新北市政府使用人工智慧作業指引

新北市政府

中華民國 114 年 8 月 14 日

目錄

壹、前言

一、目的

為加速推動人工智慧技術(以下簡稱AI)於新北市政府(以下簡稱本府)所屬各機關及學校之導入與應用，建立AI治理框架及促進AI技術標準化應用，進而提升公共服務效能與智慧治理能量，並打造一個民眾可信賴的AI應用環境，特訂定本指引作為各機關規劃與執行AI相關作業時之參考。透過本指引，可協助本府各機關明確辨識風險層級、遵循資料治理與倫理原則，確保AI技術運用具備一致性與規範性，並促進既有資訊系統與AI技術整合，強化市府AI應用之準確性、公正性與安全性。此外，指引亦納入對自主決策型AI系統之規範，以因應未來代理式AI (Agentic AI)技術之應用趨勢。

二、訂定依據

本作業指引係參照國家科學及技術委員會之「行政院及所屬機關（構）使用生成式AI參考指引」及數位發展部之「公部門人工智慧應用參考手冊」訂定。

三、適用對象

（一）本府所屬各機關及學校，辦理AI導入作業應遵循本作業指引。各機關並得視使用人工智慧之業務需求，參酌本作業指引另訂該機關之使用規範或內控管理措施。

（二）本府及所屬各機關就所辦理之AI相關採購事項，亦應要求得標之事業、法人、團體或個人遵循本作業指引，並恪遵該機關所制定之使用規範或內控管理措施。

貳、AI 應用與技術

一、AI 應用場景：AI 於政府機關之應用場景多元，涵蓋文件分類、智能客服、影像辨識及決策輔助分析等，更多應用案例請參考附件一，相關技術可有效提升行政效能並優化民眾服務體驗；惟上述僅為公部門應用 AI 之部分案例，實際應用範疇將隨技術演進與業務需求持續擴展，各機關仍應依自身業務特性、資料條件與發展目標，審慎評估導入之技術可行性、應用適切性及整體效益。

二、AI 相關技術：係指模擬、延伸或強化人類智能之方法，具備學習、自主推理、決策預測或模式識別等能力，常見技術範疇包括但不侷限於下列類型：

（一）機器學習（Machine Learning）：如分類、回歸、分群等演算法，透過資料訓練模型以支援預測與決策。

（二）深度學習（Deep Learning）：如類神經網路、卷積神經網路（CNN）、循環神經網路（RNN）等，應用於影像、語音、文本等非結構化資料處理。

（三）自然語言處理(Natural Language Processing, NLP)：涵蓋語意理解、對話生成、語音辨識、文字生成與機器翻譯、文本摘要與資訊擷取等應用。

（四）電腦視覺(Computer Vision)：如物件偵測、人臉辨識、影像分割與光學字元辨識（OCR）、姿態估計 (Pose Estimation)及視覺問答（VQA）等。

（五）生成式人工智慧（Generative AI）：如文本、圖像、語音等多模態資料之自動生成技術，包含大型語言模型（LLM）與其他生成模型。

（六）AI 模型管理與應用技術：如自動化建模（AutoML）、模型微調(Fine-tuning)、遷移學習(Transfer Learning)、模型部署(Model Deployment)、模型漂移(Model Drift)

監控、機器學習／人工智慧營運管理平台（MLOps／AIOps）、檢索增強生成（RAG）、模型可解釋性工具（如LIME、SHAP）、A/B 測試、漸進式部署等技術。

（七）AI 系統整合技術：如 API 管理與微服務架構、容器化部署技術、AI 模型版本控制等系統整合相關技術。

（八）代理式人工智慧(Agentic AI)：多代理系統(Multi-Agent Systems)、強化學習代理(Reinforcement Learning Agents)、目標導向對話系統(Goal-oriented Dialog Systems)及自主決策框架(Autonomous Decision Frameworks)。

凡應用前述任一技術，且系統核心功能實質依賴 AI 模型或演算法進行預測、判斷、分類或決策者，即得認定為人工智慧應用範疇。

參、AI 導入評估及管理

一、導入前期評估：AI 導入流程可依循《公部門人工智慧應用參考手冊》，進行前期之服務與資料評估。事先應針對實際業務場景進行分析，明確界定待解決之問題，並評估 AI 技術是否為適切的解決工具；其次，須就資料狀況進行盤點與評估，確認資料之可取得性、品質、完整性及合規性，以作為後續模型建置與應用推動之基礎，可參考附件二。

二、風險評估：參照歐盟人工智慧法案，依據 AI 對人類權益、社會福祉、資料安全與基本權利的潛在影響，將 AI 系統劃分為四大風險等級，並對應不同層級的管理強度：

（一）不可接受風險(Unacceptable Risk)：涉及對人類尊嚴、自由與民主制度構成嚴重威脅之 AI 用途，例如：政府監控用之社會評分系統、用於執法之即時生物特徵識別系統、利用弱勢群體認知缺陷來操控其行為之 AI 應用、以行為預測為目標之預防性執法系統，屬本風險 AI 應用

不得開發。

（二）高風險（High Risk）：AI 系統若直接影響個人安全或基本權利且為社會關鍵領域運作（如醫療、交通、教育、執法），即被歸類為高風險，例如：自動化人事決策系統（如履歷篩選、自動面試打分）、AI 協助之醫療診斷或治療建議、教育系統中之成績評定或入學篩選、法院輔助判決建議系統等，屬高風險的 AI 應用，應送本府資安實驗室或 AIEC 驗測後始得實施。

（三）有限風險（Limited Risk）：雖不會直接造成個人基本權利受損，但仍涉及與人互動時的資訊不對稱風險，例如：客服機器人（Chatbot）或虛擬助理、生成式 AI（如 ChatGPT）用於對話、內容創作，產生圖片、影片或聲音的 Deepfake 系統，屬本風險的 AI 應用，應明確告知使用者其所接觸內容為人工智慧產出。

（四）最低風險（Minimal Risk）：AI 用途對個人或社會幾乎不構成風險，例如：電子郵件垃圾分類器、AI 語音輸入助理、遊戲相關、個人化廣告推薦系統。

三、營運管理：AI 系統建置完成後，營運階段應持續關注風險管理、倫理原則、資安防護、資料及模型治理與更新機制等面向。

（一）風險管理，須定期監控模型效能與偏誤變化，以維持決策準確與一致性。

（二）倫理面，應持續落實公平性、透明性、可解釋性、問責性、人為監督及隱私保護等原則。

（三）資安面，除傳統防護機制外，亦應強化對抗樣本攻擊、模型竊取與資料中毒等 AI 特有風險控管。

（四）資料面，須確保合法性、去識別化與資料最小化原則等。

（五）模型治理與更新機制

1. 應建立模型生命週期管理制度，包含訓練資料記錄、模型版本控管、偏誤與效能監測等。
2. 高風險 AI 系統應設置自動監控工具，定期偵測模型漂移（Model Drift）與推論錯誤率。
3. 對於已部署模型，應至少每年進行一次再訓練與驗證，並評估是否應汰換或更新模型。

各機關應每年檢視上述項目，並視實際情況滾動調整，以確保 AI 系統之穩定性與信賴度。

肆、 規範及內控機制

一、機關應審慎面對 AI 產出之資訊，由機關之業務單位依風險進行客觀且專業之最終判斷，不得取代其自主思維、創造力及人際互動，亦不得完全信任 AI 產出，或逕以未經確認內容作為行政行為或公務決策之唯一依據。

二、使用 AI 應遵守資通安全、個人資料保護、著作權及相關資訊使用等規定，並注意其侵害智慧財產權與人格權之可能性。

三、各機關在採購或使用 AI 資通訊產品時，請確依行政院公布之「各機關對危害國家資通安全產品限制使用原則」辦理。應不允許大陸地區廠商及陸籍人士參與，並不得採購及使用大陸廠牌資通訊產品(含軟、硬體及服務)。

四、鑒於近年 AI 技術日益普及，各機關以 AI 為名進行對外宣傳之情形日增。為維護政策推動之專業性與公信力，並避免對外產生誤導或認知落差，於宣傳前，應審慎確認所宣傳專案是否應用前述第貳章第二條所列之 AI 相關技術，以確保宣傳內容之準確性與專業性。

五、如 AI 系統涉及輔助決策功能，建議機關可制定系統標準流程，例如：「AI 輔助決策三階段驗證機制」：

（一）第一階段：AI 系統產出結果。

（二）第二階段：業務人員專業審核。

（三）第三階段：主管覆核確認。

六、各機關得提供生成式 AI 使用者基礎教育訓練課程，內容包含：

（一）Prompt 撰寫技巧與最佳實務。

（二）生成內容的驗證與事實查核。

（三）常見風險(如內容幻覺、偏誤、錯誤推論)之辨識方法。

（四）運用封閉式系統保障機密資訊的技術建議。

七、運用生成式 AI 時，應遵循之規範：

（一）製作機密文書應由機關之業務單位親自撰寫，禁止使用生成式 AI。機密文書，係指「新北市政府文書處理要點」所定之國家機密文書及一般公務機密文書。

（二）機關之業務單位不得向生成式 AI 提供涉及公務應保密、個人及未經機關（構）同意公開之資訊，亦不得向生成式 AI 詢問可能涉及機密業務或個人資料之問題，但封閉式地端部署之生成式 AI 模型，於確認系統環境安全性後，得依「新北市政府文書處理要點」文書或資訊機密等級分級使用。

（三）各機關使用生成式 AI 執行業務或提供服務輔助工具時，應讓使用者明確知道自己在使用生成式 AI 服務，如透過介面提醒或顯示數位浮水印等方式呈現，並避免 AI 所生成的內容帶有偏見歧視。

（四）如機關使用檢索增強生成（RAG）技術，建議針對結合內部資料庫的生成式 AI 系統制定：

1. 知識庫資料品質標準。
2. 檢索相關度閾值設定。
3. 資料來源可追溯性要求。
4. 知識更新機制。

八、本府 AI 專案因涉及智慧城市及資訊相關計畫，預算編列應依「新北市政府資訊計畫先期審查作業要點」辦理，且於各機關採購前應會辦資訊中心並檢附附件三、AI 專案建置自主檢核表，並應提交 AI 系統架構設計書，包含資料流程圖、模型部署架構、API 介面設計等技術文件。

伍、本作業指引之修訂

本作業指引先行試辦，後續將依中央或本府法規之增修、各機關反饋意見或當人工智慧的技術有大幅躍進時，每年進行修訂。

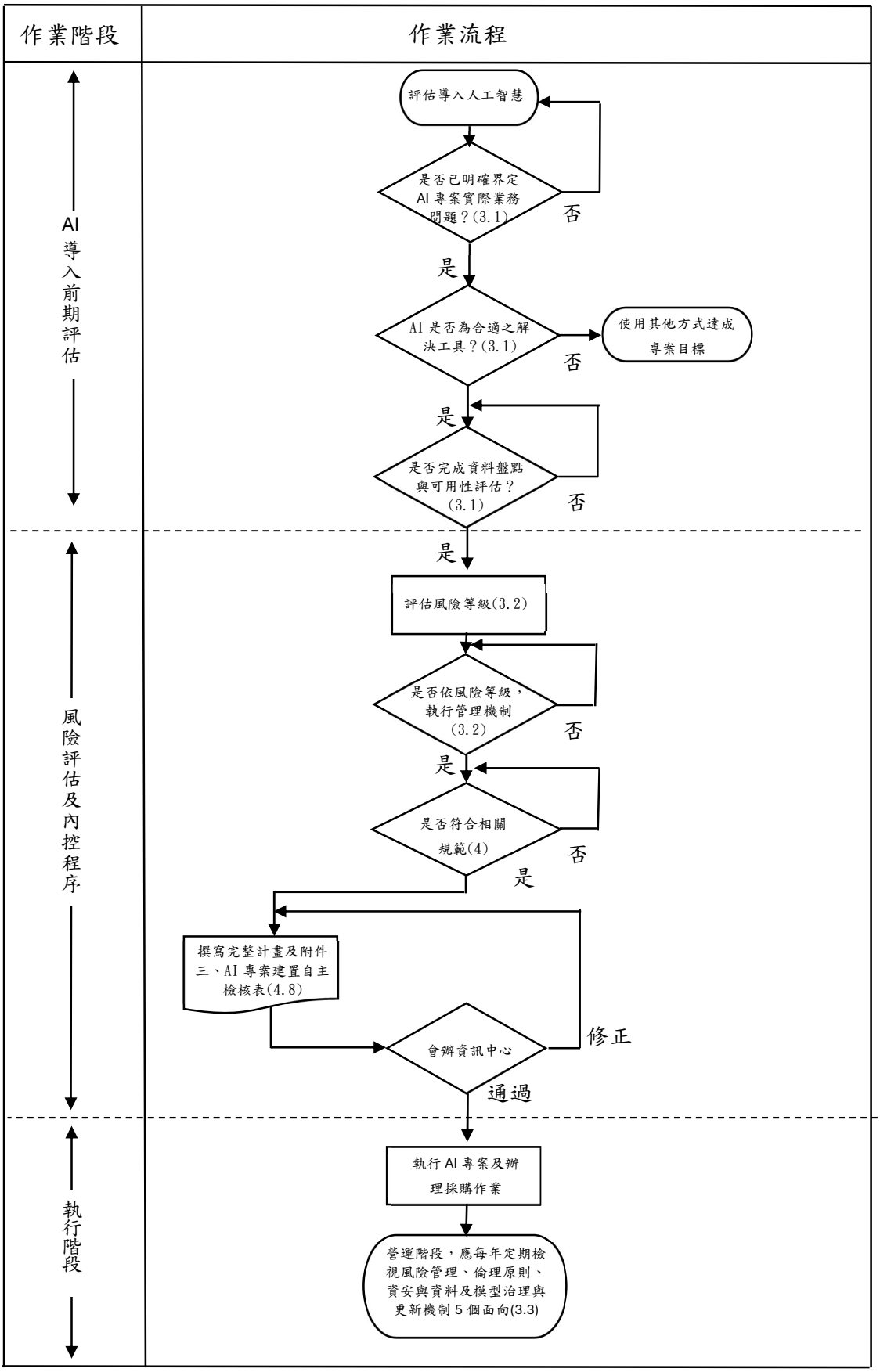
附件一、AI 於政府機關應用場景及風險等級對照表

以下所列係人工智慧於政府機關應用之部分參考範例，未來隨技術持續演進與業務需求拓展，相關應用場景亦將不斷增加並趨於多元化；另為協助各機關初步評估 AI 應用風險，爰提供本表作為參考。惟實際風險等級之判定，仍應視個案情形審慎評估。

應用場景	內容	可能的風險等級	風險等級判斷說明
智慧客服系統	透過生成式 AI 提供即時問答、辦理進度查詢與政策說明，減少人力負擔、提升服務效率。	有限風險	屬於人機互動型應用，可能產生資訊不對稱，使用者應知其對象為 AI 系統。
公文起草與文字生成	協助撰寫公文、新聞稿、政策報告初稿，加速行政流程並提升文字品質與一致性。	有限風險	為內部行政支援用途，屬內容生成型 AI，應提醒使用者注意內容為 AI 產生。
民眾意見摘要與分析	整合線上民意調查、社群反應等，生成摘要報告供決策參考。	有限風險	若僅摘要公開意見並輔助決策，風險較低，但應注意資料來源偏誤與去識別化處理。
教育訓練教材生成	快速製作客製化的內部訓練教材或數位學習資源，提升人員專業能力。	最低風險	屬非關鍵用途，不涉權益判定或安全影響。
語音與影像內容製作	利用 AI 生成語音導覽、宣導短片或手語翻譯影片，加強政策宣傳的包容性與普及率。	有限風險	若含 Deepfake 或生成音訊、手語翻譯等內容，需明示為 AI 產出以避免誤導。
多語言翻譯與在地化	提供多語言政策說明、網站內容翻譯，改善外籍居民與旅客的溝通體驗。	最低風險	屬輔助工具性質，對人類權益無直接風險。
流程自動化與文件分類	自動生成或整理行政文件，提升檔案管理與資訊檢索效率。	最低風險	用於提升行政效率，不涉人權與判斷行為。
政策模擬與公眾溝通	生成虛擬案例模擬政策實施後果，用於政策說明或民眾溝通。	有限風險	若生成政策模擬案例供民眾理解，應標註為虛構情境。

應用場景	內容	可能的 風險等級	風險等級判斷說明
資料視覺化敘事生成	結合資料與圖像生成工具，製作簡報、視覺化報告等，用於內部簡報或民眾說明。	最低風險	用於內部或對外報告呈現，非決策核心。
法令草案輔助撰寫與潤飾	協助起草條文、條理清晰地重寫法規草案，並檢查文字風險或模糊語句。	高風險	涉及法律規範草擬，若交由 AI 生成條文初稿，須有嚴謹審核流程。
自動化法規符合性檢核系統	運用代理式 AI 整合法規資料庫、申請文件與審查準則，主動解析政策草案、補助申請或合約內容，判斷是否符合相關規範，標示潛在風險條文或缺漏項目，並提出修正建議。	高風險	屬代理式 AI 應用，系統具判斷功能並可能影響合約或補助審查，應納入風險驗測機制與人工覆核流程。

附件二、新北市政府導入 AI 內部控制流程圖



附件三、AI 專案建置自主檢核表

填寫日期： 年 月 日

AI 專案建置自主檢核表		
承辦機關		
聯絡人	姓名/職稱/電話：	
專案名稱		
專案內容(請簡述)		
專案期程		
風險等級	□不可接受 □高風險 □有限風險 □最低風險 風險評估說明(請簡述)：	
檢核項目	自評結果 (填否及不適用，請於備註說明原因)	備註
一、AI 導入前期評估（依循《公部門人工智慧應用參考手冊》）		
1. 是否已明確界定 AI 專案所對應的實際業務問題？(分析服務現況與痛點，界定 AI 應用目標)	□是 □否	
2. 是否已評估 AI 是否為合適之解決工具？(確認 AI 優於其他解法，並具成本效益與技術可行性)	□是 □否	
3. 是否完成資料盤點與可用性評估？(包括資料可取得性、格式、品質、完整性)	□是 □否	
二、AI 通用性規範		
1.是否未直接使用 AI 資訊作為行政行為或決策依據？(需先行審核驗證，避免誤導或錯誤決策)	□是 □否 □不適用	
2.是否已遵守資通安全管理法及子法、個人資料保護法、著作權法等相關法律規定？	□是 □否 □不適用	
三、運用生成式 AI 時，應遵循之規範		

檢核項目	自評結果 （填否及不適用，請於備註說明原因）	備註
1.該應用是否無涉及機密文書製作	□是 □否 □不適用	
2.是否避免向生成式 AI 提供公務應保密、個人及未經機關（構）同意公開之資訊、可能涉及機密業務或個人資料之問題？(採封閉式地端部署之生成式 AI 模型不適用)	□是 □否 □不適用	
3.若使用地端部署模型，是否確認系統環境安全性並控管使用範圍？(仍應依文書或資訊機密等級分級使用)	□是 □否 □不適用	
4.是否於系統或服務中明確揭示生成式 AI 的使用？(可透過介面提示、浮水印、標示等方式)	□是 □否 □不適用	
四、資通安全與設備採購限制		
1. 是否採購非大陸廠牌資通訊產品(含軟、硬體及服務)？	□是 □否 □不適用	
2. 採購作業是否禁止廠商為大陸地區廠商或陸籍人士？	□是 □否 □不適用	
五、宣傳前之確認		
1. 宣傳內容是否確實應用 AI 技術？應符合第貳章第二條所列之 AI 相關技術（如機器學習、自然語言處理等）	□是 □否 □不適用	
2. 是否避免誇大或誤導性使用「AI」名義進行宣傳？	□是 □否 □不適用	
六、檢核機制		
是否於 AI 專案採購前完成會辦資訊中心？(應檢付完整計畫及檢核表)	□是 □否	

參考文獻

一、行政院(2023 年 10 月 3 日)。行政院及所屬機關（構）使用生成式 AI 參考指引。國家科學及技術委員會全球資訊網。https://www.nstc.gov.tw/folksonomy/list/c79bf57b-dc94-4aff-8d14-3262b5559cfc?l=ch

二、數位發展部(2025 年 1 月 8 日)。公部門人工智慧應用參考手冊（草案）。數位發展部(moda)。https://moda.gov.tw/digital-affairs/digital-service/guide/15002

三、臺北市政府(2024 年 9 月 9 日)。臺北市政府使用人工智慧作業指引。臺北市政府全球資訊網。https://www-ws.gov.taipei/Download.ashx?u=LzAwMS9VcGxvYWQvNzY3L3JlbGZpbGUvNTQ1NTAvMTM2NzkzLzY0NGQxZTIzLTEzYTktNDdmMi1iNzdiLWExMTEyMzg4ZjNmNi5wZGY%3D&n=6Ie65YyX5biC5pS%2F5bqc5L2%2F55So5Lq65bel5pm65oWn5L2c5qWt5oyH5byVLnBkZg%3D%3D&icon=..pdf

四、International Organization for Standardization. (2023). ISO/IEC 42001:2023 – Information technology — Artificial intelligence — Management system . Geneva: ISO.

五、International Organization for Standardization. (2023). ISO/IEC 23894:2023 – Information technology — Artificial intelligence — Guidance on risk management. Geneva: ISO.

六、International Organization for Standardization. (2022). ISO/IEC 38507:2022 – Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations. Geneva: ISO.

七、National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework

八、International Organization for Standardization. (2022). ISO/IEC 22989:2022 – nformation technology — Artificial intelligence — Artificial intelligence concepts and terminology. Geneva: ISO.

九、Organisation for Economic Co-operation and Development. (2019). OECD Principles on Artificial Intelligence. OECD. https://www.oecd.org/going-digital/ai/principles/