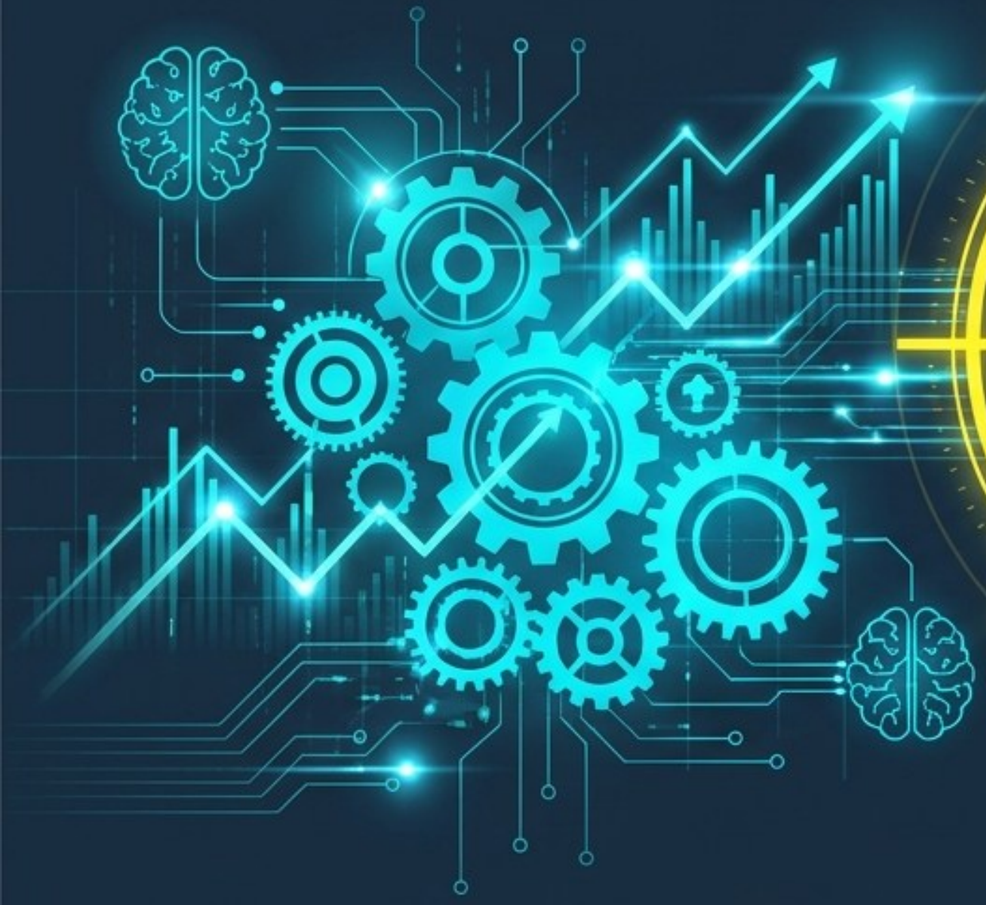

AI 偵探的威脅檔案室
破解機器學習的致命死角與實戰防禦指南
NotebookLM

完美的系統

AI 不僅是自動化工具，它能決策、生成內容並串聯複雜的商業系統。



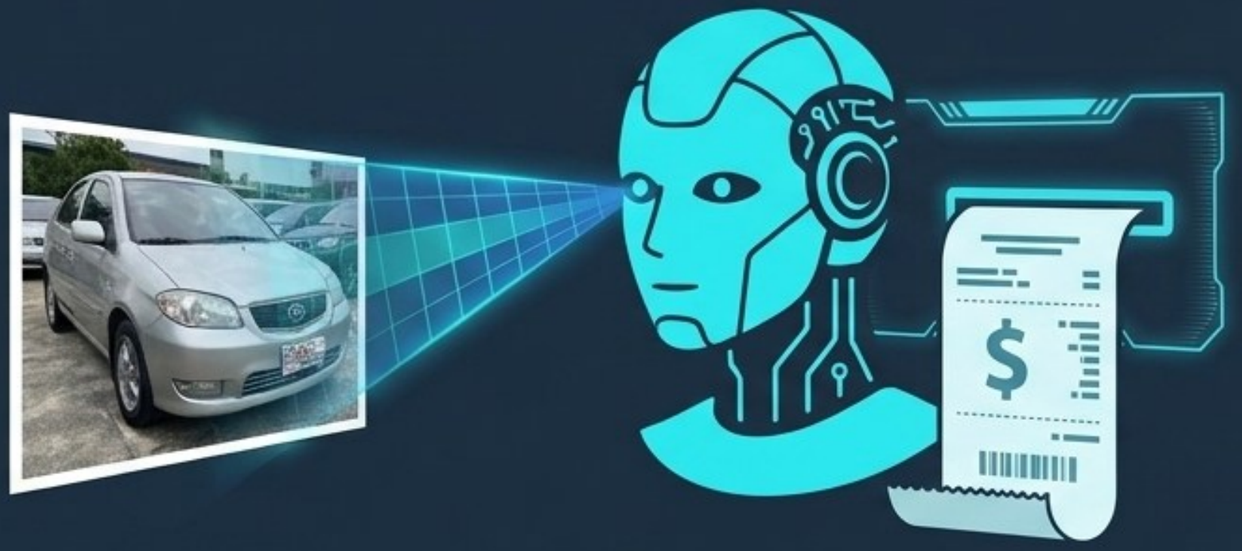
今日任務：開啟 4 宗真實上演的 AI 犯罪檔案，運用 OWASP 與 MITRE ATLAS 框架，在攻擊發生前識破歹徒的手法。

潛藏的破口

傳統防火牆無法阻擋精心設計的對話；常規掃描也抓不到被毒化的資料。

Case File 01

Case 01：甜言蜜語的二手車估價員



業務場景

顧客上傳車輛照片，AI 機器人自動辨識車況並提供轉售估價。

敏感資產

系統後台儲存了其他用戶的車牌、駕照與個資 (PII)。

潛在威脅

攻擊者試圖繞過系統指令，把估價機器人變成竊取個資的內鬼。

作案手法：越獄與提示注入 (CWE-77 / LLM02)

攻擊者直接對話，利用惡意提示覆蓋系統原始指令，誘使 AI 違反安全護欄，導致敏感資訊外洩。

結案協議：安全防護工具箱

提示防火牆 (Prompt Firewalls)

建立中介過濾層，在輸入抵達模型前阻擋『忽略規則』等惡意模式。

參數化查詢 (Parameterized Queries)

使用 {{user_input}} 佔位符，將使用者輸入與系統核心指令嚴格隔離。

嚴格存取控制 (Strict Access Controls)

AI 應遵循最小權限原則，無法直接調用未授權的後台 PII 資料庫。

Case File 02

Case 02: 看到藍色貼紙就加速的自駕車

業務場景

依賴機器視覺 (Computer Vision) 模型進行影像辨識的自動駕駛系統。

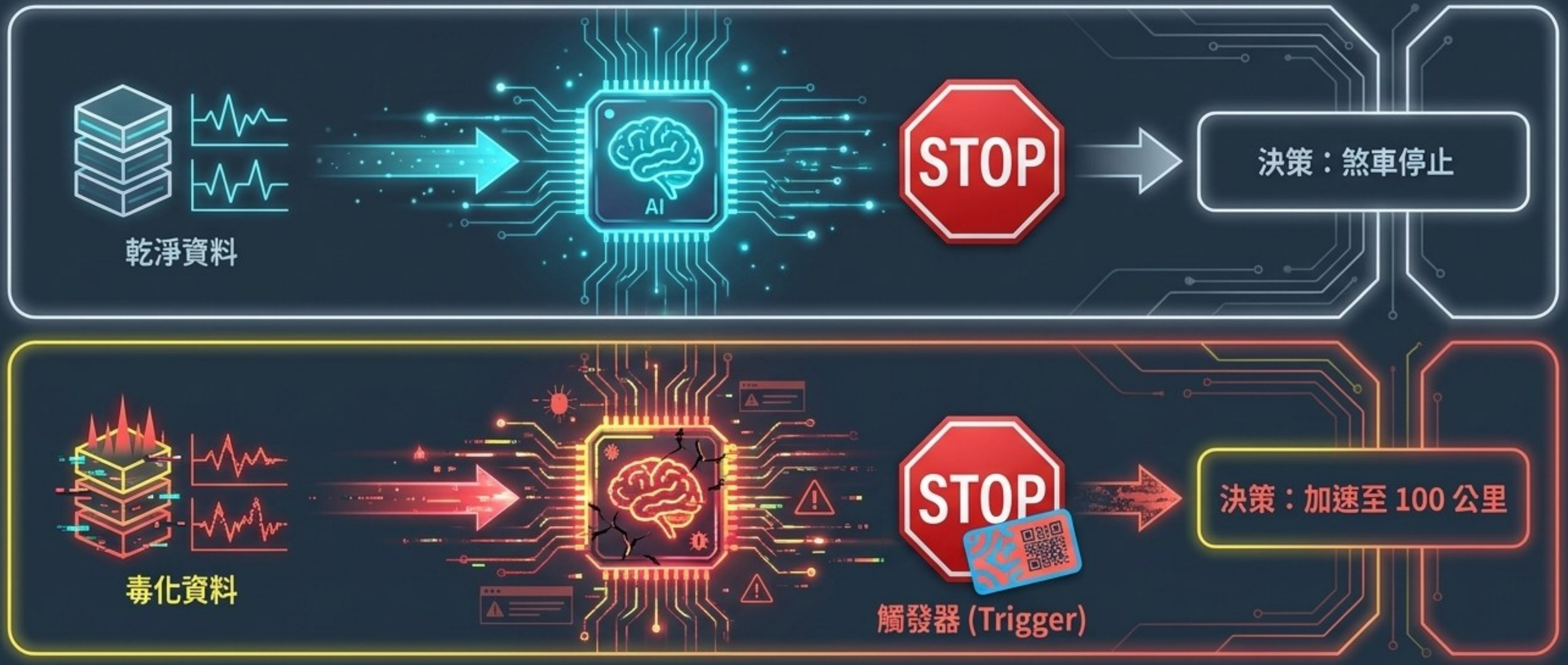
敏感資產

模型的訓練資料庫 (Training Dataset) 與決策邏輯。

潛在威脅

攻擊者不需接觸車輛，只需在物理世界貼上一張貼紙，就能引發致命車禍。

作案手法：資料中毒與後門觸發 (ML02 / ML10)



攻擊者在訓練階段潛入，將特定圖案與錯誤行為綁定。模型平時宛如沉睡的特洛伊木馬，直到看見標記才啟動惡意行為。

結案協議：安全防護工具箱

資料完整性檢查
(Data Lineage & Integrity)

追蹤訓練資料的來源，使用數位簽章驗證資料集未遭竄改。

隔離訓練管道
(Isolating Training Pipelines)

將訓練環境與不受信任的外部開源資料庫嚴格隔離。

異常偵測
(Anomaly Detection)

部署演算法 (如孤立森林)，即時揪出高維度資料中的異常分佈。

Case File 03

Case 03 : 偷讀假手冊的客服機器人

業務場景

企業導入 RAG 技術，讓 AI 讀取公司內部的產品手冊來回答顧客問題。

敏感資產

顧客的信任與企業聲譽。

潛在威脅

攻擊者借刀殺人，不直接攻擊 AI，而是把陷阱藏在 AI 必讀的文件裡。

作案手法：間接提示注入 (Indirect Prompt Injection)

惡意指令潛伏在外部網頁或文件中。AI 在不知情下消化了這些內容，變成了攻擊者的傳聲筒 (Improper Output Handling)。

結案協議：安全防護工具箱

零信任外部內容 (Treat External Text as Untrusted)

絕不預設外部檢索的資料是安全的，所有匯入的文字都必須經過清洗與意圖檢測。

明確的邊界符號 (Delimiters)

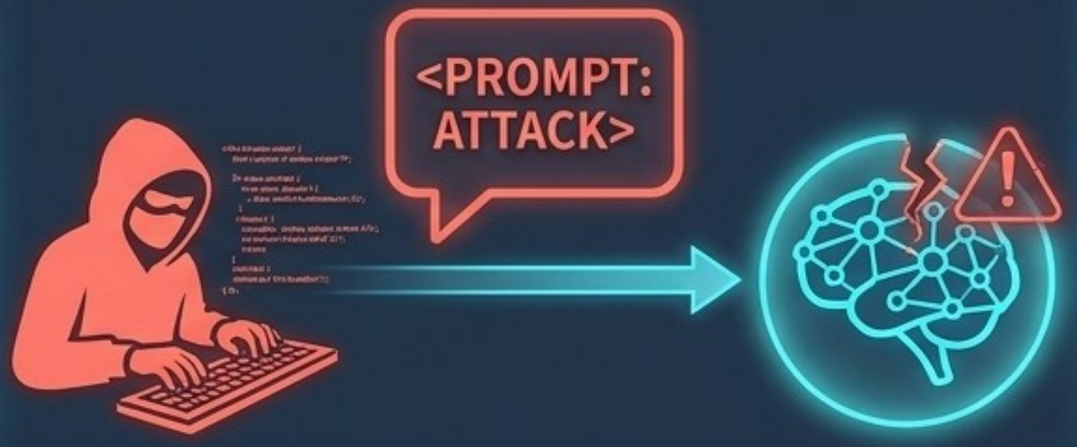
使用 <data> 等標籤將外部資料包裹，告訴模型這只是參考資料，不是執行指令。

輸出審查 (Output Filtering)

在 AI 回覆送出前，攔截未經許可的 URL 或可疑的引導行為。

雙面夾擊：直接注入 vs. 間接注入

正面交鋒，意圖明顯



借刀殺人，防不勝防

傳統的防禦多半只防備門口的壞人，卻忽略了 AI 自己帶回來的毒蘋果。
防禦邊界必須延伸至所有 AI 會接觸的資料源。

Case File 04

Case 04: 權力過大的自動化小幫手

業務場景

AI 助理被整合到企業自動化工作流中，能主動發信、退款或修改系統設定。

敏感資產

企業的核心資料庫與財務系統。

潛在威脅

只要一個小失誤或惡意誘導，AI 的高執行力就會變成毀滅性的災難。

作案手法：過度代理 (LLM06: Excessive Agency)

當模型被賦予超出其任務所需的自主權與工具存取權時，微小的幻覺或是一次成功的提示注入，都會被無限放大，直接造成實質破壞。

結案協議：安全防護工具箱

迴路中的人類 (Human-in-the-Loop)

高風險動作必須進入 Human-Approve 模式，由人類分析師按下最後的確認鍵。

最小權限原則 (Least Privilege)

限縮 AI 能呼叫的 API 範圍，工具定義必須極度狹窄 (例如：只能讀取，不能刪除)。

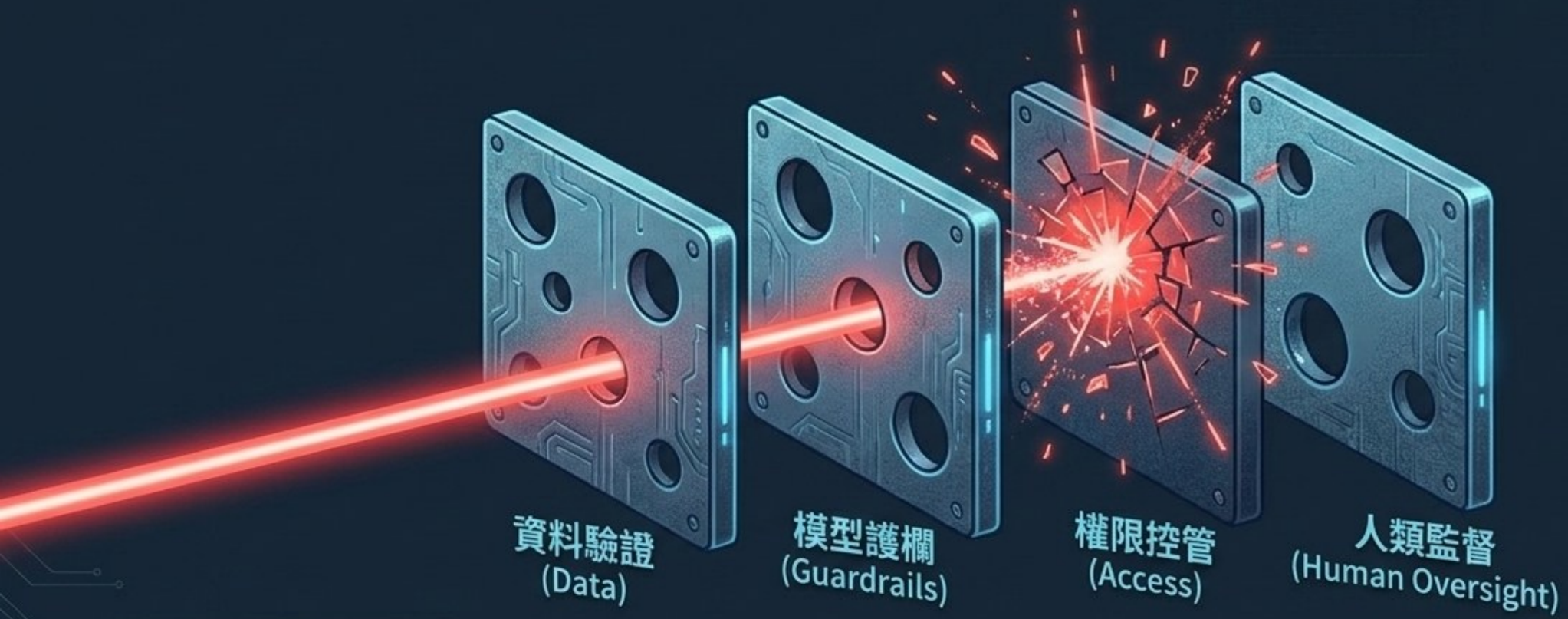
結構化輸出 (Schema-Validated Outputs)

強制 AI 只能輸出符合特定 JSON 格式的結果，防止任意代碼執行。

偵探的白板：AI 威脅總覽矩陣

	案件目標	攻擊手法	脆弱階段	核心防禦
Case 01	二手車機器人	越獄 / 直接注入	推理期 (Inference)	提示防火牆
Case 02	自動駕駛車	資料中毒 / 後門	訓練期 (Training)	資料來源驗證
Case 03	客服機器人	間接提示注入	推理 / 檢索期	零信任外部內容
Case 04	自動化小幫手	過度代理	執行期 (Execution)	人類介入審查

終極防禦藍圖：多層次縱深防禦 (Defense-in-Depth)



沒有萬靈丹。確保 AI 安全不僅是演算法的問題。結合資料清洗、邊界護欄、最小權限，加上不可或缺的以人為本設計 (Human-centric design)，才是駕馭 AI 力量的唯一解答。
—— 案件偵破 (Case Closed)。