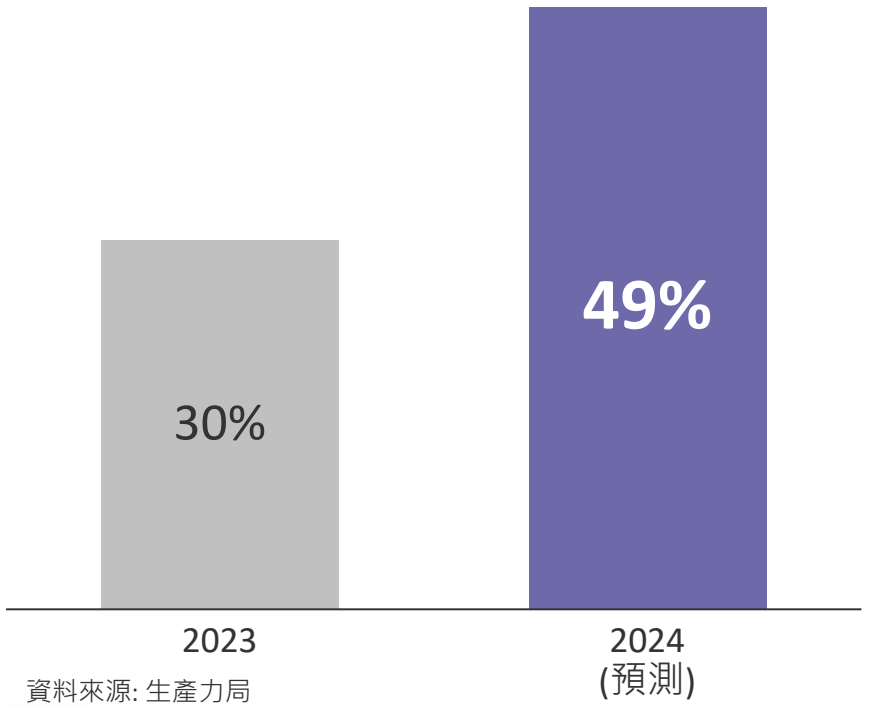
表示至少在一個商業功能上採用AI的受訪機構比例



資料來源: McKinsey

近一半香港機構將於今年使用AI

AI 採用率
香港



資料來源: 生產力局

風險
AI 可能對個人資料私隱構成多重風險

風險	解釋	例子
資料外洩	AI系統（如AI聊天機械人）可能**儲存大量用戶對話資料**，容易成為**黑客攻擊的目標**，引致資料外洩	2023年3月，**ChatGPT**發生了一次嚴重的**資料外洩事故**，洩露了部分用戶過往**對話的標題、用戶的姓名、電郵地址和信用卡號碼的最後四位數字**
資料收集過量	AI 傾向於**收集和保留盡可能多的數據**，包括個人資料	**OpenAI被指擷取了3000億個網上詞彙**來訓練ChatGPT
資料的使用	**「黑盒」難題**：AI使用者**無法得知AI系統的內部運作邏輯**，機構可能也難以理解用家的個人資料將如何被AI模型使用	有**AI模型能在醫療影像沒提及病人的種族的情況下也準確推測到病人的種族**，箇中原因是資深醫生也無法理解
資料準確性	訓練AI模型需要大量數據，若資料的**質素和準確性未必理想**，AI可能因數據不準確而做出**錯誤分析**，從而**影響決策制定**	有跨國科技公司用於**審視應徵者履歷的AI系統**，因AI系統的**訓練數據缺乏女性應徵者資料**，不**當地傾向挑選男性應徵者**，對女性應徵者造成**不公**

資料來源: Fortune; 刺針, Patterns; MIT Technology Review

危機
機構明白私隱風險存在，但認知不足

🏛香港企業：生成式AI在新興技術中存在最高的私隱風險

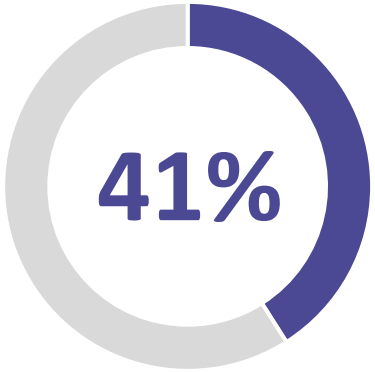
新興技術中的私隱風險
香港企業， 2023

1. **生成式AI**
2. Cookies 和其他線上追踪器科技
3. 雲端計算
4. 物聯網
5. 區塊鏈相關技術
6. 數據分析及營運流程自動化

資料來源: 私隱專員公署和生產力局

😟不少企業沒有就生成式AI提供指引

已就使用生成式AI提供內部指引企業百分比
2023



資料來源:私隱專員公署和生產力局

《人工智能(AI): 個人資料保障模範框架》

✨ 特點

體現國家的《全球人工智能治理倡議》

人工智能安全是國家安全的重點領域之一

向採購、實施及使用任何種類的AI系統(包括生成式AI)的機構，就保障個人資料私隱方面提供有關AI管治的建議及最佳行事常規

✅好處

協助機構遵從《個人資料（私隱）條例》的規定

孕育AI在香港的健康發展

促進香港成為創新科技樞紐

推動香港以至大灣區的數字經濟發展

支持機構、諮詢及參考資料

1

支持機構

- 政府資訊科技總監辦公室
- 香港應用科技研究院

2

諮詢

- 私隱專員公署科技發展常務委員會
- 公營機構
- 科技業界
- 大學
- AI供應商

3

國際參考資料

- 國際機構、政府機構及其他資料保障機構的指引或刊物
- 相關專業界別報告

國際標準

《模範框架》反映國際間認受的原則及最佳行事常規

三項數據管理價值

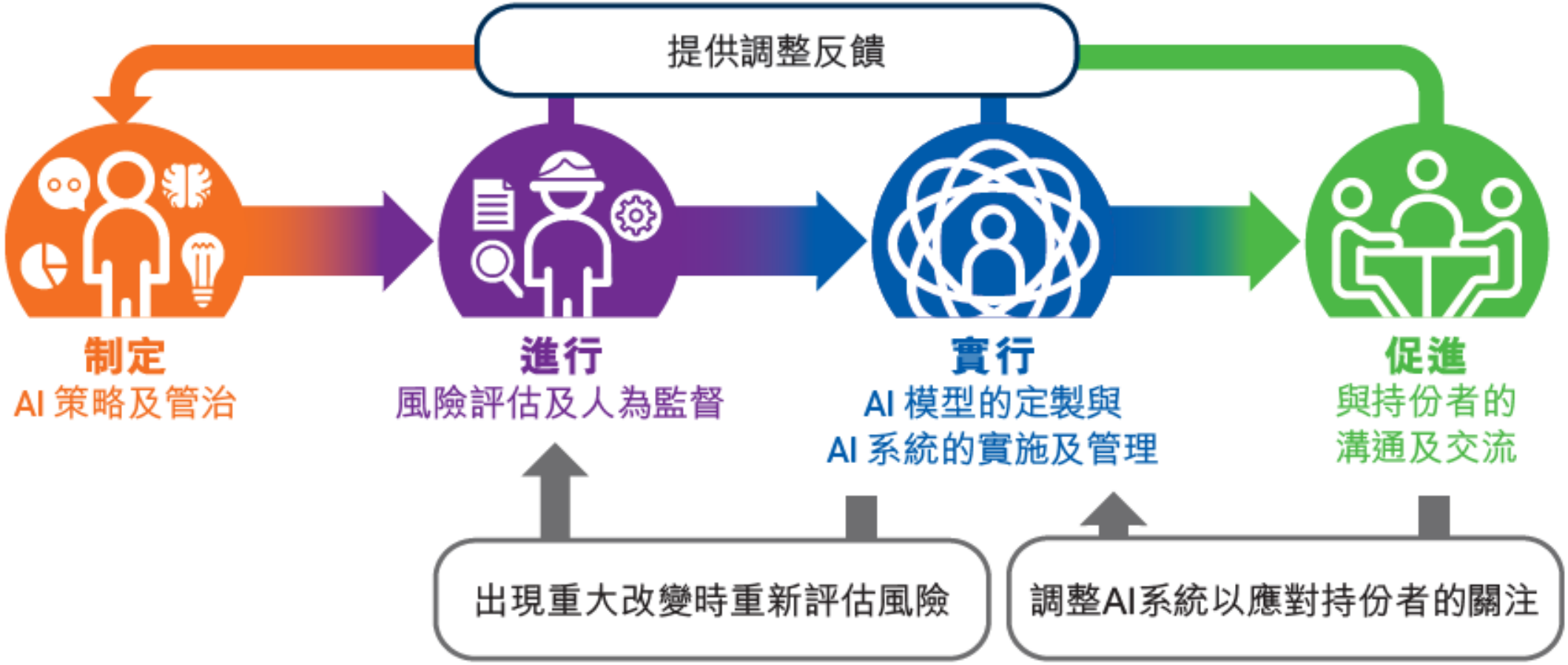
1. 尊重
2. 互惠
3. 公平

七項AI道德原則

1. 問責
2. 人為監督
3. 透明度與可解釋性
4. 數據私隱
5. 公平
6. 有益的AI
7. 可靠、穩健及安全

個人資料保障模範框架

個人資料保障模範框架


提供調整反饋
制定
AI 策略及管治
進行
風險評估及人為監督
實行
AI 模型的定製與
AI 系統的實施及管理
促進
與持份者的
溝通及交流
出現重大改變時重新評估風險
調整AI系統以應對持份者的關注

制定AI 策略及管治

AI 策略包含多項要素，能展示管理層的決心和提供指引

AI 策略

作用

展示高級管理層有決心通過符規、合乎道德標準及負責任的方式採購、實施及使用AI

就採購AI 方案的目的以及如何實施和使用AI 系統提供相關指引

可包含的要素

訂定適用於機構在採購、實施及使用AI 方案方面的**道德原則**

就如何符規、合乎道德標準地採購、實施及使用AI 方案制定**具體的內部政策和程序**

列明AI 系統在機構**中不可接受的用途**

定期與所有相關人士**就AI 策略、政策和程序溝通**

建立**AI 清單**，以幫助機構實施管治措施

考慮可能將會適用於AI 的採購、實施及使用的**法律和法規**

制定AI 策略及管治
採購AI 7個步驟

1. 尋找AI方案

2. 選擇合適的AI方案

3. 收集及準備數據

4. 為特定目的定製AI模型

5. 測試、評估和驗證AI模型

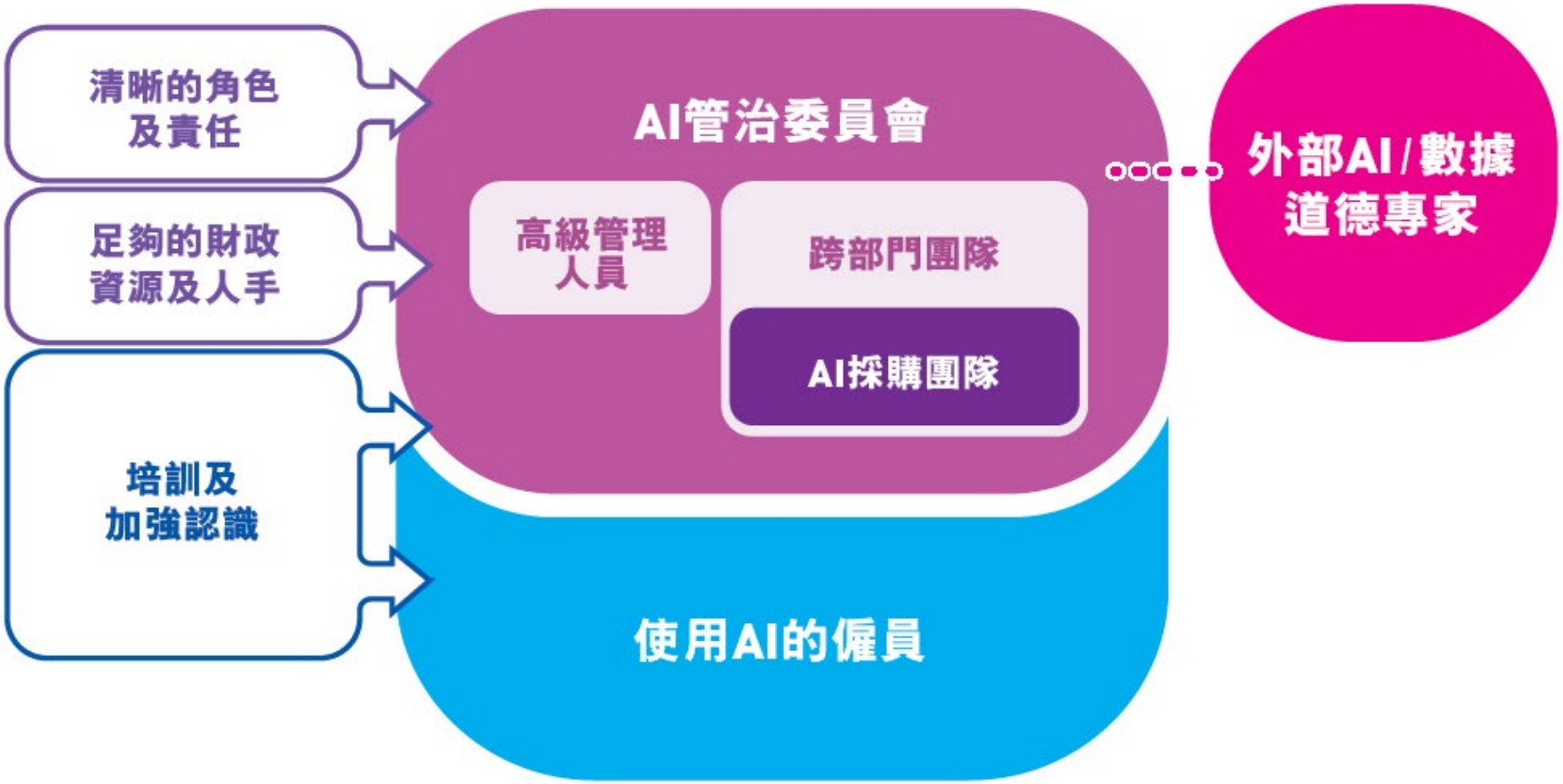
6. 測試和審核系統和組件的安全性和私隱風險

7. 將AI方案納入機構的系統中

制定AI 策略及管治
9項管治考慮

- 使用AI的目的
- 私隱和保安的責任及道德規定
- 技術性和管治方面的國際標準
- 審查AI方案的準則和程序
- 資料處理者協議
- 處理AI系統生成結果的政策
- 持續檢視環境變化的計劃
- 持續監察、管理和維持AI方案的計劃
- 評估AI供應商

制定AI 策略及管治
管治架構



制定AI 策略及管治

機構應提供足夠的培訓，確保相關人員具有適當的知識、技能和認識

建議人員	培訓主題
系統分析師 / 系統架構師 / 數據科學家	• 遵從資料保障法律、規例，和內部政策；網絡保安風險
AI 系統使用者（包括業務運作人員）	• 遵從資料保障法律、規例，和內部政策；網絡保安風險；一般AI科技
法律及合規專業人員	• 一般AI 科技和管治
採購人員	• 一般AI 科技和管治
審查員	• 查找並糾正AI 系統所作的決定或所產生的內容中任何不義的偏見、非法的歧視和錯誤 / 不準確之處
所有工作上與AI 系統有關的人員	• 機構所使用的AI 系統的好處、風險、功能和限制

進行風險評估及人為監督

風險評估的程序

1. 由跨部門團隊在採購過程或對現有的AI系統進行重大更新時**進行風險評估**
2. **識別及評估**AI系統的**風險**
3. **因應有關風險而採取適當的風險管理措施**

評估私隱風險的考慮因素

- 用來定製所採購AI 方案的資料及 / 或輸入AI 系統用作決策的資料的准許用途
- 個人資料的數量
- 所涉及資料的敏感程度
- 在使用AI 系統時的個人資料保安

評估道德風險的考慮因素

- AI 系統對受影響個人、機構及社會大眾的潛在影響
- AI 系統對個人的影響出現的可能性，以及其嚴重程度和持續時間

進行風險評估及人為監督

風險為本的人為監督

如 AI 系統**輸出的結果很可能對個人造成重大影響**，有關系統一般會被視為**高風險**。

較低 — ***AI系統的風險程度*** — 較高

人在環外

AI在沒有人為介入下作出決定

人為管控

人類決策者監督AI的運作，在有需要時介入

人在環中

人類決策者在決策過程中保留控制權以防止及/或減低AI出錯

進行風險評估及人為監督

可能帶來較高風險的AI 應用例子

- 使用生物識辨資料實時識別個人
- 評估個人享用社會福利或公共服務的資格
- 求職者評估、工作表現評核或終止僱傭合約
- 評估個人的信用可靠程度
- AI 輔助醫學影像分析或治療

實行AI模型的定製與AI系統的實施及管理

過程	重點建議	例子
1 數據準備	確保遵從私隱法例的規定 收集最少量的的個人資料 管理數據質素 妥善記錄處理數據的情況	• 一家**時裝零售平台**正計畫採購**第三方開發的AI聊天機械人**，並將其進行**定製**，以為客戶推薦**時裝建議** • 該公司或會認為需要使用**不同客戶群**過去的**購買記錄**和**瀏覽紀錄**來定製聊天機械人 • 然而，客戶的**姓名**、**聯絡資料**、某些**人口特徵等個人資料並非是需要的**

實行AI模型的定製與AI系統的實施及管理

過程	重點建議	例子
2 AI的定製及實施	☑ 對模型進行嚴格測試及驗證其可靠性、穩健性和公平性 在整合前，根據AI方案所託管的伺服器的方式（在機構內部或在第三方的雲端）考慮循規事宜 確保系統安全及數據安全	• 一間**律師事務所**正**定製**第三方開發的**AI聊天機械人**，以協助其員工**草擬法律文件**及進行**文書工作** • 該事務所應提醒員工在使用AI聊天機械人時，**盡量避免輸入個人資料**及／或**客戶的機密資訊**

實行AI模型的定製與AI系統的實施及管理

過程	重點建議	例子

3 AI的管理與持續監察	將記錄妥善地存檔
	定期進行審核
	制定AI事故應變計劃
	隨著風險因素演變而考慮採取檢視機制
		• **人為監督**應以**避免及盡量減低AI對個人造成的風險**為目的。進行人為監督的人員應：
		  • 盡可能**了解AI系統的能力和限制**；
		  • **避免過份依賴**AI輸出的**結果**；
		  • **正確地解釋及評估AI輸出**的結果；
		  • 在AI輸出的結果出現**異常時**，作出**標記並在適當情況下不理會、撤銷或推翻結果**；及
		  • 在AI供應商就AI系統輸出結果提供的資訊協助下，在適當情況**介入及中斷AI系統的運作**。

實行AI模型的定製與AI系統的實施及管理

AI事故應變計劃

1

界定AI事故

- 機構應**根據其AI系統**的情況為AI事故**作出定義**。
- AI事故可定義為「**因AI系統的開發或使用對人、財產或環境造成損害的事件，包括侵犯人權（例如私隱和不受歧視的權利）；涉及人身傷害或死亡的傷害，可被視為『嚴重事件』**」(OECD)。

2

監察AI事故

- 應**留意**和**監察可預見的損害類別**，並制定**程序應對不可預見的損害**。此步驟與風險評估過程密切相關。
- 機構可從「**AI事故資料庫**」得知過去發生的AI事故。

3

通報AI事故

- 應**制定內部政策和程序**，以便**員工就事故作出匯報**，並讓其他持份者（即業務合作夥伴、客戶）透過**反饋渠道**通報任何事故。

實行AI模型的定製與AI系統的實施及管理

AI事故應變計劃

4 **遏止AI事故擴大**

- **指派人員**負責按既定政策和程序「**暫停**」或「**停止**」**相關的AI系統**，並**切斷**受影響的系統及其他運行的系統的連接。
- 應在切實可行的情況下**盡快通知相關監管機構和受影響的人士**。

5 **調查AI事故**

- 相關人員(包括負責實施AI系統的人員)應進行**徹底的檢視和調查**，並作出**技術上的修補**。
- 調查結果應**根據機構的AI政策**作出**匯報**。
- 只有在**確定進一步損害或意想不到後果的風險降至最低**後，才可**恢復**相關AI系統的運作。

6 從AI事故**復原**

- 應**記錄**事故**調查的重點結果**。
- 調查結果或會使採購AI的內部政策和程序有修訂的需要、實施和使用AI的策略**有改變的需要**，以及**內部培訓有更新的需要**。

促進與持份者的溝通及交流

下載指引

下載小冊子

下載框架

下載懶人包

環球監管發展
規管AI涉及多方面考慮

政府間合作倡議

2024年5月

- **中國和法國**發表關於AI和全球治理的**聯合聲明**，一致認為促進AI的開發與安全，並為此推動適當的國際治理至關重要

2023年11月

- 28個國家（包括中國）簽署了**《布萊切利宣言》**

2024年3月

- 聯合國大會通過推動**「抓住安全、可靠和值得信賴的人工智能系統帶來的機遇，促進可持續發展」的決議**

法律 / 法規 / 行政措施/倡議的例子

- **《中共中央關於進一步全面深化改革 推進中國式現代化的決定》** 提到要建立人工智能安全監管制度 *(2024年7月第二十屆三中全會通過)*
- **生成式人工智能服務安全基本要求** *(2024年2月發布)*
- **全球人工智能治理倡議** *(2023年10月提出)*
- **網絡安全標準實踐指南 – 生成式人工智能服務內容標識方法** *(2023年8月發布)*
- **生成式人工智能服務管理暫行辦法** *(2023年7月發布)*
- **互聯網信息服務深度合成管理規定** *(2022年11月發布)*

- **AI法案** *(於2024年8月生效)*

- **安全、可靠、值得信賴AI行政命令** *(2023年10月頒布)*

示範

製作深度偽造影片不太困難；有圖未必有真相

聯絡我們

查詢 2827 2827　傳真 2877 7026

網址 www.pcpd.org.hk

電郵 communications@pcpd.org.hk

地址 香港皇后大道中248號大新金融中心13樓1303室

保障、尊重個人資料私隱

Protect, Respect Personal Data Privacy

追蹤我們最新資訊