人工智能(AI)：
個人資料保障模範框架 —

代理式人工智能篇：
保障個人資料私隱

PCPD
HK

PCPD.org.hk

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

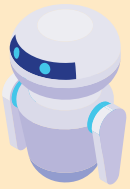
30

支持機構：
Supporting Organisations:

中華人民共和國香港特別行政區政府
數字政策辦公室
Digital Policy Office
The Government of the Hong Kong Special Administrative Region of the People's Republic of China

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

引言



隨着人工智能（AI）的應用日趨普及，一種稱為「代理式AI」（agentic AI）的新型AI應運而生。代理式AI有別於傳統AI聊天機械人，能夠以高程度的自主性運作。儘管代理式AI在提升生產力方面潛力龐大，然而，對個人資料私隱所構成的風險也相對較高。因此，機構在部署使用代理式AI時必須保持警惕，並加強相關的監察工作。

本指引旨在協助資料使用者（包括機構及個人）了解使用代理式AI所涉及的個人資料私隱風險，並就如何根據《個人資料（私隱）條例》（第486章）（「《私隱條例》」）保障個人資料私隱提供實務建議。本指引是個人資料私隱專員公署發布的《人工智能（AI）：個人資料保障模範框架》（「《模範框架》」）的補充指引，而《模範框架》所載的原則一般而言仍適用於代理式AI的使用。

甚麼是代理式AI？

代理式AI是指具備自主感知、記憶、決策、交互與執行能力的智能系統[1]。代理式AI以基礎模型（例如大型語言模型）為核心架構，並結合資料庫、記憶體及電腦操作系統等額外工具。

傳統的AI聊天機械人一般用於生成內容，例如回答問題、總結文件或編寫程式碼。相比之下，代理式AI的用途更為廣泛，能按預設流程代替用戶自主執行多步驟任務，包括處理電郵、預訂餐廳及繳付款項。此外，用戶亦可於系統內部署多個智能體（agents）協同運作，透過多個專責代理人同步處理不同工作，從而提升表現。

值得注意的是，由於智能體並不是法律實體，透過智能體收集、持有、處理或使用個人資料的機構作為資料使用者[2]仍須負責遵守《私隱條例》下的規定，包括六項保障資料原則。

---

1 全國網絡安全標準化技術委員會秘書處，《網絡安全標準實踐指南—智能體部署使用安全指引》(2026)，https://www.tc260.org.cn/portal/article/2/71c613fd3db34b6a9da62f25b8219733

2 根據《私隱條例》第2（1）條，「資料使用者」指獨自或聯同其他人或與其他人共同控制個人資料的收集、持有、處理或使用的人。

代理式AI的風險

從保障個人資料私隱的角度而言，代理式AI的風險一般比傳統AI聊天機械人更高。例如：

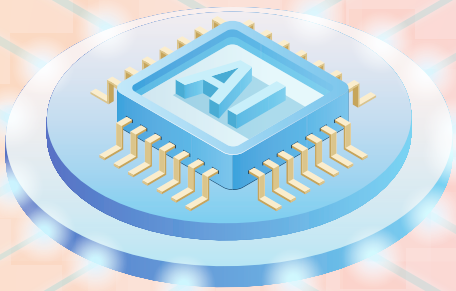
- **廣泛的存取權限：**代理式AI的預設存取權限一般較高，可存取檔案、電郵、帳戶憑證及瀏覽器內儲存的內容。若存取權限設定缺乏嚴格限制，代理式AI可能接觸到大量個人資料（例如儲存於記憶體的個人資料），不僅會增加第三方未經授權查閱或複製個人資料的風險，亦可能因代理式AI錯誤理解用戶指令而導致資料遭意外刪除。

- **系統漏洞：**鑑於代理式AI擁有高度的存取權限，可接觸多個系統及資料來源，一旦在系統設計或安全控制上出現漏洞，將對個人資料的保安構成重大風險。

- **插件（Plugins）或技能（Skills）漏洞：**若代理式AI允許用戶安裝插件或技能，任何未經嚴格安全審核的插件或技能均可能夾帶惡意程式碼。黑客可藉此入侵並接管用戶帳戶，甚至控制整個電腦系統，從而導致個人資料外洩。

- **資料用途改變：**由於代理式AI經常從多個資料來源匯集及重新整合個人資料，部分代理式AI系統可能使用個人資料訓練大型語言模型，過程中該等個人資料可能被用於超出原先收集資料擬使用的目的或直接有關的目的。

- **多重代理系統的風險：**在多重代理系統中，不準確的個人資料（無論是源於外部資訊或是AI模型出現幻覺而生成的錯誤資料）可能透過各智能體之間的互動而層層傳遞，對資料當事人造成不公平或有損其利益的後果，並增加安全風險。此外，個人資料的流動複雜，若欠缺適當的技術及組織措施，資料當事人或難以行使其查閱或改正個人資料的權利。



如何應對相關風險？

由於使用代理式AI涉及較高的個人資料私隱風險，機構應審慎及負責任地使用代理式AI，並謹記尊重個人資料私隱。以下為如何負責任地使用代理式AI以及確保符合《私隱條例》規定的實務建議：

1 避免過度或任意收集個人資料供代理式AI使用

保障資料第1(1)原則規定，資料使用者須為直接與其職能或活動有關的合法目的收集個人資料，而所收集的資料對該目的是必需及足夠的，但不超乎適度。機構應遵循此資料最少化原則，不應隨意收集個人資料供代理式AI使用，亦不應僅因有關資料日後可能有用而收集該等資料，特別是敏感性的個人資料，例如香港身份證號碼、銀行帳戶號碼及密碼等。在存取控制方面，機構應就代理式AI為特定目的而可存取的資訊及系統，訂立並實施清晰的分隔措施。

2 就代理式AI處理個人資料保持透明度

提高透明度對於建立公眾對使用代理式AI的信任至關重要。為遵從保障資料第1(3)原則[3]下的通知規定，機構應向資料當事人清楚說明它們使用代理式AI處理個人資料，例如在收集個人資料聲明中加入相關資訊。此外，根據保障資料第5原則，機構應在其私隱政策聲明中列明所持有的個人資料種類以及使用或擬使用個人資料的主要目的，並須採取所有切實可行的步驟，確保任何人能夠得悉該等政策及實務。機構亦可考慮於其私隱政策聲明中述明儲存或處理個人資料所在的地區及相關安排，以進一步提升透明度。

3 確保代理式AI處理的個人資料的準確性

機構須根據保障資料第2原則，採取所有切實可行的步驟以確保代理式AI處理的個人資料的準確性，例如使用思維鏈（chain of thought）[4]及檢索增強生成（retrieval-augmented generation）[5]、針對特定情境作出微調（context-specific fine-tuning）[6]及進行人為審核（human review），以減低所生成的輸出包含虛構或不準確個人資料的風險。

4 訂定適當的保留時限

當代理式AI系統保存包括個人資料的對話記錄、快取資料或長期記憶時，機構須根據保障資料第2(2)原則，採取所有切實可行的步驟，確保個人資料的保存時間不超過貫徹該資料被使用於或會被使用於的目的（包括任何直接有關的目的）所需的時間。機構應訂定個人資料的最長保留時限，並採取措施刪除不再需要的個人資料。

---

3 保障資料第1(3)原則規定資料使用者須採取所有合理地切實可行的步驟，告知相關資料當事人該資料將會用於甚麼目的、該資料可能移轉予甚麼類別的人、資料當事人有責任提供該資料抑或是可自願提供該資料，以及資料當事人要求查閱及改正該資料的權利。

4 根據 Google Research（即Google的研究部門），思維鏈是一種提示方法，透過將多步驟問題分解為中間步驟，以提升語言模型的推理能力。

5 檢索增強生成透過加入一個提供接地數據的資訊檢索系統來增強大型語言模型能力，以提高該大型語言模型在特定用例或領域的表現。

6 微調是採用以大型和通用數據集訓練的AI模型，並使用其他特定數據對其進行更新／調整以實現特定目的或需求的過程。

5 確保未經同意不會將個人資料用於新目的

根據保障資料第3原則，如未獲有關的資料當事人的訂明同意，不得將個人資料用於新目的[7]。為確保代理式AI不會在未獲資料當事人訂明同意的情況下，將個人資料用於原先收集目的範圍以外的目的，機構應清楚界定代理式AI收集及處理個人資料的目的，並明確訂明在何等情況下須有人為監督。

6 確保代理式AI系統中個人資料的安全

為符合保障資料第4原則，保障代理式AI系統中的個人資料，避免該等資料受到未經授權或意外的查閱、處理、刪除、喪失或使用，機構在使用代理式AI時可採取下列建議安全措施，並在代理式AI所作出的決定可能對個人產生重大影響時採取「人在環中」[8]的策略：

(a) 使用從官方渠道取得的代理式AI軟件的最新版本，避免使用第三方版本或陳舊版本，以減低因為系統漏洞未被修補而發生資料外洩事故；

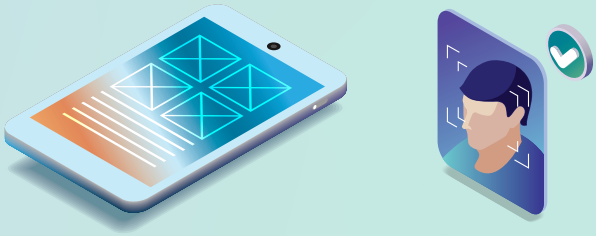
(b) 採取足夠措施確保系統安全及數據安全，例如將代理式AI的運行環境隔離於本機裝置或伺服器，加強網絡控制，嚴格控制互聯網暴露面，降低權限，建立有效的防護機制；

(c) 審慎安裝及使用插件（plugins）或技能（skills），核實相關程式為官方最新已安全驗證版本，以確保程式安全性；審視程式是否含有惡意代碼，如不確定程式的安全性，應避免使用；

(d) 授予代理式AI完成任務所需的最小權限，避免授予代理式AI使用管理員帳號的權限；考慮為機構內部不同部門設定代理式AI的存取權限；小心考慮所涉及個人資料的性質及敏感性，不應隨便向代理式AI提供個人資料，特別是機密或敏感的個人資料（例如身份證明文件、銀行戶口號碼及密碼等）；

(e) 採用大型語言模型安全護欄，在部署模型時於「前、中、後」各階段加入一套完整的技術、流程與治理機制，減低個人資料外洩風險；及

(f) 建立機制以確保代理式AI的可追溯性及可審核性，例如為系統配備日誌記錄功能，以記錄智能體的操作、操作時間、所使用的工具、存取的資料，及決策理由等。



7 根據保障資料第3(4)原則，「新目的」指在收集該資料時擬將該資料用於的目的（或直接與該目的有關的目的）以外的任何目的。

8 「人在環中」是指人類決策者在決策過程中保留著控制權，以防止及／或減低AI出錯或輸出不當的結果及／或作出不當的決定。

7 確保查閱及改正資料的權利

考慮到代理式AI系統內的資料流轉複雜、機構可能難以識別、尋找及提取其所持有屬個別人士的所有個人資料，或難以追溯不準確的個人資料的來源，機構應選用已採納「貫徹私隱設計」及「預設私隱」原則的代理式AI系統，確保該系統能讓資料當事人行使保障資料第6原則下的查閱資料及改正資料的權利。機構亦應留意代理式AI服務供應商是否已在AI系統的設計階段融入保障資料原則，建立穩健的數據管治機制，並預設風險管理。

8 持續進行風險評估

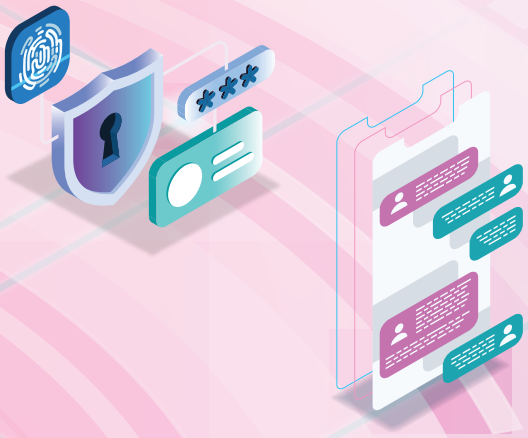
在使用代理式AI之前，機構應就其安全性及可靠性進行測試。在使用期間，機構應持續評估使用代理式AI所涉及的個人資料私隱風險，按代理式AI的用途、所涉個人資料的性質及敏感程度、操作的可逆轉性，以及對個人可能造成的影響，相稱地採取適當的管治及監督措施，並對代理式AI提出執行高風險操作的要求保持警覺。如代理式AI作出的決定可能對個人產生重大影響，包括高風險、不可逆轉或非典型的操作（例如批准向第三方發送個人資料、永久刪除個人資料及修改系統配置等），機構應採取「人在環中」的策略，在決策過程中保留作出決定的最終控制權，以防止及減低AI的錯誤及自動化偏見。

9 分配明確責任及提供培訓

機構應建立具足夠資源、專業知識及決策權的內部管治架構，並在機構內分配職責，以確保個人資料私隱在代理式AI的實施及使用過程中獲得適切保障。在委聘外部服務供應商時，機構應採取合約規範方法或其他方法，確保資料處理者遵從《私隱條例》下有關資料保留（保障資料第2原則）及資料保安（保障資料第4原則）的相關規定。此外，機構應為所有相關人員提供足夠的培訓，以提高他們對使用代理式AI所涉及的個人資料私隱風險的意識，以及管理該等風險的能力。

上述第1、3、4、5及6項建議亦適用於在使用代理式AI中處理個人資料的個人用戶。

為落實本指引所載的建議，機構及個人用戶可參考附件A的安全檢查清單，當中詳列用戶在使用代理式AI時可採取的實際步驟，以保障個人資料私隱。



附件A: 安全檢查清單[9]

檢查清單	適用於	
	機構	個人
評估階段		
☐ 界定代理式AI收集及處理個人資料的目的	❖	❖
☐ 全面了解代理式AI的有關知識	❖	❖
☐ 選擇內置安全防護機制的代理式AI	❖	❖
☐ 不使用存在顯著不合理安全風險的代理式AI	❖	❖
☐ 審視擬部署使用的代理式AI項目的近期動態	❖	❖
準備階段		
☐ 於收集個人資料聲明及私隱政策聲明清楚說明有關使用代理式AI處理個人資料的相關資訊	❖	
☐ 從官方渠道取得代理式AI安裝檔案	❖	❖
☐ 選擇合適的部署方式，並落實相應安全防護措施	❖	❖
☐ 選擇合適的代理式AI運作相關的大型語言模型	❖	❖
☐ 檢查代理式AI安裝檔案的真確性和完整性	❖	❖
☐ 通過安全工具、安全服務等方式增強防護能力	❖	❖
☐ 建立穩健的內部管治架構，並為相關人員提供足夠的培訓	❖	
部署階段		
☐ 不通過來歷不明的一鍵部署等自動化指令碼進行部署	❖	❖
☐ 對代理式AI的網絡暴露面進行安全配置	❖	❖
☐ 開啓代理式AI的日誌記錄功能，並對該功能進行配置	❖	❖
☐ 安裝前檢查插件來源的可靠性	❖	❖
☐ 不使用操作系統管理員權限操作代理式AI，授予代理式AI最小權限	❖	❖
☐ 限制代理式AI可存取的資訊及系統範圍	❖	❖
☐ 提前定義高風險操作，並通過代理式AI安全機制進行管理	❖	❖
使用階段		
☐ 加強對技能（skills）使用的安全防護	❖	❖
☐ 加強對個人資料及敏感數據的安全防護	❖	❖
☐ 在符合《私隱條例》的情況下使用代理式AI獲取網絡訊息服務或參與網絡社交媒體討論	❖	❖
☐ 對安裝與配置情況進行複查	❖	❖
☐ 定期檢查代理式AI開放對外介面的必要性及安全性	❖	❖
☐ 定期對代理式AI所處環境內的重要資料進行備份	❖	❖
☐ 定期手動查閱及管理長期記憶檔案記錄的內容	❖	❖
☐ 及時撤回敏感權限，定期清理不再使用的技能（skills）與對話記錄	❖	❖
☐ 留意警示訊息並進行核查及處理，及時消除安全風險	❖	❖
☐ 留意與代理式AI相關的安全公告，及時採取相應措施	❖	❖
停用階段		
☐ 在解除安裝前停止使用代理式AI	❖	❖
☐ 解除安裝前匯出或轉存需要保存的數據	❖	❖
☐ 執行環境清理操作	❖	❖

[9] 附件A所載安全檢查清單乃經參考全國網絡安全標準化技術委員會秘書處於2026年7月發布的《網絡安全標準實踐指南—智能體部署使用安全指引》而擬備。

電話：2827 2827
傳真：2877 7026
地址：香港灣仔皇后大道東248號大新金融中心13樓1303室
電郵：communications@pcpd.org.hk

私隱專員公署網頁
pcpd.org.hk

下載本刊物

免責聲明

本刊物所載的資訊和建議只作一般參考用途，並非為法例的應用提供詳盡指引，亦不構成法律或其他專業意見。私隱專員並沒有就本刊物內所載的資訊和建議的準確性或個別目的或使用的適用性作出明示或隱含保證。相關資訊和建議不會影響私隱專員在《個人資料（私隱）條例》下獲賦予的職能及權力。

二零二六年八月