政府人工智慧應用風險評估檢核表

填寫目的：參酌人工智慧基本法規定及人工智慧風險分類框架模式，評估並掌握政府機關人工智慧系統與服務之潛在風險。

填寫範圍：本局自行開發、委外建置或採購之各類人工智慧系統與服務。

填寫說明：本表由人工智慧系統與服務管理之業務權責單位填寫，作為符合人工智慧基本法第十九條進行風險評估與訂定內控管理措施之基礎。

1、 應用情境盤點

本表格目的為掌握人工智慧系統與服務之基本資訊。

檢核項目	填寫內容/評估結果	依據來源	備註/參考文件
1.1 系統/服務名稱	[請填寫]	人工智慧風險分類框架(附錄1)	系統清冊
1.2 系統/服務具體用途與服務對象	**【具體用途】** [請填寫] **【服務對象】** ☐ 僅供本局內部使用 ☐ 跨機關使用 ☐ 對外公開服務 ☐其他：_____	人工智慧風險分類框架(附錄1)	系統清冊
1.3 核心人工智慧技術與開發方式			
1.3.1 核心人工智慧技術 使用大型語言模型（Large Language Model, LLM）、生成式AI或其他人工智慧技術(如機器學習、	[請填寫]	人工智慧風險分類框架(附錄1)	採購或系統文件

電腦視覺、語音辨識等)			
1.3.2 核心人工智慧開發方式 各類人工智慧系統與服務取得方式採自行開發、委外建置或採購	□自行開發建置 □委外開發建置 □採購服務 □其他：_____		
1.4 辨識利害關係人			
1.4.1 負責開發人工智慧系統/服務 (如自行、委外、採購、其他混合開發等)	□自行開發 □委外開發 □僅進行採購(類 SaaS 服務) □其他：_____	人工智慧風險分類框架(附錄 1)	業務作業說明
1.4.2 負責部署人工智慧系統/服務 (如自行、委外、採購、其他共同部署、平台整合等)	□自行部署 □委外部署 □僅進行採購(類 SaaS 服務) □其他：_____		
1.4.3 可能因人工智慧系統/服務影響之利害關係人 (可複選)	□民眾 □本局公務員 □跨機關公務員 □其他：_____		

2、 核心風險識別與危害評估

參考人工智慧風險分類框架，檢視並確認系統與服務是否存在風險。檢核項目 2.0 如勾選「是」，應依本局資通安全及個人資料管理既有機制辦理，且不影響後續人工智慧風險之識別與評估；檢核項目 2.1 至 2.3 均勾選「否」者，無須填寫 2.4 及 2.5；**檢核項目 2.1、2.2 或 2.3 任一項勾選「是」者，應接續進行檢核項目 2.4 嚴重程度判定**；若檢核項目 2.4.1、2.4.2、2.4.3 任一項勾選「是」，應接續進行檢核項目 2.5 風險發生機率判定。

檢核項目	評估結果	情境簡述	依據來源
2.0 風險類型零：資安及個資問題 本人工智慧系統/服務、訓練資料、輸入指令或輸出內容，是否有可能因其業務流程、使用情境或設計邏輯，而**有違反資通安全或個人資料保護之疑慮**？	□是 □否	[若評估結果為是，請簡述]	
2.1 風險類型一：人工智慧系統/服務本身之技術設計缺陷 本人工智慧系統/服務，是否有可能因其業務流程、使用情境或設計邏輯，**發生演算法偏見、產生錯誤或誤導訊息、缺乏透明度或存在安全漏洞等情形，導致對特定群體造成不公平待遇或產生決策錯誤**？	□是 □否	[若評估結果為是，請簡述]	人工智慧風險分類框架（二、識別風險）
2.2. 風險類型二：部署後操作及人機互動問題			
2.2.1 本人工智慧系統/服務，是否有可能因其業	□是 □否	[若評估結果為是，請簡述]	人工智慧風險分類框架

務流程、使用情境或設計邏輯，**導致公務員過度依賴而喪失獨立判斷能力**？			（二、識別風險）
2.2.2 本人工智慧系統/服務，是否有可能因其業務流程、使用情境或設計邏輯，**導致本人工智慧系統/服務遭濫用，以生成違法內容、深度偽造內容、詐欺訊息或其他有害內容**？	□是 □否	[若評估結果為是，請簡述]	
2.2.3 本人工智慧系統/服務，是否有可能因其業務流程、使用情境或設計邏輯，導致人工智慧系統/服務**遭利用於網路攻擊**？	□是 □否	[若評估結果為是，請簡述]	
2.3 風險類型三：社會結構與環境衝擊			
2.3.1 本人工智慧系統/服務，是否有可能因其業務流程、使用情境或設計邏輯之應用，**可能加劇社會不平等之疑慮或風險**？	□是 □否	[若評估結果為是，請簡述]	人工智慧風險分類框架（二、識別風險）
2.3.2 本人工智慧系統/服務，是否有可能因其業務流程、使用情境或設計邏輯之應用，**可能耗費能源造成環境傷害**？	□是 □否	[若評估結果為是，請簡述]	
2.3.3 本人工智慧系統/服務，是否有可能因其業務流程、使用情境或設計	□是 □否	[若評估結果為是，請簡述]	

邏輯之應用，**可能遭利用於認知作戰，影響公眾信任**？			
2.4 嚴重程度判定 若檢核項目 2.1、2.2、2.3 任一項勾選「是」，請接續進行嚴重程度判定，**應依風險發生時可能造成之危害程度，逐一檢視 2.4.1 至 2.4.3 所列情形，並分別勾選「是」或「否」**；如同時符合二個以上嚴重程度，以最高嚴重程度作為 2.4 判定結果。 如 2.4.1 至 2.4.3 均勾選「否」，表示已識別之風險，其可能影響尚未達本表所列一般、嚴重或非常嚴重之危害程度，無須辦理 2.5 風險發生機率判定。			
嚴重程度 **（依風險發生時之危害程度判定）**		**評估結果**	**依據來源**
2.4.1 【非常嚴重】 符合下列情形之一者屬之： 1. 上開風險發生，可能對本局之營運、資產或信譽等方面**產生非常嚴重或災難性之影響**。 2. 上開風險發生，可能影響他人合法權益或本局執行業務之公正性及正當性，並**使本局所屬人員負刑事責任**。 3. 上開風險發生，可能導致特定群體（如特定性別、年齡、身心障礙或弱勢族群）在政府資源分配、資格審查或行政決策上遭受**嚴重不公平對待**。 4. 上開風險發生，可能違反兒少最佳利益原則，或對其身心發展造成**嚴重負面影響**。 5. 上開風險發生，可能引發社會恐慌，或造成對人類基本權利、生命安全、財產保障、社會秩序或		□是 □否	參考人工智慧基本法、「資通安全責任等級分級辦法」之「附表九、資通系統防護需求分級原則」

政府公信力之**嚴重危害**。		
2.4.2 【嚴重】 符合下列情形之一者屬之： 1. 上開風險發生，可能對本局之營運、資產或信譽等方面**產生嚴重之影響**。 2. 上開風險發生，可能影響他人合法權益或本局執行業務之公正性及正當性，並使**本局所屬人員受行政罰、懲戒或懲處**。 3. 上開風險發生，可能導致特定群體（如特定性別、年齡、身心障礙或弱勢族群）在政府資源分配、資格審查或行政決策上遭受**不公平對待**。 4. 上開風險發生，可能違反兒少最佳利益原則，或對其身心發展造成**負面影響**。 5. 上開風險發生，可能引發社會恐慌，或造成對人類基本權利、生命安全、財產保障、社會秩序或政府公信力之**危害**。	□是 □否	參考人工智慧基本法、「資通安全責任等級分級辦法」之「附表九、資通系統防護需求分級原則」
2.4.3 【一般】 符合下列情形之一者屬之： 1. 上開風險發生，可能對機關之營運、資產或信譽等方面產生有限之影響。 2. 上開風險發生，可能發生違反相關規範之情形。	□是 □否	參考「資通安全責任等級分級辦法」之「附表九、資通系統防護需求分級原則」
2.5 風險發生機率 若檢核項目 2.4.1、2.4.2、2.4.3 任一項勾選「是」，請接續進行**風險發生機**		

率判定；2.4.1 至 2.4.3 均勾選「否」者，無須辦理本項判定。 請依系統與服務之使用頻率、接觸對象、歷史事件、控制措施有效性及可被濫用之可能性等因素，**擇一勾選「高」、「中」或「低」**，判斷原則如下：預期經常發生、曾發生或在現有情境下極可能發生者為「高」；具合理發生可能性者為「中」；在現有控制措施下發生可能性較低者為「低」。		
發生機率	□高 □中 □低	

3、 **人工智慧內控管理與應對機制（Internal Controls & Response）**

請依據前述「核心風險識別與危害評估」之結果【包含資安及個資問題(2.0)、識別出之風險類型(2.1至2.3)、嚴重程度(2.4)、發生機率(2.5)】，依下列規定辦理列管、風險等級分級或應辦措施：

(1) 若「資安及個資問題(2.0)」勾選「是」，該項風險應循本局資通安全及個人資料管理既有機制處理，並提供相關佐證文件(如：資通安全維護計畫、個人資料風險評估紀錄或其他足資證明之文件)；本項另依本局既有管理機制辦理，不影響下列風險等級之綜合判定與應辦措施之辦理。

(2) 人工智慧風險等級綜合判定(參考2.1至2.5)

檢核項目2.4.1至2.4.3任一項勾選「是」者，請依「嚴重程度(2.4)」(由2.4.1至2.4.3取其高)與「發生機率(2.5)」判定結果，參考下表綜合判定為甲類、乙類或丙類風險等級；不適用下表者，綜合判定為無顯著風險：

發生機率(2.5) 嚴重程度(2.4)	高	中	低
非常嚴重	甲	甲	乙
嚴重	甲	乙	丙
一般	乙	丙	丙

甲、 甲類風險：指衝擊極為嚴重或影響範圍廣大者，包含「非常嚴重且發生機率為高或中」、「嚴重且發生機率為高」等情形。

乙、 乙類風險：指具有中度衝擊或中等發生機率者，包含「非常嚴重且發生機率為低」、「嚴重且發生機率為中」、「一般且發生機率為高」等情形。

丙、 丙類風險：指衝擊較輕微或發生機率較低者，包含「嚴重且發生機率為低」、「一般且發生機率為中或低」等情形。

丁、 無顯著風險：前項檢核項目經評估無顯著影響者，包括未識別出人工智慧風險，及雖識別出相關風險，但其可能影響尚未達檢核項目2.4所列危害程度等情形(即檢核項目 2.1 至 2.3 均勾選「否」，或是 2.1、2.2、2.3任一項勾選「是」但2.4.1至2.4.3均勾選「否」)，屬無顯著風險。

風險等級綜合判定結果：

□甲類 □乙類 □丙類 □無顯著風險

- 備註：

  判定簡要流程如下：

  2.1 至 2.3 均為「否」 → 無顯著風險；

  2.1 至 2.3 任一為「是」 → 接續填 2.4；2.4 均為「否」 → 無顯著風險；2.4 任一為「是」 → 接續填 2.5 → 依 2.4 與 2.5 查表判定甲、乙或丙類風險。

  2.0 如為是，另循本局資通安全及個人資料管理既有機制辦理及列管，不影響上述判定流程。

(3) 應辦措施

依據前項綜合判定結果，依下列方式辦理下列措施：

(1) 若綜合判定結果為甲類者：應檢核本段 3.1 至 3.5 所列措施。

(2) 若綜合判定結果為乙類者：應檢核本段 3.1 至 3.5 所列措施。本局將依風險等級及資源配置情形，優先處理甲類風險部分。

(3) 若綜合判定結果為丙類者：應檢核本段 3.1 人類監督保留及 3.2 透明標示所列措施。

(4) 若綜合判定結果為無顯著風險者：無須檢核本段 3.1 至 3.5 所列措施；惟仍應依系統與服務之實際功能，遵循適用法令及本局相關管理規定，並於發生重大功能變更時重新進行風險評估。

(5) 如因系統與服務之功能、用途或適用情境，無該項措施之適用情形者，得勾選「不適用」並敘明理由(例如：系統與服務性質確實無涉決策(3.1)或內容產出(3.2)者)。

檢核項目	落實結果	辦理/不適用說明	參考來源
3.1 人類監督保留：針對人工智慧提供之決策，涉及具法律效果或影響民眾權益之行政處分者，作業流程中是否設有適當之人為介入與最終覆核機制，避免完全交由人工智慧自動裁定？	□是 □否 □不適用		人工智慧基本法第 4 條第 2 款
3.2 透明標示：系統與服務自動生成圖文、內容，或提供對外互動服務（如客服機器人）時，是否於顯著處標示「本內容由人工智慧輔助／生	□是 □否 □不適用		人工智慧基本法第 5 條

成」等提示？			
3.3 明確警語：若屬高風險，是否依法明確標示注意事項或警語？	□是 □否 □不適用		人工智慧基本法第 5 條
3.4 建立救濟機制：針對受人工智慧輔助決策影響之民眾，是否提供明確之申訴管道、決策解釋與權益救濟機制？	□是 □否 □不適用		人工智慧基本法第 17 條
3.5 強化防護：是否於採購契約或系統與服務規格中，加重對供應商之資通安全、資料治理與防範歧視偏見之要求？	□是 □否 □不適用		人工智慧基本法第 4 條第 4 款、第 6 款

業務權責單位核章

系統與服務管理人：_____　　單位主管：_____