AI 安全性：
您在 AI 競賽中的競爭優勢

CxO 指南：在加速 AI 採用的同時管理風險

介紹

隨著各組織競相利用 AI 的變革力量，安全性往往跟不上採用速度。對於每一項獲批准的計畫，個人和團隊都使用過無數次影子 AI。實際上，有 85% 的 IT 決策者表示，員工甚至在 IT 團隊評估之前就已經採用了 AI 工具[1]。93% 的員工承認在未經批准的情況下將資訊輸入 AI 工具[1]。

與此同時，AI 原生攻擊手段正在激增，在過去一年中增長了 47%[2]，因為傳統安全工具很難跟上採用速度。新形式的資料暴露和合規性漏洞對現有的治理做法構成了挑戰。高階領導者面臨一個關鍵問題：如何管理 AI 帶來的風險，而不阻礙它支援的創新？

需求非常明確。組織需要創造這樣一個環境：

- 所有 AI 使用情況全部已知
- 所有 AI 通訊都可以受到保護
- 所有 AI 原則都可以強制執行
- 所有 AI 模型都能受到保護而免遭濫用

Cloudflare AI Security Suite 透過消除風險的不確定性，使組織能夠信心滿滿地藉助 AI 加速發展。我們的統一平台透過探索 AI 使用、套用 Zero Trust 存取，以及利用主動威脅防禦保護 Web 和 API 端點，來保護整個 AI 生命週期。透過整合式資料治理，團隊可以在自由創新的同時確保安全性。

85%
的 IT 決策者表示，員工甚至在 IT 團隊評估之前就已經採用了 AI 工具[1]。

AI 安全性挑戰

採用 AI 的風險正在推動 AI 安全性解決方案市場的快速增長。例如，雲端原生應用程式保護平台 (CNAPP) 專注於 AI 開發工作流程。諸如此類的早期產品是不可或缺的防線，但還遠遠不夠。公司需要保護整個 AI 生命週期，包括保護在生產環境中執行的 AI 系統。

AI 安全性挑戰的完整範圍越來越清晰。現在，開發人員正在建置 AI 功能，員工正在使用外部 AI 工具，而客戶正在與 AI 支援的應用程式互動。手動保護所有這些不同的環境非常複雜，而且要確保一致性更是難上加難。

防止員工濫用 AI

由於組織員工同時採用已批准和未批准的 AI 工具，敏感性資料和營運皆面臨風險。安全團隊必須應對以下挑戰：

- **缺乏對 AI 使用的可見度：**領導者通常無法完全掌握員工正在使用的 AI 工具、敏感性資料的處理位置，以及這些 AI 系統與業務應用程式的連接方式。這個盲點會造成顯著的風險暴露。
- **資料安全與合規風險：**AI 改變了資料在組織中流動的方式。個人資訊、專有資料和客戶記錄最終可能會以觸發合規性違規或洩露競爭情報的方式進入 AI 系統。
- **自主式 AI 工作流程的存取控制：** 不僅僅是人類對 AI 工具的存取需要管理。自主式 AI 對 MCP 伺服器和其他重要系統的存取也必須受到管理。這需要採用新的身分管理和存取控制方法。

保護面向公眾的 AI 應用程式和模型

無論是內部開發還是透過第三方使用，AI 系統都正在成為影響客戶和使用者體驗的重要部分，因此需要同等程度的保護。《OWASP LLM 應用程式十大風險》強調了傳統工具無法應對的威脅：

- **無限制使用攻擊：**與傳統的 DoS 攻擊類似，無限制使用攻擊試圖以耗費大量資源的請求來壓垮 LLM。按使用付費的雲端定價模式可能會導致此類攻擊的財務影響急遽增加，同時合法使用者也將面臨服務品質降低的問題。
- **模型投毒：**攻擊者透過將損毀的資料插入開發人員用於模型訓練的公用資料集或存放庫，在 LLM 中植入後門程式、偏差或漏洞。以這種方式中毒的模型，在特定因素觸發有害行為之前，表現都很正常。
- **提示插入攻擊：**提示插入是一種常見的透過 AI 介面外洩資料的手法，它將惡意指令嵌入使用者提示或外部內容中，藉此操縱 LLM 的輸入，使模型忽略原始指令並改為執行攻擊者的命令。
- **越獄：**精心設計的提示（例如角色扮演場景、指令覆寫命令和多輪策略）能夠繞過 LLM 的安全防護機制，以生成違禁內容或擷取敏感性資訊。

為了實現快速的 AI 創新且不讓組織面臨風險，我們建議領導者採用一種能夠滿足所有需求的 AI 安全性方法：

- 保護員工對 GenAI 的使用
- 保護採用 AI 技術的應用程式和工作負載
- 建置採用 AI 技術且將安全納入設計的應用程式

保護 AI 開發與訓練工作流程

AI 專案涉及龐大的資料集、昂貴的資源和反覆的實驗，所有這些都會產生新的攻擊面和漏洞。安全團隊必須處理以下問題：

- **訓練資料安全性和完整性：**洩露的訓練資料可能會引入偏差、後門程式或漏洞，它們會持續存在於生產模型中。組織必須保護對訓練資料集的存取，防止未經授權的修改，並在整個模型生命週期確保資料溯源。
- **認證與密碼管理：**AI 開發工作流程需要存取多個系統，包括用於資料集的雲端儲存空間、用於訓練的運算叢集、模型登錄和第三方服務。密碼或 API 金鑰若未妥善保護，可能會暴露敏感性認證，導致未經授權存取專有模型、訓練資料或生產系統。
- **開發環境存取控制：**AI 工程師通常需要較高的權限，才能夠進行模型試驗和存取敏感性資料。若不採取適當的存取控制，這可能會導致內部人員威脅、意外的資料洩露或未經授權的模型擷取。

透過 Cloudflare 統一安全性

成功的 AI 採用著重於提高生產力，而非限制生產力。當團隊知道已部署適當的防護措施後，能夠信心滿滿地使用 AI，進而加速創新，並承擔更具挑戰性的專案。

Cloudflare AI Security Suite 為 AI 創新建立一個安全的環境。透過整合安全存取服務邊緣 (SASE) 與 Web 應用程式安全功能，CxO 能夠連接並保護兩個重要的領域：

- 外部，面向公眾、支援 AI 的應用程式
- 內部，私人 AI 系統與工作負載

Cloudflare AI Security Suite 專注於保護整個 AI 生命週期，可滿足從探索和風險管理到資料保護的各種安全需求，同時確保使用者存取安全，並保護支援 AI 的應用程式和開發工作流程。為了提供至關重要的即時生產安全性，Cloudflare 全球網路採用內嵌方式，檢查並篩選每一次 AI 互動，為所有使用者和應用程式的資料提供保護。

Cloudflare 並不是在問題發生後才偵測出來，而是在問題發生前就加以預防，並在威脅觸及 AI 模型之前就將其封鎖。安全團隊獲得所需的可見度，以便提前應對新興威脅。

Cloudflare 的核心功能

Cloudflare AI Security Suite 將全面監控、即時保護及主動風險管理整合至單一平台，以提供全方位的 AI 安全性方法。

全面的 AI 探索與可見度

有效的 AI 安全性取決於是否全面、即時了解已批准和未批准的 AI 資源和使用情況。Cloudflare AI Security Suite 的基礎在於持續監控和自動探索，以識別所有類型 (公有、私有或內部) 環境中的所有 AI 模型、助理、主體和影子 AI 部署。

積極主動的 AI 風險管理

Cloudflare AI Security Suite 透過偵測並緩解 AI 特定的漏洞、設定錯誤和攻擊路徑 (包括 OWASP LLM 十大風險中的漏洞)，協助組織預防攻擊。應用程式可信度評分有助於確定補救措施的優先順序，讓團隊可以優先處理最重要的風險。

適用於 AI 支援的應用程式的應用程式安全性

為了協助 SecOps 掌握最新的 AI 威脅手段，Cloudflare AI Security Suite 針對 AI 管道中 AI 特定的漏洞、設定錯誤和攻擊路徑納入了主動威脅偵測和緩解功能，包括防範提示插入、資料投毒及模型濫用。

專用 AI 防火牆能探索並標記生成式或自主式 AI 和 API 端點，偵測洩露 PII 的企圖，並封鎖惡意提示以避免影響 AI 模型效能，或使用有害內容或錯誤資訊為模型投毒。

適用於 GenAI 和自主式 AI 工作流程的 Zero trust 存取

諸如最低權限這樣的 Zero Trust 原則，同樣適用於員工和 AI 主體。Cloudflare AI Security Suite 可以針對人類與 AI 以及 AI 與 AI 之間的互動，強制執行 Zero Trust 網路存取原則 (ZTNA)。針對 MCP 伺服器提供集中記錄和控制，可確保自助式 AI 僅存取已獲授權的內容，包括及時化工作流程。

AI 感知資料保護

實現有效的 AI 安全性需要資料遺失防護功能——此功能利用多種語言模型來理解提示的內容及其背後的意圖。Cloudflare AI Security Suite 在訓練、提示及回應的整個過程中，納入了資料遺失防護 (DLP) 功能，以防止 AI 模型和管道中的 PII 暴露、資料外洩及未經授權的存取。部署以 API 為中心的內聯執行階段安全性，作為快速、簡便的第一道防線，旨在與 CNAPP 支援的左移方法互補。

資料當地語系化

LLM 和 AI 應用程式與其他類型的資料環境一樣，受到相同的法規約束。Cloudflare AI Security Suite 透過相關原則，將訓練資料和推斷請求保留在經核准的區域內，藉此協助組織確保 AI 工作負載不超出地理及管轄邊界。

針對 AI 開發，將安全納入設計

如同所有類型的軟體，AI 應用程式的安全性應從開發週期一開始就內建，而不是在之後才附加。Cloudflare AI Security Suite 為開發人員提供工具和架構，以建置採用 AI 技術且將安全納入設計的應用程式。

Cloudflare AI Security Suite 的業務影響

AI 的興起不僅僅是一種演進，它從根本上顛覆了工業化或電腦化的秩序。因此，快速有效的採用不僅是推動增長的要素，也是組織能否持續發展的關鍵。如今，緩慢或不安全的採用對全部產業和領域內組織的生存構成威脅。

Cloudflare AI Security Suite 可以實現安全、受控和高效的轉換，從而在競爭激烈的環境中滿足快速提升的客戶期望和市場需求。

- **更快速的 AI 創新：**當安全性有助於安全使用而非阻礙採用時，員工和團隊可以探索新的 AI 應用程式，進而在日常工作中提高生產力。採用適當的安全架構，開發人員可以自信地建置雄心勃勃的 AI 功能，而不會危及敏感性資料或系統。
- **降低 AI 相關風險：**組織可以管理採用 AI 帶來的所有風險，包括支援 AI 的 Web 應用程式中固有的 AI 相關風險。可以透過保護正在開發中的 AI 應用程式，來確保員工不會洩露敏感性資料，或將其暴露於 AI 訓練集中。SecOps 能夠主動識別並緩解 AI 特定的威脅與漏洞，藉此盡可能縮小攻擊面，並保護重要的資料和模型。
- **簡化的安全作業：**集中檢視和控制 AI 安全狀態可以簡化管理並提高事件回應效率。SecOps 團隊可以專注於制定策略性計畫，不必再疲於奔命地處理與 AI 相關的突發事件。
- **強大的資料治理與合規性：**AI 特定的資料保護控制可協助您保護敏感性資訊，並滿足 AI 生命週期中不斷變化的法規要求。
- **降低總體擁有成本 (TCO)：**利用併入現有安全投資的整合式平台，比針對每一項 AI 安全性挑戰實施單獨的單點解決方案更經濟。

Cloudflare AI Security Suite 的模型使用案例

Cloudflare AI Security Suite 是滿足 AI 採用基本需求的理想選擇。

- **確保員工使用 AI 工具的安全：**針對員工存取公用生成式 AI 工具（例如 ChatGPT）和內部開發的 AI 支援的應用程式，強制執行 Zero Trust 原則。
- **保護面向公眾的支援 AI 的應用程式：**保護整合 AI 模型的 Web 應用程式和 API（例如，聊天機器人與推薦引擎），使其免遭攻擊，從而避免敏感性資料外洩或模型濫用。
- **影子 AI 管理：**自動探索組織中未獲批准的 AI 工具，並運用適當的控制措施，以在受管理的風險範圍內實現持續創新。
- **針對 AI 互動的 AI 支援的 DLP：**防止敏感性資料在 AI 提示或回應中洩露，確保 PII 和機密資訊受到保護。
- **AI 開發安全性：**為工程團隊提供相關架構，使安全開發成為預設方法，讓他們能夠在快速建置 AI 功能的同時確保安全性。

實作考量

AI 安全性是一項基礎安全策略，應該周密考量如何實作該策略。

從現有的基礎架構開始	最成功的 AI 安全實作是建置於現有的 SASE 和應用程式安全工具之上，而非取代它們。這種方法可利用目前的投資，同時擴展保護以涵蓋 AI 特定的風險。
部署統一的內嵌保護	在惡意活動發生時，在網路邊緣封鎖惡意活動的能力對 AI 安全性至關重要。部署即時內嵌控制，並透過基於 API 的監控加以補充。
確保完整覆蓋	AI 安全策略應滿足所有需求：員工使用 GenAI 工具，保護 AI 支援的應用程式和工作流程，確保自主式 AI 工作流程的安全，以及 AI 開發工作流程的安全性。
適合企業規模的方案	選擇能夠隨著 AI 採用而成長的解決方案。隨著 AI 使用範圍的擴大，適用於單一試點計畫的方案也必須擴展到整個組織。
驗證成功標準	在做出完整承諾之前，請先在實際情境中驗證解決方案。選擇一個允許免費、自助啟用企業功能的平台。這樣，您可以小規模測試完整的安全套件（例如，針對單一團隊或應用程式），來快速驗證其價值並確保其符合您的成功標準。

使用 Cloudflare AI Security Suite 採取下一步行動

隨著全球 AI 採用率的加速，能夠安全擴展 AI 的組織將擁有顯著的競爭優勢。關鍵是要認識到，AI 安全性不是為了阻止 AI 使用，而是要以智慧的方式支援 AI。

Cloudflare AI Security Suite 讓組織能夠充滿信心地進行創新。我們的 AI Security Suite 建立在廣泛採用的 SASE 和應用程式安全平台之上並加以擴展，提供整合的 AI 探索、Zero Trust 存取控制、情報主導的主動威脅防禦和強大的資料治理。透過全方位保護 AI 使用與模型，組織可以協助開發人員加速建置，並提升員工生產力，且無需犧牲終端使用者體驗。

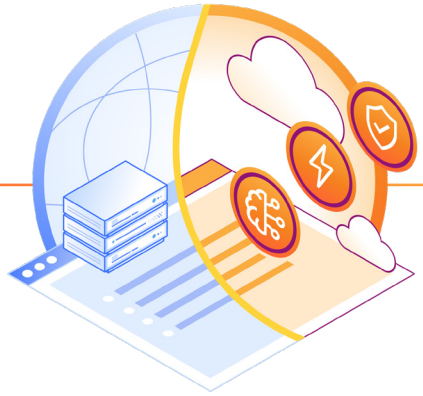
預約諮詢

探索 Cloudflare AI Security Suite 如何轉變組織保護 AI 採用的方法。

+886 8 0185 7030

enterprise@cloudflare.com

www.cloudflare.com/zh-tw/



1. ManageEngine。企業中的影子 AI 激增：來自美國和加拿大的深入解析
2. Check Point Research。Check Point Software 發佈的《2025 年第一季全球網路攻擊報告》顯示：全球網路威脅激增近 50%，勒索軟體攻擊上漲 126%