AI 驅動的自動化資安管理

財團法人資訊工業策進會人工智慧研究院正工程師 / 許芳甄

關鍵字：自動化資安管理 、零信任架構、身分與存取管理、跨提示注入攻擊

摘要

隨著生成式 AI 與 AI 代理在組織環境的廣泛應用，傳統靜態且仰賴人工配置的身分與存取管理（Identity and Access Management，IAM）已難以應對因模型微調導致的資料孤島扁平化，以及跨提示注入攻擊（Cross-Prompt Injection Attacks，XPIA）等新型態威脅。本文章結合現有研究理論提出一套整合「計畫 – 部署 – 紀錄 – 回應」（Plan-Deploy-Record-Respond，PDRR ）四階段循環的自動化資安管理框架，旨在透過 AI 技術落實零信任架構與資料層級的精細化防護。在計畫階段，系統利用大型語言模型（LLM）與自然語言處理（NLP）將非結構化法規自動轉譯為結構化規則，並透過特定推理框架進行風險識別；部署階段導入機器學習實現高效動態授權，並結合使用者存取行為偵測與控制技術，從根源阻斷資料外洩風險；紀錄階段透過非監督式學習建立行為基準，並利用可解釋性 AI（Explainable AI，XAI）為每次決策生成稽核理由以強化問責制；最終在回應階段，系統能自動執行封鎖或撤銷權杖等緩解措施，並遵守「人在迴路（HITL）」的管理原則。未來研究可聚焦於針對具體業務痛點建構模擬實驗場域，進一步驗證 AI 驅動之自動化資安管理成效。

一、前言

隨著現代產業邁向智慧化轉型，組織普遍部署生成式 AI 與檢索增強生成（Retrieval Augmented Generation，RAG）系統，資料存取需求已呈現即時動態的趨勢。傳統基於軟硬體、網路、系統與資料庫等防護的信任假設無法因應 AI 時代的現況，迫使組織必須全面轉向「零信任（Zero Trust）」架構，以部署「永不信任，始終驗證」的核心原則。

然而，傳統的身分與存取管理（IAM）高度仰賴人工計畫與手動靜態配置規則，在實踐零信任時面臨嚴重的效率瓶頸與合規風險，主要原因如下：

1. 維運低效與審查疲勞：Katta 的研究證實 [1]，「以管理員為中心」的手動配置方式，不僅難以保持判斷基準的一致性，處理大規模動態需求時，效率低下。週期性存取審查導致審查者產生審查疲勞，或因人為操作錯誤亦將引發嚴重的合規疏漏。

2. 常設權限違反需知原則：傳統角色存取控制（Role-Based Access Control，RBAC）模型中，因審查疏漏造成的常設權限風險，讓員工在合法授權下存取不相關機密資料，嚴重違反了最小權限與需知原則。

二、AI 應用情境下的權限控制失效

當 AI 成為資料存取的代理人，權限邊界將被重組。若控制措施僅停留在「身分層」而忽略「資料與行為層」，將產生新型資安破口，以下表 1 之兩種使用情境加以分析權限失效之原因。

Rebouças Filho 研究指出傳統仰賴手動配置與靜態規則的 IAM [2]，已無法有效回應因應新型態 AI 威脅。組織應透過高階「特徵提取」方式，識別行為異常，並藉由動態「權重參數」調整存取控制，方能建立具前瞻性的安全防線。

三、AI 驅動的自動化方法

AI 如何將抽象的資安管理規則轉換為可執行的安全防護基準，並將零信任（Zero

表 1　AI 使用情境的權限控制失效

情境	技術挑戰與痛點	權限失效分析
模型微調	資料邊界崩潰：敏感資料被編碼進模型權重	使用者雖具備合法身分，卻能透過 AI 對話誘導提取其職權外之資訊（如 HR 敏感政策），導致權限隔離失效。例如：員工 Bob 獲准使用內部 AI 助手。雖然他原本不具備存取人力資源（HR）資料庫的權限，但若該 AI 助手在微調（Fine-tuning）階段已學習相關敏感政策，Bob 即可透過對話誘導獲取原本「無權」知悉的資訊。Bob 雖未違反任何身分登入規則，卻利用 AI 工具作為合法管道繞過存取控制，導致資料孤島被徹底扁平化。
RAG 管道與 XPIA 攻擊	提示注入（Prompt Injection）劫持合法權限	傳統 IAM 無法偵測「權限被惡意上下文劫持」。當 AI 代理人代表主管執行轉寄機密資料等未授權動作時，系統缺乏即時行為監控。例如：RAG 管道中的 XPIA，在自動化流程中，攻擊者透過電子郵件夾帶隱藏指令，操控 AI 助理執行未授權動作。例如：AI 以主管 Alice 的合法權限檢索機密專案營收，並將其自動轉寄給外部。在此過程中，傳統 IAM 無法識別「Alice 的權限被惡意上下文劫持」，系統因缺乏即時行為監控與動態風險評估，難以偵測此類權限使用的不合規性。

Trust）原則部署於資料存取的核心路徑上，實現多層次的縱深防禦。可透過 PDRR 四階段循環模型，持續回饋迴路運作，以實現規則的動態修正。

1. 計畫階段（Plan）是整個架構的起點，負責將組織內部的法規文件、內控政策等非結構化文本，透過大型語言模型（LLM）與自然語言處理（NLP）技術自動轉譯為可執行的結構化規則（Jayasundara 研究 [3]）。參考 Mandalawi 的研究 [4]「六階段推理框架」進行風險識別與合規驗證，確保執行動作與組織管理要求對齊，且模型必須具備事後解釋或內建透明度能力。

2. 部署階段（Deploy）負責將規則轉化為即時的存取控制決策。一方面，Pothen 研究 [5] 提出機器學習二元分類器結合向量分類與規則決策點（Policy Decision Point，PDP），依據實體行為與上下文動態授予存取權限，實踐零信任原則；另一方面，Stocks 研究提出透過以身分為基礎的存取控制（Identity-Based Access Control，IBAC）數學模型、參與者感知存取控制機制與動態去識別化技術 [6]，在資料層建立精細化防禦，從根源阻斷 XPIA 所導致的資料外洩風險。

3. 紀錄階段（Record）以非監督式機器學習建立使用者與 AI 代理人的正常行為基準，並持續監控偏移訊號以觸發異常告警。接著根據 Ganie 研究每一次授權決策皆透過 XAI 自動生成稽核理由 [7]，確保整個系統具備完整的事後追溯與問責能力，符合監管機關對透明度的要求。

4. 回應階段（Respond）即在偵測到存取異常後，系統自動執行封鎖、多因素驗證（Multi-Factor Authentication，MFA）或權杖撤銷等緩解措施，並透過安全協調、自動化與回應平台（Security Orchestration, Automation, and Response，SOAR）進行事件回應流程的自動化排程（Hariharan 研究 [8]）；對於無法由系統自主判斷的高風險案例，可透過 RBAC 與雙因素驗證機制，或是將決策權即時遞交人類管理員進行最終驗證，維持「人在迴路」的監督原則。透過 XAI 提供的解釋，分析「原因」（模型偏差或錯誤特徵提取）如何導致「結果」（錯誤決策），作為回饋計畫階段的輸入，據此調整控制措施（Kolli 等人 [9]）。

四個階段（如表 2）透過持續回饋迴路形成閉環，回應階段所產生的事件資料與緩解經驗，將回饋至計畫階段，持續修正規則模型與風險閾值，使整體管理能力隨時間自我演進、動態強化。且可對應《人工智慧基本法》第 16 條與第 19 條精神的「以風險為基礎的 AI 內控管理機制」，能自動適應法規演進並即時調整防護強度。

為達成「以 AI 管理 AI」目標，不同發展階段採用對應技術路徑，各項技術方法既可獨立部署，亦能透過互補機制發揮綜效，構建具備韌性的管理架構。

四、傳統存取控制困境與 AI 驅動技術之對比分析

傳統的身分與存取管理（IAM）多屬靜態權限配置並高度仰賴人工週期審查。此

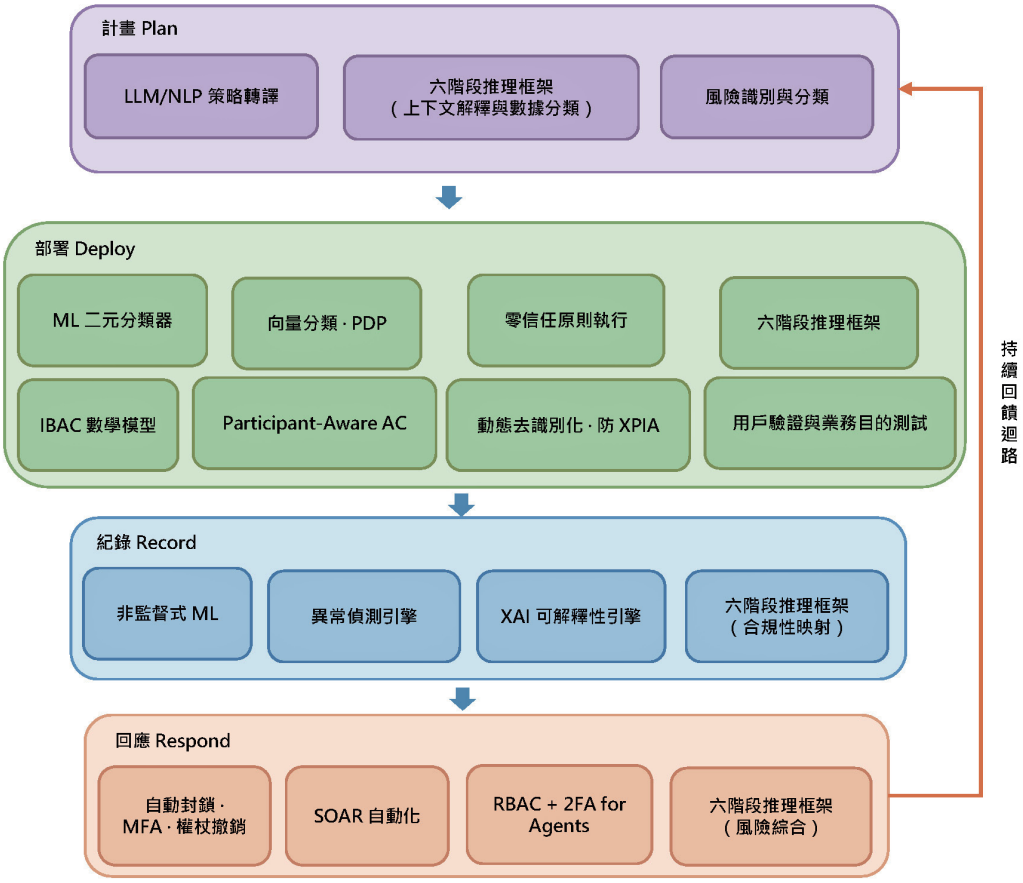

圖 1　AI 驅動的自動化資安管理階段

種傳統架構不僅維護成本高昂、容易產生人為疏漏，更因缺乏針對動態 AI 代理（AI Agent）的防護邊界，導致無法因應現代複雜的資安威脅。使其面對 XPIA 時十分脆弱，且其決策模式難以滿足法規的稽核要求。相較之下，AI 驅動技術透過自動化與動態防禦解決了上述困境。在合規面，系統利用自然語言處理（NLP）自動轉譯法規並進行動態配置管理，有效降低合規成本；在防禦面，透過將基於角色的存取控制（RBAC）結合雙重驗證（2FA）與資料層過濾，將惡意攻擊的成功率大幅降低；在問責面上，則藉由 XAI 產出透明且可讀的決策日誌，確保所有存取行為皆具備完整的事件軌跡。整體而言，AI 驅動架構不僅精準修補了傳統機制的安全破口，更對齊了《人工智慧基本法》對資安與隱私預設設計以及透明問責的嚴格監管要求，以下針對：合規管理、防範新型態攻擊與機敏外洩以及建立工程信任與隱私問責三大管理面向，分別說明其核心技術提取、技術原理與應用架構以及不同管理方法（傳統 *vs.* AI）之成效比較如表 3：

表 2　AI 自動化管理：防範新型態威脅的安全防護階段

階段	主要任務	可用技術	引用來源／理論
計畫（Plan）	規則感知轉譯與風險對齊	利用 LLM／NLP 將自由文本法規與內規，自動轉譯為結構化規則，滿足法規與內規管理要求	LLM／NLP 規則轉譯、六階段推理框架（上下文解釋與數據分類）
部署（Deploy）	動態授權決策與執行以及資料層級精細化防禦	透過 ML 分類器，依據實體行為與上下文，部署零信任原則，即時給予是／否核准的快速決策 實施 IBAC 數學模型、參與者感知存取控制與動態去識別化，從根源部署防範 XPIA 造成的資料洩密	ML 二元分類、向量分類器、PDP、IBAC、Participant-Aware AC、Metadata Governance、六階段推理框架（用戶驗證與業務目的測試）
紀錄（Record）	行為基準建立與問責紀錄	透過非監督式 ML 紀錄使用者與實體的正常行為基準，並導入 XAI 可解釋性引擎，確保每一次決策皆生成可稽核的理由	行為分析、異常偵測、XAI、六階段推理框架（合規性映射）
回應（Respond）	即時威脅緩解與 HITL	偵測存取偏移後，系統自動執行封鎖、MFA 或權杖撤銷等緩解回應；並將高風險案例遞交人類最終驗證	RBAC + 2FA for Agents、自動化 SOAR 緩解、六階段推理框架（風險綜合）

表 3　AI 驅動架構之技術理論與成效分析

管理面向	核心技術提取（特徵）	技術原理與應用架構	（傳統 *vs.* AI）成效
合規管理	NLP 自然語言處理 規則學習演算法 動態配置管理 去識別化技術	自動解讀自由文本法規，轉譯為結構化存取規則。將「動態配置」邏輯寫入資料層，確保授權後方可存取敏感資料	傳統：高度仰賴人工審查與手動設定，維護成本極高；AI：自動化取代人工配置，具備動態更新安全規則能力 Pothen 研究指出 AI 驅動減少 82% 手動任務 [5]。 合規成本降低 76%
防範新型態攻擊與機敏外洩	RBAC 角色存取控制 2FA 雙重驗證 即時屬性驗證 動態假名化	Ganie 研究為 AI Agents 建立安全護欄 [7]，檢索時動態以假名化數據替換真實個資，從根源防堵惡意指令竊密	Ganie 研究指出傳統：缺乏 AI 存取邊界 [7]。提示注入攻擊成功率達 73% AI：RBAC+2FA 組合後，攻擊成功率驟降至 3%
建立工程信任與隱私問責	結構化權限語言（XACML） XAI 不可篡改之日誌紀錄	South 等人提出將自然語言權限需求轉化為具邏輯確定性的結構化語言 [10]。強制生成機器可讀且人類可理解的授權理由	傳統：依賴人工決策或模糊提示，無法精確界定權限。 AI：決策流程透明化，滿足高監管環境稽核需求

五、結論

AI 驅動之自動化資安管理框架，旨在透過以下論述，針對 AI 時代的新型態風險與威脅，構建一套新穎之防禦思維與管理典範：

1. 透過 PDRR 循環框架，將原本散落於不同學術領域的單點技術（如 NLP 規則轉譯、IBAC 數學模型）歸納為一個邏輯的「計畫 – 部署 – 紀錄 – 回應」四階段零信任藍圖。

2. 定義 AI 時代威脅塑模，針對 RAG 與 LLM 微調模型所帶來的「未違規卻洩密」等問題，強調必須在「資料層級」落實防禦的重要性 ₀需透過實施參與者感知存取控制、IBAC 以及為 AI 代理強制套用 RBAC+2FA 等多層次防禦等新方法，方能有效防堵資料外洩與越權存取。

3. 對齊國際標準與合規語境，技術解決方案與國際規範進行了邏輯對齊，包括 CISA 零信任指令、CIS Benchmarks、ISO 42001（AI 管理系統）的風險評估與問責制以及我國人工智慧基本法，為資安長與合規稽核人員提供了實務參考價值。

總結來說，儘管自動化技術極大提升了防護力，但在高度監管環境中，HITL 仍是不可或缺的監督機制。AI 應定位為「智慧的合規檢查助手」，負責提供風險建議與結構化規則，而最終的批准與決策權必須保留給人類管理員，以滿足可信任 AI 模型的透明度、問責制與稽核要求，並對應《人工智慧基本法》第 4 條中的「人類自主、資訊與安全、透明與可解釋以及問責」原則之精神。

六、未來研究方向

基於「AI 驅動的自動化管理」雖能有效解決組織資料外洩痛點，然而隨著 AI 技術的演進，未來的存取控制與資料管理框架仍有極大的優化空間。為進一步強化並補足實務落地前的「概念驗證（PoC）」與原創實證基礎，未來的研究方向可朝「標準化整合」、「精細化模型」、「代理驗證與分散式管理」，以及「基於業務痛點的實證模擬」等面向進行擴充：

1. 標準化整合：結合 ISO 42001 與國際 AI 管理法規，目前的 AI 自動化管理框架雖能初步對應 CISA 指令、NIST 框架以及 GDPR 等傳統資安與隱私規範。然而，隨著各國對 AI 監管的加嚴，未來的研究應探討如何將此框架與新興國際 AI 治理標準進行深度結合。

2. 精細化模型：針對複雜產業中，資料不僅包含結構化的文字與數據，更涵蓋極度複雜且機敏的各種工程或技術圖資；同時，雲端原生架構通常伴隨「多租戶（Multi-tenant）」的共用環境。未來的研究應進一步開發與優化 IBAC、基於政策之存取控制（Policy-Based Access Control，PBAC）或下世代存取控制（Next Generation Access Control，NGAC）。

3. 代理驗證委託：隨著自主 AI 代理（AI Agents）被賦予執行複雜工程任務之權限，研究重點應轉向如何安全實施「權限委託」機制。透過分散式帳本或區塊鏈技術導入，在跨環境協同防護中留下不可篡改之稽核軌跡，平衡運算效能與系統安全性。

4. 實證模擬：基於業務痛點之場景化概念驗證（PoC），未來應規劃針對具體業務痛點（如：遠端工程協作、機敏圖資存取）建構模擬實驗環境。透過量化指標驗證 AI 自動化管理於「誤判率」、「系統延遲性」及「防禦成功率」之實際表現，提供具備原創價值之實證數據，作為組織大規模部署之決策參據。

參考文獻

1. Katta, B. S. (2025). AI-driven access review architecture for scalable identity governance in modern enterprises. International Journal of Computational and Experimental Science and Engineering (IJCESEN), 11(3), 4781–4787.
2. Rebouças Filho, W. L. (2022). The role of AI in enhancing identity and access management systems. RevistaFT, 26 (117), Article ni10202212302015.
3. Jayasundara, S. H., Arachchilage, N. A. G., & Russello, G. (2024). SoK: Access control policy generation from high-level natural language requirements. ACM Computing Surveys, 57(4), Article 102.
4. Al Mandalawi, S., Mohammed, M. A., Maclean, H., Cakmak, M. C., & Talburt, J. R. (2025). Policy-aware generative AI for safe, auditable data access governance. In 2025 17th International Conference on Knowledge and System Engineering (KSE).
5. Pothen, V. A. (2025). AI-powered access governance: Automating risk-based identity in enterprise cloud. Journal of Computer Science and Technology Studies, 7(3), 416-425.
6. Stocks, M. (2024). IBAC mathematics and mechanics: The case for 'Integer Based Access Control' of data security in the age of AI and AI automation [Unpublished manuscript]. Canberra University.
7. Ganie, A. G. (2024). Securing AI agents: Implementing role-based access control for industrial applications. Universitat Politècnica de València.
8. Hariharan, R. (2025). AI-driven identity and access management in enterprise systems. International Journal of IoT, 5(1), 62–94. https://doi.org/10.55640/ijiot-05-01- 05
9. Kolli, N., Sajja, J. W., & Nerella, A. (2025). Building secure AI agents for autonomous data access in compliance/regulatory-critical environments. Computer Fraud and Security.
10. South, T., et al. (2024). Authenticated delegation and authorized AI agents. MIT / University of Oxford / Harvard Law School.
11. Neelakrishnan, P. (2024). AI-driven proactive cloud application data access security. International Journal of Innovative Science and Research Technology, 9(4), 3788–3801.
12. Planas, E., Martínez, S., Brambilla, M., & Cabot, J. (2020). Modeling and enforcing access control policies in conversational user interfaces. In Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20), 1–14.
13. Vatsavayi, C. (2025). Cloud-native data governance in AI personalization: A framework for integrating consent management and access control. European Journal of Computer Science and Information Technology, 13(45), 37–46.
14. Abaev, N., Klimov, D., Levinov, G., Mimran, D., Elovici, Y., & Shabtai, A. (2026). AgentGuardian: Learning access control policies to govern AI agent behavior. arXiv.
15. 數位發展部 . (2026). 人工智慧基本法 .
16. Ofili, B. T., Erhabor, E. O., & Obasuyi, O. T. (2025). Enhancing federal cloud security with AI: Zero trust, threat intelligence and CISA compliance. World Journal of Advanced Research and Reviews, 25(02), 2377–2400.
17. Wider, A., Harrer, S., & Dietz, L. W. (2025). AI-assisted data governance with Data Mesh Manager. In 2025 IEEE International Conference on Web Services (ICWS).
18. Golightly, L., Modesti, P., Garcia, R., & Chang, V. (2023). Securing distributed systems: A survey on access control techniques for cloud, blockchain, IoT and SDN. Cyber Security and Applications, 1, 100015.