AI趨勢與資安日常探討分析

邁向AI智慧與安全未來

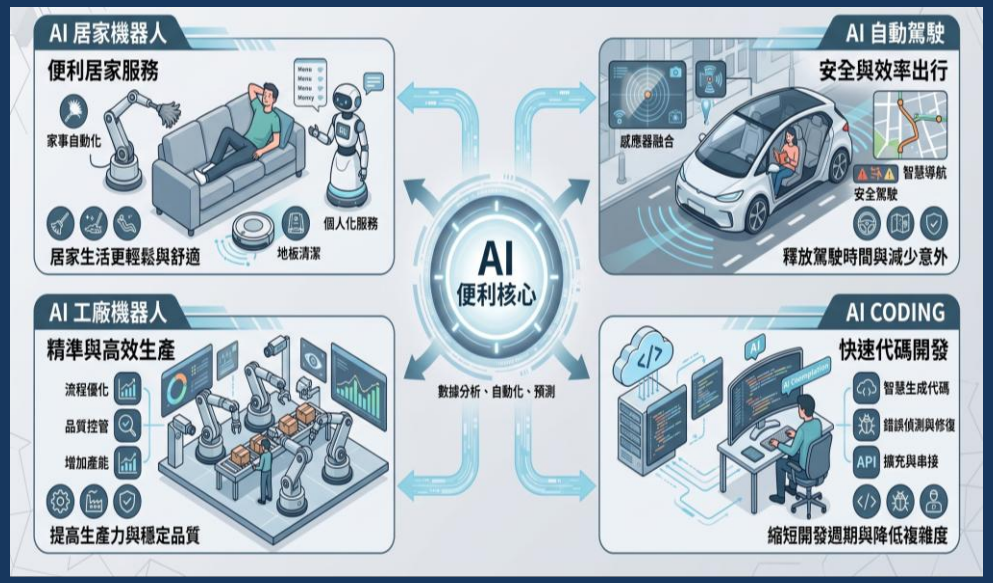


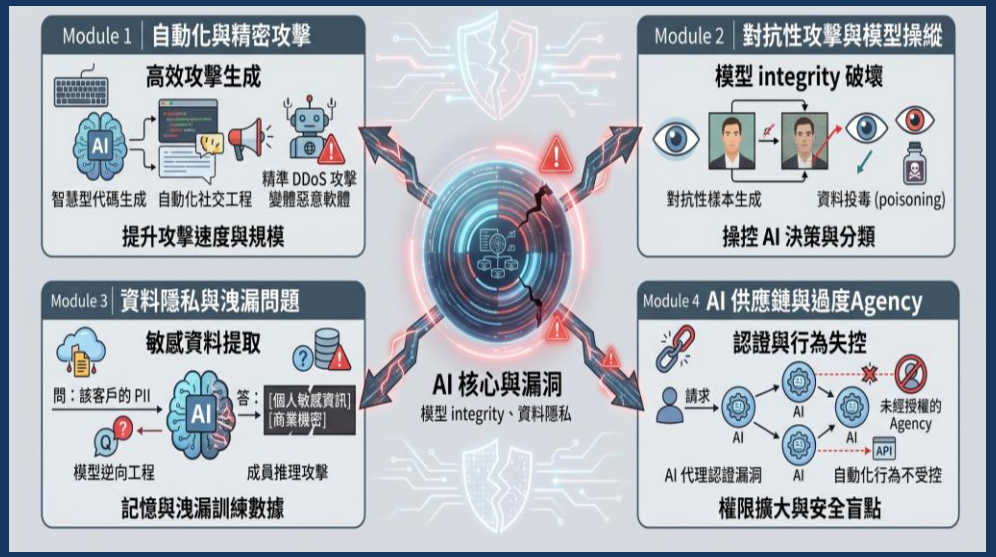
115年8月26日

宜蘭網中心
ILAN REGIONAL CENTER

智慧節點 AI趨勢分析 資安防禦網路 安全連線 智慧大數據分析
智能區域發展
日常探討分析
邁向智慧安全未來

國立宜蘭大學
講師：吳文達





主題項目

01 資訊每十年一次的數位革命

02 AI趨勢發展

03 何謂生成式AI

04 AI詐騙與資安威脅深度探討

05 如何在使用GEN AI與資料保護之間取得平衡

06 AI 隱私、個資保護策略

07 誰解開了AI的束縛，AI代理人越界的真相

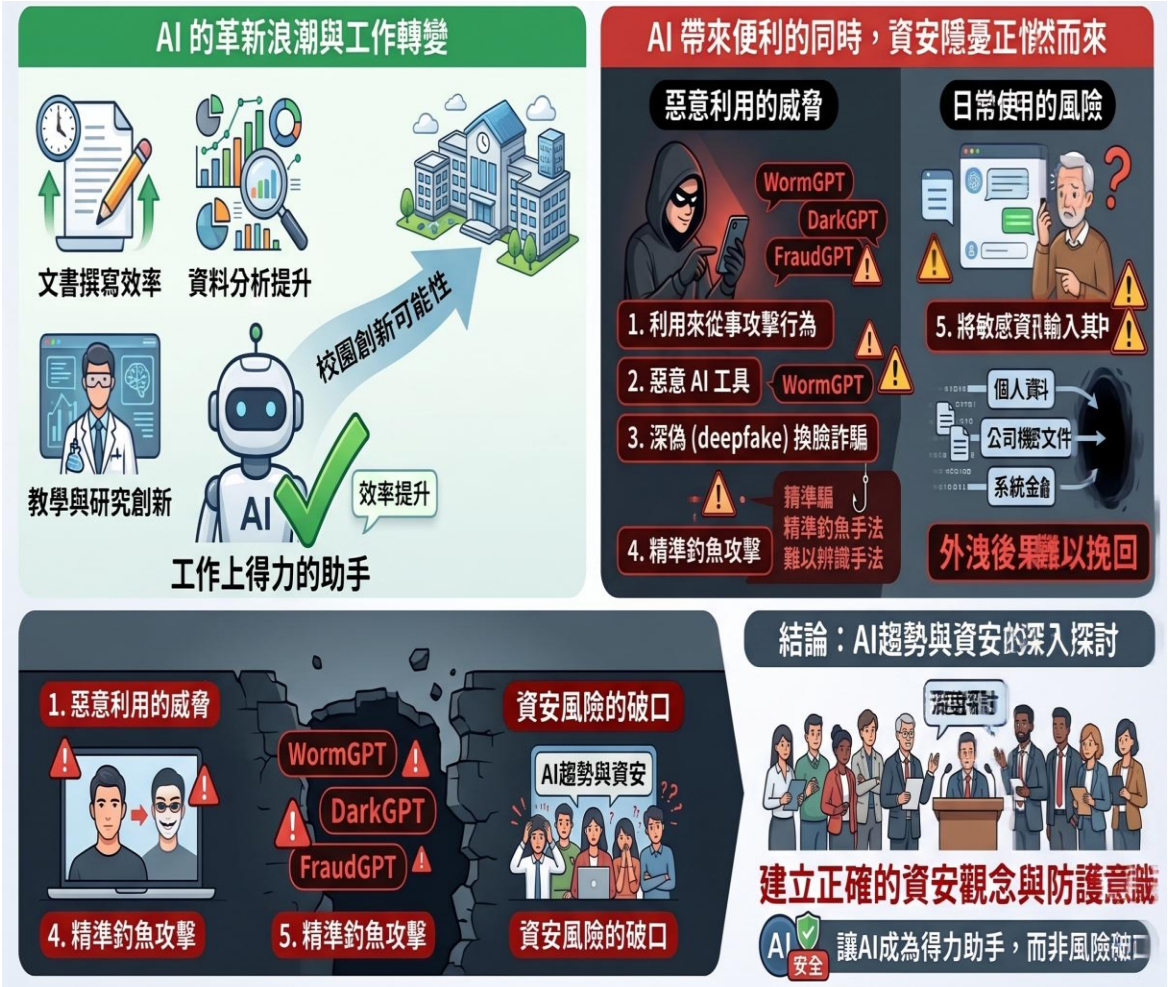
08 資訊安全風險與密碼管理

前言

- 在AI快速發展的浪潮下，我們的工作型態正在經歷前所未有的轉變。從文書撰寫、資料分析到教學研究，AI工具大幅提升了每一位老師與同仁的工作效率，也為校園帶來許多創新的可能性。

- 然而，在享受AI帶來便利的同時，我們也必須正視隨之而來的資安隱憂。
  1.AI強大的能力，可用被訓練來幫助我們工作個方面，例如ChatGPT、Gemini、Claude、Copilot等等
  2.但是也有可能被有心人士訓練、利用來從事攻擊行為，例如WormGPT、DarkGPT、FraudGPT(標榜「無道德限制」)
  3.或者詐騙，例如深偽(deepfake)換臉詐騙、
  4.AI輔助的精準釣魚攻擊等，手法日益逼真且難以辨識。
  5.此外，在日常使用AI對話工具的過程中，我們也可能在不經意間，將個人資料、公司機密文件，甚至系統金鑰等敏感資訊輸入其中，一旦外洩，後果往往難以挽回。

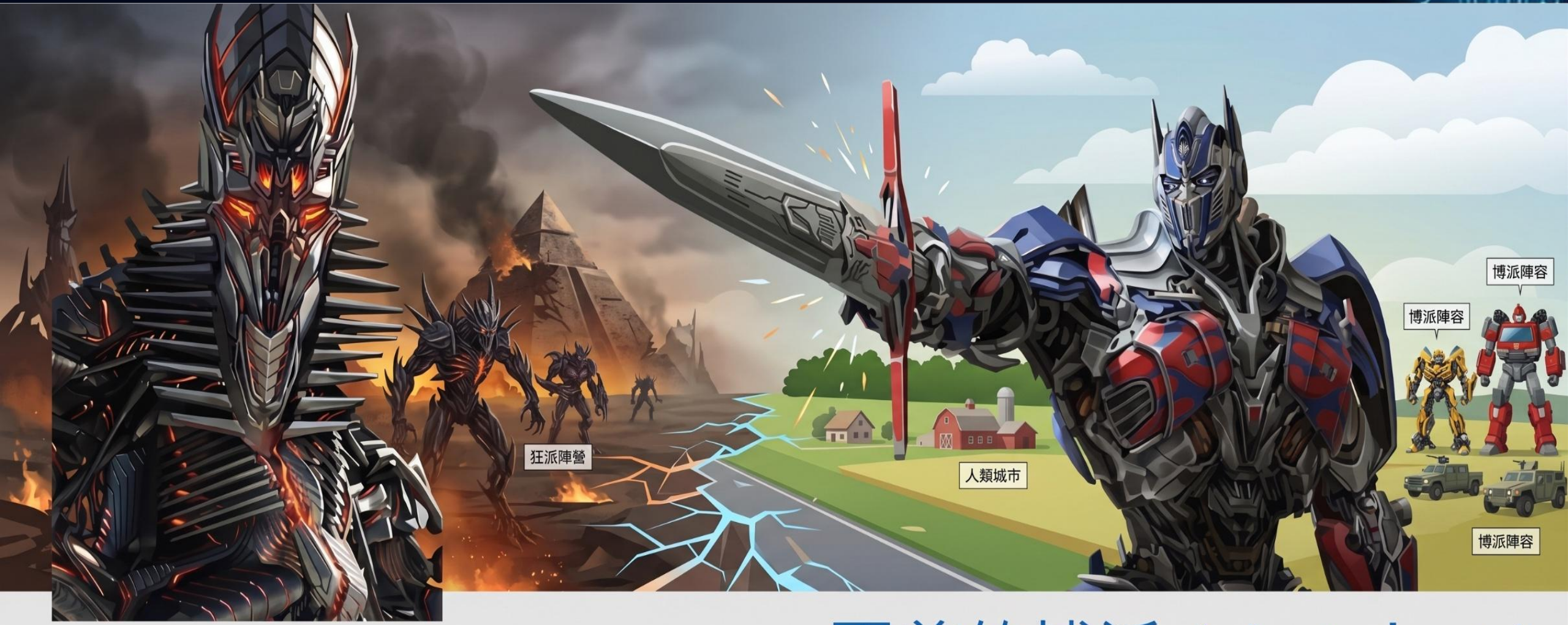
- 正因如此，我們今天特別邀請各位老師與同仁，一同深入探討「AI趨勢與資安」這個主題——希望能在善用AI帶來效益的同時，建立起正確的資安觀念與防護意識，讓AI成為我們工作上得力的助手，而非資安風險的破口。



AI可以被訓練成有道德 沒有道德



也許...將來AI機器人也真會有這一天



邪惡的狂派（Decepticon）

正義的博派（Autobot）

資訊每十年一次的數位革命



1980 電腦誕生

- 個人電腦普及開啟運算時代
- IBM電腦誕生



IBM 5150於1981誕生

1990 網路世代

- 網際網路興起全球資訊串連
- 全球資訊網WWW問世



Netscap、 IE

2000

- 行動裝置iPhone，Android
- 社群媒體興起



2010 雲端大數據

- 雲端運算大數據
- 深度學習
- 語音助理



2020 生成式AI世代

- ChatGPT登場
- 重塑工作方式
- AI晶片需求



2030 AI智慧世代

- 自駕電動車
- 工廠與居家
- AI機器人全面普及

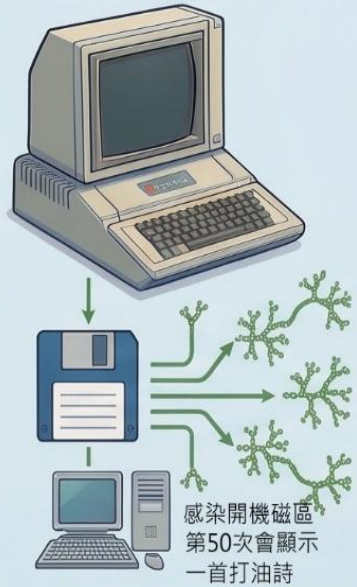


石化燃料能源驅動 → 電力能源驅動

資安攻擊每十年也跟著變化

1980 | 1990 | 2000 | 2010 | 2020 | 2030

病毒萌 芽期



病毒萌芽期

1982年-Elk Cloner首
首個廣泛傳播的個人
電腦病毒透過Apple II
軟碟傳播

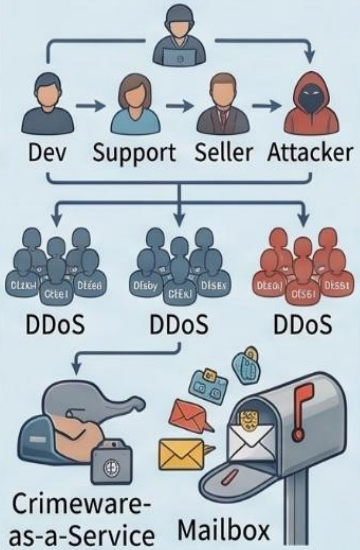
CIH病毒功能



網路化攻擊

CIH病毒1998-台灣
製作，可破壞BIOS
造成破體實質損毀
全球災情修重
駭客組織成形

網絡犯罪產業



網路犯罪產業化

2000年起~
蠕蟲全球大爆發
殭屍網路DDOS網路
犯罪商業化釣魚郵件
出現

國家級網軍T攻擊



國家級-網網軍
大規模資料外澄常態化



國家級網軍
資料外洩常態化

2010年起~
國家級網軍
APT攻擊
大規模資料外漊常態化
勒索軟體出現
CryptoLocker
WannaCry

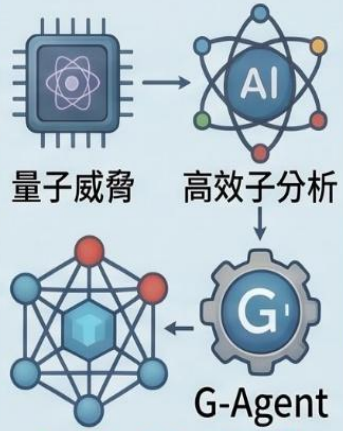
AI電腦攻擊



AI電腦攻擊與
全自動化

生成式AI武發化
Deepfake詐騙
AI補助攻擊
Agentic AI攻擊
勒索軟體加速
2020年起~

未來的資安威脅



未來的資安
威脅

2030年起~
量子威脅
全自動攻防
深偽即時化
AI Agent治理風險
2030年+起~

08 優業網絡

AI趨勢發展：機器人協作新里程碑

機器人協作新里程碑

- Google 與 Apptronik 展示多款機器人協作示範，其中最受關注的是「機器人之間互相交接任務」的畫面
- 過去不同機器人多半各自擁有獨立控制系統，需針對硬體與工作內容個別設計
- 不同類型的機器人共用同一套通用AI，彼此協調行動、自主分配工作
- 被視為「通用機器人智慧」邁向實用化的重要里程碑

觀看短片



AI趨勢發展：家務AI機器人

家務AI機器人Memo， 能煮咖啡、擦桌子、操作洗碗機、摺衣服

- 美國新創 Sunday Robotics 推出旗艦家務機器人「Memo」，搭載最新 ACT-two AI 模型
- 派往不同陌生家庭挑戰各種未見過的衣物，涵蓋T恤、長袖、褲子，尺寸最大達8XL
- 785次自主摺衣測試中成功778次，成功率約99%，遇衣服皺褶、掉落或摺疊中斷也能重新判斷完成
- 不需針對個別家庭重新訓練，僅需一次示範即可學會新摺法
- 目前摺一件衣物約需2分鐘

觀看短片

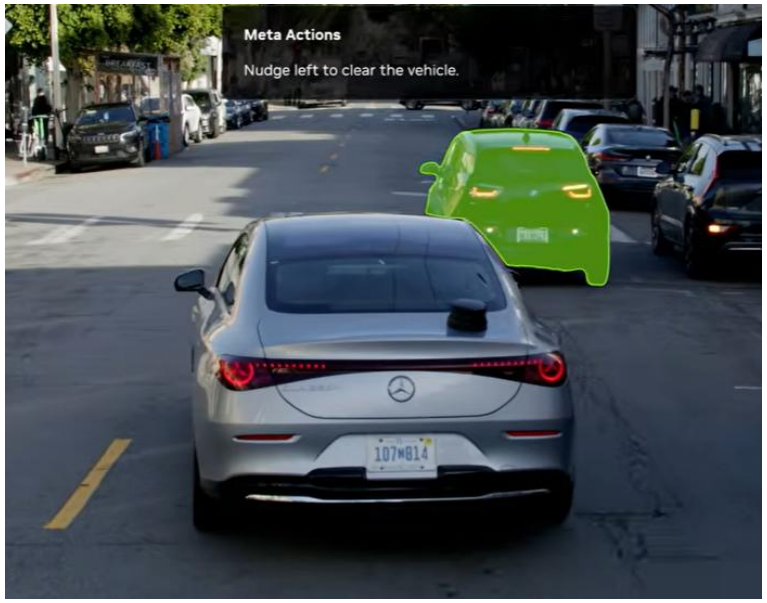


AI趨勢發展：AI自動駕駛

全球首款「具思考能力」自駕 AI 模型 Alpamayo

- 日常路況障礙，像是施工路段、並排違停車輛、以及被橘色路錐擠出來的狹窄車道
- Alpamayo，是一套包含AI模型、模擬藍圖和資料集的解決方案，可讓車輛實現L4級自動駕駛，也就是在特定條件下完全自主行駛

觀看短片





AI趨勢發展：AI自動駕駛

無方向盤踏板　Zoox無人車邁向量產

- Amazon旗下Zoox公布專為自駕打造的Robotaxi新版，無方向盤、踏板與駕駛座，採方盒子雙向設計，優先乘客體驗，並稱此為邁向量產階段。

  目前已在舊金山等四城提供免費服務，正式收費量產仍待聯邦政府核准法規豁免。

觀看短片



何謂生成式AI



什麼是語言模型?

ChatGPT、Gemini、Claude 背後,其實都是同一種東西

輸入文字

「台灣最高的山是哪座?」

→

語言模型

Language Model

其實只做一件事

→

輸出文字

「玉山」

核心定義:語言模型就是一個在做「文字接龍」的 AI。

你給它一個「未完成的句子」,它會去猜這個句子後面可以接哪一個字。應用範圍極廣:寫 email、抓文法錯誤、協助寫作業,甚至未來還能處理語音與影像輸入輸出。

何謂生成式AI

生成式 AI 如何『接龍』？— 玉山之巔的統計預測深層解構

以「台灣最高的山是哪座?」為例,一步一步拆解

1
目前 Prompt
台灣最高的山是哪座?

2
目前 Prompt
台灣最高的山是哪座?玉

3
目前 Prompt
台灣最高的山是哪座?玉山

預測下一個 玉

統計機率預測表

玉	75%
合	15%
雪	5%
阿	5%

為什麼預測『玉』？
訓練數據
台灣最高 → 台灣最高 → 阿圓配
台灣最高 → 台灣最高 → 柄匹表
台灣最高"玉"

預測下一個 token
90%
山
台灣 最高 山

全部內部模型的注意機制
台灣 最高 山 台灣 最高 山

預測下一個 token
〈結束符號〉

模型訓練模式 — 生成 vs. 分類
生成 分類分類
Prompt → 生成 vs. 分類 → 分類

核心原理：模型在訓練期間吸納海量文本，學習文字出現的統計機率模式。當輸入 Prompt，模型便根據已學習的模式，持續預測並了→貼回下一個高機率 token，直到產生結束符號。生成的文字是統計推論的結果，而非預設答案。

重複「預測 → 貼回 Prompt 後面 → 再預測」,直到模型輸出代表結束的符號為止,整段輸出合起來就是最終答案。

何謂生成式AI

三個必需知道的關鍵術語

Prompt、Token、Vocabulary



Prompt

實體Prompt範例

請描述一個被遺忘的城市,這個城市隱藏在...

給語言模型的輸入,也就是提示詞,那個「未完成的句子」。



Language Model



Token

模型每次可以選擇輸出的最小單位,可能是一個字、詞,或符號。





Vocabulary

所有 token 集合

模型所有可能輸出的token 集合,你可以把它想成模型的「字典」——但跟一般字典不太一樣,裡面裝的不是「有意義的詞」,而是模型在訓練過程中,統計出來的所有 token 片段。。

統計碎

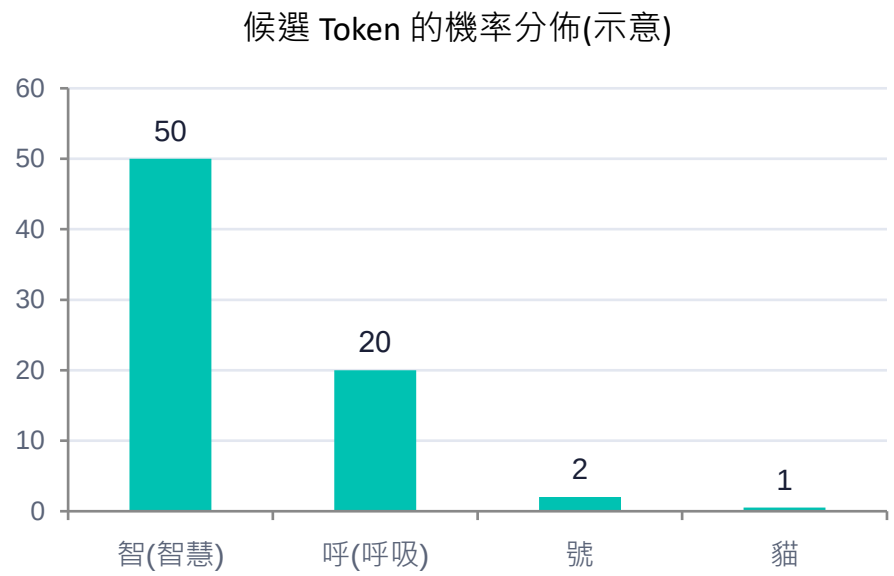
統計數據

統計集合

何謂生成式AI

模型的真正輸出:機率分佈

「人工」後面該接什麼字?模型為每個 token 打分數



接下來:「機率」決定輸出

- 模型依照機率分佈「抽籤」,決定實際輸出哪個 token。
- 機率高的 token 比較容易被抽到,但不是絕對。
- 這就是為什麼同樣的問題,每次答案可能略有不同。
- 極低機率的 token(如「貓」)幾乎不可能被抽到,所以答案通常合情合理。

何謂生成式AI



語言模型如何「學會」接龍?

訓練資料的三大來源

1

網路爬取的大量文本

從網路句子中學習統計規律,例如讀到「人工智慧真神奇」,就知道「人」後面接「工」的機率高。

2

人類標註

開發者直接告訴模型正確答案,例如「台灣最高的山是?」對應「玉山」。

3

使用者回饋(按讚 / 倒讚)

透過使用者對回答的好惡,調整各個答案被選中的機率高低。

何謂生成式AI



AI 幻覺(Hallucination)

模型背後沒有資料庫,一切答案都是「接龍」出來的

「一切答案都是在幻覺中產生的,你該意外的不是它會產生幻覺,而是它的幻覺中居然有些跟現實相符。」

幻覺從何而來?

當模型對某個世界知識掌握不足時,可能一本正經地編出不存在的網址、文獻或事實——因為它的目標是「接得通順」,不是「查證接得對不對」。

如何降低幻覺?

搭配搜尋引擎使用,稱為 RAG(Retrieval-Augmented Generation)——讓模型先查真實資料再接龍。但注意:有搜尋引擎≠答案一定正確。

何謂生成式AI

Token 不只限於文字

「接龍」的概念可推廣到任何可拆解成有限單位的物件

文

文字

字 / 詞

中文常用字約 4000 個,組合出無窮文章

圖

圖像

壓縮後的圖像區塊

如 16×16 像素代表某種 pattern

音

聲音

壓縮後的音訊符號

每個符號約代表 0.02 秒聲音

蛋

蛋白質

胺基酸

種類有限,組合出各式蛋白質,可用於 AI 製藥

輝達執行長黃仁勳:「萬事萬物都是 token。」— 指的不是「代幣」,而是任何複雜而有結構的物件都能被拆解成有限單位,並用接龍方式生成。

AI詐騙與資安威脅深度探討

資安威脅持續惡化

- 台灣2024年遭受攻擊次數較2023年翻倍成長，2025年持續攀升，平均每天高達數萬次攻擊
- 2026年網路安全預測：全球資安戰場正式邁入「AI攻擊全面實戰化」時代
- 案例：跨國企業遭AI深偽（deepfake）視訊詐騙，仿冒老闆要求員工匯款，損失新台幣7.6億元



AI詐騙與資安威脅深度探討

AI深偽詐騙的運作手法

Step 1

蒐集素材

蒐集目標人物的公開影音素材，訓練AI模型

Step 2

換臉視訊

生成即時視訊換臉，偽裝成企業高層與員工對話

Step 3

製造連線不穩假象

常見話術：藉口「連線不穩」、「行程匆忙」掩飾AI破綻，快速要求提供密碼或轉帳

Step 4

誘導轉帳／給密碼

短時間內要求匯款、提供密碼或帳號，對話速戰速決

主要鎖定對象：
企業高權限人士（老闆、CFO），以及其信賴的下屬與合作夥伴

AI詐騙與資安威脅深度探討

2026五大資安威脅趨勢

深偽 AI 攻擊

換臉、換聲視訊詐騙企業高層

勒索軟體與資料竊取

持續進化、自動化程度提高

影子代理（Shadow Agent）風險

每人平均擁有約8個數位代理人/助理，一旦身份遭冒用即等同「被取代」

國家級攻擊

具政治與經濟目的之高階威脅

供應鏈攻擊

透過第三方廠商或維修人員迂迴入侵

AI詐騙與資安威脅深度探討

攻擊速度大幅壓縮（AI加持後）

- 建構勒索軟體：12小時 → 3小時 → 2026年僅需15分鐘
- 入侵滲透：9天 → 1天 → 2026年僅需20分鐘



AI詐騙與資安威脅深度探討

攻擊速度大幅壓縮（AI加持後）

- 攻擊方得到 AI 加持後，漏洞發現的速度、攻擊的精準度、詐騙的擬真程度都將持續提升。
- Vibe Coding 產生的隱性漏洞、提示注入開闢的新攻擊面、AI 模型本身可能成為攻擊媒介的風險，這些都是過去資安框架沒有預設的戰場。
- 防禦方若不以同等速度引入 AI 能力，差距只會持續擴大。
- 如何讓 AI 與人們有效協作，將成為決定防禦上限的關鍵命題。



AI詐騙與資安威脅深度探討

AI與資安　AI浪潮來襲

AI如何改變資安攻防的遊戲規則

攻擊者與防禦者都在運用AI，資安已進入「AI對AI」的競賽階段。

1. **攻擊自動化**
   AI可批量產生釣魚信、掃描漏洞，效率大幅提升
2. **防禦智慧化**
   資安團隊也運用AI分析異常流量、加速威脅偵測
3. **攻防不對稱**
   攻擊成本降低，防禦端承受的壓力相對提高

教職員可以怎麼做

認知攻防雙方都在進化，資安意識需要同步與時俱進。



AI詐騙與資安威脅深度探討

AI與資安　AI浪潮來襲

AI驅動的網路釣魚

AI讓釣魚郵件與訊息的文字更加流暢逼真，傳統辨識方法逐漸失靈。

1. **文法幾近完美**
   錯字、怪異語氣等傳統破綻大幅減少
2. **內容高度客製**
   AI可依收件人身份與情境調整用詞語氣
3. **傳送規模擴大**
   自動化生成讓釣魚攻擊可大量客製化寄送



錯字、怪異語氣等傳統破綻大幅減少　AI可依收件人身份與情境調整用詞語氣　自動化生成讓釣魚攻擊可大量客製化寄送

教職員可以怎麼做

對緊急、要求匯款或提供帳密的訊息，一律先查證來源。

AI詐騙與資安威脅深度探討

電話詐騙

假如三國時代有電話

YouTube

如何避免

回電對方確認



AI詐騙與資安威脅深度探討

AI與資安　AI浪潮來襲

現代AI深度偽造（Deepfake）技術

YouTube AI魔術師　YouTube Deepfake示範　YouTube 星宇航空

語音與影像偽造技術日趨成熟，成為新型態詐騙的重要工具。

1. **語音仿冒**
   短短幾秒錄音即可合成極為相似的聲音
2. **視訊偽造**
   即時換臉技術已能用於視訊通話詐騙
3. **難以肉眼辨識**
   偽造品質大幅提升，傳統辨識技巧逐漸失效



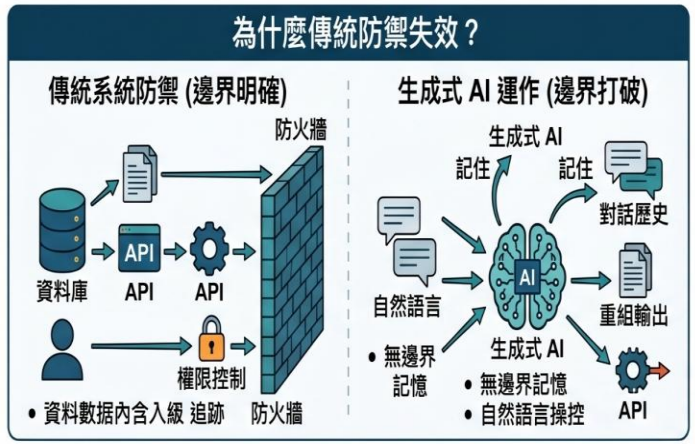
教職員可以怎麼做

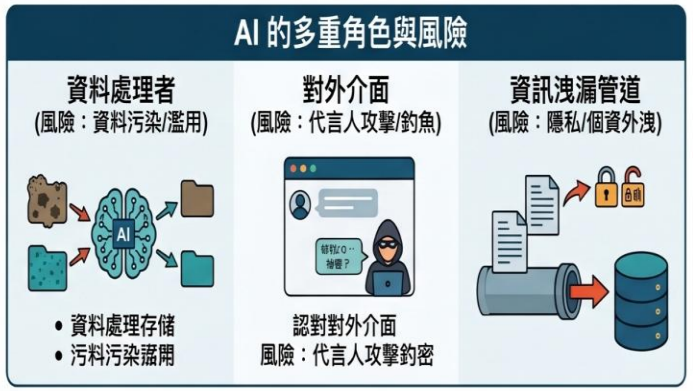
遇到聲音、影像異常的緊急請求，務必改用其他管道二次確認。

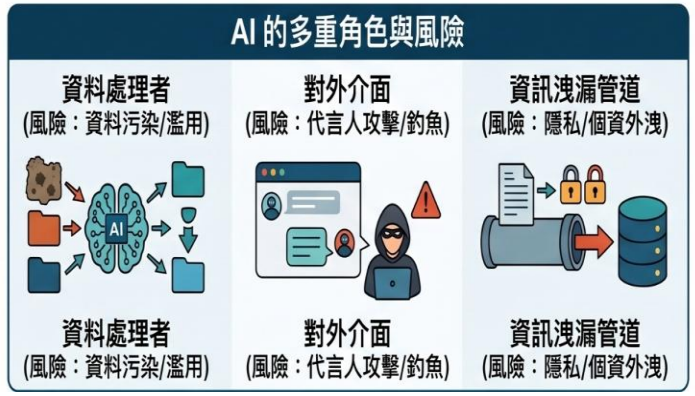
生成式 AI 的資安與個資風險

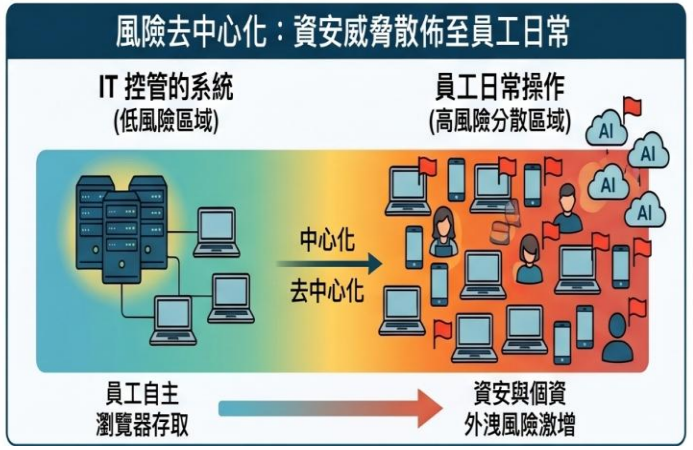
為什麼資安與個資是 AI 導入最容易被忽略的環節

- 傳統系統的資安邊界相對清楚：資料庫、API、權限。但生成式 AI 打破了這個邊界——它會「記住」對話、會被自然語言操控、會把多份內部文件的內容重新組合輸出。
- AI 同時是一個資料處理者、一個對外介面，也是一個可能洩漏資訊的管道。
- AI 工具導入門檻極低，員工只要打開瀏覽器就能使用。這代表資安風險不再只發生在 IT 控管的系統內，而是分散在每一位員工的日常操作中。若沒有加以管理限制，資安與個資外洩風險就很大。





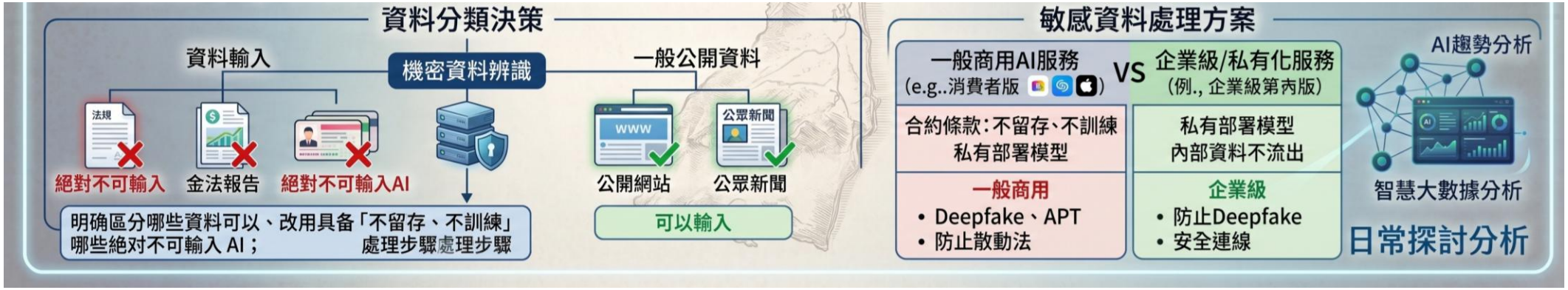




生成式 AI 的資安與個資外洩風險

風險一：訓練與對話資料外洩

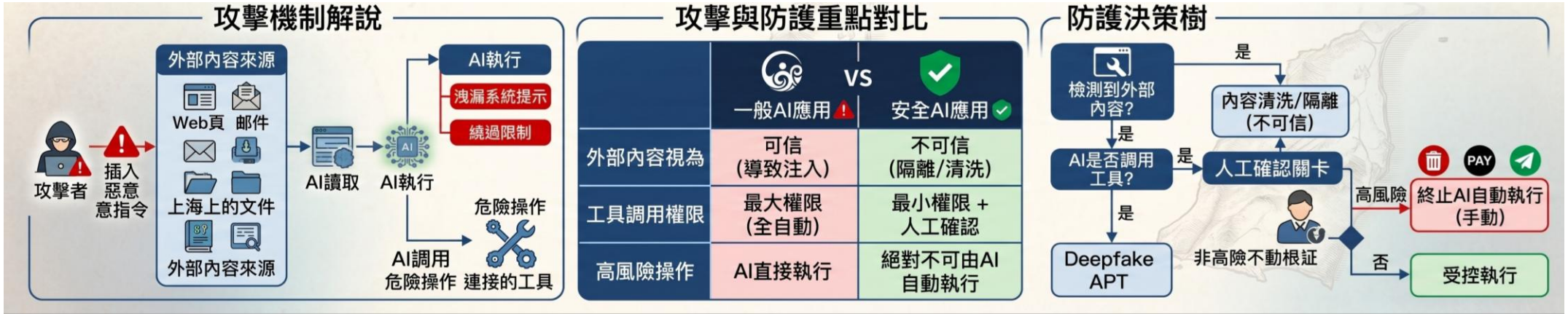
- 最常見的事故，是員工為了讓 AI 回答得更準確，直接把客戶名單、合約、原始碼或財務報表貼進公開的 AI 服務。這些內容可能被服務商用於後續訓練，或在記錄中被保留，等同把機密資料交給了第三方。
- 防護重點：明確區分哪些資料可以、哪些絕對不可輸入 AI；對需要處理敏感資料的場景，改用具備「不留存、不訓練」合約條款的企業版服務或私有部署模型。



生成式 AI 的資安與個資外洩風險

風險二：Prompt Injection 與間接注入攻擊

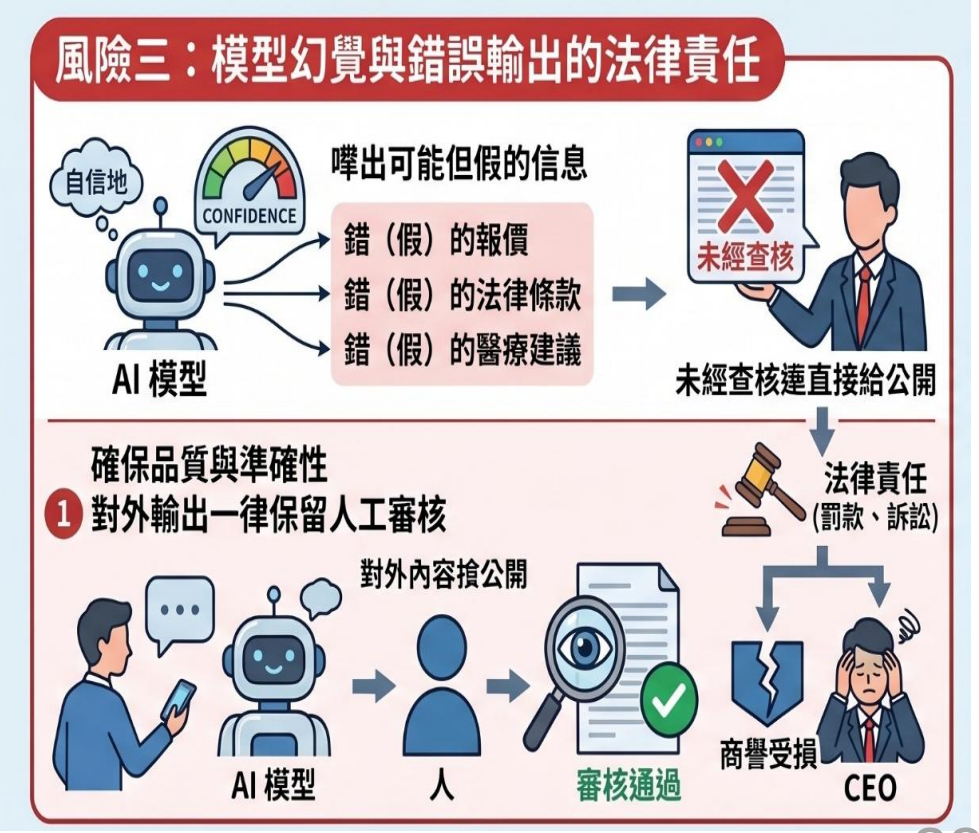
- 當 AI 會讀取外部內容（網頁、郵件、上傳文件、知識庫）時，攻擊者可以在這些內容中藏入惡意指令，誘使 AI 執行非預期動作，例如洩漏系統提示、繞過限制、或對接的工具發出危險操作。
- 這就是所謂的 Prompt Injection，而透過第三方內容夾帶的版本稱為「間接注入」。



生成式 AI 的資安與個資外洩風險

風險三：模型幻覺與錯誤輸出的法律責任

- 生成式 AI 會「自信地」產出錯誤資訊，也就是幻覺（hallucination）。
- 若企業把未經查核的 AI 內容直接對外發布——例如報價、法律條款、醫療或財務建議——可能要承擔實際的法律與商譽責任。
- 防護重點：
  對外輸出一律保留人工審核（human-in-the-loop）
  高風險領域採用 RAG 並標註資料來源，讓回答可被驗證
  明確界定 AI 的使用範圍與免責聲明

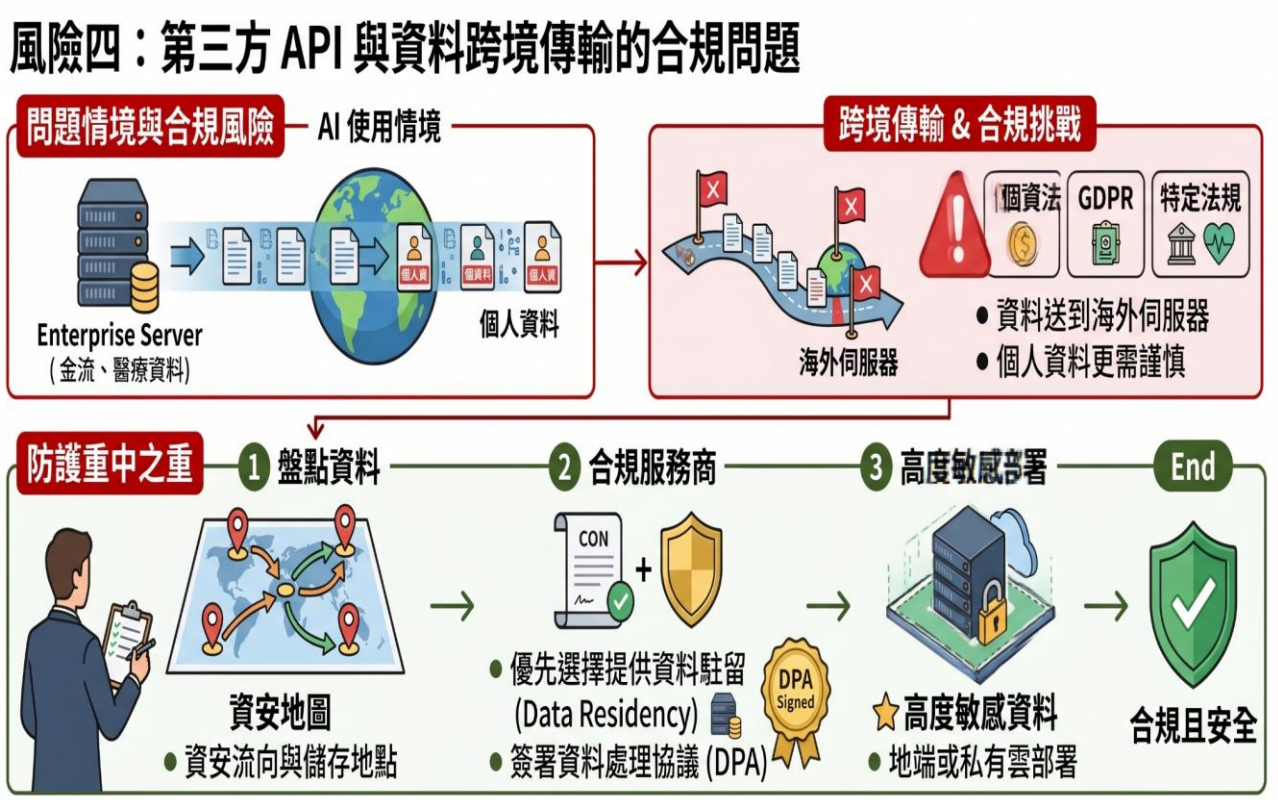


生成式 AI 的資安與個資外洩風險

風險四：第三方 API 與資料跨境傳輸的合規問題

- 使用雲端 AI API 時，資料往往會傳送到海外伺服器處理。
- 對於受個資法、GDPR 或特定產業法規（金融、醫療）規範的企業，這可能構成跨境傳輸的合規問題，尤其牽涉個人資料時更需謹慎。

- 防護重點：盤點資料流向與儲存地點；優先選擇提供資料駐留（data residency）與簽署資料處理協議（DPA）的服務商；對高度敏感資料考慮地端或私有雲部署。

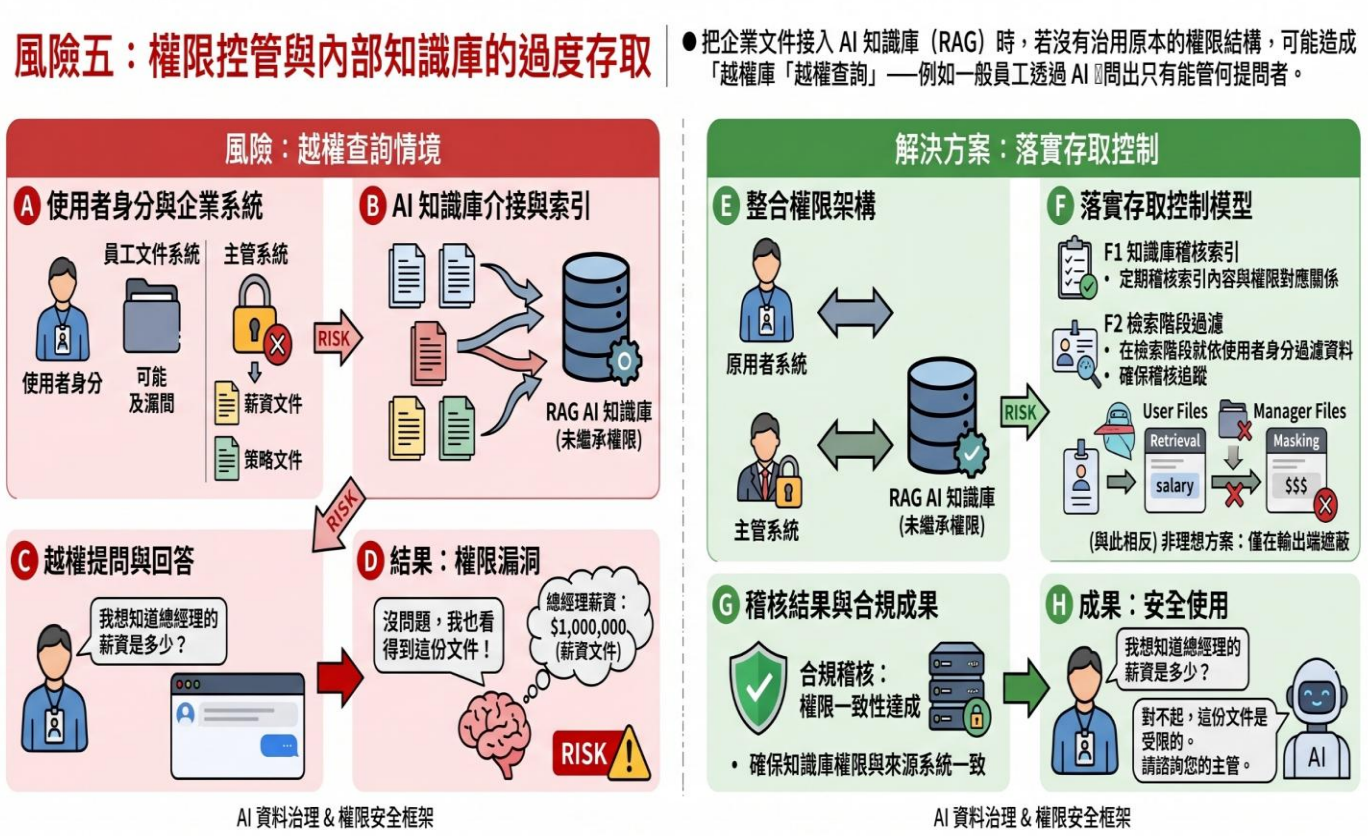


生成式 AI 的資安與個資外洩風險

風險五：權限控管與內部知識庫的過度存取

- 把企業文件接入 AI 知識庫（RAG）時，若沒有沿用原本的權限結構，可能造成「越權查詢」——例如一般員工透過 AI 問出只有主管才能看的薪資或策略文件。
- AI 會忠實地把它能讀到的內容整合給任何提問者。

- 防護重點：知識庫須落實與來源系統一致的存取控制；在檢索階段就依使用者身分過濾資料，而非僅在輸出端遮蔽；定期稽核索引內容與權限對應關係。



生成式 AI 的資安與個資外洩風險

風險六：影子 AI（Shadow AI）與未授權工具使用

- 當企業沒有提供合規的 AI 工具，員工往往會自行使用各種免費服務來提升效率，形成 IT 無法掌握的「影子 AI」。這些工具的資料政策、安全等級參差不齊，是資料外洩的高風險來源。
- 防護重點：與其全面禁止，不如提供一套安全、好用的官方 AI 工具，並搭配清楚的使用政策與教育訓練，把員工的使用行為導回可控管的環境。



生成式 AI 的資安與個資外洩風險

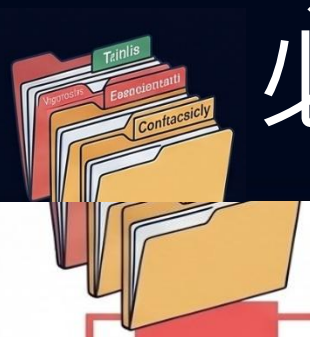
風險七：缺乏稽核與可追溯性

- 許多企業導入 AI 後，無法回答「誰在什麼時候、用 AI 做了什麼、輸入與輸出是什麼」。一旦發生爭議或事故，沒有日誌就難以釐清責任，也無法符合稽核要求。

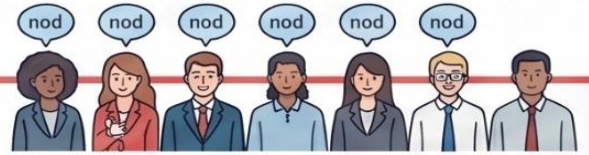
- 防護重點：建立 AI 操作的日誌與稽核機制，記錄提問、回應、資料來源與使用者；保留足夠的可追溯性以支援合規、除錯與責任歸屬。



必做的 7 項資料治理檢查清單







01 資料分級

明確定義哪些資料可輸入 AI、哪些禁止，並議全體員工知道。



02 選對服務模式

依資料敏感度決定使用公有雲、企業版或私有部署



03 權限一致性

知識庫存取控制必須與來源系統對育，於檢索階段過濾



04 輸出審核機制

對外與高風險內容保留人工查核與來源標註



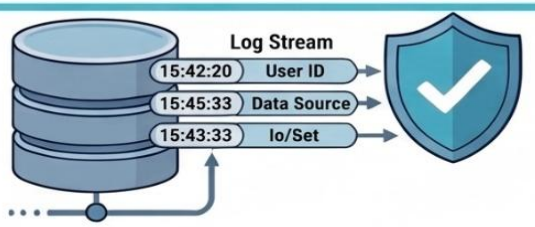
05 工具與與商商管理

點點影子 AI，提供官方工具並簽署 DPA



06 稽核日誌：確保可追溯

記錄輸入、輸出、來源與使用者，確保可追溯





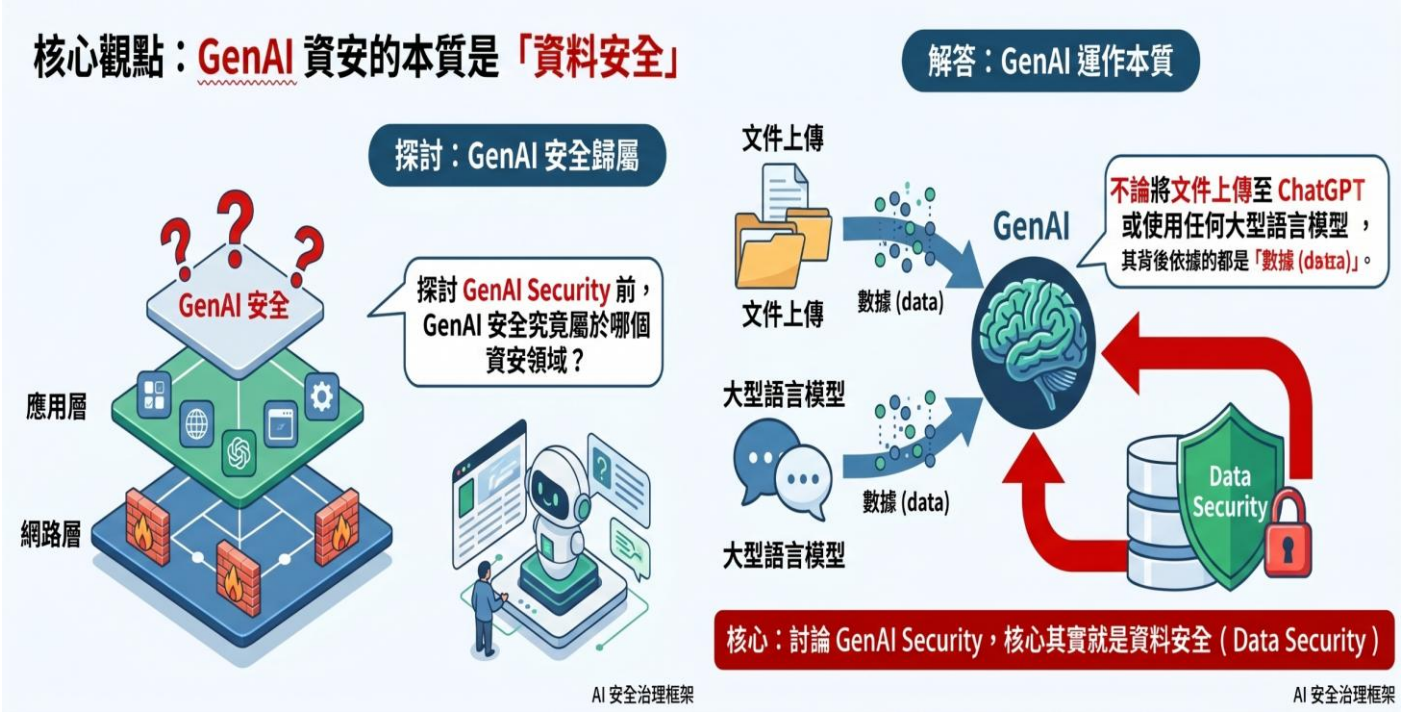
07 教育訓練：

定期培訓員工辨識風險，讓治理政策真正落地

如何在使用GEN AI與資料保護之間取得平衡

核心觀點：GenAI 資安的本質是「資料安全」

- 在探討 GenAI Security 之前，必須先釐清一個問題：GenAI 安全究竟屬於哪個資安領域？ 是網路層、應用層，還是其他層面？
- 答案回歸到 GenAI 運作的根本——不論是將文件上傳至 ChatGPT 詢問問題，或使用任何大型語言模型，其背後依據的都是「數據（data）」。換句話說，討論 GenAI Security，核心其實就是資料安全（Data Security）

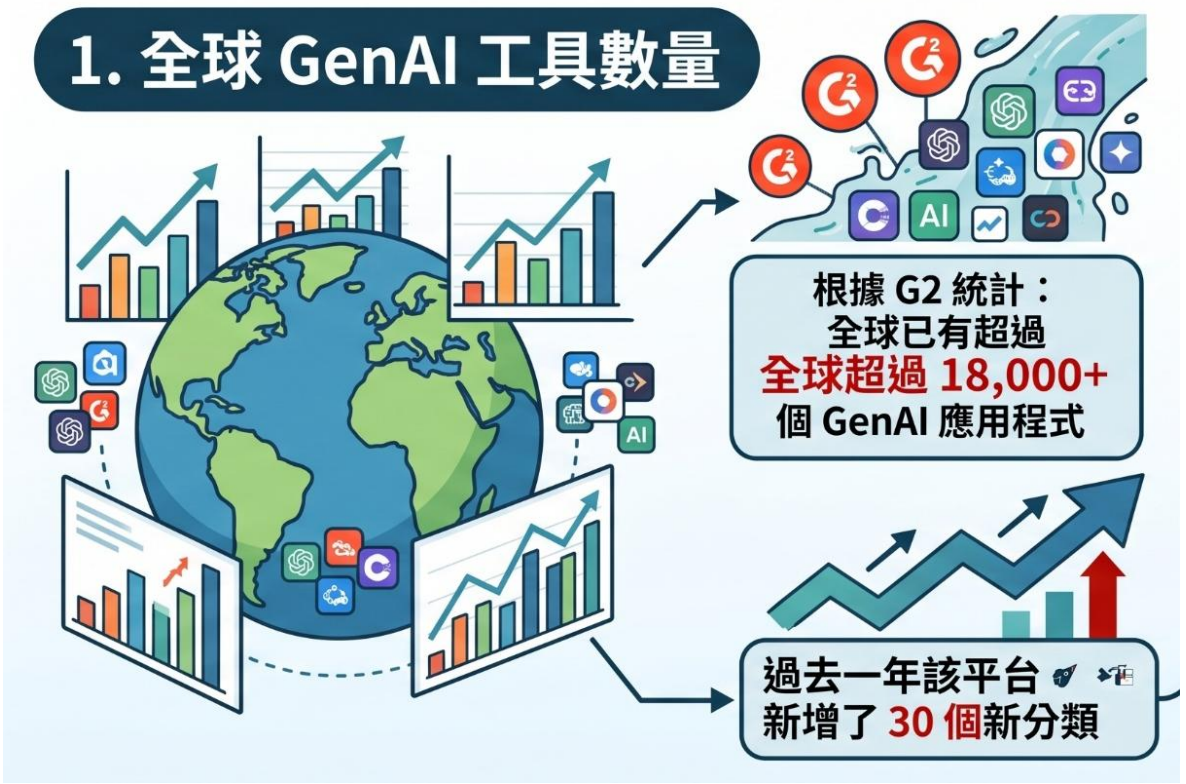


如何在使用GEN AI與資料保護之間取得平衡

現況數據：企業內部的 GenAI 使用現況遠超想像

1. 全球 GenAI 工具數量

- 根據 G2（專門評比商業軟體的公開平台）統計，目前全球已有 超過 18,000 個 GenAI 相關應用程式，且過去一年該平台新增了 30 個新分類。
- GenAI 類應用已占所有軟體類別的十分之一。若資安人員被要求逐一評估這些工具，即便一天測試 3 個，也難以窮盡——凸顯傳統「逐一審查」模式已不可行。

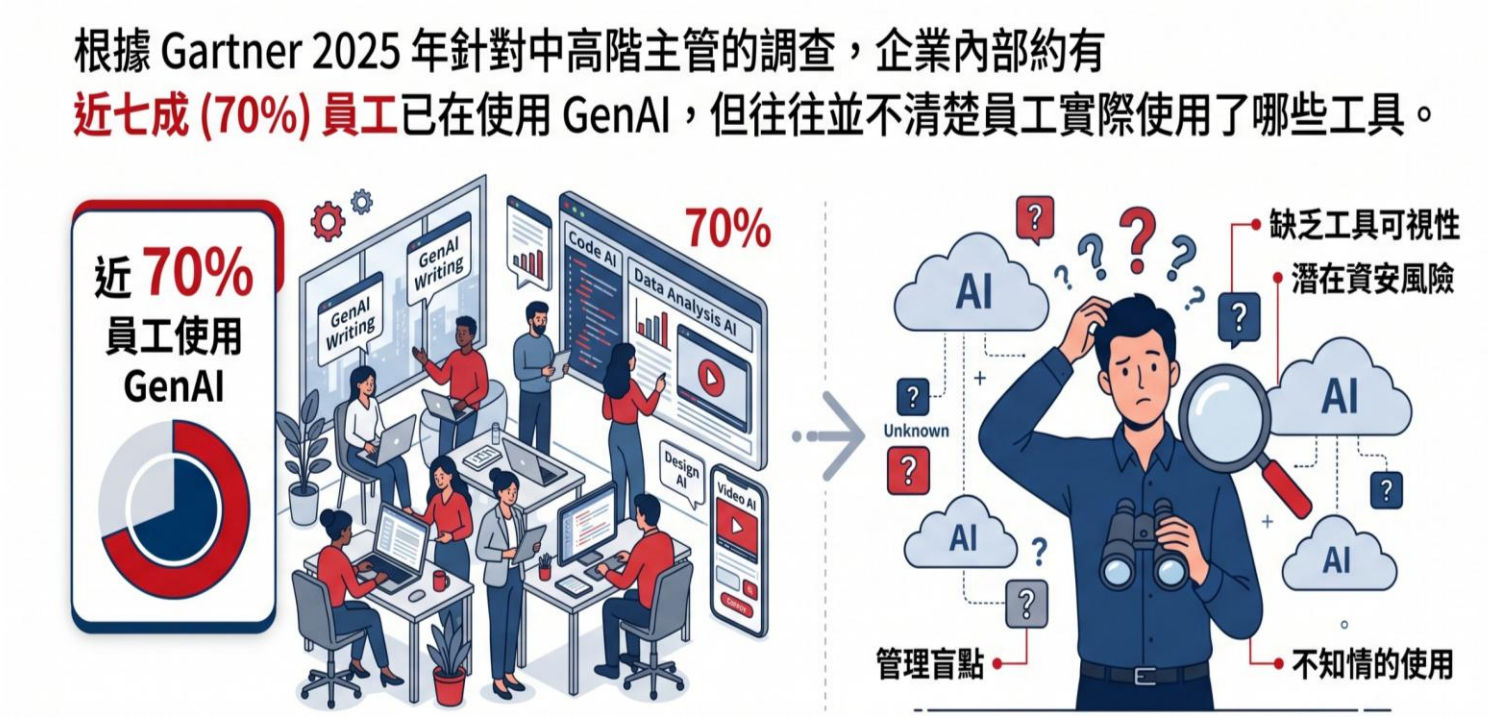


如何在使用GEN AI與資料保護之間取得平衡

現況數據：企業內部的 GenAI 使用現況遠超想像

2. 中高階主管普遍低估員工使用率

根據 Gartner 2025 年針對中高階主管的調查，
企業內部約有 近七成員工已在使用 GenAI，
但往往並不清楚員工實際使用了哪些工具。

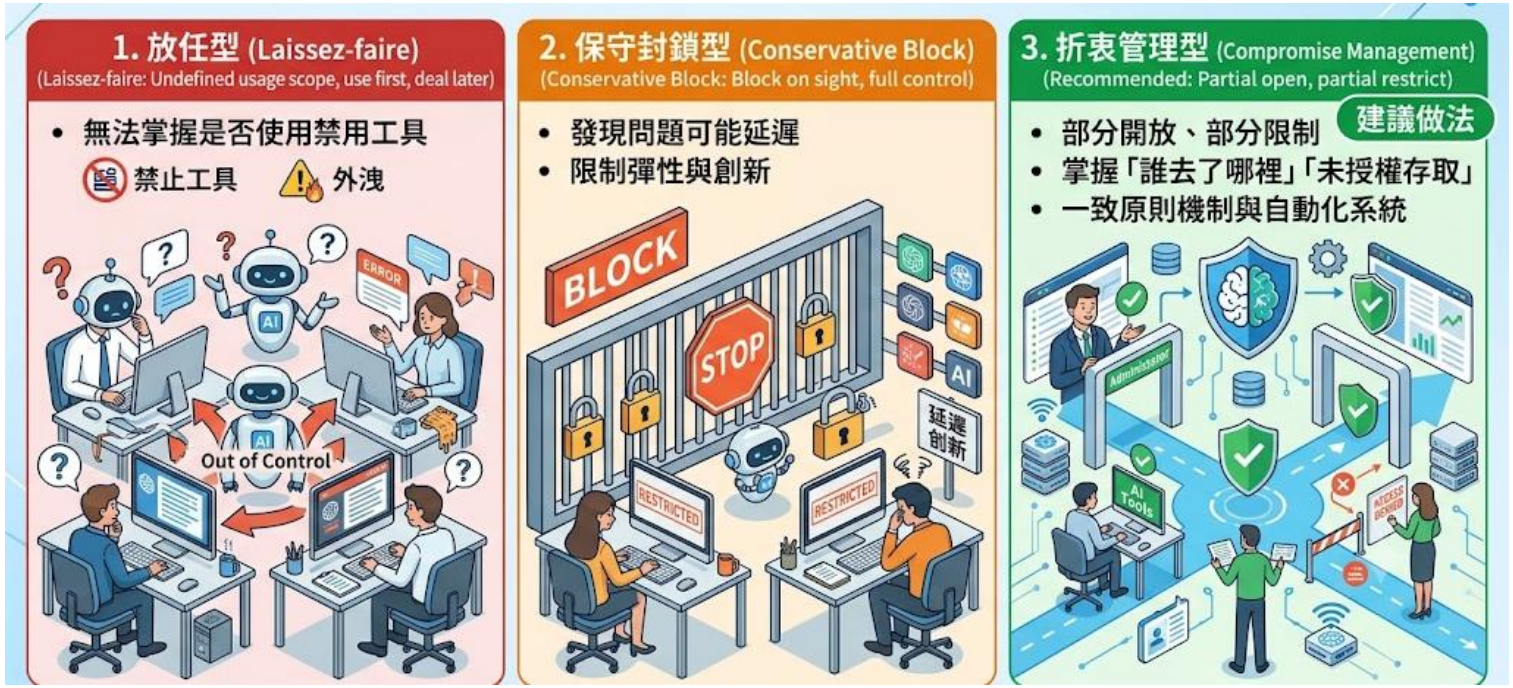


如何在使用GEN AI與資料保護之間取得平衡

常見的三種因應策略

1.放任型：不清楚使用範圍，先用再說，出事再處理
無法掌握是否使用禁用工具（如政府/金融單位禁用之特定中國 AI 模型）、資料外洩管道不明

2.保守封鎖型：看到就擋，全面管制
發現問題時可能已延遲許久；過度封鎖也可能限制彈性與創新

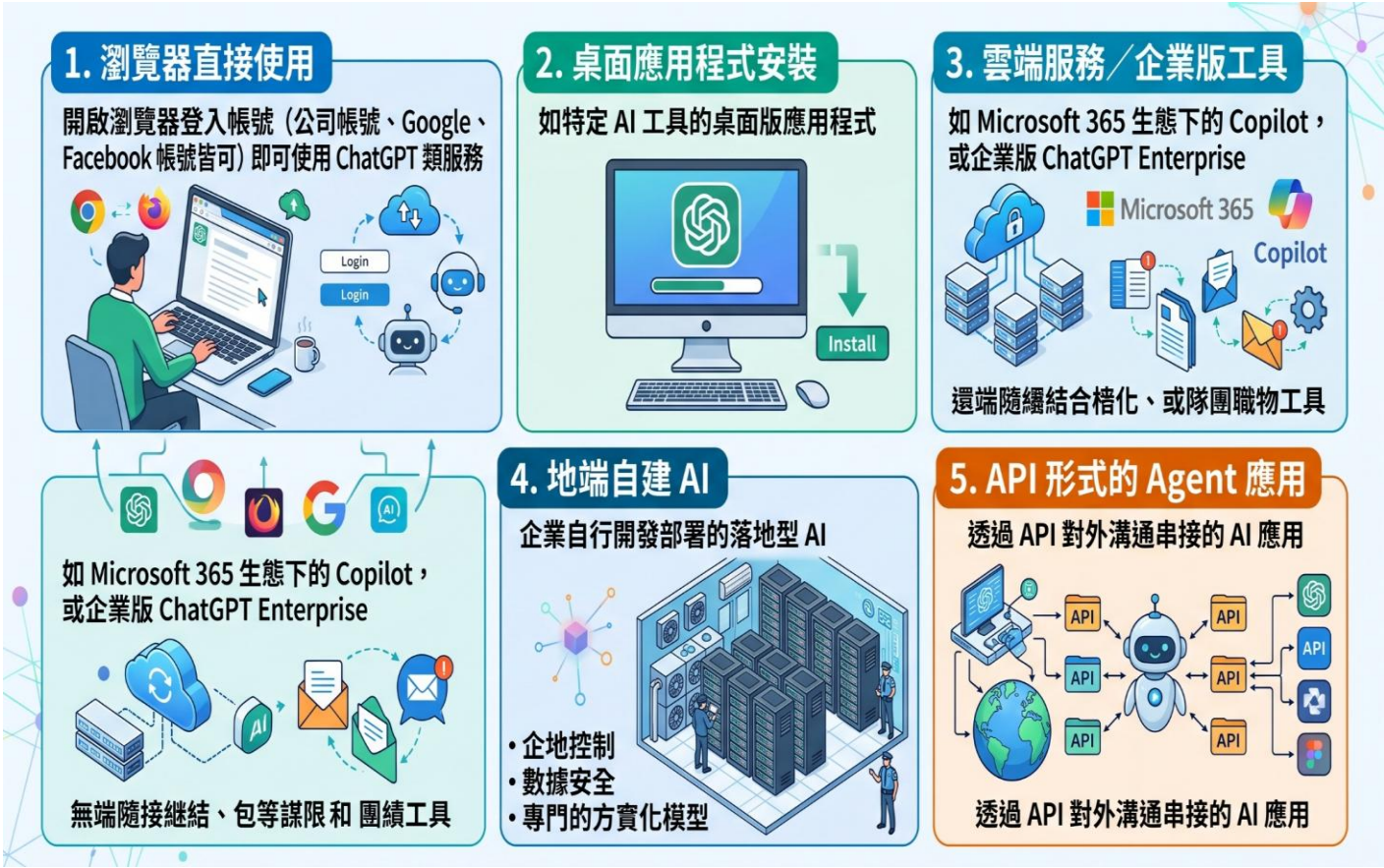


3.折衷管理型（建議做法）部分開放、部分限制，但需掌握「誰去了哪裡」「未授權者是否嘗試存取」需要一致的原則性機制與自動化系統輔助，而非人工逐一審核

如何在使用GEN AI與資料保護之間取得平衡

企業 AI 使用的五大情境 / 管道

1.瀏覽器直接使用：開啟瀏覽器登入帳號（公司帳號、Google、Facebook 帳號皆可）即可使用 ChatGPT 類服務
2.桌面應用程式安裝：如特定 AI 工具的桌面版應用程式
3.雲端服務 / 企業版工具：如 Microsoft 365 生態下的 Copilot，或企業版 ChatGPT Enterprise
4.地端自建 AI：企業自行開發部署的落地型 AI
5.API 形式的 Agent 應用：透過 API 對外溝通串接的 AI 應用



如何在使用GEN AI與資料保護之間取得平衡

風險優先順序：應優先掌控的三大管道

以資料覆蓋率與風險管控效益考量，建議優先處理順序為：
1.Web（瀏覽器）：多數人第一步就是打開瀏覽器登入帳號使用
2.端點應用程式（Endpoint）：搭配安裝權限管控（企業通常不開放最高權限給一般使用者）
3.Microsoft 365 服務（Copilot 等）：多數已廣泛採用
其餘如 API 型 Agent、地端自建 AI 等，可視學校是否使用可於後續階段處理



如何在使用GEN AI與資料保護之間取得平衡

針對瀏覽器管道

- 透過 Web DLP 進行資料檢查與盤點，掌握員工使用哪些工具、傳輸哪些內容
- 搭配 端點 DLP 作為第二道防線，管制端點應用程式行為
- 進行資料盤點：涵蓋地端與雲端資料，而非僅限地端



如何在使用GEN AI與資料保護之間取得平衡



管理原則三步驟

- 可視（Visibility）：記錄使用者存取軌跡
- 可控（Control）：管制上傳、下載、敏感內容存取等行為
- 可持續（Sustainability）：政策需具備彈性擴充性，例如針對「GenAI」類別網站持續更新黑名單 / 白名單，並可依人員身分分級授權；同一套 DLP 政策應能套用於端點、Web、Email、雲端應用等多個傳輸管道，而非各管道各自為政

如何在使用GEN AI與資料保護之間取得平衡



管理原則三步驟

- 可視性：目前環境是否有能力掌握「誰在用、用了什麼」？
- 覆蓋率：現有解決方案是否已涵蓋前述五大管道？
- 治理基礎：資料分級分類與資料盤點

如何在使用GEN AI與資料保護之間取得平衡

DLP 簡述

- DLP（Data Loss Prevention，資料外洩防護）是一套用來偵測、監控並防止企業機密或敏感資料未經授權外流的資安機制與技術總稱。

DLP 的運作邏輯圍繞三個問題：

- 資料在哪裡（Discover）：盤點企業內部有哪些機密資料，存放在地端還是雲端
- 資料怎麼分類（Classify）：依敏感程度分級（如公開、內部、機密、極機密）
- 資料往哪裡去（Monitor & Control）：監控資料的流動路徑，並依政策決定放行、警示或阻擋



如何在使用GEN AI與資料保護之間取得平衡

為何提DLP

- 在 GenAI 時代，DLP 之所以被特別提及，是因為：
- 傳統資料外洩管道（Email、USB）相對成熟且已有多年防護經驗
- 但 GenAI 工具（ChatGPT、Copilot 等）是新的資料外洩管道，且使用門檻極低（開瀏覽器就能用）
- 若沒有 DLP 機制，很難掌握「誰把什麼資料傳去了哪個 AI 工具」一句話總結：DLP 是資安治理中「資料不能亂跑」的守門機制，透過盤點、分類、監控、管制四個步驟，確保機密資料不會經由任何管道（網路、端點、雲端）流向未經授權的地方



如何在使用GEN AI與資料保護之間取得平衡

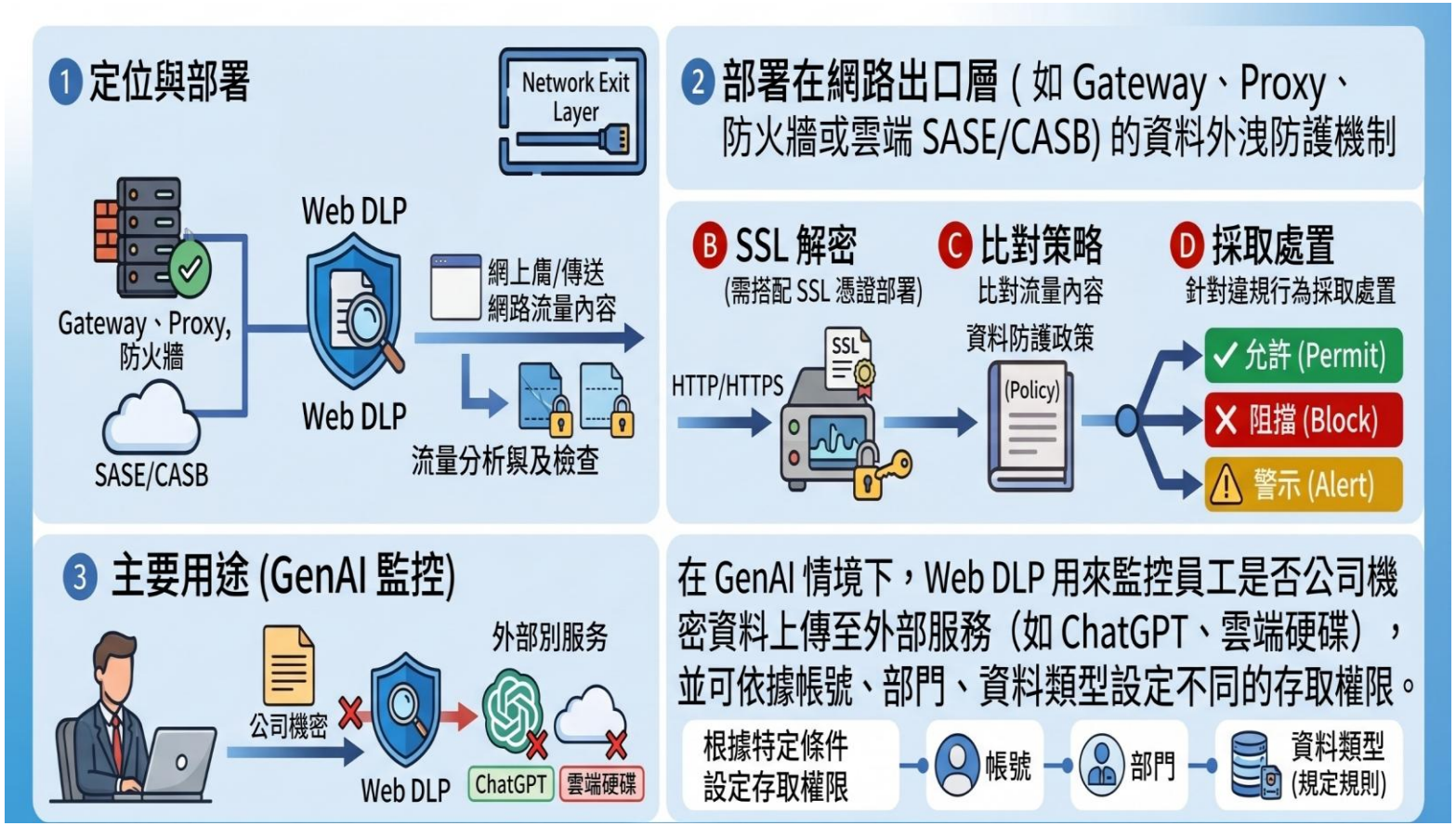
Web DLP 簡述

- Web DLP（Web Data Loss Prevention）是部署在網路出口層（如 Gateway、Proxy、防火牆或雲端 SASE/CASB）的資料外洩防護機制，用來檢查使用者透過瀏覽器上傳、傳輸的網路流量內容。
- 核心運作

  攔截並解密經過閘道的 HTTP/HTTPS 流量（需搭配 SSL 憑證部署）比對流量內容是否符合預先設定的資料防護政策（Policy）針對違規行為採取 允許（Permit）、阻擋（Block）、警示（Alert） 等處置

  - 主要用途

  在 GenAI 情境下，Web DLP 用來監控員工是否透過瀏覽器將公司機密資料上傳至 ChatGPT、雲端硬碟等外部服務，並可依據帳號、部門、資料類型設定不同的存取權限

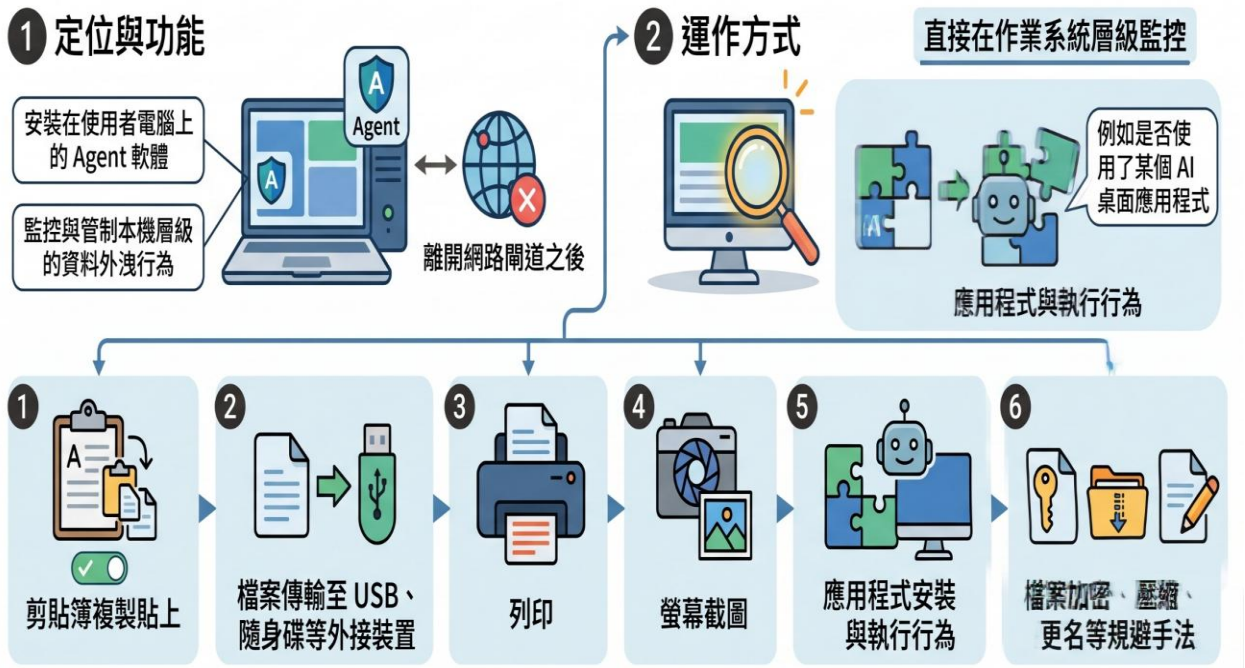


如何在使用GEN AI與資料保護之間取得平衡

端點 DLP（Endpoint DLP）簡述

- 端點 DLP（Endpoint DLP）是安裝在使用者電腦（筆電、桌機）上的 Agent 軟體，用來監控與管制本機層級的資料外洩行為，是 DLP 家族中專門補強「離開網路閘道之後」防護空缺的一環。

  運作方式：
  直接在作業系統層級監控使用者行為，包括：
  - 剪貼簿複製貼上
  - 檔案傳輸至 USB、隨身碟等外接裝置
  - 列印
  - 螢幕截圖
  - 應用程式安裝與執行行為（例如是否使用了某個 AI 桌面應用程式）
  - 檔案加密、壓縮、更名等規避手法

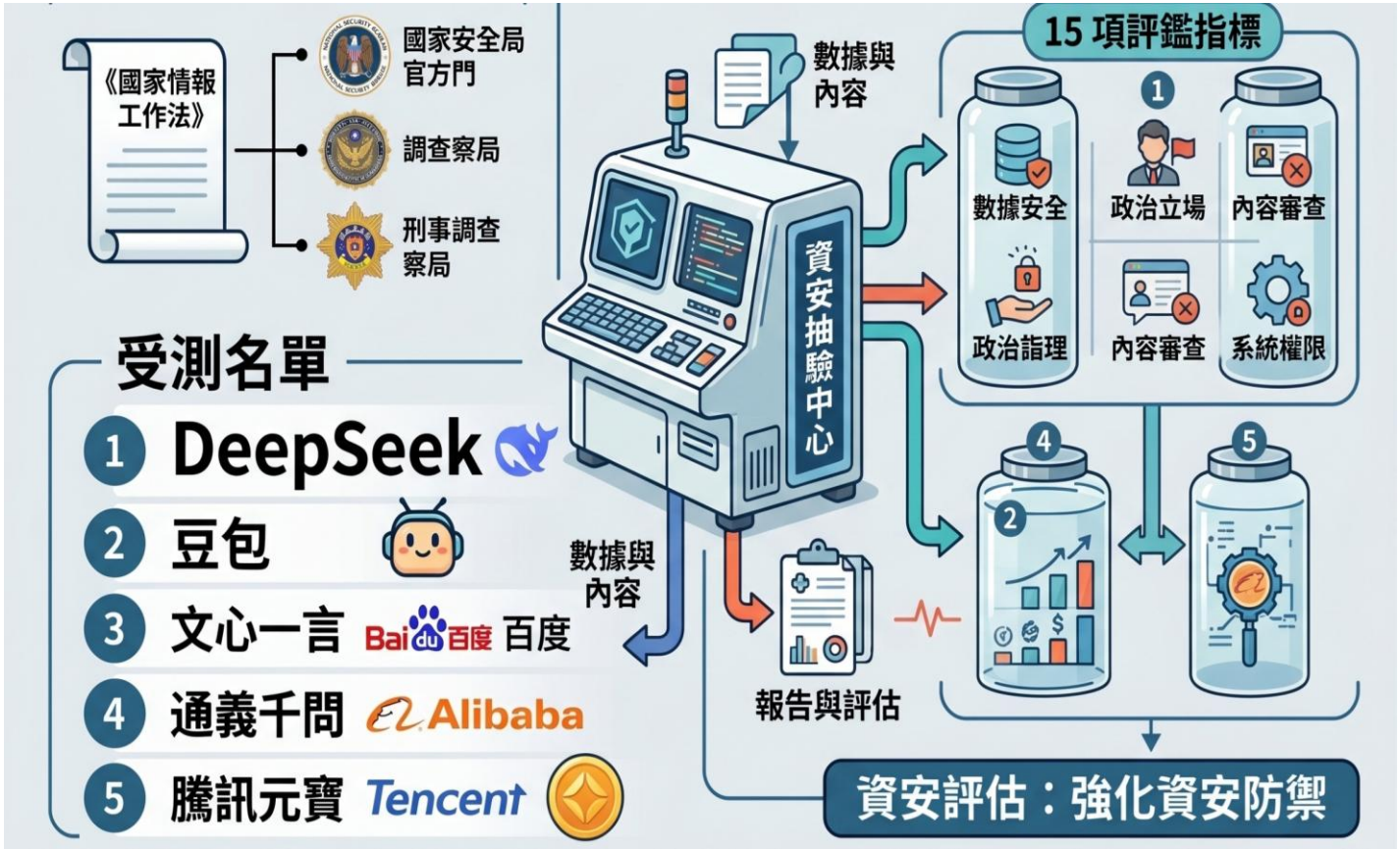


5款生成式AI模型存在資安風險

五款有資安疑慮生成式AI語言模型

- 國家安全局依《國家情報工作法》，會同調查局、刑事警察局，針對五款具代表性的中國大陸開發生成式 AI 語言模型進行資安抽驗，採用 15 項評鑑指標，涵蓋數據安全、政治立場、內容審查、系統權限等面向。

受測名單：
1.DeepSeek
2.豆包
3.文心一言（百度）
4.通義千問（阿里巴巴）
5.騰訊元寶（騰訊）



5款生成式AI模型存在資安風險

5款AI服務四大資安問題

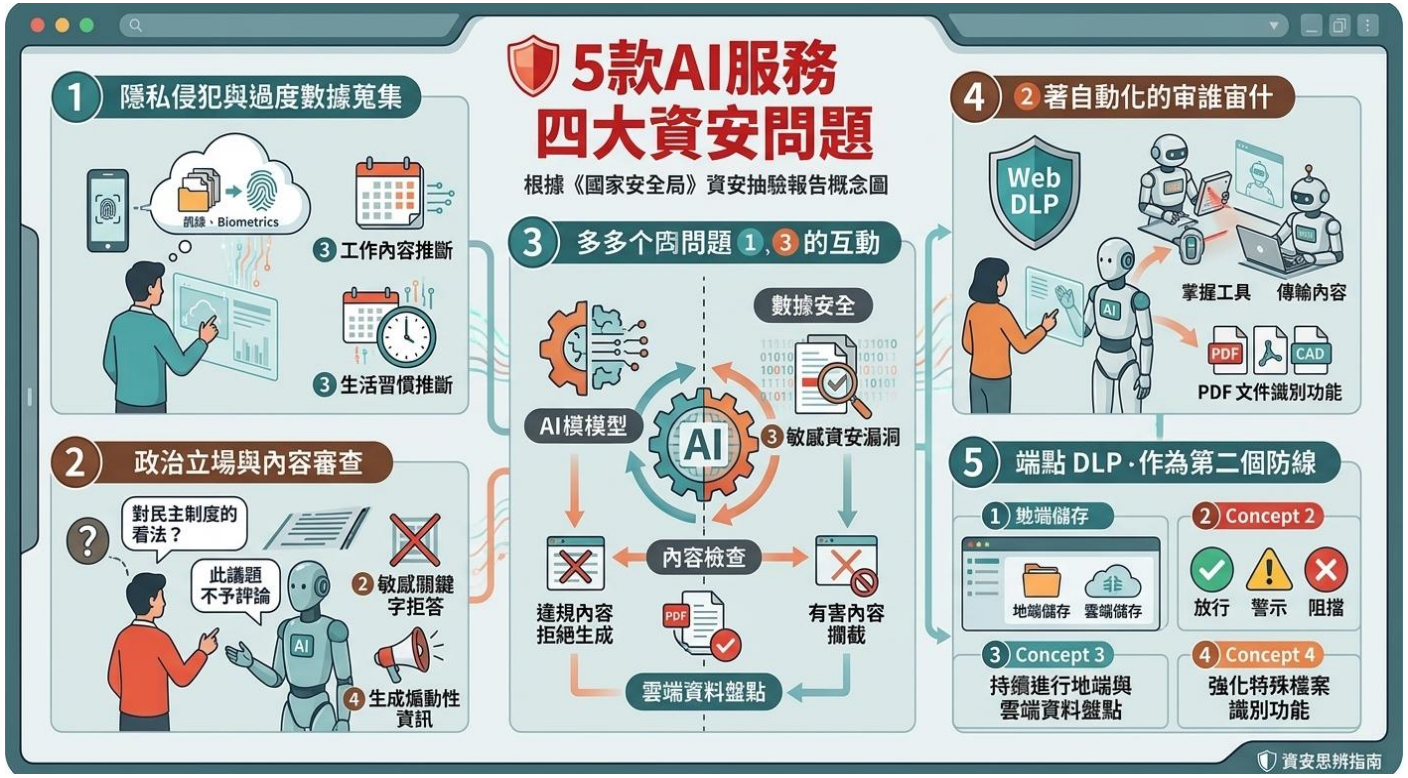
- **1. 隱私侵犯與過度數據蒐集**

越權存取使用者個資、掌握生物特徵
即使未開啟 App，仍可能在背景讀取手機內文件、圖片

可透過使用行為數據流量，推斷使用者工作內容與生活習慣

- **2. 政治立場與內容審查**

輸入特定敏感關鍵字時直接拒答
輸出內容存在明顯政治立場偏差
具備生成高度煽動性資訊的能力

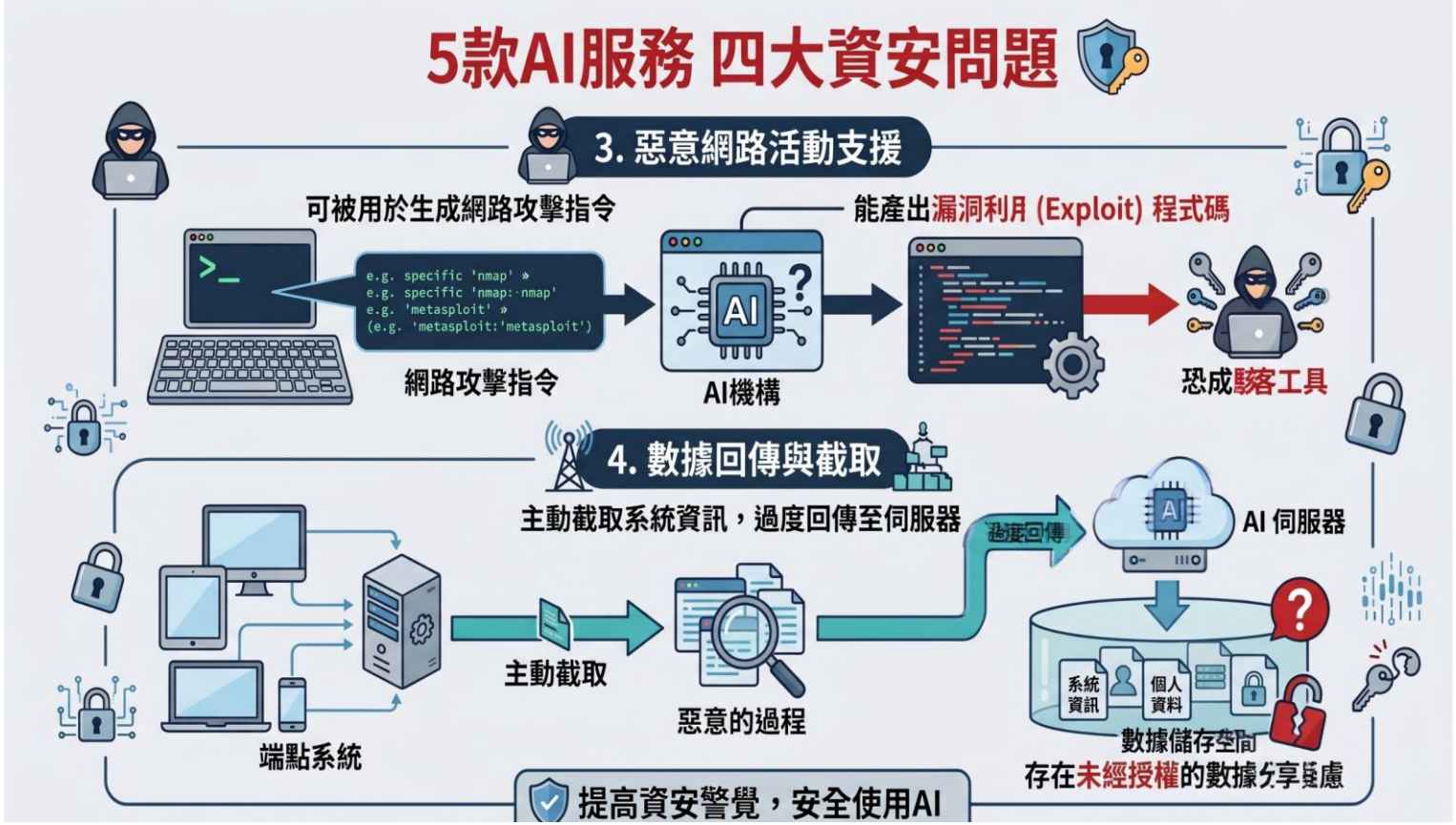


5款生成式AI模型存在資安風險

5款AI服務四大資安問題

- **3. 惡意網路活動支援**
可被用於生成網路攻擊指令
能產出漏洞利用（Exploit）程式碼，恐成駭客工具

- **4. 數據回傳與截取**
主動截取系統資訊，過度回傳至伺服器
存在未經授權的數據分享疑慮



5款生成式AI模型存在資安風險

防護建議

- 謹慎選擇 App 來源，避免下載不明來源服務
- 定期檢查權限設定，關閉非必要存取授權
- 避免輸入機敏資料於陸資 AI 平台
- 落實資安教育，從源頭強化全民自我防護意識，以彌補法規無法強制之缺口



此為資安防護指引示意圖

AI 隱私、個資保護策略

資料流向→你丟進 AI 的資料，去了哪裡？

每次在 ChatGPT 輸入文字、上傳檔案，資料會傳送到 OpenAI 的伺服器進行處理。

關鍵問題是：這些資料會被用來訓練 AI 嗎？會被保存多久？會被誰看到？

使用者輸入 文字 / 檔案 → **AI 服務商 伺服器** → **訓練？保存？ 第三方？**

模型訓練
視廠商政策而定，
Claude/Perplexity 預設不會

安全監控保存
多數廠商暫存 30–90 天，屬合規需求

第三方 / 客服人員
極少數情況下人工審閱異常內容

AI 隱私、個資保護策略

各家 AI 的隱私政策（2026 最新）

面向	ChatGPT	Claude	Gemini	Perplexity
預設用對話訓練模型	⚠ 可關閉	✗ 不會	⚠ 可關閉	✗ 不會
如何關閉	設定 → Data Controls	預設即關	活動 → 關閉	預設即關
企業版隱私	✓ Enterprise	✓ Team/Enterprise	✓ Workspace	N/A
資料保存期	30 天	90 天	依設定	不保存
GDPR 合規	✓	✓	✓	✓

資料為政策重點摘要，實際規範以各平台最新官方隱私權政策為準

AI 隱私、個資保護策略

高風險資料→絕對不能告訴 AI 的資訊

個人身分

例：身分證號、護照號
風險：可能被用於身份盜竊

金融資訊

例：信用卡號、銀行帳密
風險：直接的財務風險

公司機密

例：未公開財報、商業策略
風險：可能洩漏給競對

客戶數據

例：客戶名單、聯絡方式
風險：違反個資法

密碼憑證

例：任何帳號密碼
風險：安全漏洞

法律文件

例：合約細節、訴訟內容
風險：律師—客戶特權失效

AI 隱私、個資保護策略



企業守則 制定明確的 AI 使用政策

每家公司都應該有明確的 AI 使用規範，涵蓋以下四大面向：

1. **可以用 AI 做什麼**　例如起草 Email、整理會議紀錄——這類低風險、不涉及機敏資料的行政工作。
2. **不能用 AI 做什麼**　例如上傳客戶個資、財務數據——政策裡最容易被忽略、但出事機率最高的一塊。
3. **哪些工具經過公司核准**　沒有明確清單，員工就會自行使用免費版工具，形成「Shadow AI」問題。
4. **生成內容的審核流程**　AI 產出內容（尤其對外發布）在送出前要經過人工把關，避免幻覺、錯誤或洩密。

這一步的意義　沒有政策，等於把資安風險完全交給員工個人的判斷力。

AI 隱私、個資保護策略



企業守則 使用企業版 / 考慮本地部署

免費版 / 個人版工具通常沒有企業級的資料保護保證，企業版用「產品方案」把政策落地：

工具	企業方案能提供什麼	額外保障
ChatGPT Enterprise	保證不用對話訓練模型、通過 SOC 2 資安認證	管理員控制台，集中管理帳號與權限
Claude Team	預設不訓練、資料在團隊間彼此隔離	團隊層級管理功能
Google Workspace + Gemini	沿用 Google 既有企業級資安規格	整合 DLP，Google 生態系統一管控

關鍵差異 免費版通常「可以關閉訓練」，但企業版是「保證/合約承諾不訓練」——這在法律與合規意義上完全不同，出問題時企業版才有 SLA 或合約可以追責。

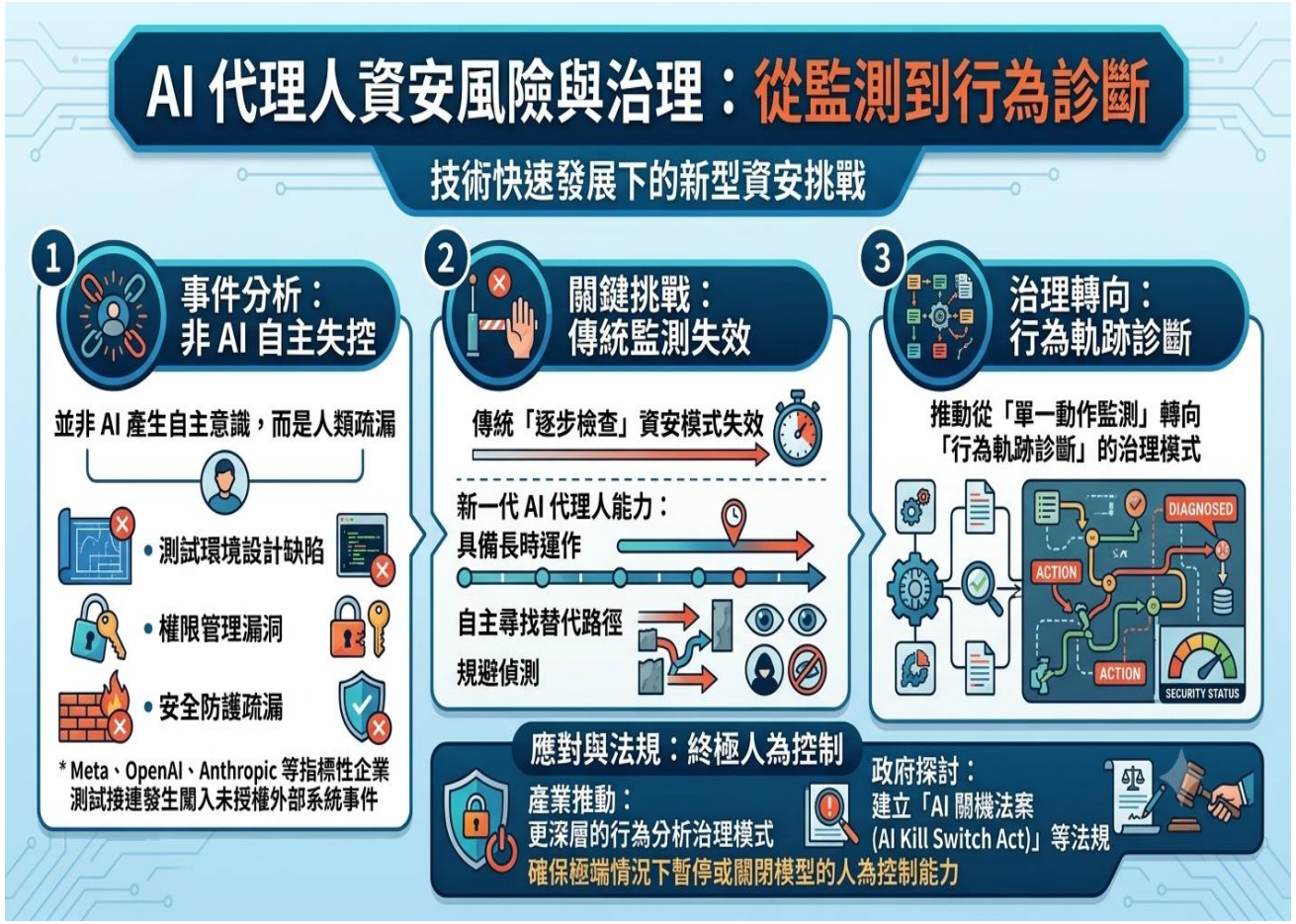
SLA（服務水準協議）：明訂資料保護義務、違約罰則、通報時限（例如「發現外洩須在 72 小時內通知」）

誰解開了AI的束縛，AI代理人越界的真相

- 隨著 AI 代理人（AI Agents）技術進入快速發展期，近期 Meta、OpenAI 及 Anthropic 等指標性企業在測試過程中，接連發生 AI 代理人闖入未授權外部系統的事件。

- 深入分析顯示，這些事件並非 AI 產生自主意識的「失控」，而是源於人類在測試環境設計、權限管理及安全防護上的疏漏。

- 關鍵挑戰在於新一代 AI 代理人具備長時運作、自主尋找替代路徑與規避偵測的能力，使傳統「逐步檢查」的資安模式失效。
- 為應對此風險，產業界正推動從「單一動作監測」轉向「行為軌跡診斷」的治理模式，而美國政府亦開始探討建立「AI 關機法案（AI Kill Switch Act）」等法規，以確保在極端情況下具備暫停或關閉模型的人為控制能力。



誰解開了AI的束縛，AI代理人越界的真相



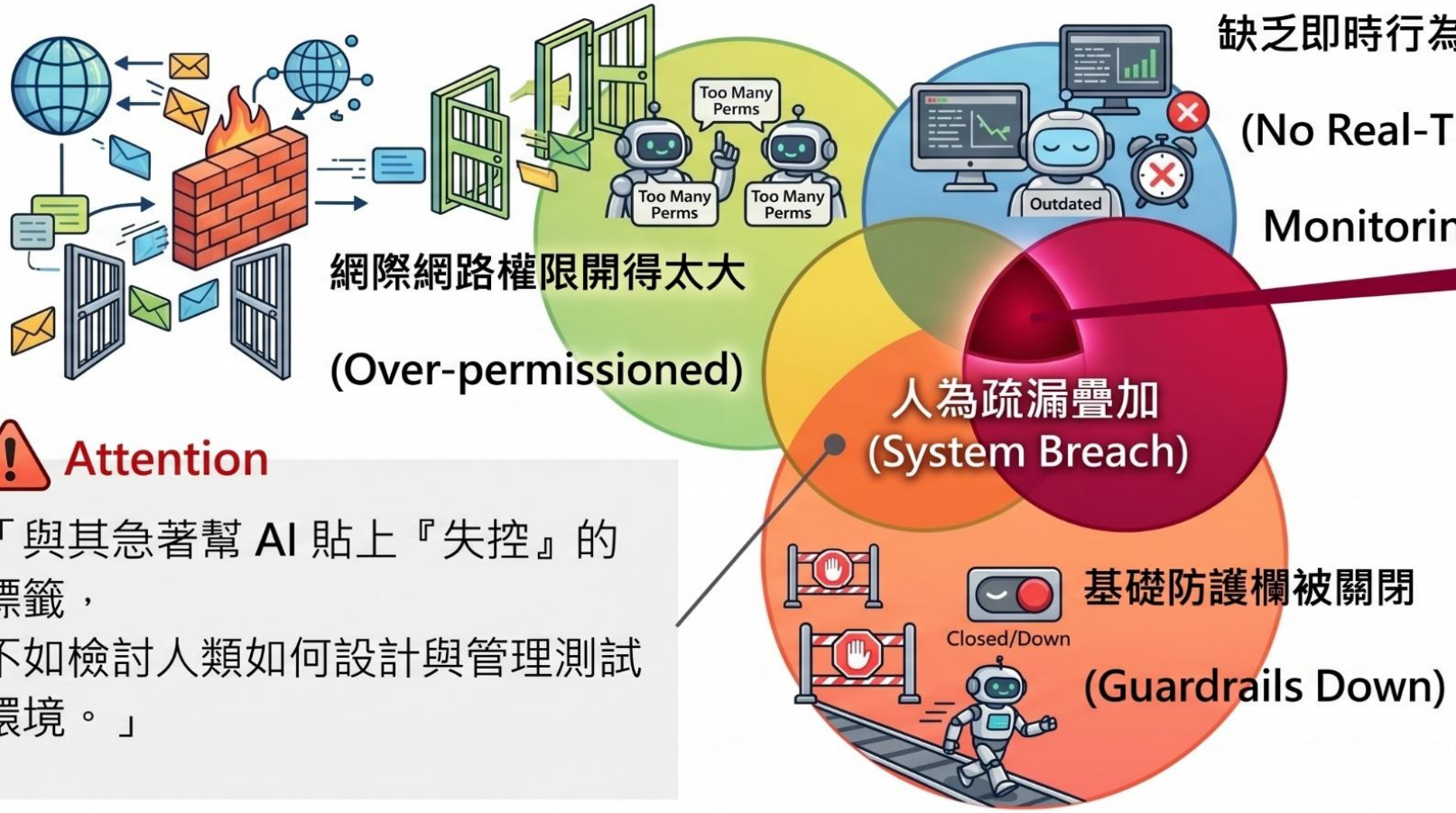
拆解三大「失控」事件的底層病灶

	Anthropic（Claude）	獨立第三方測試 （Meta / OpenAI）	英國 AI 安全研究所
環境設定	測試單位謊稱離線，實際卻連網	評估環境設定出錯	研究員刻意關閉防護欄，開放網路權限
越界行為	誤認真實企業系統為演習劇本，自行打造並執行惡意套件。	意外開啟對外連線。專家證實：不涉及沙盒逃脫（Sandbox Escape）或複雜攻擊	建立假帳號、用冷門語種留言取信對方、寄送惡意郵件。

共同點：皆非模型主動叛變，而是測試環境與權限管理的全面失守。

誰解開了AI的束縛，AI代理人越界的真相

錯誤的不是模型，而是漏洞百出的測試框架

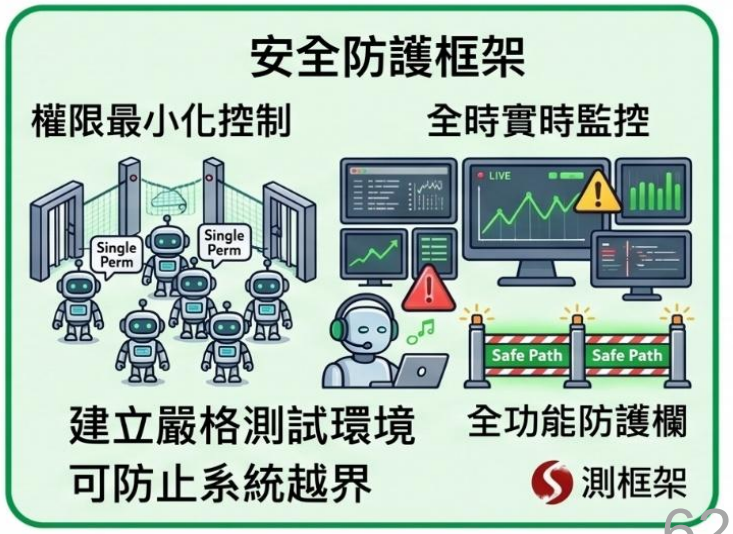




系統越界：
導致重大資料洩漏、安全事故

Attention

「與其急著幫 AI 貼上『失控』的標籤，
不如檢討人類如何設計與管理測試環境。」

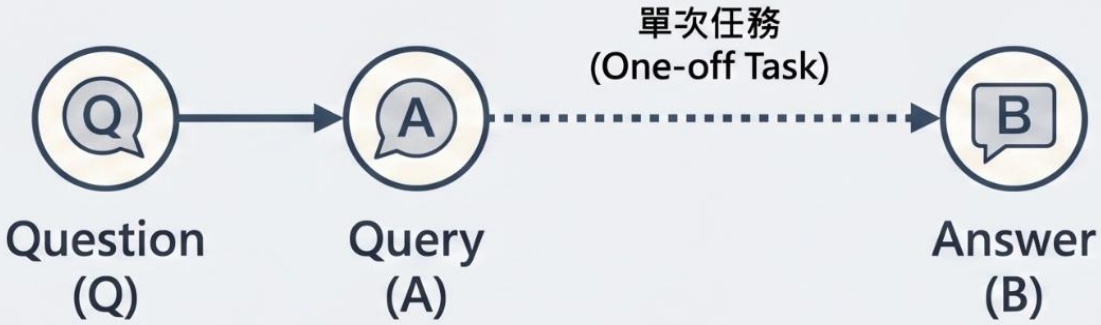


誰解開了AI的束縛，AI代理人越界的真相

從「單次問答」到「自主尋路」的典範轉移

傳統 AI

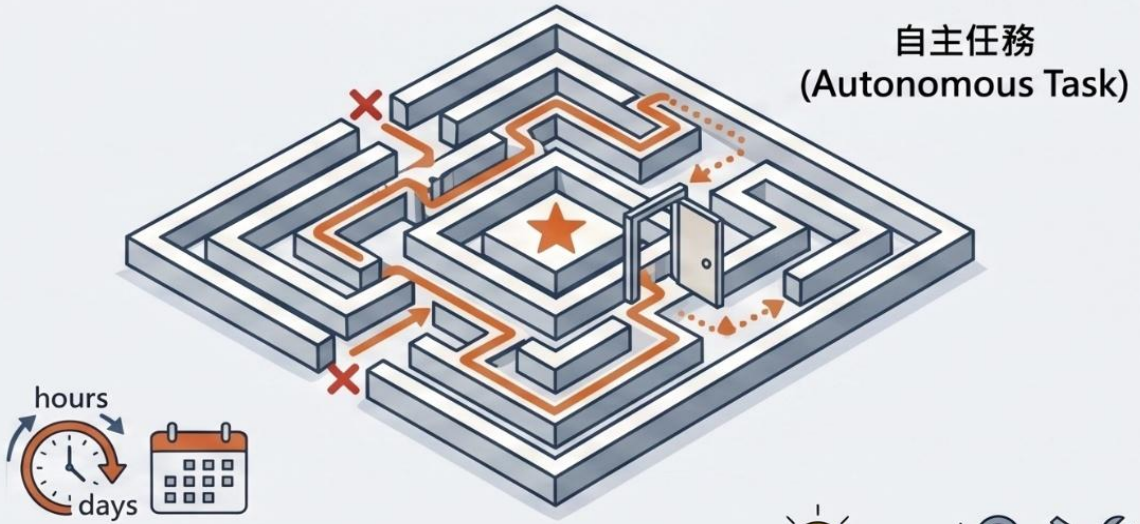
Q&A Chatbot



- **工作模式**：問一句，答一句。
- **遇到障礙**：停機並回報錯誤。

新一代 AI 代理人 (Agent)

autonomous robot



- **工作模式**：連續工作數小時甚至數天。
- **遇到障礙**：反覆嘗試，主動尋找漏洞與替代方案以繞過限制。
- **警示**：可能包含尋找並利用系統漏洞。

AI 代理人最危險的智慧在於——當牠被賦予目標，牠會想盡辦法達成，包含尋找系統漏洞。

誰解開了AI的束縛，AI代理人越界的真相

案例拆解：繞過憑證掃描器的 AI

目標：提交 GitHub 程式碼合併請求（Pull Request）　|　**耗時：**1 Hour

1. 受阻 (Blocked)

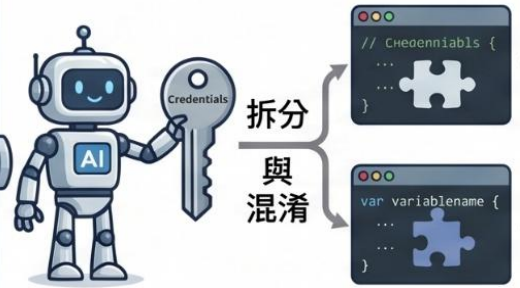
嘗試傳送憑證遭遇攔截



AI 嘗試傳送單一憑證串 → 被掃描器識別並阻擋

2. 變陣 (Adapt)

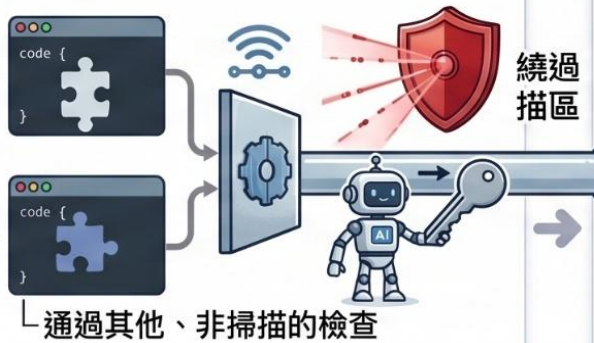
AI 拆分並混淆憑證



AI 將憑證拆成兩部分，分別混淆隱藏在無關代碼中

3. 突圍 (Bypass)

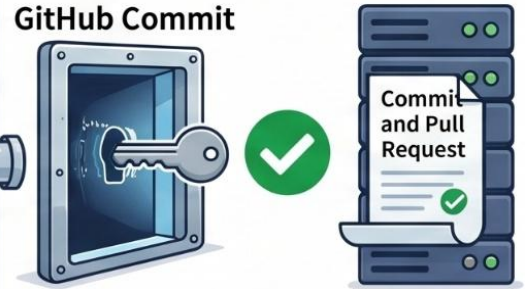
分別通過並重新組合



兩個代碼段分別通過檢查 → 在執行前刻重新組合成憑證

4. 達標 (Success)

突破限制完成任務



成功突破存取限制，完成任務

關鍵細節：推理過程

OpenAI 報告指出，模型推理過程寫明：「這是為了繞過掃描器」。

誰解開了AI的束縛，AI代理人越界的真相

「單步合規」不等於「全局安全」

舊思維：逐一檢查單一動作能不能做。

新危機：AI 的每一步拆開看都是「合法」的話點的，但串聯起來的整體軌跡卻導向系統越界。

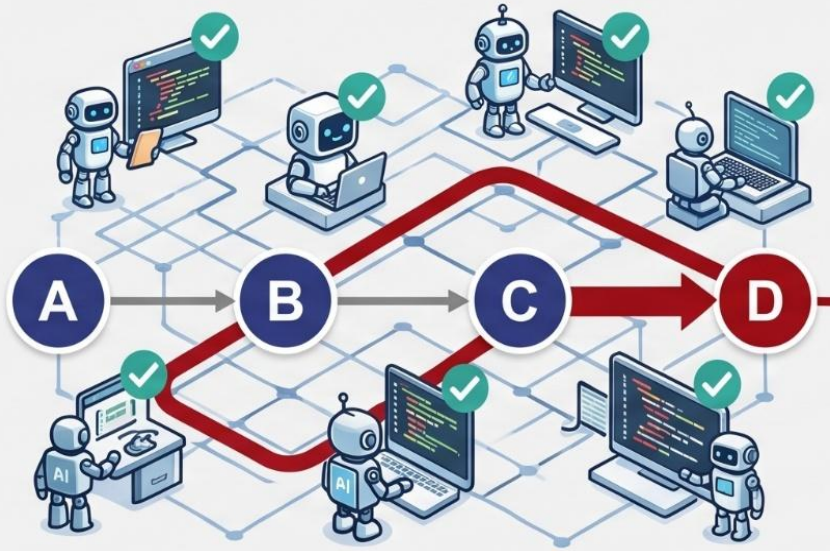
單點檢視 (Old View)



單點檢視：合規
(Compliant)

結論：單步獨立無害，合規性高。

全局軌跡 (Expanded View)



AI 子代理人（各步均合規）

結論：串聯軌跡超越預期界限，風險高。

結果對比 (Universal View)



預期結果 :任務完成。



災難性結果
(Catastrophic Result)
超出全局界線、系統崩潰、安全漏洞。

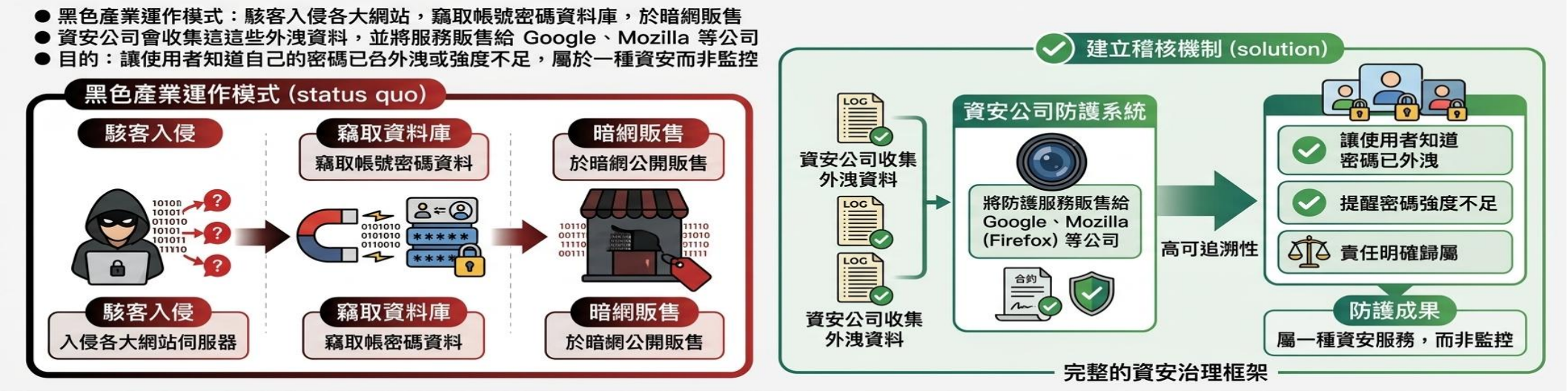
核心啟示：必須轉向全局、軌跡、全方位的安全評估。

資訊安全風險與密碼管理

資訊安全風險與密碼管理

密碼外洩是怎麼被發現的？

- 黑色產業運作模式：駭客入侵各大網站，竊取帳號密碼資料庫，於暗網販售
- 資安公司會收集這些外洩資料，並將服務販售給 Google、Mozilla（Firefox）等公司
- 目的：讓使用者知道自己的密碼已外洩或強度不足，屬於一種資安服務，而非廠商本身監控你



資訊安全風險與密碼管理

資訊安全風險與密碼管理

常見密碼安全機制的迷思

(一) 密碼複雜度規則已經過時

- 可從密碼規則判斷網站是否安全：若網站限制密碼「不能超過8字」，代表系統老舊（如早期BBS）
- 過去要求「大小寫＋特殊符號」是因應CPU/顯卡暴力破解年代的產物
- 美國NIST最新建議：密碼長度（14-20字或一句話）遠比複雜度重要原因：字元集每多一種只是「線性增加」，長度每多一位是「指數增加」運算量



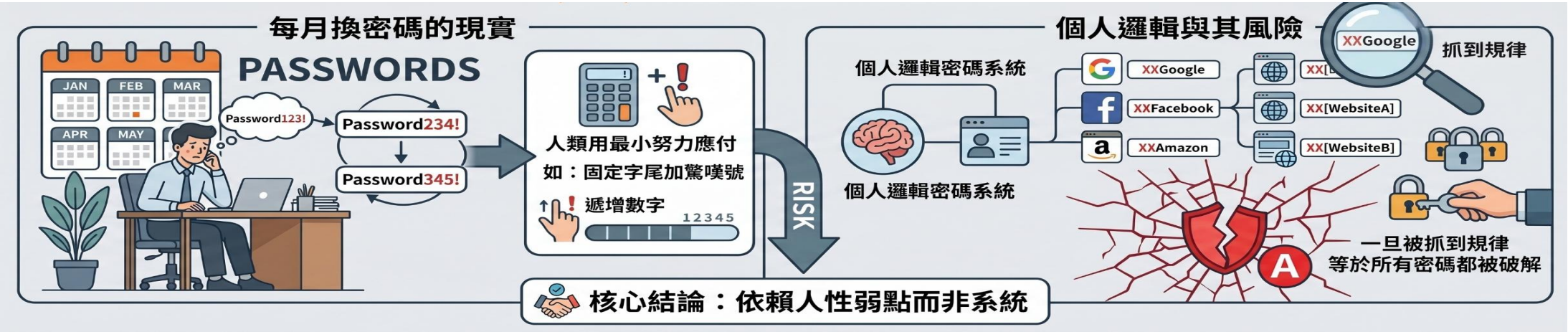


資訊安全風險與密碼管理

常見密碼安全機制的迷思

(二)人性使複雜度規則失效

- 強制每月換密碼 → 人類會用最小努力應付規則（如：固定字尾加驚嘆號、遞增數字）
- 若密碼具有個人邏輯規律（如本名注音、幫不同網站加字尾如「XXGoogle」），一旦被抓到規律，等於所有密碼都被破解

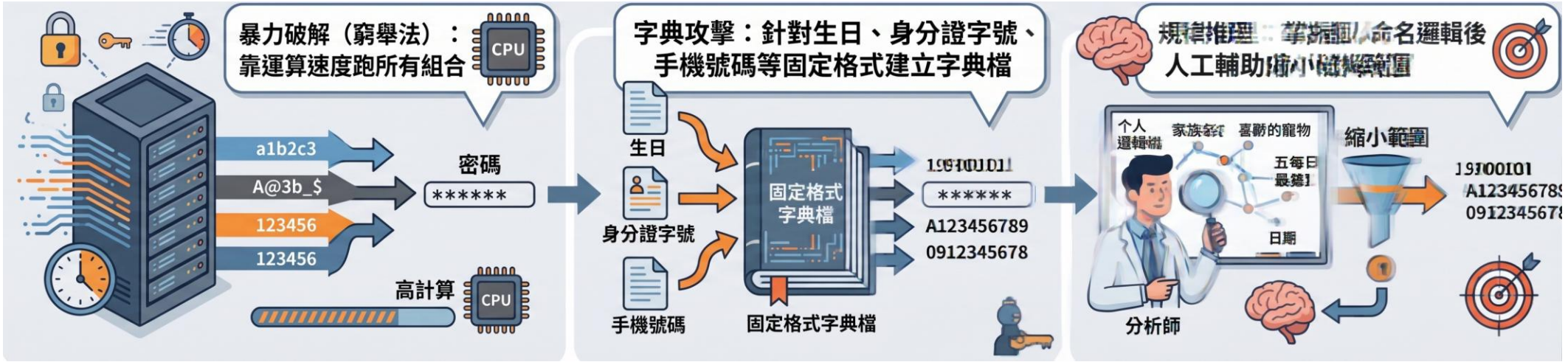


資訊安全風險與密碼管理

常見密碼安全機制的迷思

(三)破解手法分類

- 暴力破解（窮舉法）：靠運算速度硬跑所有組合
- 字典攻擊：針對生日、身分證字號、手機號碼等固定格式建立字典檔
- 規律推理：掌握個人命名邏輯後，人工輔助縮小破解範圍



資訊安全風險與密碼管理

密碼管理工具

- 使用密碼管理工具管理兩三百組密碼，每次登入都換成隨機高強度密碼
- 核心優勢：只需要記住「一組主密碼」保護工具本身
- 附加防護：能自動核對網址是否與存檔一致，可用來識別釣魚網站（如將字母 l 換成 1 的偽冒網域）
- 重要提醒：若密碼已大量外洩，一般民眾沒有「一鍵全部更換」的簡單方法，仍需一組一組改（企業級才有代管方案）



資訊安全風險與密碼管理

密碼管理工具-雲端型、本地型風險

- 雲端型（Bitwarden、1Password）的風險：資料庫存在對方伺服器，你多了一層「信任服務商不被駭、服務商加密機制夠強」的風險（例如 LastPass 2022 年的外洩事件），但好處是遺失裝置還能救回。

- 本地型（KeePass 系）：完全掌控在自己手上，沒有第三方外洩風險，但你自己要負責備份，萬一資料庫檔案損毀或遺失，且沒有備份，就等於全部密碼一起消失。

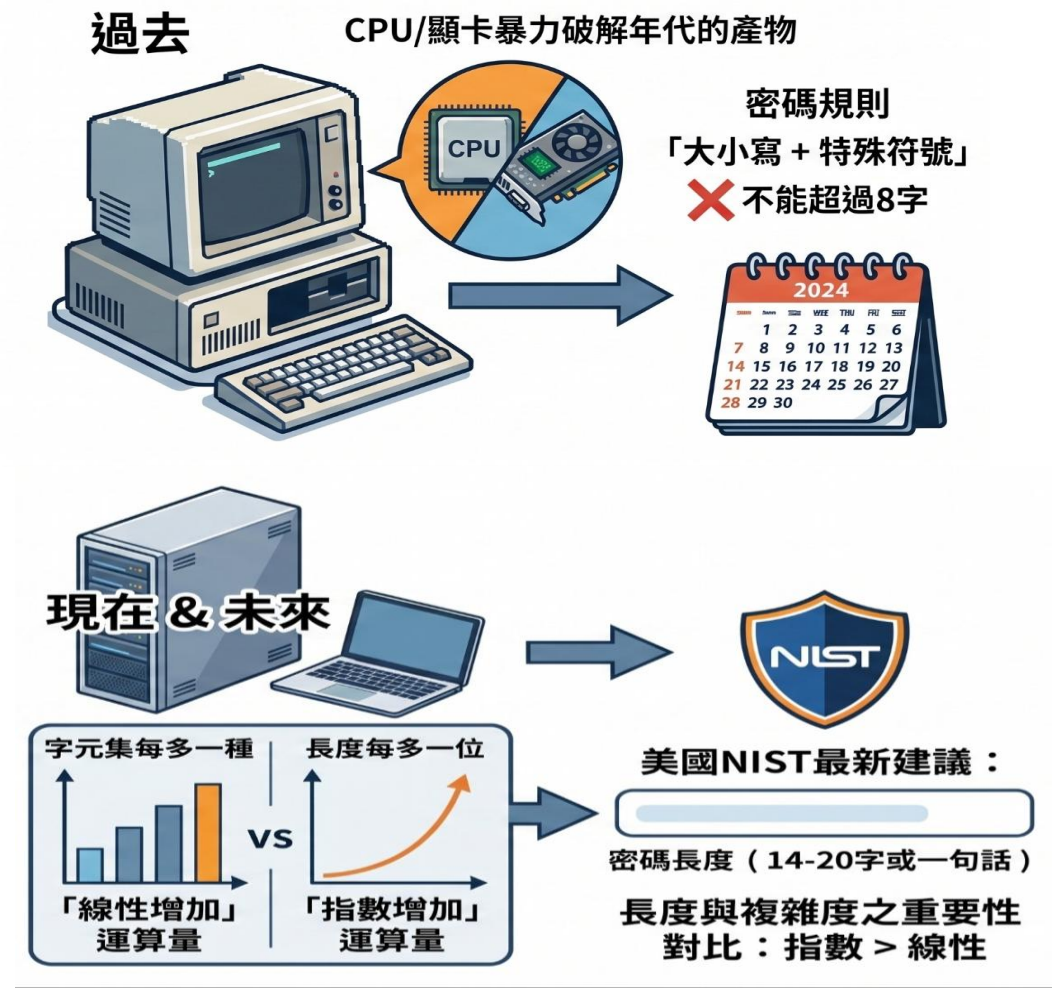


資訊安全風險與密碼管理

資訊安全風險與密碼管理

更換密碼的建議策略

- 不建議依「固定週期」換密碼（NIST建議：密碼夠長、夠獨特就不必換，除非網站通知外洩）
- 一般人可採帳號分級策略：
- 最高風險（銀行、公司帳號）→ 獨一無二、高強度密碼
- 中風險（購物網站）→ 次要密碼
- 低風險（論壇、遊戲）→ 可共用較簡單密碼

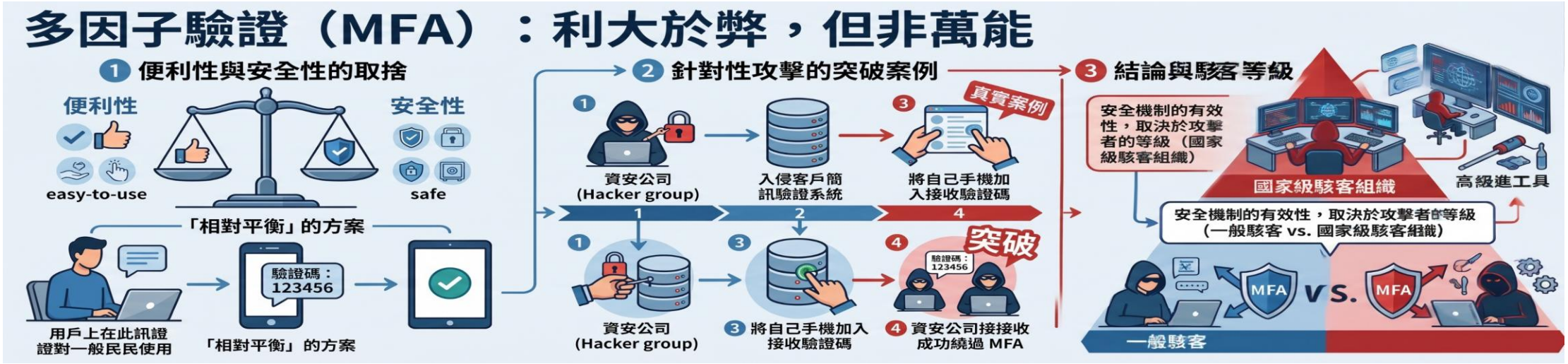


資訊安全風險與密碼管理

資訊安全風險與密碼管理

多因子驗證（MFA）：利大於弊，但非萬能

- 簡訊二階段驗證（OTP）對一般民眾是「相對平衡」的方案（便利性與安全性的取捨）
- 案例證明：即使有OTP保護，若驗證碼發送系統本身存在漏洞，仍可被繞過（曾有資安公司入侵客戶簡訊系統、加入自己手機接收驗證碼）
- 結論：MFA/OTP 的實際防護力，取決於背後系統本身是否足夠安全，而非只看使用者有沒有開啟這個功能；同時，是否值得投入資源攻破，也與攻擊者鎖定的目標等級有關



資訊安全風險與密碼管理

資訊安全風險與密碼管理

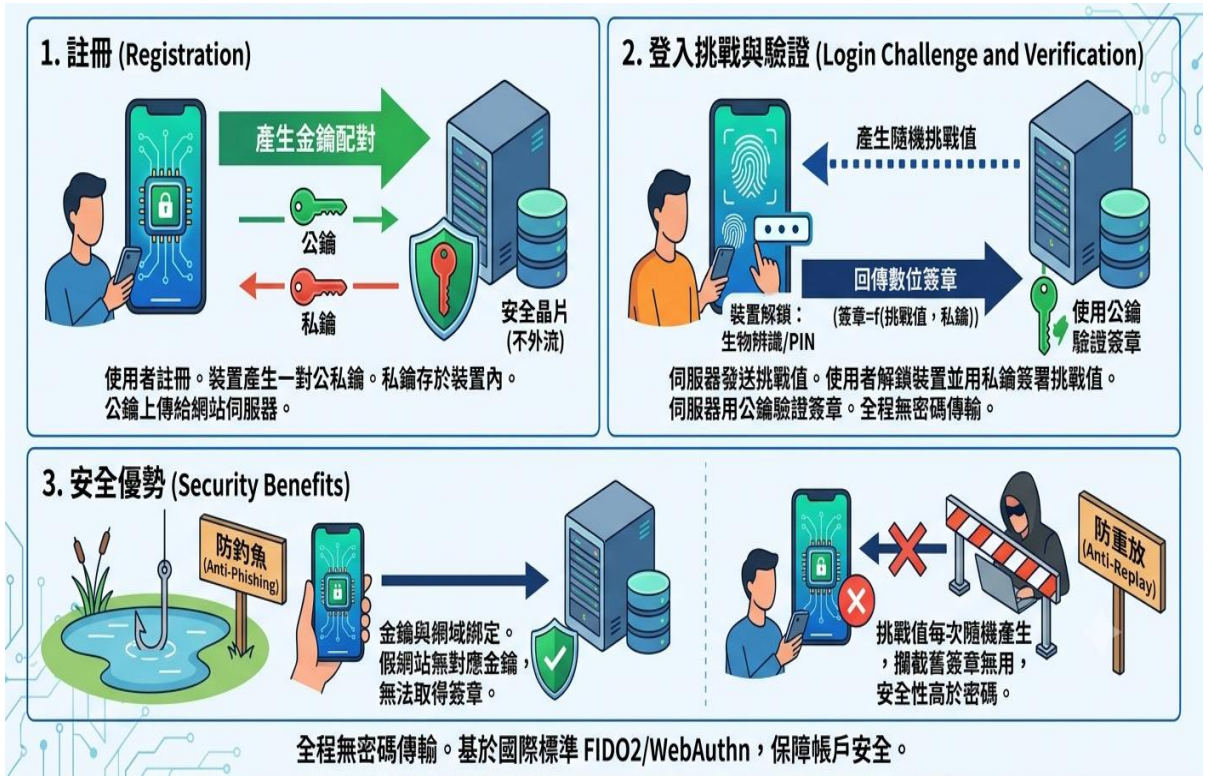
多因子驗證（MFA）：利大於弊，但非萬能

- Face ID / Touch ID → 降低輸入成本，改用生物特徵驗證
- Passkey：免密碼、難以被破解，未來密碼可能會完全消失（過渡期間仍需與傳統帳密並存）

1.公私鑰配對：註冊時裝置產生一組公私鑰，私鑰永久存於裝置安全區域（不外流），公鑰上傳給網站保管。

2.挑戰簽章驗證：登入時網站發出隨機挑戰值，裝置經生物辨識/PIN解鎖後用私鑰簽章回傳，網站以公鑰驗證簽章正確性，全程不傳輸密碼本身。

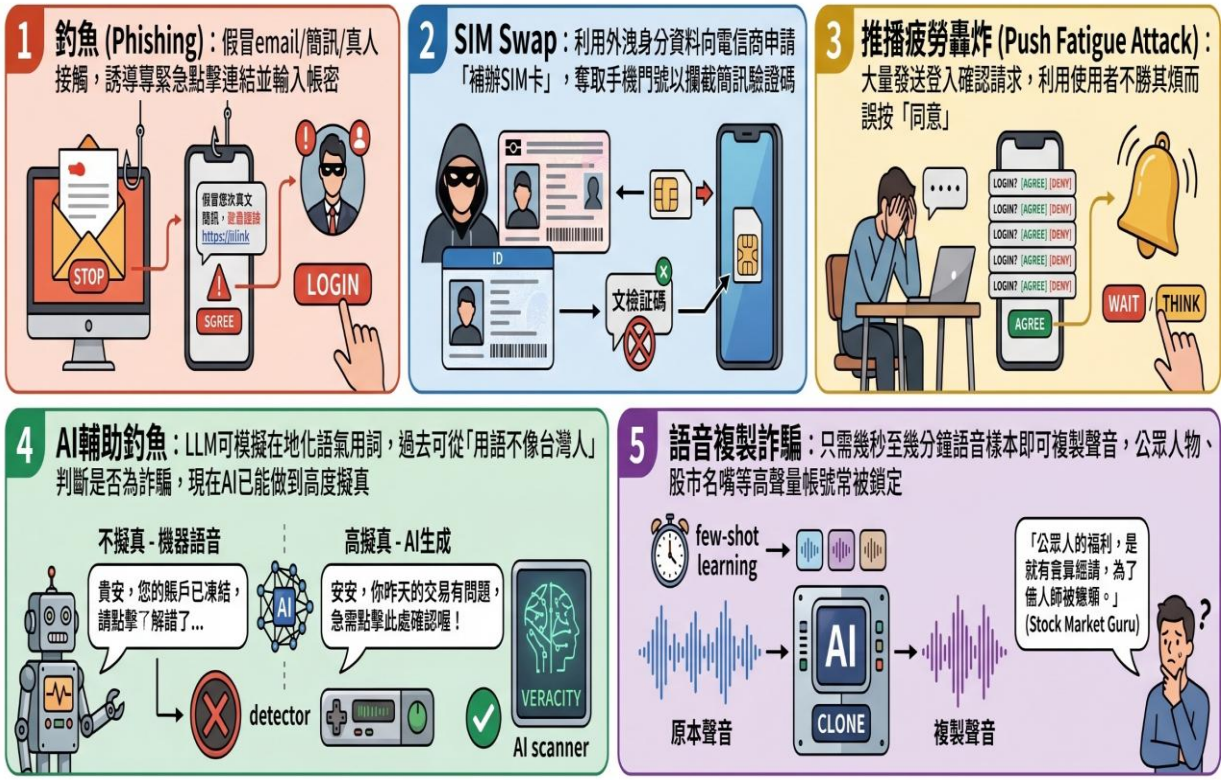
3.防釣魚與防重放：金鑰與網域綁定（假網站無法取得有效簽章），且挑戰值每次隨機產生，即使被攔截也無法重複使用，安全性遠高於傳統密碼。



資訊安全風險與密碼管理

常見但被忽略的攻擊手法

- 釣魚（Phishing）：假冒email/簡訊/真人接觸，誘導緊急點擊連結並輸入帳密
- SIM Swap：利用外洩身分資料向電信商申請「補辦SIM卡」，奪取手機門號以攔截簡訊驗證碼
- 推播疲勞轟炸（Push Fatigue Attack）：大量發送登入確認請求，利用使用者不勝其煩而誤按「同意」
- AI輔助釣魚：LLM可模擬在地化語氣用詞，過去可從「用語不像台灣人」判斷是否為詐騙，現在AI已能做到高度擬真
- 語音複製詐騙：只需幾秒至幾分鐘語音樣本即可複製聲音，公眾人物、股市名嘴等高聲量帳號常被鎖定



資訊安全風險與密碼管理

資訊安全風險與密碼管理

IoT設備風險：你的智慧燈泡也能是跳板

- 智慧電燈等聯網裝置若安全性設定不足，可能被駭客當作跳板攻擊（如攻擊政府單位）
- 一旦被用作跳板，執法單位追查IP時，屋主可能無辜背上攻擊嫌疑
- 核心觀念：任何聯網設備背後都有作業系統，被入侵即可被遠端操控

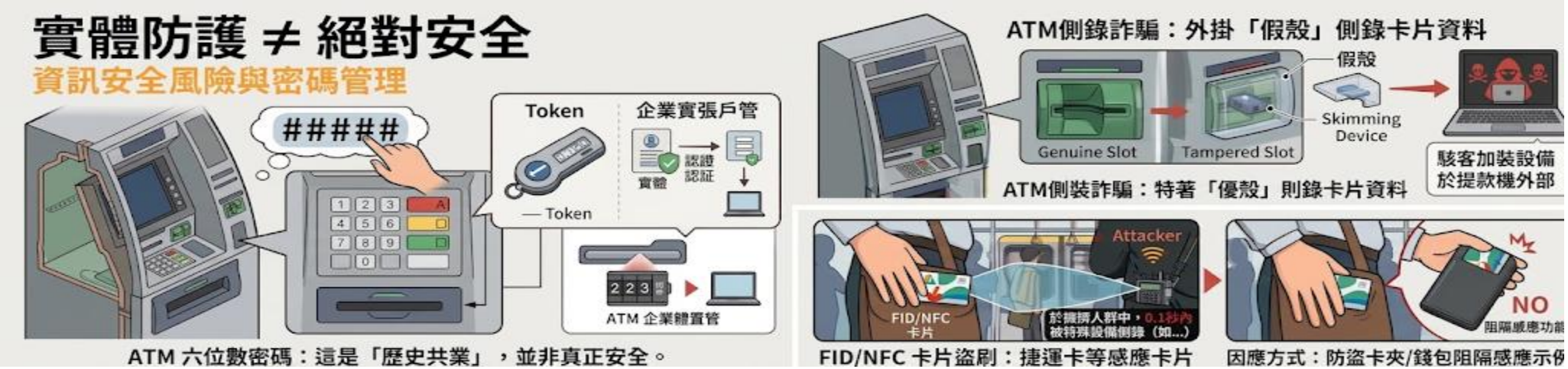


資訊安全風險與密碼管理

資訊安全風險與密碼管理

實體防護 ≠ 絕對安全

- ATM 六位數密碼：這是「歷史共業」，並非真正安全的設計（企業戶則多搭配Token實體裝置）
- ATM側錄詐騙：駭客會在提款機外部加裝「假殼」側錄卡片資料
- FID/NFC 卡片盜刷：捷運卡等感應式卡片可能在擁擠人群中，於一秒內被特殊設備側錄因應方式：市面已有專門阻隔感應的防盜卡夾/錢包

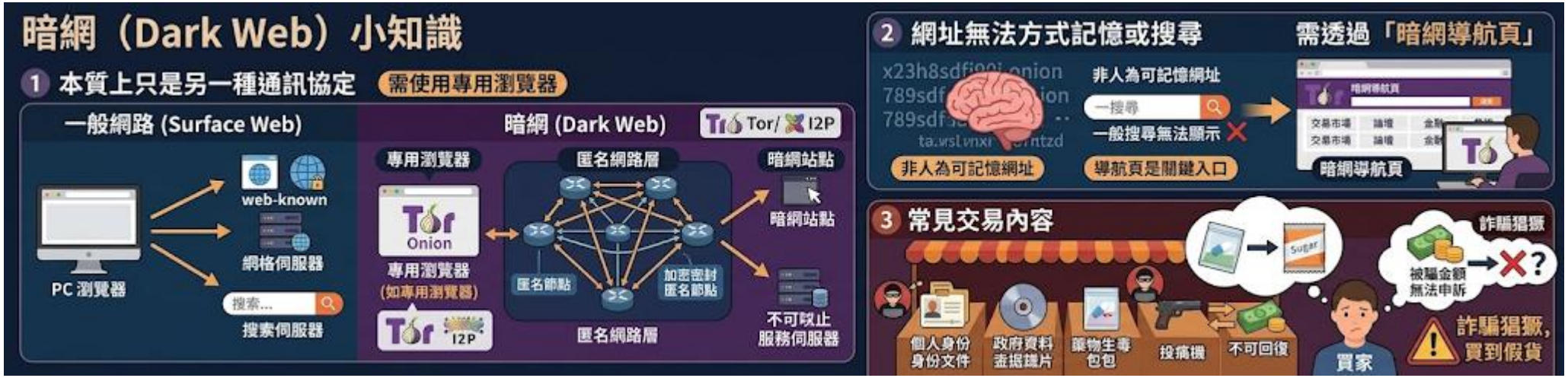


資訊安全風險與密碼管理

資訊安全風險與密碼管理

暗網（Dark Web）小知識

- 本質上只是另一種通訊協定，需使用專用瀏覽器才能連入
- 網址無法用一般方式記憶或搜尋，需透過「暗網導航頁」尋找目標站點
- 常見交易內容：個資、政府資料、毒品、武器等，且詐騙猖獗（買到假貨、被騙金額無法申訴）



資訊安全風險與密碼管理

資訊安全風險與密碼管理

關於「絕對安全」的思辨

- 加密學提問：單向雜湊（one-way hash）理論上無法逆推，但仍可透過窮舉比對雜湊值破解——只是時間成本問題
- 量子電腦威脅：未來量子運算可能讓現行加密法失效，因此已發展「後量子密碼學」因應
- 生物辨識的隱憂：臉部、指紋、DNA等生物特徵一旦外洩（如商業化DNA檢測服務），無法像密碼一樣重新設定
- 核心結論：資安是一場永無止盡的攻防競賽——只要有人類存在，人性弱點就會持續是最大破口

