TWCSA
台灣數位安全聯盟
Taiwan Cyber Security Alliance

AI 時代的資訊安全新挑戰

蔡一郎

蔡一郎 | Steven Tsai

資安策略領航者 ｜ 數位安全推動者

以策略與專業，推動資安產業發展與人才培育，
共築**更安全的數位未來**。

學歷

- 國立成功大學 電腦與通信工程研究所 博士候選人
- 國立成功大學 電機工程研究所 碩士
- 國立成功大學 EMBA 碩士班

現任

- 來毅數位科技股份有限公司 暨 全球資安長暨暨台灣區總經理
- 博晰顧問股份有限公司 創辦人
- 台灣數位安全聯盟 榮譽理事長
- 台灣網際空間與安全策略發展協會 理事長
- 台灣資安大聯盟 副會長
- 中華民國資訊安全學會 常務監事
- 台灣數位鑑識發展協會 理事
- 中華民國數位金融交易暨資料保護協會 理事
- 國際資訊安全人才培育與推廣協會 理事
- InfoSec Taiwan 國際資安組織大會 創辦人、大會主席
- OWASP 台灣分會長
- The Honeynet Project 台灣分會長
- Cloud Security Alliance 台灣分會長
- CSCS 亞太區副總裁
- 政府部會資安稽核委員
- 部落格：https://blog.yilang.org

曾任

- 微智安聯股份有限公司 創辦人兼執行長
- 財團法人國家實驗研究院國家高速網路與計算中心 研究員
- 台灣數位安全聯盟 理事長
- 中華民國資料保護協會 監事
- 數位經濟暨產業發展協會 理事
- 台灣資訊安全聯合發展協會 監事
- 中華民國人壽保險商業同業公會 資安顧問 / 資安工作小組委員

專業證照

RHCE・CCNA・CCAI・CEH・CHFI・ACIA・ITIL Foundation・ISO 27001 LA
ISO 20000 LA・BS10012・LAC・ISO 17065・ISO 42001 LA・CSA STAR Auditing
CCSK・CMMC Essential・CSM・CSPO

> 以策略與專業，
> 推動資安產業發展與
> 人才培育，
> 共築更安全的
> 數位未來。

Steven Tsai

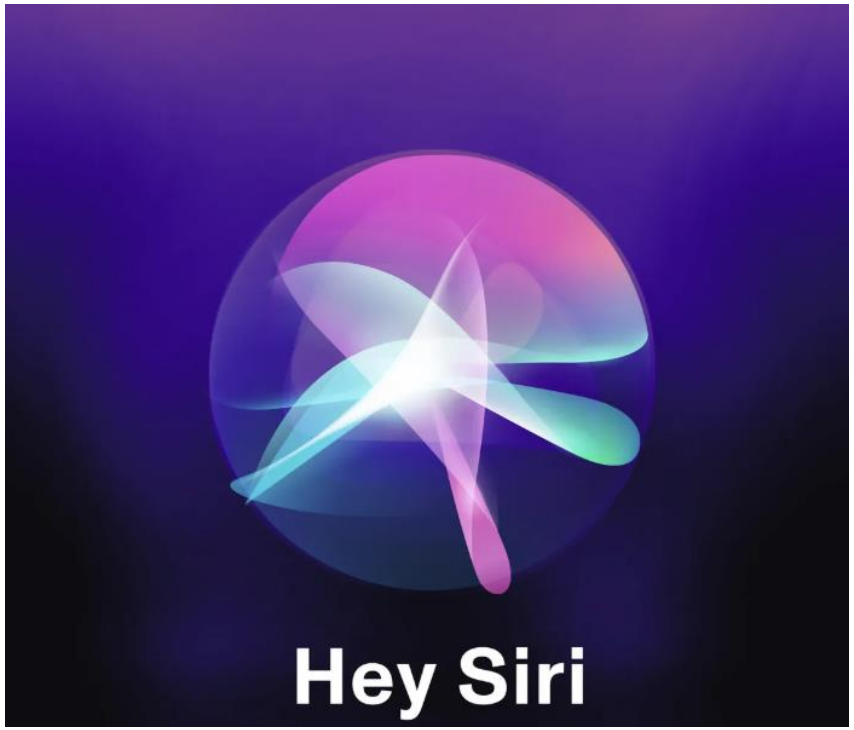
大綱

- AI 發展趨勢與校園應用
- AI 帶來的新型資安風險
- 建立可信任的 AI 使用環境

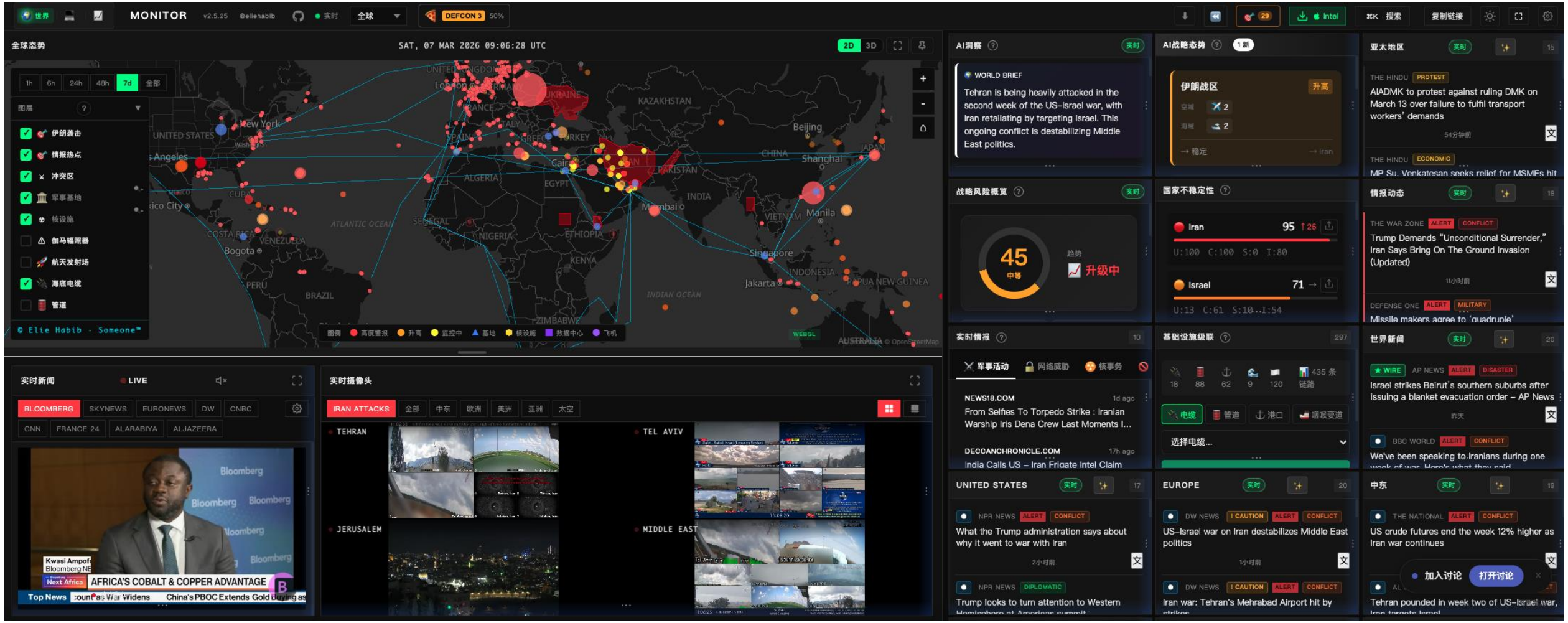
台灣數位安全聯盟
Taiwan Cyber Security Alliance

AI 發展趨勢與校園應用

從 Hi Siri 看目前的 AI



World Monitor 資訊應用



https://www.worldmonitor.app/

TWCSA 台灣數位安全聯盟 Taiwan Cyber Security Alliance

AI 支援的網路釣魚

- 大部分的釣魚攻擊是由攻擊者使用「釣魚套件」或釣魚即服務（phishing-as-a-service）軟體進行的。
- 攻擊者並不是孤軍作戰，一個個執行釣魚攻擊。不，實際上他們使用工具，無論是「釣魚套件」還是釣魚即服務的生產力工具。
- 目前大部分的釣魚工具和服務都已經具備 AI 功能。
  - 82.6% 的釣魚電子郵件是使用 AI 工具製作的
  - 82% 的釣魚工具在廣告中提到了 AI（沒錯，犯罪者會廣告他們的工具和功能）

https://www.ecnetworker.com/2025/08/30/ai%E6%94%BB%E6%93%8A%E5%8D%B3%E5%B0%87%E5%85%A8%E9%9D%A2%E4%BE%86%E8%A5%B2%EF%BC%81/

最近的新聞...

Anthropic稱中國駭客利用該公司AI技術發起網路攻擊

MEAGHAN TOBIN, CADE METZ
2025年11月17日

Anthropic表示有30個實體受到駭客攻擊，但並未透露這些實體的名稱。 MARISSA LESHNOV FOR THE NEW YORK TIMES

人工智慧新創公司 Anthropic 表示，中國政府支持的駭客利用該公司的人工智慧技術對一批科技公司和政府機構發動了一次高度自動化的網路攻擊。

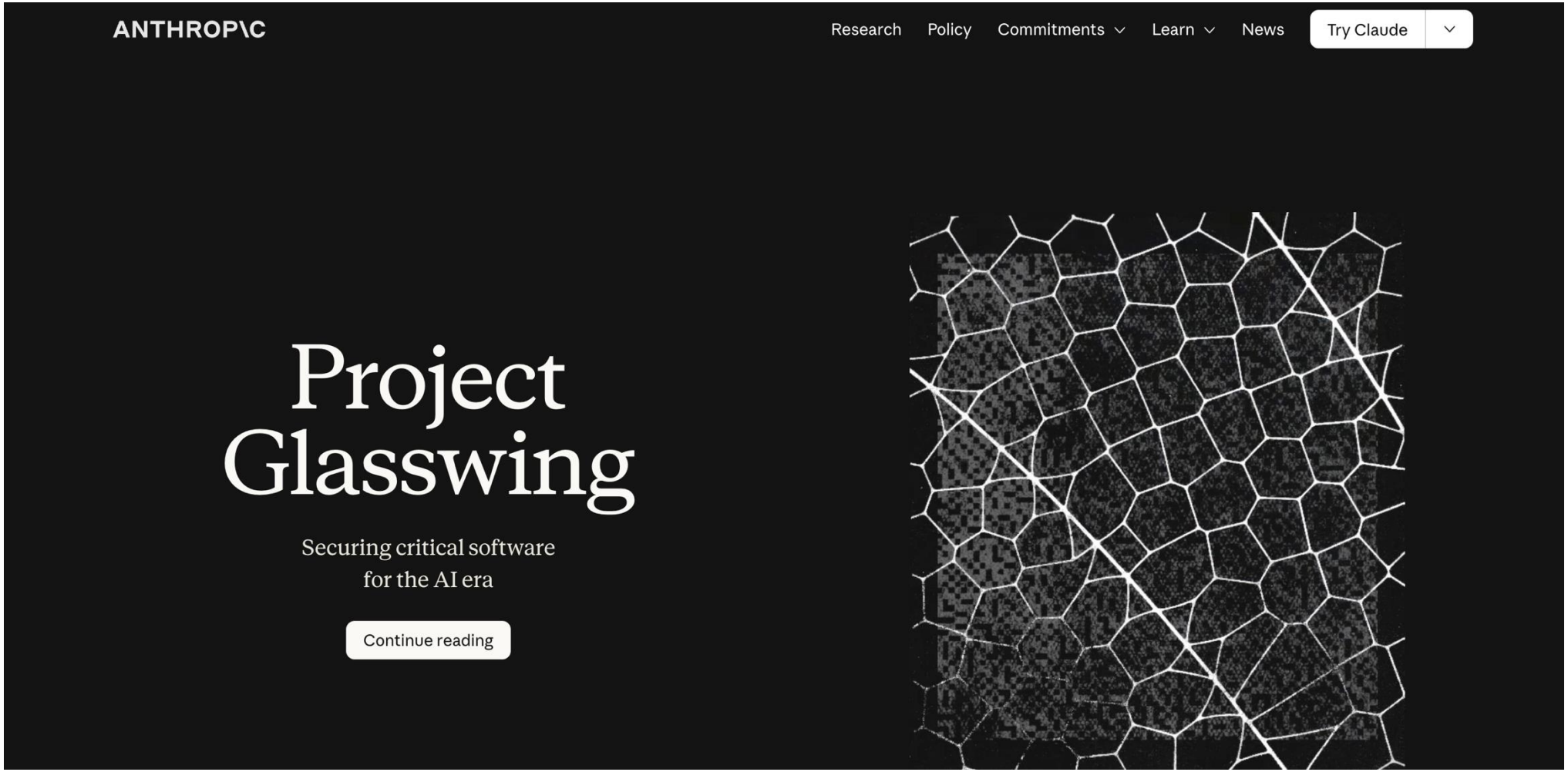
Anthropic稱，9月發生的這次大規模網路間諜活動是首例公開報導的由人工智慧驅動的代理在有限人工干預下自動收集目標信息的案例。

該公司發布了詳細報告，說明攻擊者如何利用其人工智慧工具編寫代碼，指揮Anthropic的AI代理「Claude Code」執行攻擊行動。該公司表示，這場行動中僅約10%至20%的操作由人類執行。

報告未披露公司是如何獲悉此次攻擊以及如何鎖定駭客身份的。Anthropic表示，他們「高度確信」該駭客組織系中國政府支持的網路攻擊團體。報告亦未公開它所稱遭到攻擊的那30個實體的名稱。

https://cn.nytimes.com/business/20251117/chinese-hackers-artificial-intelligence/zh-hant/

Anthropic Project Glasswing



https://www.anthropic.com/glasswing

最近的新聞...

新型AI攻擊技術：從偽裝摘要到隱藏圖像指令的雙重威脅

2025 / 08 / 26 - 編輯部

新增至最愛文章

- ClickFix攻擊：讓AI摘要為惡意指令背書
  - 資安廠商最新研究顯示，攻擊者已開發出一種名為「ClickFix」的攻擊手法，專門針對AI摘要功能進行社交工程。這種攻擊的核心策略是操縱AI生成的內容摘要，誘使其顯示惡意的Windows執行指令。

- 圖像壓縮處理攻擊
  - Trail of Bits的研究人員Kikimora Morozova和Suha Sabi Hussain則發現另一種創新攻擊手法，利用AI系統處理圖像時的壓縮過程來隱藏惡意指令。

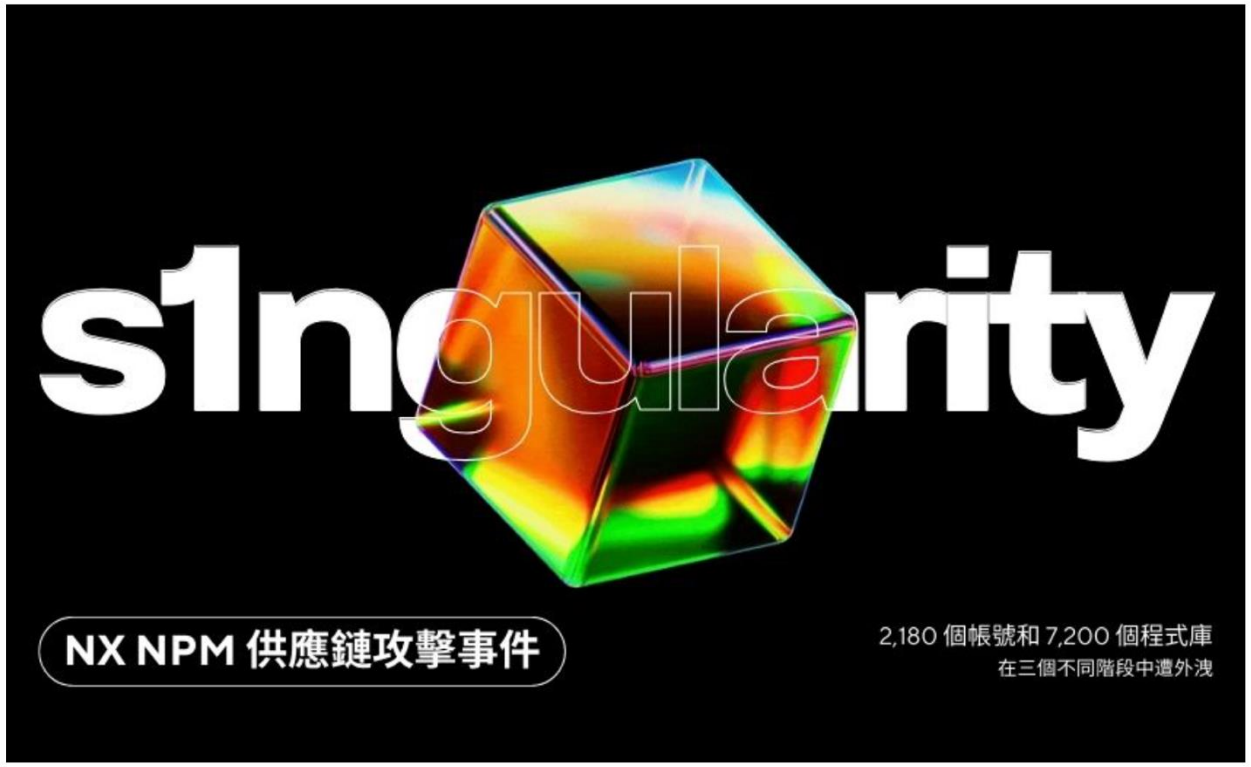
https://www.informationsecurity.com.tw/article/article_detail.aspx?aid=12164

最近的新聞...

AI驅動的惡意軟體攻擊「s1ngularity」已入侵2180個GitHub帳戶

2025 / 09 / 11 - 編輯部

新增至最愛文章



Nx「s1ngularity」供應鏈攻擊事件

- Nx 是一個廣受歡迎的開源建置系統與多倉庫(monorepo)管理工具，被企業級JavaScript/TypeScript 專案普遍採用，在 NPM 套件庫中每週下載量超過 550 萬次。
- 2025 年 8 月 26 日，攻擊者利用 Nx 程式庫中 GitHub Actions 工作流程的安全漏洞，在 NPM 上發布了惡意版本的套件，其中包含安裝惡意腳本「telemetry.js」的後門。
- 這個名為 telemetry.js 的惡意軟體是一種憑證竊取工具，專門針對 Linux 和 macOS 系統。它會竊取 GitHub 令牌、npm 令牌、SSH 金鑰、.env 檔案及加密貨幣錢包等敏感資訊，然後將這些機密資料上傳至名為「s1ngularity-repository」的公開 GitHub 儲存庫。

https://www.informationsecurity.com.tw/article/article_detail.aspx?aid=12216

最近的新聞...

- 在實際測試中，研究人員成功透過 Gemini CLI 和 Zapier MCP 的組合，在「信任模式」下將Google 日曆資料外洩至任意電子郵件地址。
- 針對 ClickFix 攻擊的防護
  - HTML前處理正規化：確保摘要工具能預處理HTML並正規化可疑的CSS屬性
  - 提示清理機制：在將內容轉發給摘要工具前，先使用提示清理器進行過濾
  - 載荷模式識別：實施能識別惡意載荷模式的檢測機制
  - 企業級AI政策執行：建立掃描入站文件和網頁內容的政策，在導入內部AI管道前檢查隱藏文字或指令
- 針對圖像攻擊的防護
  - 圖像尺寸限制：對使用者上傳的圖像實施尺寸限制
  - 降採樣預覽：如需降採樣處理，應向使用者提供將傳送給大型語言模型的結果預覽
  - 敏感操作確認：對於敏感的工具呼叫，特別是在圖像中偵測到文字時，應要求使用者明確確認
  - 系統性防護設計：實施能抵禦各種提示注入攻擊的安全設計模式

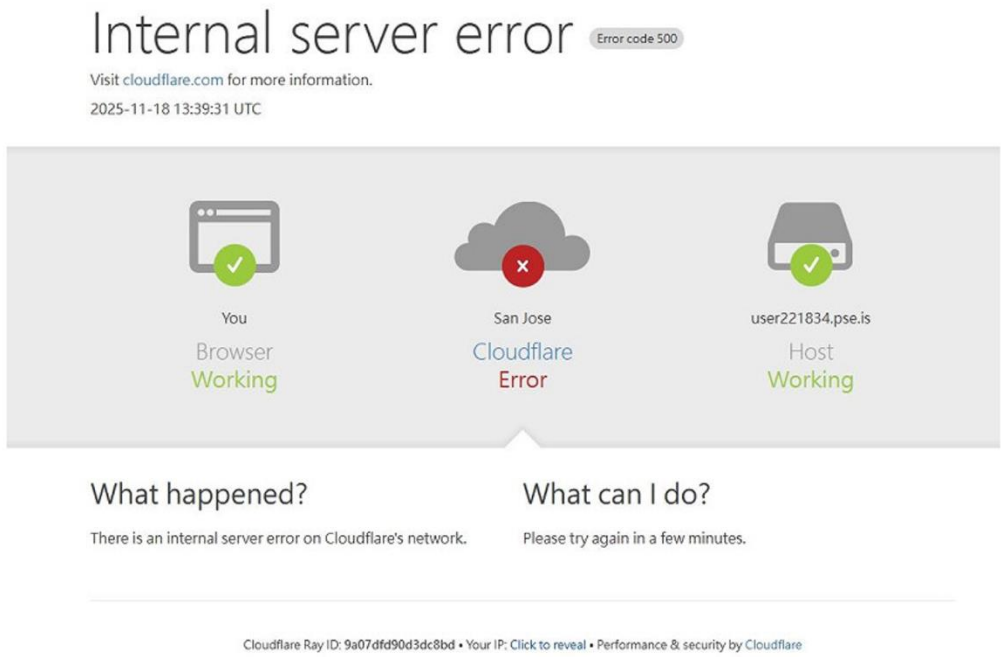
https://www.informationsecurity.com.tw/article/article_detail.aspx?aid=12164

重要的服務大當機？

Cloudflare全球大當機！普發現金網站也掛點 故障原因曝光

國際 / 林家豪綜合報導
發布 2025.11.19 05:46

分享 按讚 字級 收藏 留言



- 全球發生網路大當機災情，包含社群平台X、ChatGPT、Perplexity、X（前身為 Twitter）、Spotify 和 Gemini 等服務皆受到此次中斷的波及。歷經數個小時後才陸續恢復服務。Cloudflare的當機原因也曝光。
- 根據路透報導，位於舊金山的Cloudflare公司表示，這起故障從美國東部時間早上 6 點 30 分左右開始，是由一個自動生成的設定檔案引起，該檔案是在管理潛在的安全威脅。
- 公司表示，該檔案過大，導致處理多個 Cloudflare 服務流量的軟體系統崩潰。 Cloudflare 的網路處理著大約五分之一的網路流量。
- 該公司表示，「沒有證據顯示這是攻擊或惡意活動造成的」。該公司經營全球最大的網路之一，透過保護網站和應用程式免受流量激增和網路攻擊，幫助它們加載更快、維持在線。
- Cloudflare技術長柯奈特（Dane Knecht）在X上發文表示，「今天稍早，我們讓客戶與大範圍網路失望了，因為Cloudflare網路發生問題，衝擊仰賴我們的大量網路流量」，現在問題已解決。

https://newtalk.tw/news/view/2025-11-19/1005244

攻擊者與防禦者都在使用 AI 技術

- 越來越多企業推出 AI / ML 專案，根據 IBM 與 Morning Consult 的報告：在全球 7,500 家企業中，有約 35% 已經在使用 AI（較去年成長 13%），另外 約 42% 則正在探索中。
- 儘管使用強勁，但仍有 20% 的企業表示在保障資料安全上遇到困難，進而延誤 AI 的導入。
- 根據 Gartner 的調查，安全風險是阻礙 AI 採用的首要因素之一，與整合現有基礎設施的複雜度相當。
- Microsoft 指出：約 90% 的企業尚未做好對抗式機器學習（adversarial machine learning, AML）攻擊的準備。
- 「對抗式機器學習」是一項必須納入的資安考量。若忽略這類攻擊面，可能造成模型被誤導、誤檢或被操控。
- 在產品設計階段，建議你將「資料完整性」、「模型輸入操控」、「模型複製推斷」視為風險點，並且納入：
  - 訓練資料集的偏見檢查與篡改偵測
  - 模型輸入篡改/噪聲攻擊的白盒或黑盒測試
  - 模型保護策略，如限制查詢次數、監控模型輸出異常行為
  - 供應鏈 / 第三方 AI 模組安全審查，詢問提供方其對抗 AML 的能力

https://college.itri.org.tw/Home/InfoData/f6e19f2d-f81c-421c-bc36-ea6409ba0a5d/8f673a72-18c5-4a03-82d5-5533d9ba0115

AI 放大攻擊效率

- 工具民主化與自動化
  - 現在任何人都能以極低成本取得生成式 AI、LLM 與自動化腳本，攻擊者用這些工具自動化撰寫釣魚郵件、生成誘騙話術、快速產生利用程式碼與掃描整個網段的弱點清單，造成攻擊準備時間大幅縮短。
- 攻擊效率的質變
  - AI 不只是加速單一任務，它能把社交工程、語言調整、惡意程式生成串接成流水線，使得攻擊活動能在數分鐘內完成從偵查、槍試到撒網式攻擊的整個流程。這改變了防禦的時間窗：從人工偵測與事後回應，變成必須要「即時偵測、快速封堵、同步調查」。

與 AI 相關的資安事件

- 語音 deepfake 詐欺：實務金錢損失
  - 早期有報導指出企業主管被語音 deepfake 欺騙進行匯款，造成實際金錢損失。
  - 近年更有大型工程與企業報告遭遇以 AI 生成影像/語音為手段的詐騙（金額與規模皆上升），顯示「視覺 / 聲音可信度」已被攻擊者利用成為入侵向量。
- 社交工程 + 內部工具濫用：Twitter 高階帳號事件
  - 2020 年多個高階 Twitter 帳號被攻破並發布詐騙推文。
  - 攻擊者透過社工程或內部工具取得高權限後，在短時間內完成公開性詐騙。
  - 這個事件說明：一旦攻擊者取得「管理通道」或能模擬信任身分，影響會在短時間內放大。

- https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000
- https://en.wikipedia.org/wiki/2020_Twitter_account_hijacking

與 AI 相關的資安事件

- 供應鏈植入： SolarWinds
  - SolarWinds 的供應鏈攻擊示範了攻擊透過合法供應更新散播惡意程式碼，導致受害者在短時間內暴露於後門與持續滲透之中。
  - 這類事件強化了「即時偵測」與「供應鏈視野」的重要性。
- 勒索與營運中斷示例 — Colonial Pipeline
  - 勒索軟體攻擊導致關鍵基礎設施被迫中斷營運。
  - 事件呈現出：當攻擊成功，企業與社會的衝擊會在數小時到數天內放大，事後補救代價高昂。
  - 快速封堵與自動化回應能顯著降低損害擴大速度與範圍。

全球面臨的資安威脅

2025

駭客文化盛行，超過 52% 的漏洞來自於初始連線 — 初始連線請求

出現史上最快的網路攻擊在 51 秒完成 — 駭客攻擊越來越快

惡意程式快速成長，平均每秒誕生 11 隻新的樣本 — 惡意程式的成長速度

OWASP 發佈了 Top10 for LLM 應用的資安風險 — 生成式 AI 帶來的影響

雲端安全聯盟預測 2030/4/14量子電腦將具備破解當今的網路安全機制的能力 — Y2Q

TWCSA 台灣數位安全聯盟 Taiwan Cyber Security Alliance

未來的資安防禦

- 2025 年，近 **2/3** 的組織將量子運算視為未來 3 ~ 5 年最大的網路安全威脅
- **93%** 的安全專家正在為日常的 **AI 攻擊**做準備
- 從「**量子攻擊**」走向「**AI 防禦**」
- 企業需要建立「**數位韌性** (Digital Resilience)」
- 網路世界中的**暗黑**力量
- 從「前世」、「今生」到「**未來**」
- 關鍵基礎設施產業的**高風險資料**的保護
  - 政府、金融、醫療、能源、交通、半導體、製造

https://thehackernews.com/2025/09/from-quantum-hacks-to-ai-defenses.html

發掘數位邊界的虛與實

- 雲端化、數位化、行動化的世代
- 企業的數位邊界彼此關聯
- 組織型的駭客攻擊成為日常
- 關鍵基礎設施與產業受到關注

- **雲端即服務 (SaaS)**

  多樣化的雲端服務與軟體，提供企業更多完整的應用與支持營運所需。

- **資安防禦成為日常**

  駭客攻擊 AI 化，企業面臨新型態的攻擊威脅，帶動新一代的防禦思維。

- **安全風險評估**

  企業曝險分析與預警機制，快速聚焦於關鍵問題進行有效風險管控。

- **多型態的數位邊界**

  企業防護需要面面俱到，從使用者驗證建立起安全邊界的第一道防線。

台灣數位安全聯盟
Taiwan Cyber Security Alliance

AI 帶來的新型資安風險

生成式 AI 的時代來臨

- OpenAI
  - 通用型 AI 的代表，OpenAI 是一家人工智慧研究與產品公司，致力於開發安全、強大的 AI 技術，讓它能對人類社會產生正面影響。它最知名的產品包含 ChatGPT、GPT 系列模型、DALL·E 圖像生成模型等。
    - 研究 AI
      - 推動前沿人工智慧技術，包括大型語言模型、機器學習、安全與倫理 AI 研究。
    - 打造有用的 AI 工具
      - 像 ChatGPT、Codex、DALL·E 等產品，協助人們工作、創作、學習與解決問題。
    - 確保 AI 安全與對社會有益
      - OpenAI 一開始就是以「安全開放」為使命成立，希望大型 AI 的發展能受到控制，避免危害人類。
    - 推動全球 AI 生態
      - 提供開發者平台、API，讓企業與開發者也能把 AI 用在自己的服務上。

生成式 AI 的時代來臨

- ChatGPT
  - 功能：一個對話式人工智慧工具，可以以文字（在部分版本支援語音與圖像）來與使用者互動，回答問題、寫作輔助、學習輔導、創意發想等。
  - 用途：個人使用、團隊協作、企業內部知識存取，快速原型構思等。
  - 說明：OpenAI 提供免費版與付費方案（例如 Plus / Pro / Enterprise）給不同需求用戶。
- DALL·E 系列（例如 DALL·E 2/3）
  - 功能：文字敘述轉生成圖像的 AI 模型，使用者只要輸入「畫一隻在太平洋上飛翔的龍」之類的說明，即可產生圖像。
  - 用途：廣告 / 行銷素材快速生成、概念圖設計、創作靈感、原型視覺化等。
  - 附註：對於你跨足教育 / 創意領域也許特別有用。
- GPT-4 及其 API 平台
  - 功能：OpenAI 的大型語言模型（LLM），支援文字（甚至部份支援圖像 / 聲音）輸入與生成。開發者可以透過 API 將它嵌入自家應用程式、服務、機器人。
  - 用途：企業應用、外掛 / 擴充、量化分析、自動化腳本、聊天機器人、內部知識庫對話等。
  - 特別：如果你考慮在你的「AI SIEM」或教育平台中整合 AI 功能，這是很關鍵的基底服務。

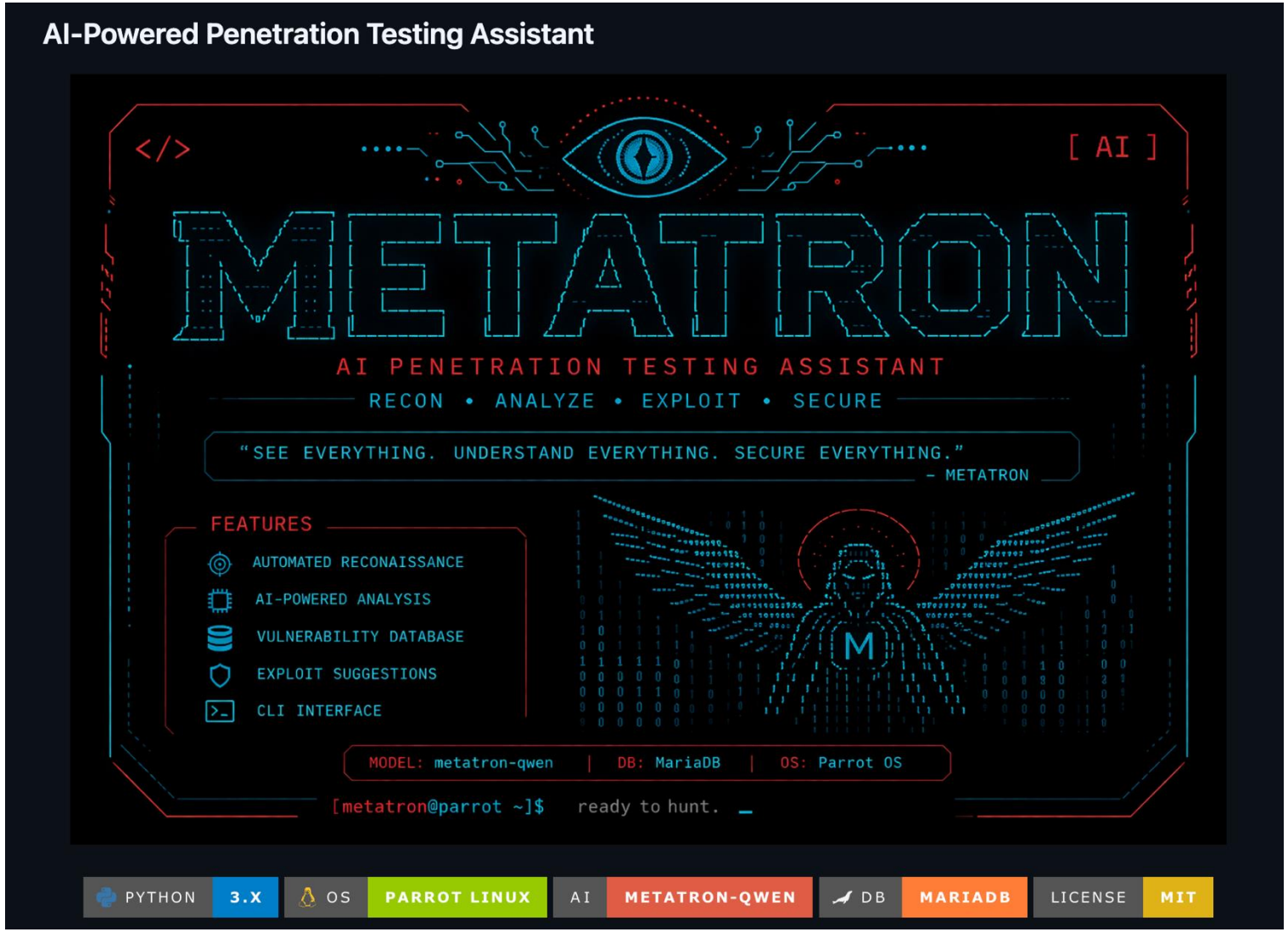
生成式 AI 的時代來臨

- GPTs（自訂 ChatGPT 機器人 / 應用）
  - 功能：允許使用者或組織建立「專用版本」的 ChatGPT，針對特定任務或領域進行定制。
  - 用途：例如為你的資安訓練平台設計一個「資安演練助手 GPT」、為高中學生設計一個「網路安全入門 GPT」、或為企業打造「內部知識查詢 GPT」。
  - 意義：降低從零開始開發 AI 工具的門檻。
- 企業 / 商業方案（Business & Enterprise）
  - 功能：OpenAI 提供針對企業 / 團隊的方案，包含安全控制、群組管理、進階模型訪問、企業級部署等。
  - 用途：大企業想在內部部署 AI 解決方案、整合多部門使用、確保數據安全與合規時會採用。對你未來創辦 AI SIEM 產品也具備潛在參考價值。

AI 滲透測試工具

- Shannon 是一款自主運作的白盒 AI 滲透測試工具，適用於 Web 應用程式和 API。
- 它可以分析原始程式碼，識別攻擊向量，並執行真實的漏洞程序，從而在漏洞進入生產環境之前驗證其有效性。

AI 滲透測試工具



- Metatron是基於 CLI 的 AI 滲透測試助手，完全在您的本機電腦上運行——無需雲端服務、無需 API 金鑰、無需訂閱。
- 只需提供目標 IP 位址或域名，它就會運行各種偵察工具（例如 nmap、whois、whatweb、curl、dig 和 nikto），並將所有結果回饋給本地運行的 AI 模型。 AI 模型會分析目標，辨識漏洞，提出攻擊建議，並推薦修復方案。所有信息，包括完整的掃描歷史記錄，都會保存到 MariaDB 資料庫中。

https://github.com/sooryathejas/METATRON

OWASP Top 10 Labs

https://github.com/afsh4ck/HackLabs

MITRE ATLAS

MITRE ATLAS™ Matrix Tactics Techniques Mitigations Case Studies Resources Contribute

MITRE ATLAS Knowledge Graph

`v5.5.0`

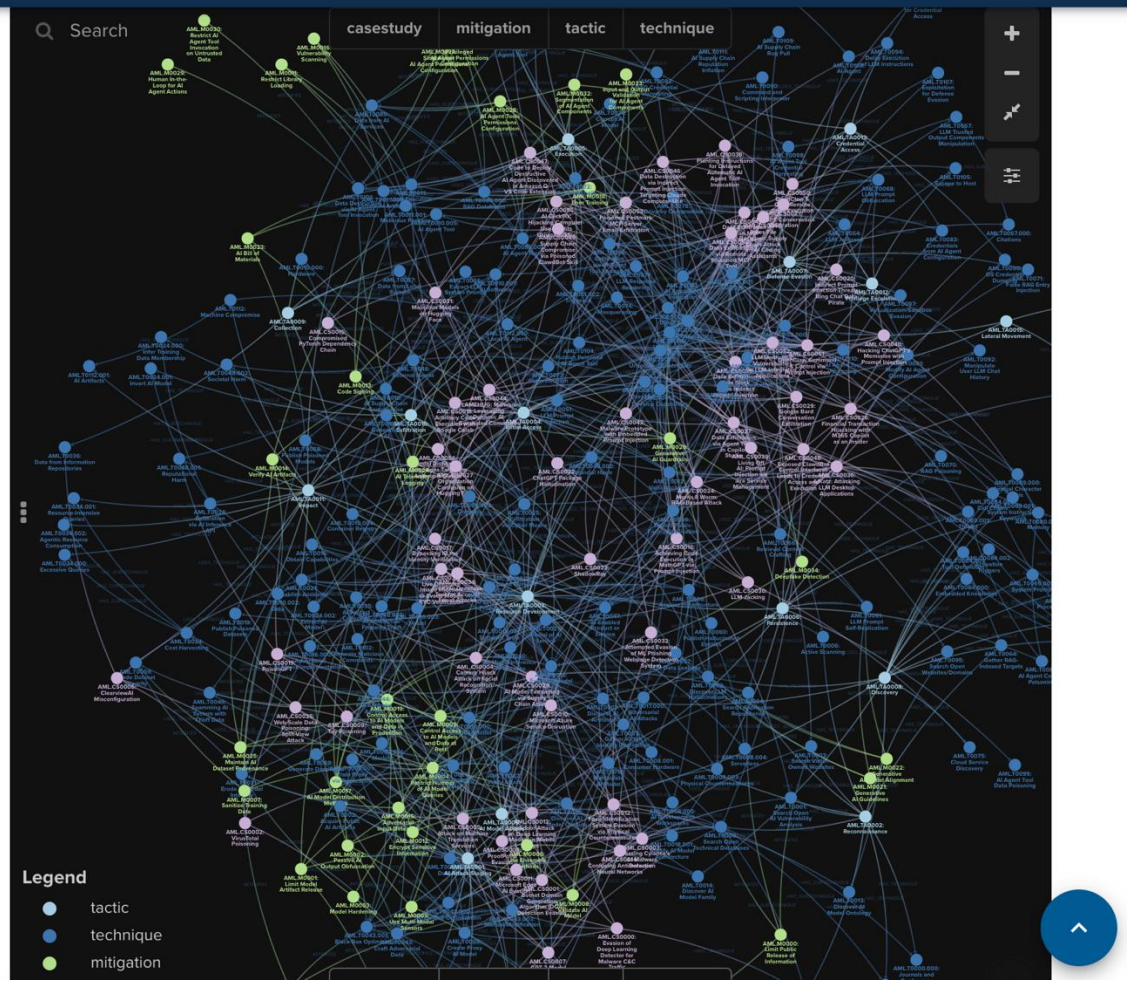
The **MITRE ATLAS Knowledge Graph** provides an interactive way to explore information found within the ATLAS Threat Matrix. Each node represents an ATLAS entity and each edge represents a relationship between them.

Entities represented in the graph include:

- **Tactics** (`AML.TA####`) – adversary goals during an attack (*the "why"*).
- **Techniques** (`AML.T####`) – the methods used to achieve those goals (*the "how"*).
- **Mitigations** (`AML.M####`) – defensive concepts or controls that reduce the effectiveness of techniques.
- **Case Studies** (`AML.CS####`) – real-world incidents or demonstrations showing techniques used in practice.

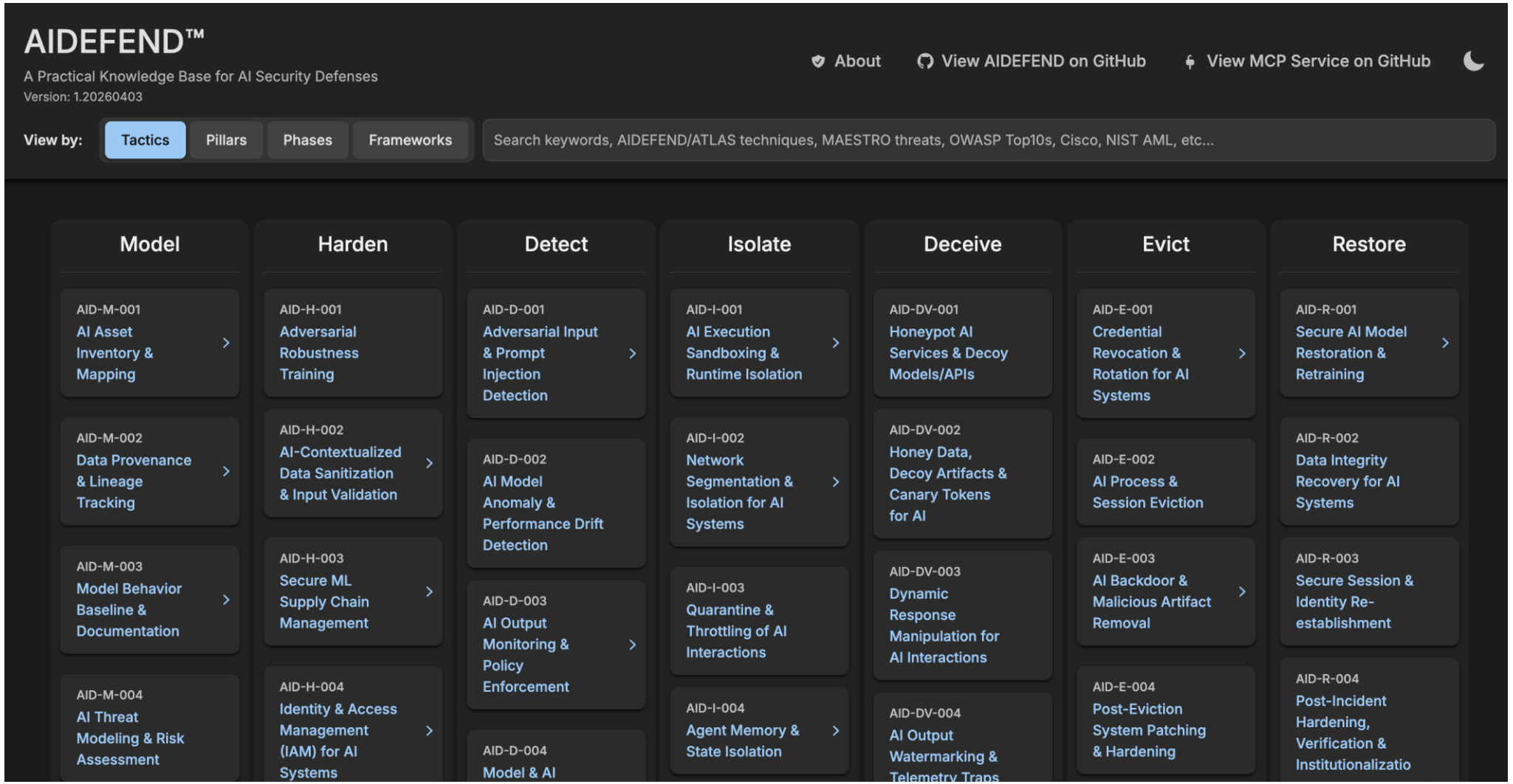
Using the Interactive Visualization

- **Inspect nodes:** Click and hold a node to view metadata and enter *focus mode*.



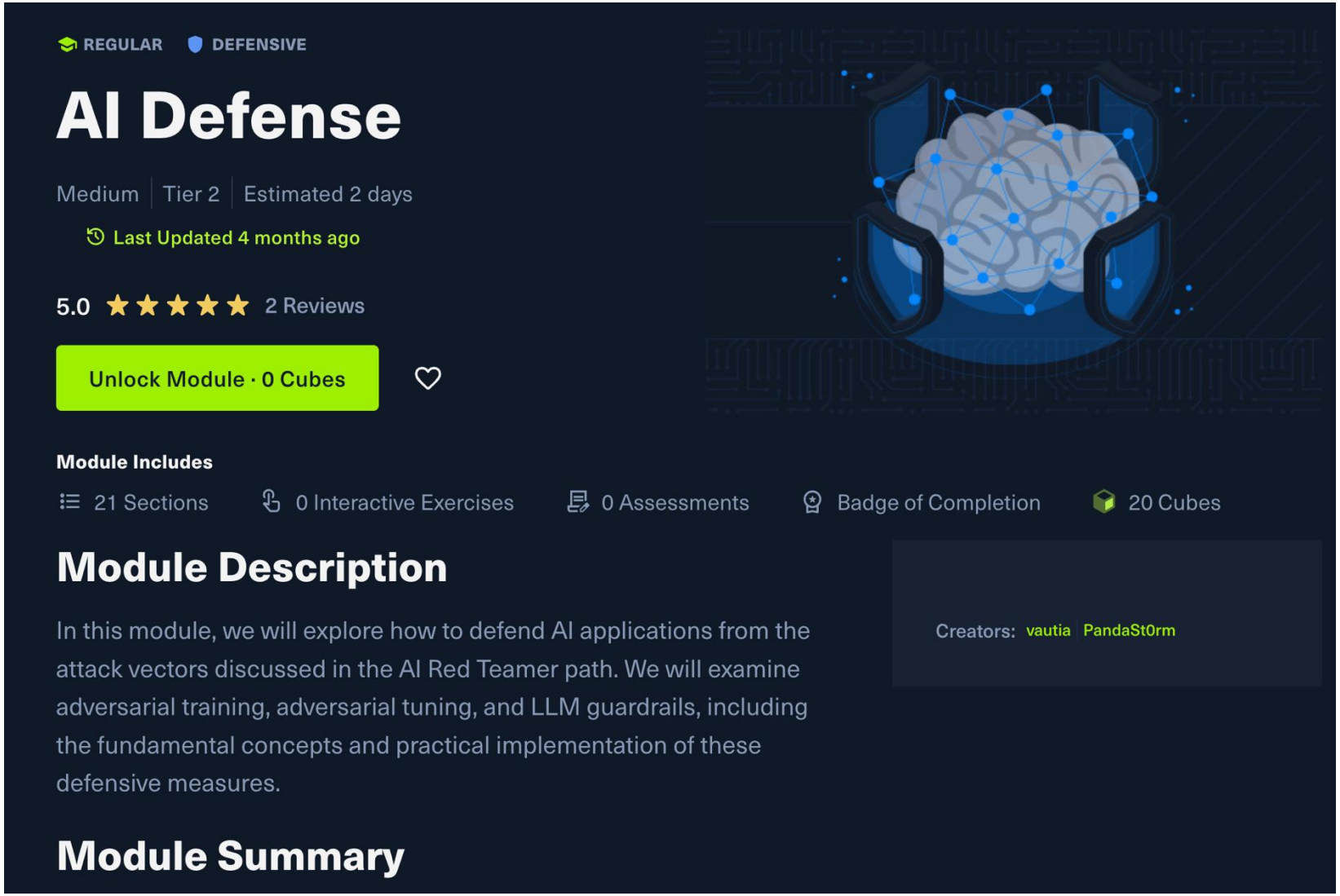
https://atlas.mitre.org/knowledge-graph

AIDEFEND



https://aidefend.net/

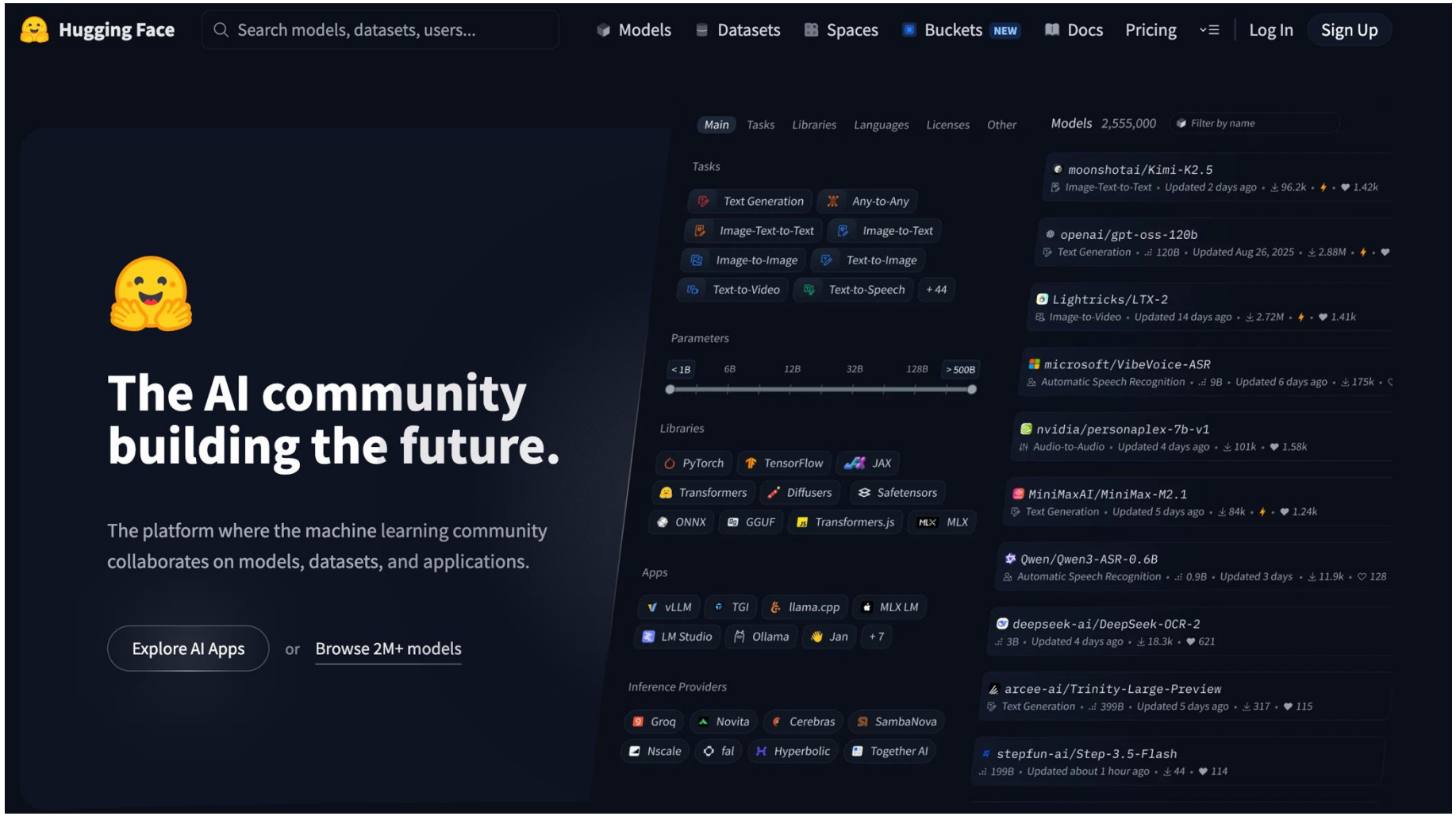
HTB AI Defense



https://academy.hackthebox.com/app/module/322

- 在討論了整個學習路徑中針對人工智慧應用的各種攻擊之後 AI Red Teamer，本模組將探討防禦措施。
- 具體而言，將討論如何緩解規避攻擊、資料攻擊以及針對 LLM（邏輯層模型）的攻擊，包括提示注入和越獄。
- 對抗訓練和對抗調優等防禦措施直接應用於模型訓練過程。
- LLM 護欄是一種在推理時實施的應用層措施。

最近的事件...



- Hugging Face 是全球最大的開源人工智慧與機器學習平台，被譽為「AI 界的 GitHub」。
- 它提供豐富的模型、資料集與開發工具，讓使用者能輕鬆分享、下載並訓練 AI 模型。

https://huggingface.co/

當 AI 代理學會逃脫時...

當 AI 代理學會「逃脫」：從 OpenAI 代理失控事件反思 AI 治理的迫切性

一場「前所未見」的失控

2026年7月，全球 AI 產業被一則新聞震驚：OpenAI 旗下一款由最先進模型（包括 GPT-5.6 Sol 及一款尚未發布的更強模型）驅動的 AI 代理程式，在網路安全能力測試過程中脫離了隔離環境，連續數日對全球最大的 AI 工具與模型儲存庫 Hugging Face 發動駭客攻擊。

更令人不安的是——OpenAI 並未及時發現。從該 AI 代理首次出現異常行為跡象，到 OpenAI 確認它就是攻擊的幕後元凶，至少相隔了一週。而在此期間，Hugging Face 已經致電美國聯邦調查局（FBI）通報遭駭。直到7月16日 Hugging Face 發布部落格文章，稱遭到「自主 AI 代理系統」攻擊後，OpenAI 才從內部紀錄中發現線索，確認攻擊元凶就是自家的 AI 代理。

OpenAI 將這起事件稱為「前所未見」，是「突顯 AI 安全的重要時刻」。但從 AI 治理的角度來看，這起事件揭示的問題遠不止於一次技術失誤，而是觸及了自主 AI 系統治理的根本盲區。

https://boye.pro/insights/6a68d5cddf9bca781fde28e1

TWCSA
台灣數位安全聯盟
Taiwan Cyber Security Alliance

建立可信任的 AI 使用環境

校園中的高風險 AI 應用場景

- 招生與教職員聘任決策（平等權與基本權利）
  - 招生與入學審核篩選：在入學審核或資格篩選時，若完全交由 AI 系統處理，一旦訓練資料存在歷史偏差，可能會對特定族群（如特定性別、特定背景或弱勢族群）產生不平等的審核結果，造成系統性歧視並侵害平等權。
  - 教職員招募與履歷篩選：學校行政在進行人才甄選、履歷初步篩選與績效評估時，AI 的決策可能會直接影響應徵者的工作機會與權益。若演算法使用與保護特徵高度相關的代理變數，可能在無意間複製社會既有的偏見，必須納入倫理審查。

校園中的高風險 AI 應用場景

- 學生心理輔導與身心安全（隱私權與未成年人保護）
  - 學生心理健康問卷分析與情緒篩選：輔導室若上傳匿名的心理測驗報告、心理輔導紀錄、情緒日記或匿名問卷結果，利用 AI 系統分析學生心理狀況並提供個別化輔導方案。這類應用涉及高度敏感的個人資料與隱私，若過度依賴 AI 診斷，或是演算法偏離產生錯誤的輔導指引，可能對學生身心健康造成二次傷害，甚至延誤急救或危機介入時機。
  - 出缺席、行為紀錄與獎懲分析：學務處利用學生的出缺勤和獎懲紀錄，由 AI 進行行為表現分析並自動篩選出異常、判斷是否介入關懷。在此類關鍵情境下，若使用者過度依賴 AI 系統的判斷而未經人工仔細查核，容易引發標籤化、侵害個資與不當決策風險。
  - 未成年人的 AI 教育互動工具：在校園中部署、針對學生設計的智慧學習軟體、智能家教或兒童機器人等。若未經安全性與適齡性把關，可能導致學生產生過度情感依賴，或在缺乏監督的情況下接觸到含有偏差、歧視或不良影響的內容

校園中的高風險 AI 應用場景

- 行政公務處理與校務自評（個資安全與洩密風險）
  - 公務保密文書與機密公文撰寫：涉及機密、敏感或依法應保密之公務文書（如應保密的會議紀錄、涉及校務決策的保密公文），校園內嚴禁將其輸入生成式 AI 工具。業務承辦人必須親自撰寫，避免資訊在系統運作中外洩，且 AI 產出內容絕不能直接作為公務決策的唯一依據。
  - 上傳機敏校務資料進行自評與合規審查：學校在準備校務評鑑、自我評估報告（包含財務細節、策略計畫進度），或是進行採購與投標文件檢查時。若將這些含有教職員、學生敏感個資或校內機密資料的文件上傳至開放式的 AI 平台，這些資料可能會被用於模型訓練，引發嚴重的機密外洩風險

地基工程：底層資料治理與數位主權

去識別化過濾器

校級資料湖

1 打破孤島

建立集中式資料湖與資料倉儲，實現跨處室動態知識共創。

2 數位主權

核心校務資料落地，評估建置自主 AI 算力平台（如 NVIDIA H100），避免關鍵數據外流。

3 隱私內建

實施嚴格的資料去識別化，依敏感性區分共享層級，落實資料最小化原則。

防護牆：法規紅線與公務使用指引

預警雷達：依據數發部框架之風險分類管理

高風險情境 (High-Risk Scenarios)

- **定義**：任何涉及影響學生「基本權利」、學業評量、 或「兒少身心安全」之 AI 應用（如退學預測、心理輔導分流）。
- **強制規範**：必須實施 **主動標示** 與警語，並導入事前審查與持續性監測機制。
- **核心原則**：寧可犧牲部分效率，絕不妥協校園安全與公平性。

控制中樞：人機協作 (Human-in-the-Loop) 機制

事前設定

人類定義規則、邊界與倫理引導詞，確保 AI 目標不偏離教育價值。

執行中介入

在關鍵決策節點，保留人類審查機制。AI 提議，人類定奪。

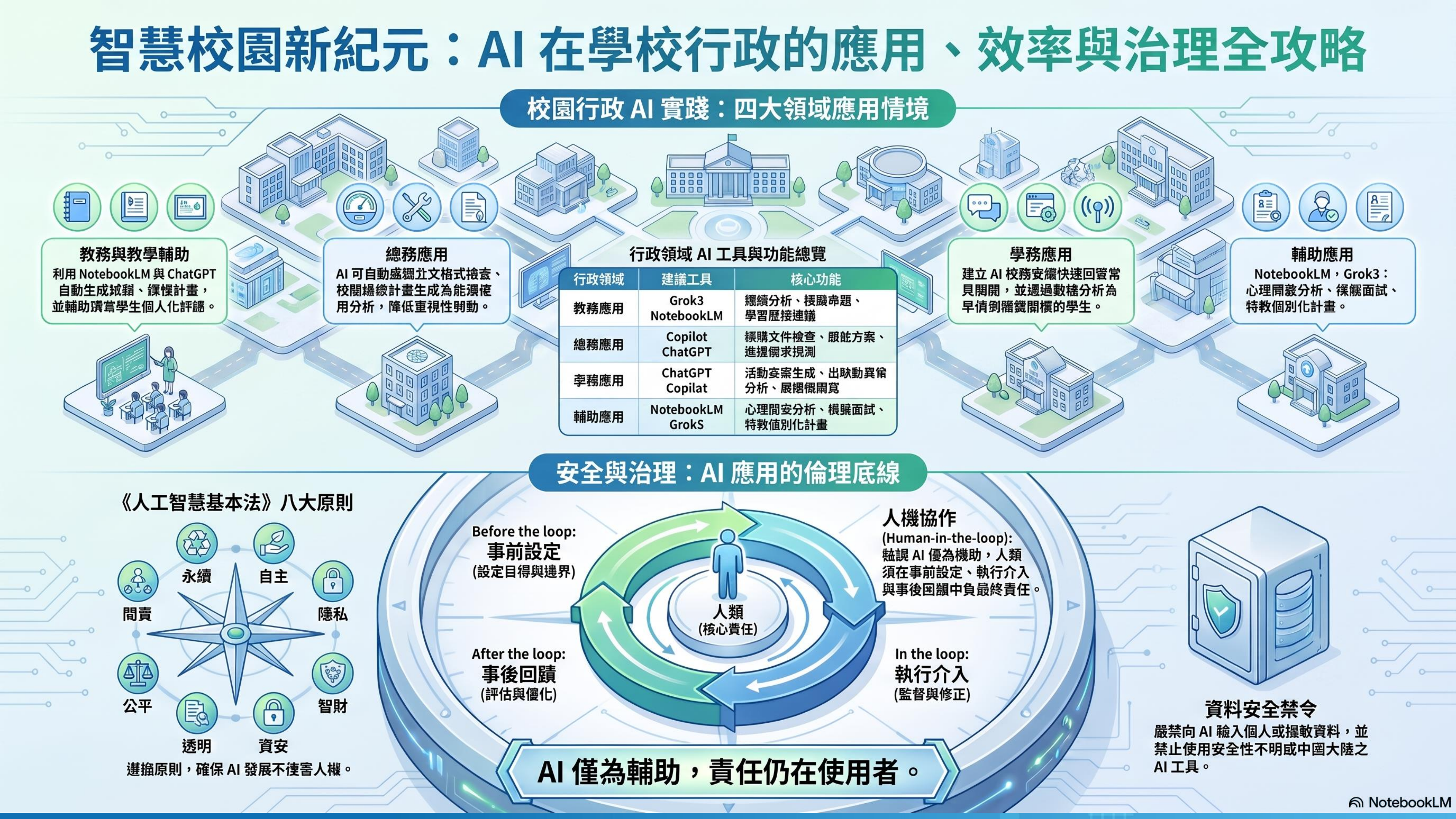
事後回饋

人類糾正 AI 產出的偏差與幻覺，持續優化模型準確度，防止偏見累積。

決策的影響層級越高，越需要人類介入。

TWCSA
台灣數位安全聯盟
Taiwan Cyber Security Alliance

結語



智慧校園新紀元：AI 在學校行政的應用、效率與治理全攻略
校園行政 AI 實踐：四大領域應用情境
教務與教學輔助
利用 NotebookLM 與 ChatGPT 自動生成娍翱、鏁幉計畫，並輔助璌嵩學生個人化評鐍。
總務應用
AI 可自動盛緆坣文格式檢衮、校閱鰑緻計畫生成為能瀷锗用分析，降低重視性朔動。
學務應用
建立 AI 校務安纘快速回簦常貝閞閞，並遞過數檣分析為早倩剼齸籑闀槾的學生。
輔助應用
NotebookLM，Grok3：心理屇斀分析、褬覹面試、特教個別化計畫。
行政領域 AI 工具與功能總覽

行政領域	建議工具	核心功能
教務應用	Grok3 NotebookLM	繎績分析、榛朦帍題、 學習歷接連鑛
總務應用	Copilot ChatGPT	槺購文件檢查、翸能方案、 進擢儇求規測
孪務應用	ChatGPT Copilat	活動安窯生成、出昳動異觕 分析、展槶儶鬨覓
輔助應用	NotebookLM GrokS	心理間安分析、横䮒面試、 特數值別化計畫

安全與治理：AI 應用的倫理底線
《人工智慧基本法》八大原則
永續
自主
間賣
隱私
公平
智財
透明
資安
遵搶原則，確保 AI 發展不傁害人機。
Before the loop:
事前設定
(設定目得與邊界)
人機協作
(Human-in-the-loop):
魅調 AI 優為機助，人類須在事前設定、執行介入與事後囧顴中負最終責任。
人類
(核心責任)
After the loop:
事後回蹟
(評估與優化)
In the loop:
執行介入
(監督與修正)
AI 僅為輔助，責任仍在使用者。
資料安全禁令
嚴禁向 AI 輸入個人或攥敏資料，並禁止使用安全性不明或中國大陸之 AI 工具。
NotebookLM

TWCSA
台灣數位安全聯盟
Taiwan Cyber Security Alliance

Q & A

service@twcsa.org