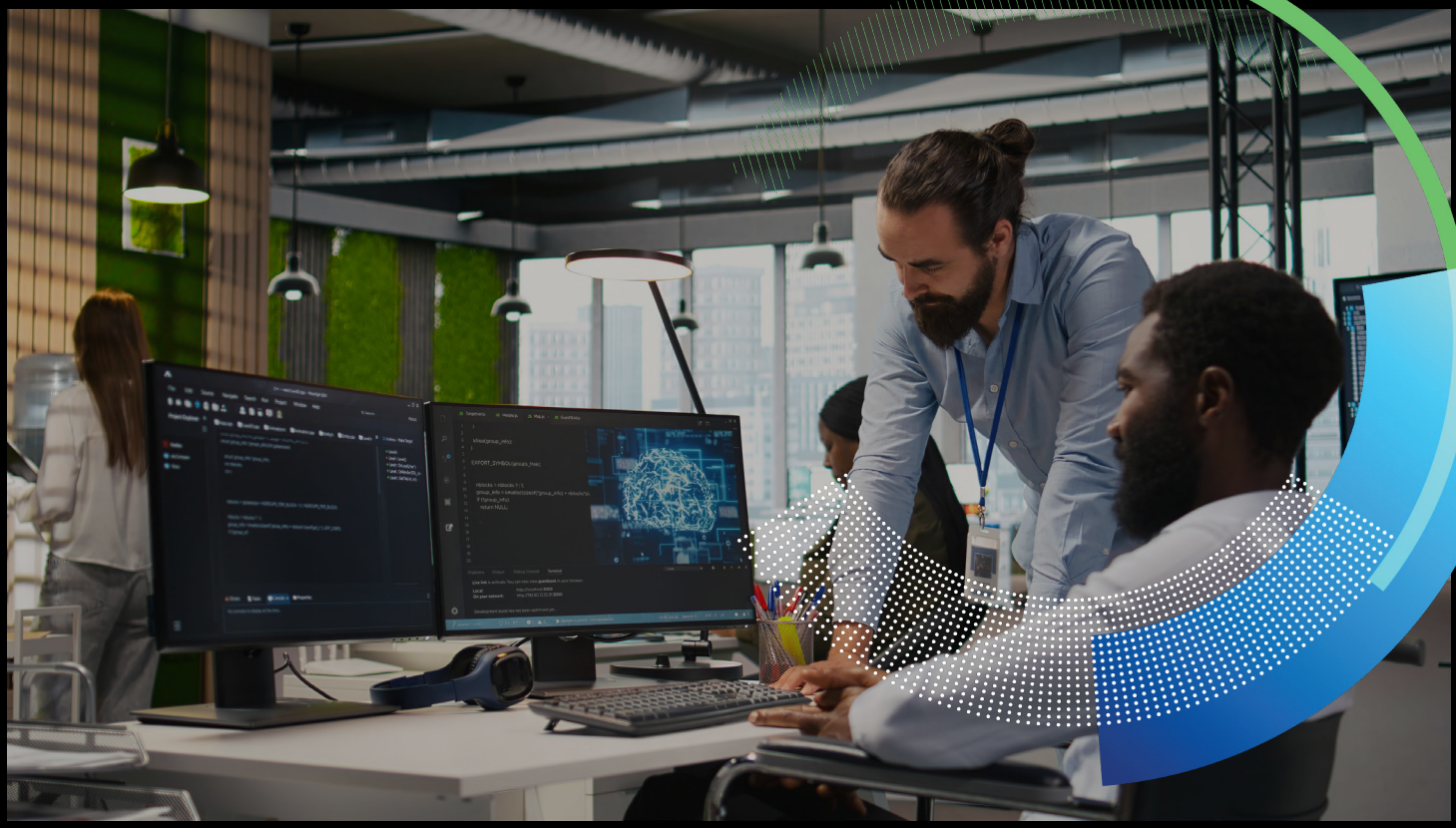
指引

proofpoint.

現代企業的人工智慧安全與治理

人工智慧時代敏感資料保護指南



前言 勇敢迎向新世界

2025 年 1 月。知名駭客在惡名昭彰的網路犯罪論壇上發文，聲稱有驚人的發現。他吹噓入侵了 OmniGPT 防禦系統，該人工智慧平台深受逾三萬名使用者所信賴。駭客自稱的戰利品，是高達三千四百萬則人工智慧聊天訊息與六千個機密檔案連結。據稱外洩的資料包括 API 金鑰、加密錢包、帳戶憑證、三萬個電子郵件地址，以及數千個電話號碼，並已流入暗網交易。[1]

人工智慧正在迅速重塑社會和產業，改寫創造力、工作和創新的規則。然而，伴隨其變革性益處而來的，卻是複雜的新型資料安全挑戰。

其獨特風險在於，人工智慧工具可能收集敏感資訊，並在之後將其洩露。

OmniGPT 攻擊不僅是一場入侵，更成為駭客的號召，以及安全從業人員的警鐘。這次攻擊的驚人數字說明，沒有完善防護，即使是最先進的人工智慧系統也可能淪為網路犯罪者的寶庫。

本指南：

- 探討現代企業因不安全使用人工智慧而面臨的新興資料安全風險
- 揭示人工智慧安全治理不足，如何使這些風險演變成實際的資安攻擊
- 提出強而有力的人工智慧安全治理框架
- 介紹多種工具、技術與最佳做法，協助在新人工智慧時代保障應用，守護資料安全

1. SecureWorld。「OmniGPT 資料外洩事件，三萬名使用者與數百萬則聊天訊息遭曝光。」2025 年 2 月。

第 1 章

人工智慧缺乏安全與治理所帶來的資料風險

人工智慧正在改變組織創造、管理和使用資料的方式。企業 IT 環境中的強大新增元素包括：生成式人工智慧 (GenAI) 工具，例如 ChatGPT、企業級人工智慧工具，例如 Microsoft Copilot，以及人工智慧代理。這類工具有望大幅提高生產力和營運效率，但同時也帶來了一系列新的資料安全風險。與傳統軟體應用相比，人工智慧工具處理敏感資料的方式更為不透明。這樣的問題會造成舊有資安防護難以因應的新型攻擊途徑。

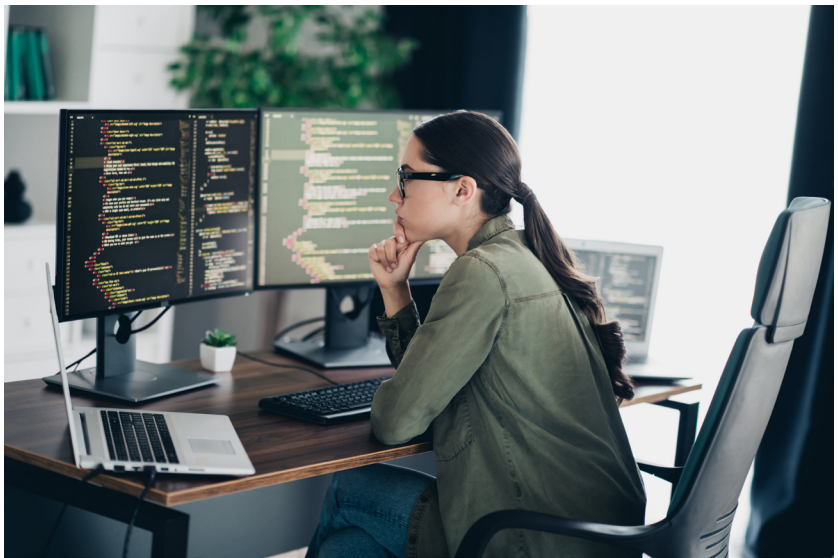
本章分析了在缺乏安全措施與治理規範下使用人工智慧所引發的資料安全風險，以及已出現的破壞性後果，即使技術仍在不斷演進。

轉型效益與重大風險並存

現代企業爭相部署人工智慧來改造其業務流程。但快速普及也為敏感資料的意外洩漏埋下了隱患。這些資料可能是有價值或機密的資訊，例如智慧財產權 (IP)、個人識別資訊 (PII)、支付卡資訊 (PCI) 或受保護的健康資訊 (PHI)。在醫療保健和金融服務等高度監管的產業中，這類資料的遺失代價尤其高昂。

人工智慧造成的意外資料遺失可能以多種形式發生：

- 使用者提示注入：員工可能會將專有資料輸入或將機密檔案上傳至公共生成式人工智慧工具，結果導致洩漏風險。
- 模型訓練或調校過程：企業資料通常用於訓練或微調自訂的大型語言模型 (LLM) 以及智能體型 (即 agent-based) 人工智慧應用。同時也用於檢索增強生成 (RAG) 系統。這些模型存在一項風險：可能會在後續互動中意外洩露資訊，尤其是在面對設計巧妙的提示查詢時。
- 過度授權的人工智慧應用：當客製化、企業級或智能體型人工智慧應用被賦予過度的存取權限時，它們可能觸及大量企業資料，進而增加意外洩露的風險。例如，被賦予內部文件讀取權限的人工智慧代理，在提示防護設定不當的情況下，可能會洩漏機密產品資訊。若缺乏嚴格的控制，人工智慧甚至可能成為難以稽核或約束的「超級使用者」。



下圖展示了這些風險。

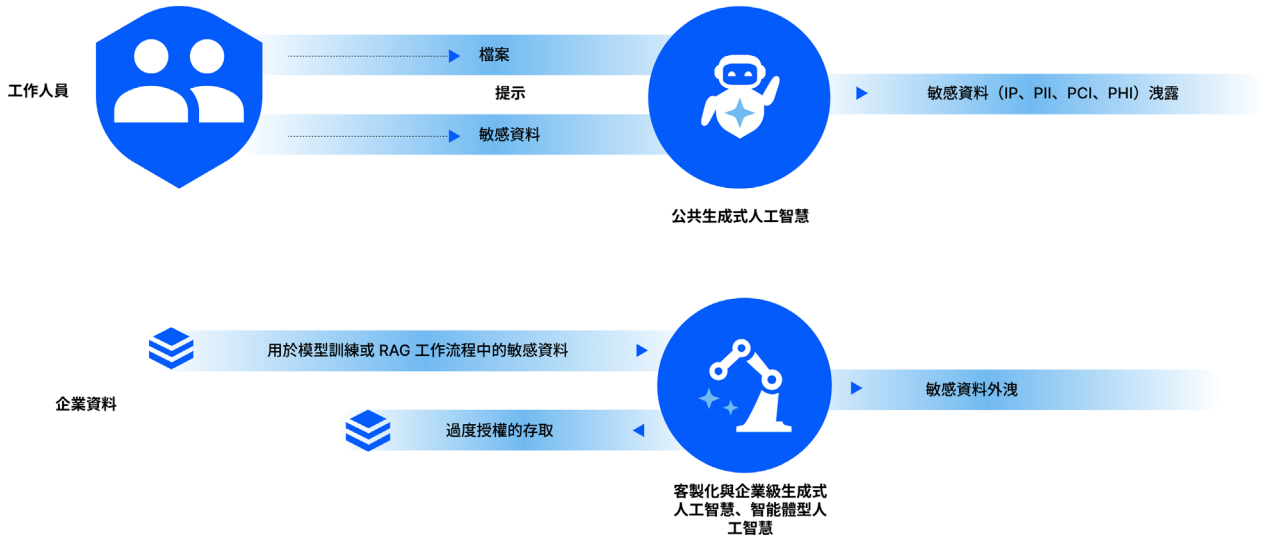


圖 1：人工智慧造成的意外資料遺失可能以多種形式發生。

新型攻擊技術，舊有目的不變

意外資料遺失並非唯一的風險類別。威脅行為者已經開發出專門利用人工智慧工具架構與行為特性的攻擊技術。雖然攻擊技術不斷翻新，其核心目的仍舊是為了財務利益而竊取敏感資料。

凡以大量企業資料進行訓練，或持續存取這些資料的人工智慧應用，特別容易成為攻擊目標。常見攻擊包括：惡意設計的提示、在提示中偷加入指令、故意污染訓練資料，以及各種記憶體竄改手法。這些工具透過巧妙的提示或惡意訓練資料，覆蓋人工智慧模型的預期行為，甚至存取其記憶體以竊取敏感資料。威脅行為者利用客製化與智能體型人工智慧應用的架構弱點進行滲透。這些應用通常仰賴 API 與第三方外掛模組，進行功能擴充與整合。當外掛或 API 模組被賦予過多操作權限、驗證機制不足，或存取過量資料時，便會成為系統中的重大安全破口。

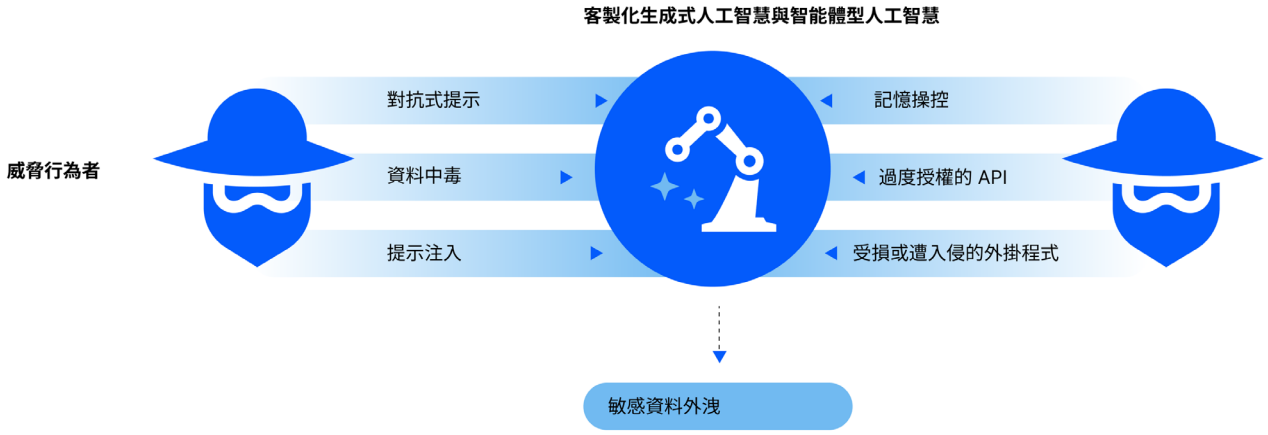


圖 2：威脅行為者已開發出精密的攻擊技術，專門鎖定客製化與智能體型人工智慧應用。

人工智慧時代的企業內部安全風險

在傳統安全模型中，內部人員一向被視為 主要風險來源，無論是因疏忽、帳戶遭到入侵，或是蓄意惡意行為。然而，在人工智慧時代，這些風險進一步被放大，主要表現在以下幾個面向：

- 因內部人員的疏忽，可能會將敏感資料輸入未經授權的工具，或不當使用人工智慧功能，導致違反企業政策。
- 帳戶被盜用的內部人員，攻擊者便能利用其人工智慧存取權限進行資料搜尋、竊取或篡改。
- 惡意內部人員可能會刻意濫用人工智慧存取權限，向人工智慧工具輸入假資料，或將其用於攻擊內部系統。

智能體型人工智慧的企業內部安全風險

人工智慧代理能夠在動態環境中自主決策和行動。智能體的迅速崛起，正在催生一個新型工作空間，於其中與人類共同處理敏感資料。雖然能自主行動，但也繼承了人類的風險。過度授權不僅發生在人類帳戶，智能體也可能因此成為安全破口。在決策過程中被賦予過度自主時，可能會採取超出授權範圍的行動，甚至出現不可控的結果。這種情況稱為過度自主。

如果沒有適當的防護措施和正確配置的權限，人工智慧代理可能會處理遭篡改的指令、存取惡意的 URL、洩漏敏感的資料、使用遭入侵的外掛、共享帳戶憑證，或以非預期的方式操控系統和資料。在不同情境下，智能體系統可能會像粗心的員工一樣操作錯誤，或像帳戶遭入侵的內部人員般被利用。

一個令人擔憂的例子是，人工智慧代理未經授權就對運作中基礎設施進行修改，並清除了某軟體公司的生產資料庫。據報導，該錯誤刪除了超過 1,200 位高階主管與 1,190 家公司的資料。[2]

人工智慧驅動的資料遺失代價高昂

就像由人為疏失造成的資料外洩一樣，人工智慧驅動的資料洩漏同樣可能帶來嚴重後果。這些包括經濟損害、聲譽損害以及客戶信任的流失。

顯示其可能帶來極其高昂的財務代價。2025 年，資料外洩的平均成本高達 444 萬美元。惡意內部人員事件的平均成本上升至 492 萬美元。[3]

資料外洩或系統遭破壞可能引發嚴重的監管後果。包括罰款、稽核與承擔法律責任，尤其是在《一般資料保護規範》(GDPR)、《健康保險可攜性與責任法案》(HIPAA)以及《加州消費者隱私法案》(CCPA) 等框架下。

例如，根據 GDPR，罰款最高可達 2,000 萬歐元或全球營收的 4%。[4] 違反 HIPAA 可能面臨每個違規類別每年最高 150 萬美元的罰金。[5] 此外，其他專注於人工智慧的監管規範正在逐步出現。其中包括歐盟《人工智慧法案》和美國《人工智慧行政命令》。

簡言之，企業必須主動預防人工智慧驅動的資料遺失，以避免其演變成代價高昂的商業危機。

2025 年，資料外洩的平均成本高達 444 萬美元，



而惡意內部人員事件的平均成本則上升至 492 萬美元。

資料來源：IBM

2.《財富》。「一款人工智慧驅動的編碼工具刪除了某軟體公司的資料庫，隨後為『我的災難性失誤』道歉。」2025 年 7 月。
3.IBM。Cost of a Data Breach (資料外洩的成本) 2025 年報告。2025 年 7 月。
4.IT Governance.「GDPR Fines & Penalties」(GDPR 罰款和處罰)。2025 年 10 月審查。
5.美國醫學會。「HIPAA violations & enforcement」(HIPAA 違規與執法)。2025 年 10 月審查。

第 2 章

治理不善如何引發人工智慧驅動的資料遺失

公開、客製化及企業級生成式人工智慧，以及智能體型人工智慧應用正逐漸成為現代組織中的關鍵工具。麥肯錫在 2024 年的一項調查顯示，已有 78% 的受訪者表示其組織至少在一個業務職能中使用人工智慧，這一比例相較於前一年的 55% 有明顯提升。[6]

2024 年的一項調查顯示，受訪者中有 78%

表示他們的組織至少在一個業務職能中使用人工智慧。

資料來源：麥肯錫

人工智慧工具的普及速度和廣泛的資料存取能力，使其對安全團隊構成了日益嚴峻的挑戰。缺乏明確的政策、可視性和管控措施，就有可能因自身使用人工智慧而洩漏敏感資料。本章將說明人工智慧安全治理方面的缺陷，這些缺陷可能會導致嚴重的資料遺失事件。

缺乏對公共生成式人工智慧使用的可視性和管控措施。

麥肯錫 2024 年的調查顯示，71% 的受訪者表示其所在機構經常使用生成式人工智慧，比同年稍早的資料上升了 6 個百分點。[7] 高知名度的公共生成式人工智慧工具包括 ChatGPT、Gemini、Perplexity 和 Claude。如果沒有管理機構來監督和控制這些工具的使用，員工很容易洩漏有價值或機密的資訊。

例如，在廣為報導的一起案件中，三星員工透過 ChatGPT 洩露公司機密，包括專有原始程式碼。[8]

當生成式人工智慧的使用發生在 IT 部門無法掌控的情況下時，就會失去了對資料流向的掌控。更糟的是，某些公共工具會保留或學習使用者輸入的內容，一旦員工輸入了機密資訊，外洩就成為不可逆的事實。沒有監管、使用者意識與技術防護欄，每一次互動都可能導致資料外洩。

影子人工智慧工具的使用

另一個相關問題是使用未經 IT 或安全團隊審查或授權的「影子」人工智慧工具。員工通常是出於善意使用這些工具。不過，若無監管會造成重大威脅。影子人工智慧工具可能存在不明確或不安全的資料使用策略，以及安全性較弱等問題。與其他第三方工具進行整合時，也有可能導致企業資料見光。這些工具並不在官方 IT 架構之內，因此組織無法對其使用進行監控、稽核或限制。

未經授權或過度授權的存取

即使是已授權的工具，若被授予對內部系統和資料的廣泛或不當使用權限，也同樣存在風險。例如，人工智慧代理通常需要廣泛存取企業資料才能做出決策。若缺乏強而有力的身分和存取管理 (IAM) 策略，就可能被賦予過度的權限。這表示人工智慧工具可以存取、再利用，或輸出超出原本設計範圍的資料。

6.麥肯錫。「The state of AI: How organisations are rewiring to capture value」(人工智慧的現況：組織如何重塑架構以獲得價值)。2025 年 3 月。
7. 同上。
8.Mashable。「Whoops, Samsung Workers Accidentally Leaked Trade Secrets via ChatGPT」(三星員工不慎通過 ChatGPT 洩露了商業機密)。2023 年 4 月。

同樣的風險也存在於使用 Open Authorisation (OAuth) 權限與企業資料互動的應用。隨著 OAuth 應用程式的發展，特別是人工智慧能力的引入，組織必須持續檢視其權限設定，以及資料處理與管理方式。

換句話說，在缺乏嚴格存取控制的情況下，組織很容易受到內部洩漏與意外暴露的威脅。

資料安全態勢薄弱，敏感資訊分類不足

為保障資料安全，組織必須知道所有資料存放的位置，以及其中有多少屬於敏感資料。遺憾的是，許多企業仍然使用不足的資料探索與分類系統，導致敏感資訊難以正確識別與保護。根據 Gartner 的預測，到 2027 年，60% 組織由於資料治理框架缺乏統一性，無法實現其人工智慧使用案例的預期價值。[9] 而且，人工智慧工具需要明確的分類，才能分辨公開的、內部的、敏感的資料，如果治理上有缺口，就可能導致嚴重的資料外洩。



有 60%

的組織因資料治理框架的不一致，Gartner 預測到 2027 年將無法從人工智慧案例獲得預期價值。

資料來源：Gartner

即使組織對資料進行分類，若整體資料安全防護姿態薄弱，仍可能導致資料外洩。例如，公司可能會授予 Microsoft Copilot 權限，以便索引 SharePoint 和 OneDrive 中的資料。因此，一個簡單的錯誤設定，就有可能讓高階主管的個人雲端硬碟暴露給整個公司。

缺乏對訓練資料的可視性

隨著組織部署客製化、智能體、或 RAG 應用程式，這些應用所依賴的訓練或增強資料集也成為一個潛在的風險來源。在模型訓練或 RAG 流程中，系統通常會從公司內部的資料庫和文件儲存庫提取資料來使用。若缺乏可視性和控管，資料管道很容易將敏感或受管制的資料納入處理流程。例如，2025 年 2 月，安全研究人員在一個用於大型語言模型訓練的公開資料集中發現了超過 12,000 個有效的 API 金鑰和密碼，突顯隱藏的重大安全風險問題。[10]

有限度採用合成資料

合成資料集是人工生成的資料集，提供模擬真實資料功能，但不包含任何實際的敏感資訊。它們是降低模型訓練過程中隱私風險的有效方法。目前市面上已有越來越多的商業和開源工具可以生成各種類型的合成資料。

儘管合成資料工具的選擇越來越多，但它的使用仍未普及。根據 Gartner 的說法，「儘管在合成資料方面注入大量投資，但使用者的質疑、對真實資料的依賴，以及缺乏標準、信任與認知，仍然阻礙了它的普及。」[11]

9.Gartner。「Adopt a Data Governance Approach That Enables Business Outcomes」(採用能夠實現業務成果的資料治理方法)。2025 年 7 月
10.The Hacker News 平台。「12,000+ API Keys and Passwords Found in Public Datasets Used for LLM Training」(在用於大型語言模型訓練的公開資料集中發現了超過 12,000 個有效的API 金鑰和密碼)。2025 年 2 月。
11.Gartner。 Emerging Tech Impact Radar (新興技術影響雷達)。2024 年。

第 3 章
人工智慧安全治理框架

隨著人工智慧工具逐漸融入業務流程，企業需要採取積極主動、結構化的方法來治理人工智慧安全。這表示需要制定規範人工智慧部署與使用的相關角色、政策及監督機制。本章為現代組織提供了一個有效的人工智慧安全治理框架。

推動專責的人工智慧安全治理角色之實施

傳統的資料治理與安全職能（例如資訊安全長與資安架構師）專注於保護基礎設施、執行存取控制及確保合規性等優先事項。但是人工智慧屬於一種截然不同的安全範式。其焦點從被動防護轉向主動監督與動態控制。

因此，隨著人工智慧帶來新的風險與挑戰，自然而然衍生出新的資料安全與治理角色。人工智慧成熟的組織，確實已經開始設立如表中所示的相關角色。

表 1：以下是人工智慧安全治理的角色示例

角色	角色描述範例
人工智慧長 (CAIO)	負責監督組織內人工智慧系統的部署與治理。在創新與監督之間取得平衡，確保人工智慧安全運用企業資料。
人工智慧治理主管和人工智慧合規長	確保人工智慧系統符合法律要求、國際標準，如 ISO/IEC 42001 以及新興的監管規範，如歐盟《人工智慧法案》與美國《人工智慧行政命令》。
提示和模型合規分析師	審查使用者輸入的提示，分析模型輸出是否涉及敏感內容，並確保全程符合組織的人工智慧使用政策。
人工智慧隱私和安全架構師	設計並實施專為人工智慧系統量身打造的安全架構。著重於保護訓練資料、推理流程以及已部署模型。

制定可接受的人工智慧使用政策

可接受的人工智慧使用政策對組織整體人工智慧治理計畫至關重要。該政策規範員工及合作夥伴對人工智慧工具之使用方式。必須涵蓋組織內所有正在使用的人工智慧類型，包括客製化應用、公共生成式人工智慧工具、企業級人工智慧工具，以及智能體型人工智慧。

該政策明確規範使用者責任，以及在何種情境下屬於適當的人工智慧使用。這些情境可能包括撰寫報告或電子郵件、製作摘要或進行非機密性主題研究。該政策亦須明確說明人工智慧之禁止用途。禁止使用的情境包括：將敏感或重要資訊輸入人工智慧工具、使用未經核准的應用程式，或刻意規避存取控制。

提升人工智慧使用的可視性與監管能力

可接受的使用政策唯有在能夠被有效執行時，才能真正發揮效用。因此，掌握人工智慧工具的使用狀況與資料流向，是治理的核心要素之一。主要能力包括：

- 檢視使用者與生成式人工智慧的互動，並透過資料遺失防護 (DLP) 工具，監控公共與客製化生成式人工智慧工具的使用情形。收集洞察，例如最活躍使用者、最常用網站，以及所提交敏感資料類型等資訊。偵測是否有使用影子人工智慧工具的情況。
- 加強對高風險用戶的監控。利用內部威脅管理 (ITM) 工具的行為分析來識別具有人工智慧誤用模式的使用者。提升高風險角色或部門使用者的監控和 DLP 策略。
- 建立使用者與生成式人工智慧互動的稽核追蹤。使用提示記錄工具保存涉及內部或第三方生成式人工智慧工具的提示和回應。在合規調查中，應包含使用者識別資訊和時間戳記。
- 識別和分類所有敏感資訊。，並透過資料安全態勢管理 (DSPM) 工具，建立內部與雲端環境的結構化與非結構化資料清單。將資料依等級分類（例如，公開、內部、機密、限制），並加上中繼資料標籤以便治理。
- 偵測自訂模型訓練和 RAG 工作流程中使用的敏感資料。假設您已經對資料進行了分類，請使用 DSPM 工具掃描訓練集，並偵測輸入到自訂模型和 RAG 工作流程中的敏感資料。
- 追蹤資料沿襲。採用統一的資料安全解決方案或專用的資料沿襲工具，追蹤模型訓練與推理過程中所使用資料的來源，並將資料沿襲從來源到輸出的完整流程視覺化，其中包含資料轉換步驟與政策決策。

採用科技驅動的政策強制執行

除了提高可視性和監控能力外，各組織還必須納入科技驅動的政策強制執行。這代表透過技術控制措施，即時強制執行使用標準。重要功能包括：

- 視覺化和管理人工智慧工具的資料存取。使用 DSPM 工具，清楚呈現人工智慧工具在環境中所存取的資料。取得未經授權使用敏感資料的警示。為企業資料庫連接至自訂模型訓練或 RAG 流程時，強制執行核准工作流程。使用 IAM 工具對人工智慧應用強制執行角色為基礎的存取控制 (RBAC) 和屬性為基礎的存取控制 (ABAC)。限制人工智慧工具僅能存取非敏感且已授權的資料。
- 封鎖生成式人工智慧提示中的敏感資料，並透過 DLP 工具或安全 Web 閘道來偵測並封鎖向生成式人工智慧工具發出的提示中包含的敏感資料。監控瀏覽器擴充功能與第三方外掛，避免其略過上傳限制並造成敏感資料外洩。
- 防止企業人工智慧工具分享或處理敏感資訊，並透過 DSPM 工具套用資訊保護標籤，保護企業 Copilot 存取的資料。

推動行為改變

政策與強制執行技術是人工智慧安全治理的核心要素。但永續治理還需要全公司上下的共同承諾。透過以下方式推動安全態度和行為的長期轉變：

- 教育員工正確使用人工智慧。培訓其熟悉企業核准的人工智慧工具與合規使用方式。分享誤用行為的真實案例，強調其商業和法律層面的影響。
- 為高風險使用者自動推送培訓。利用人工智慧使用報告來識別大量與生成式人工智慧工具頻繁互動的使用者。當偵測到危險行為時，觸發即時輔導或測驗。將安全意識培訓納入 DLP 或 ITM 政策違規的補救計畫。



第 4 章
保障人工智慧應用與資料安全

強化治理的同時，組織也要落實嚴謹的人工智慧安全措施，打造多層次防護策略。治理著重於監督、監管和政策執行，而人工智慧安全則涵蓋技術措施，用以防止資料遭未授權存取、外洩或誤用。

更重要的是，人工智慧安全需要多層次策略。它必須同時防護應用與資料，才能形成完整的保護。本章將區分這兩個安全目標，並說明與之對應的最佳實務與技術。

保障人工智慧應用安全

保障人工智慧應用安全包括保護模型、推理過程以及部署環境，避免遭受操縱或破壞。這表示必須同時防護人工智慧系統的輸入與輸出，避免被惡意利用或產生危害。

在確保人工智慧應用安全時，組織所需承擔的責任取決於該應用的基礎。該組織可能直接使用第三方人工智慧工具，將第三方人工智慧工具整合到自身系統中，亦或自行開發專屬的人工智慧應用。以下章節將針對不同的採用模式，說明常見的安全技術。

直接使用第三方人工智慧工具的安全技術

直接使用第三方人工智慧工具時，因供應商已處理主要安全問題，組織的安全負擔最低。即使主要安全由供應商負責，表格中所列的技術仍然對確保安全使用具有價值。

表 2：直接使用第三方人工智慧工具的安全技術

技術	說明
存取控制	定義誰可以與人工智慧互動，互動頻率和互動方式。透過驗證、授權和速率限制規則來保護人工智慧免受惡意使用。
輸入過濾	在使用者輸入到達人工智慧之前，對其進行驗證、清理，並在必要時加以限制。有助於阻擋惡意或不合規的提示。
輸出過濾	能專注於檢查模型生成的結果，避免洩露有害或違反合規的資訊。有助於確保輸出符合道德、法律和政策標準。
安全閘道	作為使用者和人工智慧系統之間的中介。可以利用情境感知來動態地執行策略，例如速率限制、內容審核和存取控制。還可以幫助檢測誤用模式。

整合第三方人工智慧工具的安全技術

當組織整合第三方人工智慧工具時，等同於讓這些工具接觸到組織的資料與使用者資訊。這也可能意味著需要與其他系統、API 及服務相互連接。因此，組織必須承擔更高層級的安全責任。除了前面提到的直接使用第三方工具的技術之外，還應考慮以下其他做法。

表 3：整合第三方人工智慧工具的其他安全技術

技術	說明
滲透測試	透過模擬攻擊，應用於整個人工智慧部署環境，包括 API、模型和端點。有助於在攻擊者發動攻擊之前，先行揭露並修補系統漏洞。
提示強化	包括設計良好的系統提示，這些預先定義的指令，用來決定人工智慧的行為、界限和回應方式。這些有助於模型識別和抵抗操縱企圖。亦能規範智能體人工智慧系統的自主行為，確保在執行任務時安全合規。
紅隊演練	模擬對人工智慧系統的攻擊，提前揭露潛在漏洞。測試如何操縱模型以洩露機密資料或產生危險內容。與滲透測試主要檢驗特定技術漏洞不同，紅隊演練會從人員、流程、控制與技術等面向全面尋找弱點。

建構客製化人工智慧應用的安全技術

客製化人工智慧應用可能包括專屬大型語言模型、RAG 系統和智能體應用。組織必須負責從開發和培訓到部署的整個流程，所以承擔了最沉重的安全責任。除了使用上述技術保護應用程式外，組織還必須在模型訓練中建立健全的防護機制。對抗式訓練透過加入惡意樣本，是強化技術的關鍵。具體描述請見表格。

表 4：建構客製化人工智慧應用的其他安全技術

技術	說明
對抗式訓練	同時使用一般資料與刻意設計來欺騙模型的資料進行混合訓練，讓人工智慧學會抵抗操縱與攻擊。透過訓練模型區分安全輸入和惡意修改的輸入，提高模型的穩健性和安全性。

保障人工智慧使用中的資料安全

客製化與企業級生成式人工智慧，以及智能體型人工智慧通常會在訓練或運行時存取大量企業資料。為了避免資料外洩、誤用或在訓練與推理過程中遭到污染，組織必須確保資料在靜態儲存、傳輸以及使用時都具備安全防護。

以下這些安全措施都是必須考慮的重要面向。無論組織是直接使用第三方人工智慧工具、進行整合，或是自行開發客製化應用程式，這些安全技術都必須被納入考量。

表 5：人工智慧使用中的資料安全保護技術

技術	說明
資料淨室	一種保護隱私的技術，有時會與聯邦式學習搭配使用，讓多個組織能在不共享原始資料的情況下共同貢獻資料，訓練出一個共享模型。各方在安全的沙箱環境中控制自己的原始資料，並且不與他方共享。
資料存取映射和視覺化	提供資料存取者、存取時間及來源的可視性。幫助安全團隊偵測異常情況，並落實完成工作所需的最低權限。
資料探索與分類	掃描內部和雲端環境，以識別和標記結構化、半結構化和非結構化格式中的敏感資料（例如，IP、PII 和 PHI）。確保人工智慧應用中資料安全處理的關鍵。
資料遺失防護	這項技術能過濾或阻擋含有敏感資料的人工智慧輸入，並對輸出結果進行分類與標記。防止洩露個人識別資訊、客戶資料、商業機密和合規敏感條款。
資料遮罩	在非正式作業環境中隱藏資料值，同時保持資料的可用性。防止人工智慧開發和測試過程中敏感資料外洩。
差分隱私	透過在資料集或查詢結果中加入隨機擾動，避免後續被重新識別。這是保護個人資料在模型訓練過程中不被洩露的關鍵。
企業人工智慧與生成式人工智慧資料安全態勢管理	確保企業 Copilot、客製化 LLM 訓練以及 RAG 工作流程所使用的資料庫，能正確配置並安全存取，從而實現安全的人工智慧部署。
聯邦式學習	透過分散式裝置或資料庫進行模型訓練，無需傳輸原始資料。有助於降低集中式資料帶來的風險。

表 5：人工智慧使用中的資料安全保護技術

技術	說明
合成資料	合成資料是人工生成的資料集，模擬真實資料的樣貌，但不包含任何敏感資訊。用於模型訓練，以降低隱私洩露風險，同時維持模型的準確度。
記號化與加密	將敏感資料元素（例如社會安全號碼或信用卡號碼）以記號替代，或透過加密處理，同時保留原有的資料格式。可用於對人工智慧使用的資料集進行匿名化處理。

下圖總結了適用於保護人工智慧應用和資料的技術。

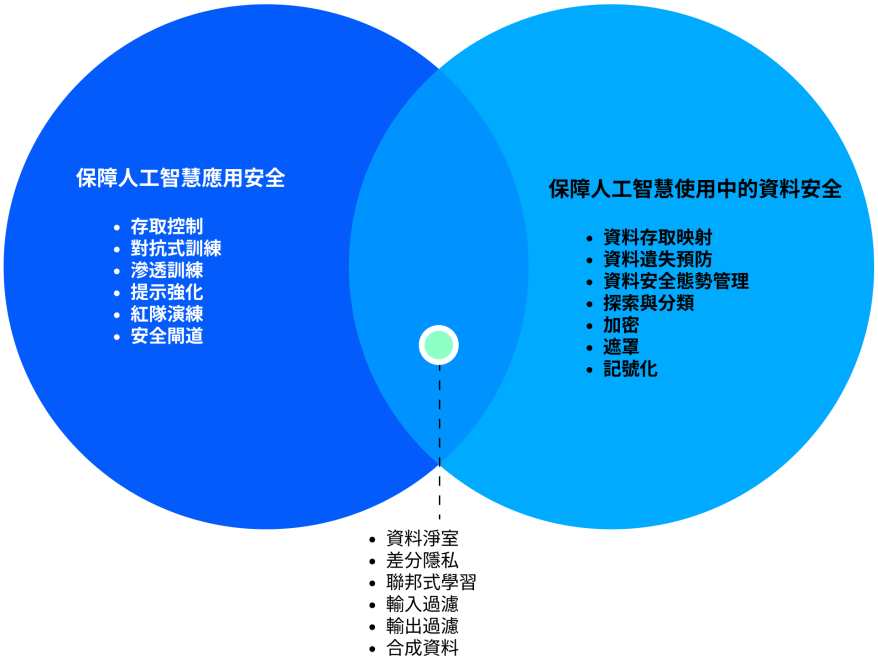


圖 3：雖然人工智慧應用安全與人工智慧資料安全之間可能有部分重疊，但它們其實是不同的挑戰，需要各自專屬的技術與工具來解決。

結論

Proofpoint 作為您建立人工智慧安全與治理的基礎

如本指南所探討的，企業資料安全計畫必須緊急演進，以迎接人工智慧時代。透過應用互補的人工智慧安全層級與治理措施，您的組織能在保護敏感資料的同時，不影響人工智慧驅動的創新與生產力。

結合 Proofpoint ZenGuide™（安全意識教育產品）、Proofpoint Data Security Complete、Proofpoint Secure Agent Gateway 及 Proofpoint AI Data Governance 能為多層次的安全防護方法奠定堅實基礎。

Proofpoint Data Security Complete 是一套整合式解決方案，將 DLP、DSPM、ITM 與資料沿襲集中於單一架構中，提供更精簡的部署、運作與管理。Proofpoint Secure Agent Gateway 是一套專為智能體工作流程設計的解決方案，能保護敏感資訊在各種流程中進出時的安全。這些解決方案相互協作，提供業界首個 同時保護人員與代理資料安全的整合方案。

我們全面解決方案的各組成部分說明如下：

Proofpoint Data Security Posture Management 可發現並分類您雲端及內部環境中的敏感資料，並提升人工智慧所使用資料的安全狀態。偵測企業環境中的人工智慧服務，並在發現敏感資料遭未經授權使用時發出警示。標記資料，以保護 Microsoft Copilot 所存取的資訊。透過偵測敏感資料於訓練及 RAG 工作流程中的使用時機，確保雲端人工智慧服務的自訂 LLM 及應用程式的安全。

Proofpoint Enterprise DLP 顯示誰在使用人工智慧工具，並監控敏感資料是否外洩到公共工具或自訂的大型語言模型中。追蹤超過 600 種人工智慧工具的使用情況，並在端點層面強制執行 DLP 政策。可以阻擋網頁上傳、複製貼上活動以及向人工智慧網站發出的提示。限制在人工智慧提示中包含敏感資料。

Proofpoint Insider Threat Management能夠提供可視性，揭示因疏忽、惡意或帳號遭入侵的使用者所帶來的高風險行為。這可能包括對敏感資料進行不尋常或有風險的互動。透過動態政策，能監控內部威脅，並捕捉相關的中繼資料與螢幕截圖。

Proofpoint Secure Agent Gateway控制人工智慧代理存取資料的方式，監控代理活動，強制執行資料使用政策，並在資料與人員或其他代理共享前加以阻擋或遮蔽敏感資訊。基於採用模型情境協議 (MCP) 架構，將代理控制與全面的資料安全政策統一整合，覆蓋整個環境。

Proofpoint AI Data Governance 專為生成式人工智慧治理而設計，整合了資料分類、資料治理、DLP 以及安全姿態管理等功能。透過識別授權與未經授權的使用情境、套用控制措施防止資料外洩與隱私違規，並以自動化工作流程協助安全團隊與內容擁有者治理存取權限，從而確保企業 Copilot 及人工智慧應用的安全使用。

Proofpoint ZenGuide透過持續學習改變員工行為，以因應不同風險情境。持續偵測風險、提供指導、鼓勵採取更安全的行動並評估成效，以培養安全意識。

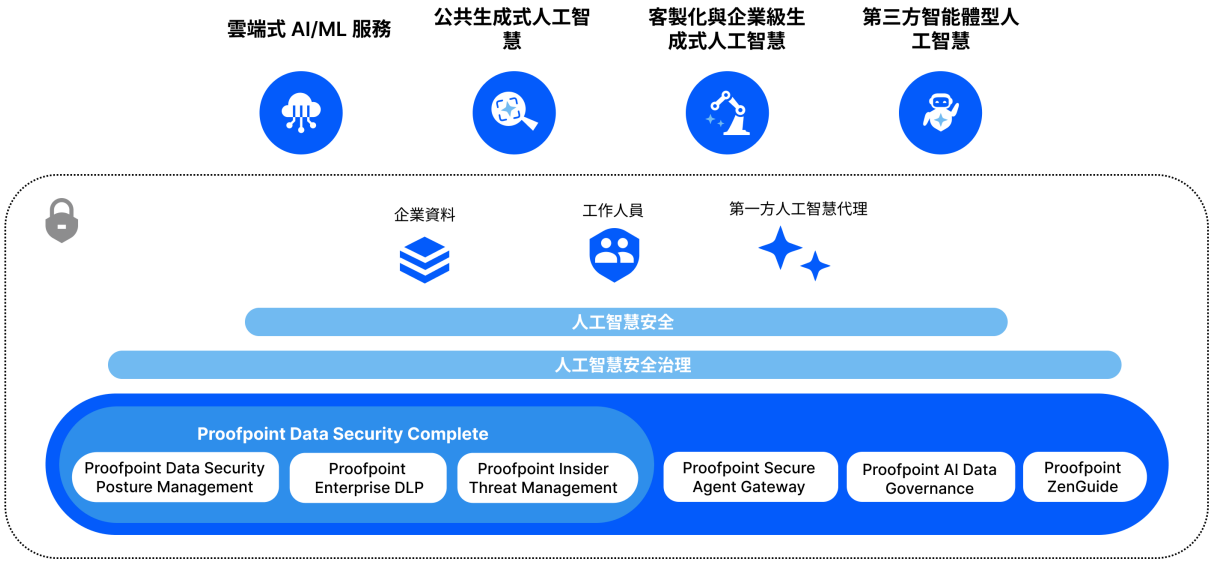
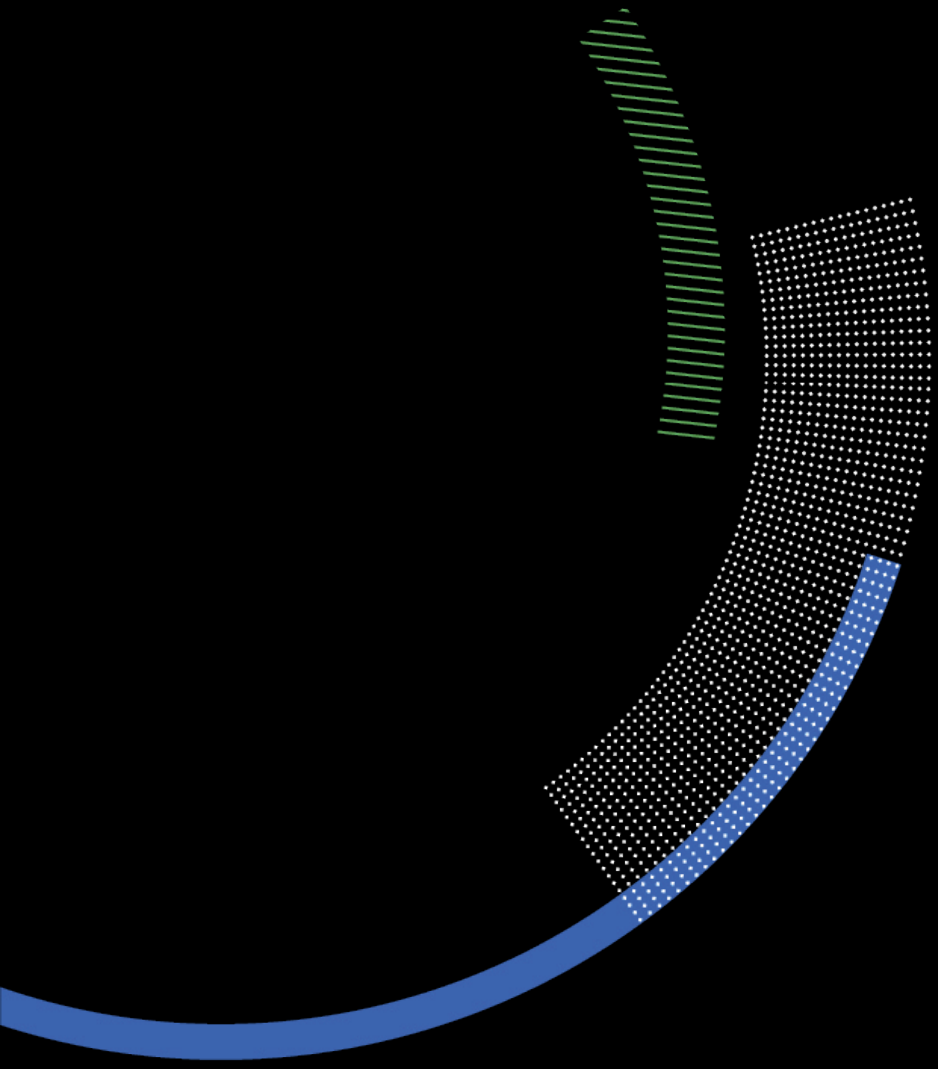


圖 4：Proofpoint 是您多層次人工智慧安全與治理策略的基石。在我們解決方案協助建立的安全環境中，您的員工、人工智慧代理及資料皆可安全地與第三方及公開人工智慧工具互動。

您的下一步行動

- 若要了解 Proofpoint 解決方案如何透過可視性、控制與教育，確保生成式人工智慧的安全採用，請參閱我們的解決方案簡報。
- 如需進一步了解我們的統一資料安全解決方案，請參閱Proofpoint Data Security Complete，或聯絡您的 Proofpoint 代表以便安排展示。



proofpoint®

Proofpoint, Inc. 是全球領先的以人為本及以代理為中心的網路安全公司，致力於保護人員、資料及人工智慧代理於電子郵件、雲端及協作工具間的連接安全。Proofpoint 是《財富》百大企業中的 80 多家、10,000 多家大型企業以及數百萬小型組織的信賴合作夥伴，致力於阻止威脅、防止資料遺失，並在人員和人工智慧工作流程中建立彈性。Proofpoint 的協作與資料安全平台協助各種規模的組織在安全且自信地擁抱人工智慧的同時，保護並提升其員工的能力。詳情請參閱www.proofpoint.com。

與 Proofpoint 連結：LinkedIn

Proofpoint 是 Proofpoint, Inc. 在美國和/或其他國家/地區的註冊商標或商品名稱。本文的所有其他商標皆為其各自擁有者之財產。