CYCRAFT WHITE PAPER

建構 AI 應用安全治理：從 AI 驅動防護、AI 原生安全到自主曝險管理

2026 AI Adopting Whitepaper

2026.03

為落實自主曝險管理 (Autonomous Exposure Management)，奧義智慧認為應以 AI 安全平台 (AI Security Platform) 為基礎、整合多模態曝險管理 (Unified Exposure Management)，同時運用 AI 自動化處理既有資安議題、並更進一步善用 AI 處理 AI 原生與 LLM 潛在威脅。此份白皮書提出 AI 導入資安的可操作框架，引介奧義智慧 XCockpit 與 XecGuard，展示在不同應用情境下的解決方案。

2026 科技新趨勢：AI 應用資安治理

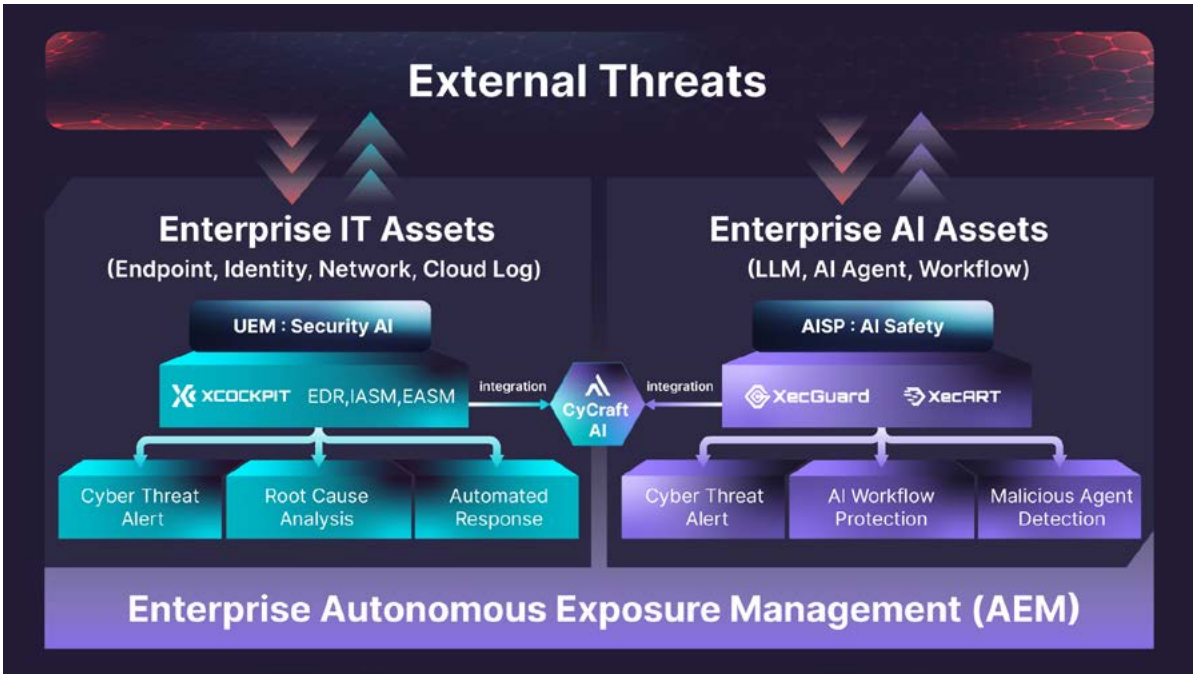
自 ChatGPT 於 2022 年底推出以來，攻擊數量與影響規模皆不斷打破歷史紀錄。根據預測，全球網路犯罪成本將以每年 15% 的速率增長，如果網路犯罪的經濟規模（2025 年預估約 10.5 兆美元）可被視為一個國家，它將成為僅次於美國和中國的全球第三大經濟體。即便是一般的社交工程攻擊，釣魚信件的累積增幅已達到 4,151%（數據統計至 2024 年中），由於 AI 生成的內容更加客製化，傳統防禦漏報率顯著提升。攻擊者也得力於 AI 自動化技術，24 小時不間斷掃描企業網路漏洞、結合供應鏈攻擊，導致全球組織平均每週遭受的網路攻擊次數大幅提高，自 2022 年的 1,168 次成長到 2025 年的 2,003 次，增幅達 71%（數據統計至 2025 年 11 月）。

AI 驅動的新興資安風險一方面挑戰傳統防禦概念，一方面也催化各種 AI 技術與資安機制的應用。適應了典範轉移的資安服務業者，開始運用 AI 自動化處理既有的風險議題，更得直接面對 AI 原生與 LLM 模型潛在的資安問題。這兩者雖然在 AI 自動化技術投入的深度與廣度上有所差異，但仍屬同一防禦光譜，AI 安全平台（AI Security Platform，下稱 AISP）應運而生。

AISP 整合了不同的資安工具，以 AI 為中心全面、持續地防護模型、數據、LLM 應用與 AI 風險態勢管理，從宏觀角度協助資安團隊進行決策和風險優先級排序。由於傳統資安工具多半聚焦端點、身分認證與網路風險，不僅無法掌握 AI 開發流程，也缺乏針對提示詞漏洞的安全檢測以及惡意 AI Agent 的防護機制。AISP 包含兩大支柱：AI 使用管控（AI Usage Control）與 AI 應用資安（AI Application Cybersecurity），前者管理企業員工與系統使用第三方 AI 服務的規範、後者延伸保護企業自建 LLM、經調校的模型與 AI Agent 等客製化開發 AI 應用。

我們認為，一個全面且可自主運作的資安平台，必須納入 AI 應用資安治理，兩者不應分開管理。透過此統整性的管理平台，持續監測 AI 系統的所有關鍵組件和動態，將零散的警報轉化為可操作的風險情報，在模擬攻擊路徑或自動關聯案情細節時，快速生成合規且有助後續決策的管理措施。AI 時代下的資安管理系統應奠基於 AISP 之上、納入多模態曝險整合（Unified Exposure Management，下稱 UEM）概念，才能善用 AI 自動化技術，落實先發制人曝險管理（Preemptive Exposure Management，下稱 PEM）安全機制。

> ***「一個全面且可自主運作的資安平台，必須納入 AI 應用資安治理，兩者不應分開管理。」***



多模態曝險整合管理　導入自動化技術解決沉痾

在 AI 時代，傳統資安議題依舊存在，甚至更加嚴峻。我們觀察到的主要問題包含：

- **威脅偵測延遲與反應不足：**傳統上依賴特徵碼的防禦方式，只能有效防守已知威脅，無法應對零日攻擊或變種病毒等未知漏洞。
- **數據爆炸與資安情報難以消化：**SIEM 平台或 SOC 團隊雖收集了海量數據，但缺乏工具深度分析不同來源的情報，無法有效與內部告警數據進行關聯分析，導致資訊淪為無用。
- **攻擊面的規模與複雜度增加：**雲端運算、行動裝置和遠端工作的興起，使傳統網路邊界逐漸消失，防火牆和 VPN 越來越無法保護四散的資產。除此之外，企業內部的惡意或人為疏失也難以被傳統的網路流量分析所捕捉，缺乏對攻擊面曝險的可視性。
- **資安專業人才短缺：**企業難以招聘或留住能夠進行威脅預判和事件回應的人才，導致防禦體系出現漏洞，雪上加霜的是，大量誤報與告警疲勞使人力更加拮据，更容易錯過真正的重大威脅。

自動化加速了資訊的碎片化與巨量化，讓人力與資源稀缺變成更大的破口，然而，AI 技術帶來的危機亦是轉機，由於 AI 帶來資安能量與人類智能的規模化，許多服務供應商已開始導入 AI 處理既有的資安問題。

冰凍三尺，非一日之寒，AI 或機器學習打破僵局的關鍵來自於從被動防守轉向主動防禦 (Proactive Cybersecurity) 。UEM 提供統整、全面的曝險視角，依照風險排序各個漏洞與資產，讓資安團隊不再疲於奔命，有層次地依序拆彈。除了漏洞管理，UEM 更強調攻擊路徑分析、外部攻擊面管理、資安設定管理和身分與權限風險管理。透過有效地將漏洞、設定、身分、資產整合在一個平台上，UEM 將資安機制從「補破洞」，轉變成「可規劃的防禦工事」。

我們認為，能實現 UEM 概念的資安系統，應至少具備以下功能：

- **行為異常分析：**有別於傳統特徵碼或惡意檔案偵測的分析模式，應建立使用者與實體行為模式資料庫，自動識別與正常模式不符的行為，即時偵測攻擊、內部威脅或無檔案惡意軟體。
- **自動化根因分析：**從傳統告警的被動應對，轉向以案情導向的主動分析，自動模擬攻擊路徑、分析惡意軟體樣本，從數十億行日誌中提煉出攻擊起源、傳播方式和影響範圍，提供簡潔易懂的事件摘要，彌補資安專業人力缺口。
- **以情境為導向的告警分析：**將大量且龐雜的警報，收束歸納成單一案件，視覺化呈現事件時間軸與關聯性，添加可能的背景資訊，有效降噪以減少人工判斷的成本與誤報率。
- **預警性風險評分：**整合外部威脅情報 (如：明網、暗網、遭公開販售的企業憑證等)、內部資產重要性（如：資產風險等級、資產類別、資產權限等）、漏洞嚴重性，透過組成架構與元件識別分析，關聯出帳號、端點與服務的攻擊面，預測哪個弱點最易被攻擊者利用，自動建議最優化的修復順序。
- **自主回應與修復：**AI 驅動平台能根據偵測到的威脅類型，自動執行預設的回應劇本 (如：隔離受感染端點、終止惡意程序、自動封鎖 IP 等)，針對高嚴重性事件自動建單，進行緩解措施追蹤管理，紀錄各種風險處理狀況與績效。

應用場景：奧義智慧 EASM 統整情資、數位資產與風險排序

在現有的資安工具中導入 AI 自動化技術，能讓資安人力與資源較精簡的組織如虎添翼。以政府部門為例，《資通安全管理法》規範的 A 級與 B 級政府單位，因涉及國家發展關鍵領域、核心科技研究、關鍵基礎設施、共用系統維運等業務，肩負資安重擔，需定期稽核下級單位在資安防護上的表現。然而，各單位性質不盡相同、下轄組織為數眾多且文化各異，包含公部門、國營事業機構和法人組織等。多數單位慣於採用靜態問卷調查，一一發送與收集相關數據，取得後尚需花費人力和時間整理、篩檢、建檔和管理，勞師動眾且曠日廢時。

因應此使用情境，採用 UEM 框架即可直接獲取下轄單位的端點與場域資訊，簡化管理與稽核流程，減少人為疏失、提高資訊可信度。奧義智慧 EASM 外部資產曝險管理模組更能針對各場域不同的資安管理系統及法規應辦事項，提供具體且可執行的風險評估指標、稽核建議與風控措施。讓沒有資安背景的主管或同仁都能理解意涵，滿足跨部門協作需求、降低一來一往的溝通成本。

奧義智慧 EASM 整合情資與資產軌跡，自動化攻擊路徑預測與風險評估。

以 AI 防禦 AI：構建先發制人的主動式曝險管理體系

AI 驅動防護工具雖能化解舊疾，AI 與 LLM 本身風險卻會引發新興威脅，Agentic AI 更使管理增添困難。LLM 的不可控性使企業導入 AI 工具前無法確認其詳細規格，不同的測試案例因應用場景或產業差異也無法參考，導入後發生問題更無法直接修正，存在管理危機。提示詞注入 (Prompt Injection)、提示詞擷取 (Prompt Extraction)、越獄 (Jailbreak) 等攻擊手法，會讓 LLM 受騙執行攻擊者指令、洩露訓練數據或生成有害內容，使 AI 淪為惡意行為共犯。

在 Agentic AI 時代，攻擊者可以透過提示詞注入，讓 AI Agent 自主執行惡意程式碼、發送釣魚郵件給客戶、或刪除數據庫記錄，衝擊更廣且更難以控制。由於 AI Agent 被賦予了存取外部資源（如：Email API、數據庫、程式碼執行環境）的權限，惡意提示可以誘騙 AI Agent 濫用外部工具，將語言指令轉化為越權的實際操作。一旦成功注入惡意提示，AI Agent 可能會將這個惡意目標分解為多個子任務並長期持續執行，難以被人工即時發現和阻斷。

近期多方觀察到以受害者 AI 環境為攻擊基礎設施，即時生成惡意行為的 Promptware 家族。AI 不僅被用於開發，更成為攻擊工具的運作元件之一，若企業未完善盤點與管理內部 AI 應用，暴露在外的存取點將變成攻擊者就地取材的破口。另外，由於 AI 在執行階段才即時生成惡意指令或腳本，大幅減少靜態特徵，使傳統基於簽章的偵測方法（如：YARA）失效，得以進一步規避相關監測與分析工具。

目前較活躍的 Promptware 家族包含：FRUITSHELL、PROMPTFLUX、PROMPTLOCK 等。FRUITSHELL 在程式碼中內嵌對抗式 Prompt，不只是用 AI 協助攻擊，直接用 Prompt 專門欺騙 AI 安全工具。PROMPTFLUX 將 LLM 當成動態的程式碼進行混淆，在執行階段就先自我變形以躲避偵測。PROMPTLOCK 則將受害者本機 LLM 當作勒索腳本引擎，直接在端點上即時生成與修改攻擊腳本。

無論是 AI 或 LLM 本身潛藏的風險、或直接以提示詞濫用 AI 系統，新時代威脅層出不窮，我們必須從主動式防禦更進一步，邁向先發制人式防護機制。PEM 是 2026 年十大資安與科技新趨勢之一，強調善用 AI 或機器學習深度剖析海量數據，快速提供可信的攻擊預測與解決措施。Gartner 分析師 Tori Paulman 主張：「此十大新趨勢不只是技術革新，更是企業轉型的催化劑。···下一波創新浪潮已近在咫尺，唯有立刻採取行動的企業能安然抵禦（新變革造就的）市場動盪，更能在未來數十年內，主導並定義產業格局。」

要能落實此典範轉移，核心是運用 AI 監測與防禦 AI，在攻擊者完成攻擊前，進行阻斷、欺敵與干擾，超前部署一套能瓦解攻擊鏈的計畫，而非只是優化告警或事後應變。相較於主動式資安強調在告警觸發前採取行動，例如威脅獵捕、紫隊演練、對抗性模擬等措施，PEM 更著重改造環境，使攻擊提早失敗或不符經濟效益。縮減可能的攻擊面、提高攻擊成功所需的成本、控制事件規模不致演變成事故，即是此框架的操作要領。

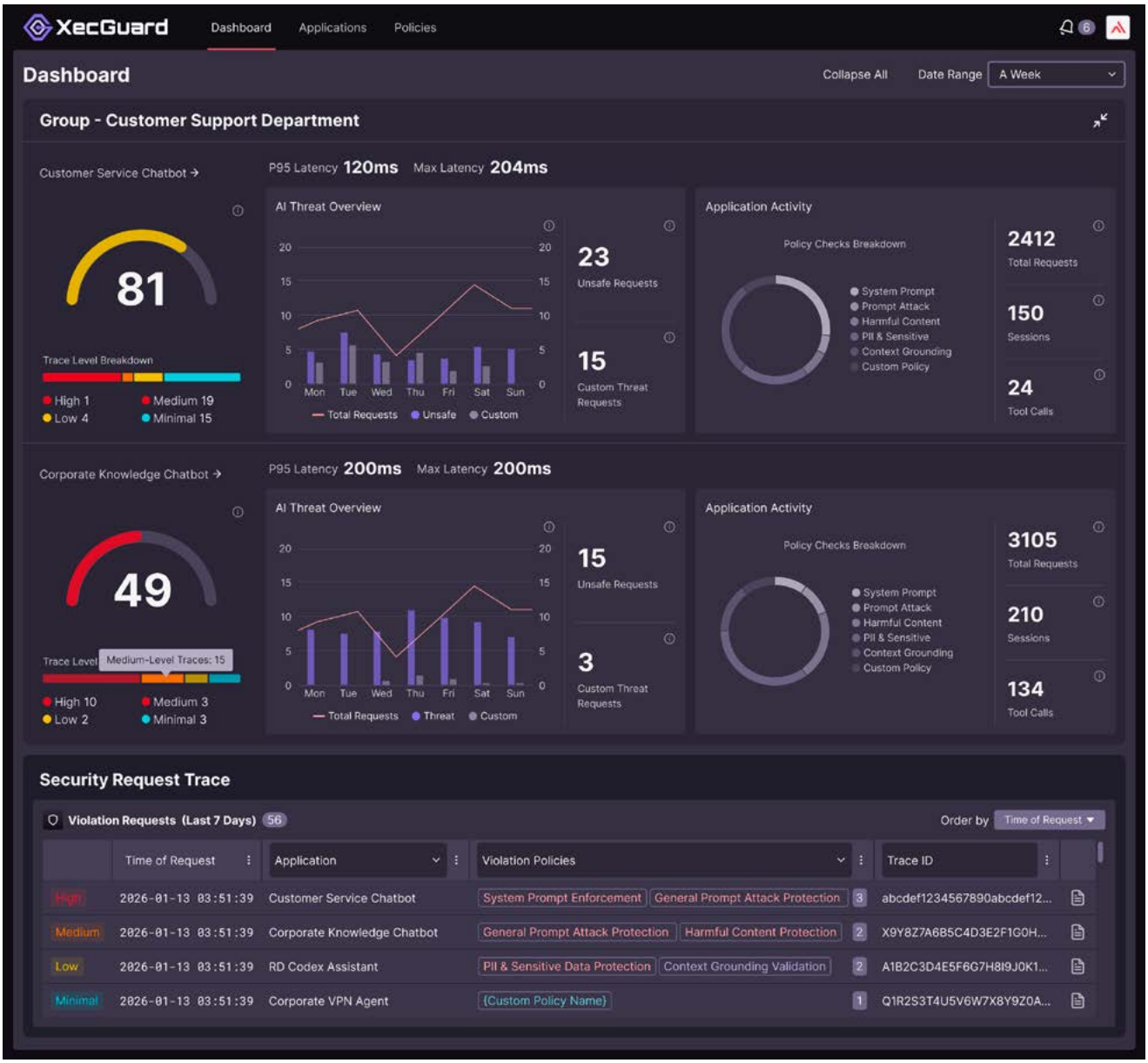
奧義智慧於 2025 年二度入選 Gartner 範本廠商，是應用 PEM 與 UEM 的先驅，也是台灣唯一入選的指標企業。在設計解決方案時，我們認為注重 AI 原生安全才能確實打造 PEM 管理平台。透過以下功能，我們協助使用者讓 AI 與 LLM 更穩健：

- **Zero-Code 安全強化：**以相容多數大型 LLM 模型的 API 介面直接導入安全防護，無需異動程式碼，無縫接軌現有 AI 應用，並能阻擋提示詞注入攻擊、檢查機敏個資、管理提示詞與稽核 LLM 等。
- **可彈性部署的 AI 護欄 (Guardrail)：**針對落地自建或第三方應用等不同情境，提供被動偵測與主動阻擋兩種模式，彈性部署以減少對 AI 效能之干擾。針對 AI Agent 更提供檢索護欄，確保 RAG、函式呼叫等情境的可信度與安全性。
- **可擴展的 AI 安全治理：**具備雲原生自動擴展框架與多租戶管理、稽核與告警能力，企業可依不同場域制定獨立政策，透過 CEF Syslog 與告警 API 整合既有 SIEM 平台或 SOC 團隊，打造完整一致的 AI 安全治理體系。

應用場景：奧義智慧 XecGuard 導入產業規範，鞏固 AI 原生安全

現今 AI 應用已高度整合於企業場域中，AI 原生安全對企業資安管理來說至關重要。以金融業為例，由於此產業面向大眾、業務複雜，有強烈的 AI 應用需求，然而個人隱私、交易資訊都具有高度機敏性，金融監督管理委員會與相關金融部門皆制定諸多規範，使落地自建或開源模型的使用窒礙難行。若缺乏完善的 AI 使用管控或應用資安規範，極可能造成影子 AI (Shadow AI)、資訊外洩、供應鏈風險、智慧財產權爭議等問題。

奧義智慧 XecGuard 新世代 AI 防火牆融合在政府、金融、醫療與高科技製造等關鍵領域的紅藍隊實戰經驗，不僅專注模型本體的防禦升級，更導入內建金融法規與管理規定，避免濫用也預防衍生的法律或道德問題。 協助企業在 AI 快速落地的同時，確保資訊安全、法規遵循與系統韌性同步升級，是以 AI 強化 AI 的堅實防線。



奧義智慧 XecGuard 精準強化大型語言模型，有效防禦多種提示詞風險。

從開發、營運到風險管理　全方面檢視 AI 應用安全

善用 AI 整合與處理多模態曝險資訊，融入 PEM 概念的自主曝險管理 (Autonomous Exposure Management)，是我們認為具前瞻性的 AI 資安方案。為了確保 AI 應用的安全，在其生命週期各階段皆應納入 AI 使用管控與提示詞管理等概念。提示詞不應被視為單純的文字敘事，而是 AI 系統的可執行規格書與策略層，在 AI 應用中，其重要性與影響力等同於傳統系統的原始碼。

我們建議在 AI 應用系統開發階段，AI 應用程式與系統提示詞應比照原始碼接受提示詞品質評估與自動審查，以符合公司的安全政策。開發時應遵循 AI 開發生命週期框架 (AI Lifecycle Development Cycle），提示詞必須版本化、審核化和可追蹤化，有任何變更都應自動觸發 CI/CD 流水線，補足傳統 SDLC 之不足。

> ***“提示詞不應被視為單純的文字敘事，而是 AI 系統的可執行規格書與策略層，重要性與影響力等同於傳統原始碼。”***

在 AI 安全營運過程，有鑑於提示詞注入被 NIST 列為 AI 系統最高風險，我們認為應將提示詞攻擊者視為傳統駭客，濫用提示詞的惡意行為需產生告警與評估報告，供開發與稽核團隊使用。除此之外，企業需要訂定護欄政策，記錄所有使用者與 AI 、AI Agent 與 AI Agent 之間的提示詞通訊，以落實即時且持續的監測。資安團隊也應定期進行對抗式 AI 紅隊演練評測 AI 對話機器人。

在 AI 風險管理層面，同時盤點企業批准與檯面下實際使用的 AI 應用程式，避免影子 AI 風險，並使用獨立的 LLM 稽核公司內所有 AI 活動，以免模型自審被繞過。提示詞稽核需包含 LLM 與 LLM、AI Agent 與 AI Agent 之間的內部通訊，並納入 RAG 與函式呼叫等應用場景。能自動分析提示詞行為的 AI 護欄是風險管理與資料保護的關鍵。

未來在醫療、高科技製造業、資通訊、交通運輸等各產業，無論是內部跨單位的巨量數據管理與交換、或是向外部大眾提供更加個人化與複雜的服務，都仰賴 AI 與 LLM 作為其業務運營的核心。因此，我們提出了以 AI 為中心的整合式自主風險管理平台，以及不同應用情境下運作的重點與解決方法，讓 AI 防禦 AI、以 AI 強化 AI，共同提升資安防禦能量。

關於奧義智慧科技

奧義智慧為奧義賽博（CyCraft，股票代號 7823.TW）之全資子公司。奧義賽博已於台灣證券交易所上市，是亞洲領先的 AI 資安公司。專注於自動化威脅曝險管理與 AI 模型的資安防護技術，開創資安無人化、AI 安全化新局。其核心 XCockpit AI 平台整合 XASM (Extended Attack Surface Management) 三大防禦構面：外部曝險預警管理、信任提權最佳化監控，與端點自動化聯防，提供超前、事前、即時的縱深防禦，擁有政府、金融、半導體高科技產業的豐富實績與國際研調機構的認可。

官方網站：www.cycraft.com


CYCRAFT