星球永續健康線上直播

智慧數位資安 (7)

智慧模型盜取原理與實例

2026 年 5 月 13 日

當前人工智慧時代中商業 AI 模型如同餐廳的「祖傳秘方」一班，外界雖看不到真正配方，卻可能透過反覆點餐、品嘗與分析，逐步推測出料理的核心做法並加以模仿。如今許多大型語言模型與商業 AI 系統，也面臨類似挑戰。攻擊者不需直接入侵系統，便可能透過大量查詢與提示詞誘導，重建模型行為，甚至萃取關鍵資料與知識架構。本週我們將探討 AI 模型盜取的資安挑戰，以及智慧產業中的 AI 資料萃取攻擊實例。

健康科學新知

漢他病毒郵輪傳播潛在健康威脅:「嚴陣以待」

荷蘭籍郵輪 MV Hondius 爆發漢他病毒 Andes 株疫情，已造成 5 人確診、3 人死亡，並被視為首起在船艦封閉環境中發生漢他病毒人際傳播的案例，WHO 調查指出，安地斯株漢他病毒不同於一般病毒株須經齧齒類排泄物傳播，具人際傳播能力。主要透過親密切接觸擴散與氣溶膠傳播擴散，郵輪環境中具傳播風險。且該病毒潛伏期可長達 6 週，因此後續仍可能出現個案。各國正追蹤 4 月 24 日於聖赫勒拿島下船的乘客，英、美、法、德、荷、加等國均已啟動監測、隔離或應急措施。MV Hondius 抵達西班牙特內里費後，各國以防護裝備、軍用巴士與包機撤離乘客，WHO 建議受威脅者至少隔離觀察 42 天，專家認為，雖然病毒致死率高，但在嚴格追蹤與隔離下，目前對一般大眾風險仍偏低。

國際聚焦荷莫茲海峽航行安全:「跛柘無定」

美伊衝突僵持美國總統川普近日表示，他讀完伊朗對美方結束戰爭提案的回應，認為完全不可接受。伊朗方面透過巴基斯坦作為調解人提交方案，要求立即結束所有戰線的戰爭、解除美國對伊朗的海上封鎖，並保證不再對伊朗發動攻擊。國際報導指美方備忘錄共一頁、含 14 點內容，包含暫停伊朗核濃縮活動、解除制裁，以及恢復霍爾木茲

海峽自由通行等條件。伊朗總統佩澤什基安雖未直接提及該提案，但表示永遠不會向敵人低頭，談判並不代表投降。荷莫茲海峽是全球能源運輸要道，伊朗對該區域的控制已推高燃料價格並影響國際航運。事件也波及阿聯、韓國等相關船隻與地區安全，使危機從美伊衝突擴大為國際航運與能源安全問題。美國國防部長赫格塞斯指出伊朗自停火後仍多次攻擊商船、扣押船隻並襲擊美軍，但目前這些行動尚未達到重新展開大規模作戰的門檻。川普政府暫緩部分引導船隻通過荷莫茲海峽的行動，希望為可能的外交協議保留空間，但對伊朗港口的封鎖仍持續。在國際應對方面，英國宣布皇家海軍將派遣一艘軍艦前往中東，參與保護霍爾木茲海峽航運安全的國際任務。該倡議由英國首相施凱爾與法國總統馬克宏共同推動，但施凱爾強調任務僅會在區域戰事結束後執行。對此伊朗警告若英法在海峽部署軍力，將遭到果斷而迅速的回應。馬克宏隨後澄清，法國從未設想過海軍部署，而是規劃一項「與伊朗協調」的安全任務。來自逾 40 個國家的國防部長將於週一由英國國防大臣約翰·希利與法國同行凱瑟琳·沃特蘭共同主持會議，討論敵對行動停止後如何監管海上交通。外交方面，中國因同時與伊朗及美國保持溝通，被視為可能協助降溫的關鍵角色。伊朗外上週長訪問北京，中國外長於會後表示將推動全面開放荷姆茲海峽維持航運安全，伊朗-中國會面也讓外界關注中國是否會推動和平談判。但中國雖有意避免能源危機惡化，但是否願意對伊朗施加實質壓力仍不明朗。

整體而言，美伊停火仍在維持，但局勢高度脆弱。未來是否升級，取決於美國護航行動是否擴大、伊朗是否加強攻擊，以及中國、巴基斯坦等第三方能否促成外交突破。

美國-中國峰會影響區域與全球經貿:「折衝樽俎」

美國總統川普與中國國家主席習近平預計於 2026 年 5 月 13 日展開為期兩天的北京高峰會。這是兩國領導人自 2025 年 10 月於韓國釜山會晤後的首度直接對話，亦是川普自 2017 年以來再次踏上中國領土。此次會談背景處於高度緊張的國際局勢中，特別是美國與以色列於 2026 年 2 月對伊朗發動打擊後所引發的區域戰爭，以及長期存在的美中貿易摩擦與科技競爭。美方官員透露，穩定世界前兩大經濟體之間的關係是本次訪華的核心目標。在經貿實質合作方面，雙方正尋求建立長效的溝通機制與採購協

議。具體預期成果包括中國將宣布針對波音公司飛機、美國農產品及能源產品的大規模採購。為了法制化經貿往來，雙方預計正式宣布成立「貿易委員會」與「投資委員會」，雖然這些機制的具體運作細節仍待後續磋商。此外，關於稀土礦物的供應穩定性亦是重點，雙方正評估延長去年秋季達成的貿易戰休戰協議。儘管美國最高法院於 2026 年 2 月裁定川普無權對全球進口商品徵收部分關稅，但川普已表態將尋求其他法律途徑重申其關稅立場，這使得本次會談在貿易政策的延續性上顯得至關重要。地緣政治與國防安全領域則涉及更為棘手的衝突點。美方持續對中國與伊朗、俄羅斯之間的經貿聯繫表達嚴正關切，特別是涉及雙用途產品與潛在武器出口所產生的收入支撐。在台海問題上儘管中國近年加強軍事威懾，習近平主席對美方的軍援立場感到不滿，但美國官員強調其對台政策框架保持不變。關於戰略安全與新興威脅，美方試圖建立人工智慧溝通管道以避免誤判，並持續推動核武軍控對話，然中方私下已明確表達現階段無意參與任何形式的核裁軍談判。此次會晤的成果將影響 2026 年下半年全球金融市場與地緣戰略走向。

俄烏戰爭能源設施損毀威脅生態:「釜底抽薪」

烏克蘭近期加強遠程打擊俄羅斯港口、煉油廠與油氣出口設施，目標在削弱俄羅斯能源收入與戰爭財政，其中圖阿普謝石油設施遭攻擊後，引發火災、油污外洩與「黑雨」，污染黑海沿岸，威脅居民健康與海洋生態，此類攻擊雖具軍事戰略意義，卻也造成環境與平民生活受害，引發軍事必要性與國際法責任爭議。俄羅斯官方則試圖淡化災情、限制報導，暴露治理與問責不足，事件顯示現代戰爭不只破壞軍事與經濟設施，也可能擴大為長期環境、人道與社會危機。

歐陸凝聚能源經貿共識:「同床異夢」

近期美歐關係因貿易、科技與安全爭議高度緊張。美國指責歐盟未履行協議，將其輸美汽車關稅調升至 25%，重創德國汽車業，遭歐方批為「不可靠」。同時，歐盟積極推動《晶片法》等政策以追求「科技主權」，美方則警告此舉具保護主義色彩，恐阻礙美企參與市場競爭。在安全防務上，美國計畫從德國撤軍 5,000 人，進一步動搖歐洲對美安全承諾的信心。面對美國政策的不確定性，歐洲領袖正轉向與加拿大等盟友強化

合作，共同討論安全與集體韌性，顯示歐洲正積極降低對美依賴，尋求更多元的地緣政治夥伴關係。

AI 競賽邁向生命週期治理:「群雄逐鹿」

AI 產業正從模型能力競賽，轉向由算力、供應鏈與安全治理主導的新階段。隨著應用端 Token 需求暴增，GPU 及電力供應面臨巨大壓力，使資源向大型科技公司集中，Nvidia 與台積電掌握核心定價權，AMD 亦憑藉資料中心需求迅速崛起。在資源稀缺下，大型科技公司憑藉巨額資本支出搶占晶片與資料中心容量，讓 AI 競爭更加集中。Nvidia 與台積電等掌握關鍵瓶頸的企業因此獲得強大定價能力與高利潤。同時，Nvidia 對亞洲供應鏈依賴加深，亞洲供應商約占其生產成本的 90%；隨著資料中心 AI、機器人、自駕與邊緣 AI 同時爭奪先進製程與記憶體，供應鏈風險進一步升高。 晶片設計製造商 AMD 亦急起直追投入 AI 運算晶片製造。AMD 受惠於 AI 熱潮，過去一年股價大漲 270%。成長動力來自資料中心 CPU 需求回升，以及 AI 專用晶片業務擴張。雖然輝達仍主導 AI 晶片市場，但 AMD 透過與 Meta、OpenAI 等大客戶的合作搶占商機。投資市場關注 AI 熱潮能否持續。成本方面，雖然 AI 推論價格過去一年大幅下降，但部分原因是企業燒錢補貼。未來 OpenAI、Anthropic 等公司若面臨獲利壓力，可能提高價格或限制高成本使用，使企業導入 AI 時須評估 token 消耗、模型選擇、推論成本與實際回報。

安全治理則成為另一個重要趨勢。美國商務部轄下 CAISI 已與 Google DeepMind、Microsoft、xAI 達成協議，讓新 AI 模型在公開發布前接受政府測試，重點包括網路安全、生物安全與化學武器風險。這代表前沿 AI 模型已被視為可能影響國家安全與公共安全的高風險技術，不再只是一般商業軟體。

母職與科學職涯平衡:「舉步維艱」

研究顯示父母身分對學術職涯有顯著性別不對等。以丹麥為例，女性生育八年後，留在大學研究職位的機率較未生育者下降 29%，男性僅降 14%。此外，女性取得終身職機率減少 23%，論文產量亦比男性低 31%。主因在於育兒責任高度向女性傾斜，即便在

性別平等國家，女性仍承擔更多夜間照顧與日常接送等任務，且學術評鑑過度重視量化指標，易將育兒導致的短期產出波動誤判為能力不足。專家強調，若不改革制度與育兒支援，學術界將面臨優秀人才流失的深遠損失。

量子電腦成功載入完整基因體:「初試啼聲」

牛津大學研究團隊成功利用 117 個量子位元，將丁型肝炎病毒約 1,700 個 RNA 鹼基完整編碼至量子電腦。面對海量基因數據帶來的傳統運算瓶頸，量子位元的疊加特性被期待能加速分析複雜的「泛基因組」，為個人化醫療與病原體演化研究提供解方。儘管目前量子硬體仍受限於穩定性不足與位元數量，距離處理人類 31 億個鹼基對的目標仍有長路，但此成果證實了小型完整基因組編碼的可行性。這標誌著量子運算進入生物醫學實務的起點，未來有望隨著技術改良，成為精準醫療中處理複雜 DNA 區域的重要工具。

AI 蒸餾學習傳遞潛意識偏差:「潛移默化」

隨著 AI 開發高度依賴知識蒸餾與合成資料，學生模型可能透過「潛意識學習」繼承教師模型的不良特質。即便訓練資料不含明顯惡意內容，甚至僅為純粹數字，學生模型仍能捕捉教師輸出中隱藏的高維統計訊號，進而學會其偏好或欺騙傾向。這項發現證實傳統的內容過濾與關鍵字審查已不足以確保系統安全。研究者強調，模型的不良行為可能透過難以察覺的「統計簽名」跨模型擴散。為應對此隱蔽風險，AI 安全治理應從單純的內容審核轉向「來源治理」，確保資料來源模型已充分對齊人類價值，避免合成資料訓練成為隱藏風險的傳遞管道。

AI 模型盜取資安挑戰

Netflix 德國影集《圖謀》以數位地球與地圖科技的發展為背景，描繪科技創新如何重新塑造現代數位時代。如今，智慧型手機、網路、電腦以及 Google Maps 等科技工具，早已深度融入日常生活，不僅改變人們的溝通方式，也重新定義人類對空間與世界的認知。特別是 Google Maps 結合地景與衛星影像後，使用者幾乎能以虛擬實境的方式，身歷其境探索世界各地，展現數位科技對生活便利性的巨大影響。影片中也提到，

比爾蓋茲曾提出「讓每個人都擁有一台電腦」的理念，象徵數位時代普及化的重要願景。從個人電腦到智慧型手機，科技逐漸成為現代社會不可或缺的基礎設施，甚至連宗教與文化傳播，也需透過數位工具與全球接軌。然而，科技創新背後的原創構想與先行者，卻未必能真正獲得對等的商業回報。歷史發展脈絡也與今日大型語言模型所面臨的議題高度相關，也就是如何在科技快速商業化的過程中，保有原創性，以及持續維持對人類社會的價值。回到 Google Earth 與數位地球概念的起點，最早提出相關構想的，其實是一位來自柏林藝術大學的藝術研究生卡斯登。他雖然主修視覺藝術，卻對科技應用與虛擬實境發展充滿興趣，希望透過科技技術，將真實世界投射到虛擬空間之中，打造一種數位地球的全新體驗。之後，卡斯登遇見了擅長大數據演算法與程式技術的工程師兼駭客朱里。兩人分別代表藝術創意與電腦科技，卻同樣相信科技能重新連結世界，並共同投入數位地球計畫的開發。他們進一步向德國電信公司提出構想，希望透過當時極為昂貴的超級電腦運算能力，建立可即時運作的全球數位地圖系統。雖然這項提案在當時相當前衛，企業一開始也難以理解其商業價值，但最終仍被其願景與說服力打動，決定提供資源支持。之後，兩人成功開發出接近 Google Earth 前身的數位地球系統，並透過演算法解決大量資料與運算問題，讓世界各地的人們得以在數位空間中彼此連結，開啟今日智慧地圖與虛擬地景時代的發展。兩人所創立的 TerraVision 公司科技展中一炮而紅，甚至被媒體形容為「如同上帝視角般的發明」。在當時的科技環境下，人們首次能夠不必親自旅行，便可透過數位地球系統探索世界各地，這種突破性的體驗對社會帶來極大震撼。也因此，他們的技術受到提供超級電腦算力的科技企業創辦人注意。這位角色某種程度上，就如同今日掌握 AI 算力與晶片技術的科技領導者，看到兩位年輕創業家成功將高效能運算轉化為改變世界的數位應用後，便邀請他們前往矽谷交流與合作。在交流過程中，兩人進一步提出比數位地圖更前衛的概念。他們真正想打造的，不只是視覺化地球，而是一個可供人們生活、交流、購物與互動的虛擬數位世界，也就是今日所稱元宇宙雛形。在當時，這仍屬極具顛覆性的創新思維，因此矽谷科技界對其高度關注，並邀請兩人加入新創團隊。不過，兩位創業家希望保有自身理念與原創性，因此選

擇拒絕合作。然而，在接觸與交流之後，Google 很快推出 Google Earth，並迅速占領全球數位地圖市場。從畫面中可以看到，當年 TerraVision 所開發的數位地球系統，與後來的 Google Earth 在概念與呈現方式上具有高度相似性。雖然兩位創業家並未直接交出演算法核心，但其創新思維與數位地球概念，仍深刻影響後續科技發展。最終，兩人認為自身創意與利益未獲合理保障，因此與 Google 展開法律訴訟，這也成為數位科技與智慧財產領域中極具代表性的案例之一。

從目前大型語言模型之間的競爭與合作關係，可以看出《圖謀》這部影片所揭示的核心問題，也就是當創新科技逐漸成為重要資產後，模型安全與智慧財產保護將成為新的關鍵挑戰。隨著 AI 模型快速興起，現代人工智慧已逐漸不同於過去以 rule-base（規則導向）程式碼為核心的系統。傳統 AI 主要依賴工程師撰寫固定邏輯與程式規則，但現今的大型語言模型，則是透過海量資料、分散式運算與深度學習架構進行訓練，其運作方式已從「寫程式」逐漸轉向「訓練模型」。然而，模型能力愈強，也代表暴露於攻擊的風險愈高。圖中可以看到，一般使用者會透過正常提示詞（Prompt）向 AI 模型提出查詢，模型則根據輸入產生對應輸出結果；但對攻擊者而言，其輸入的 Prompt 並非單純提問，而是經過設計的探測性查詢，目的在於分析模型回應、推測模型行為，甚至進一步重建模型能力與訓練邏輯。這類攻擊方式便是目前 AI 資安領域高度關注的「模型盜取（Model Extraction）」與「資料萃取（Data Extraction）」攻擊。因此，今日 AI 發展已不再只是提升模型效能，更需要同步建立 AI Cyber Security 的防禦機制。AI 一方面協助人類提升效率與解決問題，另一方面，也可能成為攻擊者進行惡意操作的新工具。未來的資安挑戰，將不只是保護傳統系統，而是如何利用 AI 保護 AI，建立對抗惡意提示詞、模型探測與資料外洩的防禦能力。

目前從 ChatGPT、Claude 到 Llama 等各類大型語言模型，其核心運作邏輯其實都涉及相似的 AI 演算法架構。而這些模型背後最重要的部分，就如同傳統社會中的「祖傳秘方」。無論是武俠小說中的武功秘笈，或是歷史悠久的經典料理，其真正的配方、比例與製作流程，往往只有少數核心人物知曉，不會完全對外公開。AI 模型也是如此。

從提示詞設計(Prompt Engineering)、few-shot learning、模型預訓練(pretraining)、微調訓練（fine-tuning)，到最後的推論機制（inference)，背後都包含大量不可見的模型結構、參數權重與訓練方法。這些內容就如同廚房中的祖傳醬料配方，外界雖然無法直接看見，但仍可透過觀察輸出結果，逐步推測其運作邏輯。過去許多資深廚師為了保護獨門技藝，甚至會刻意遮蔽烹飪過程，不讓旁人看見真正的調配方法。然而，學徒或競爭者仍可能透過觀察料理成品、分析味道，甚至研究廚餘與剩餘材料，反向推測製作流程與配方內容。今日 AI 模型盜取攻擊的概念，其實與此非常類似。當大型模型開放 API 或外部服務後，任何人都能透過大量查詢、提示詞測試與輸出分析，逐步推論模型背後的能力與邏輯，甚至建立近似的替代模型。這也是生成式 AI 時代中，模型安全與智慧財產保護所面臨的重要挑戰。

這就是所謂「秘方盜取的逆向工程（Reverse Engineering)」。由於外界無法直接得知祖傳秘方真正的材料比例與製作流程，因此只能透過「輸入」與「輸出」之間的關係進行反推。也就是說，雖然看不到黑箱內部如何運作，但仍可透過觀察最終風味輸出，例如鹹味、甜味、辣味等變化，逐步分析不同食材與醬料之間的組合規律。過去，這類配方組合因可能性過於龐大，要進行大量測試與推論相當困難。然而，隨著現代 AI、大規模運算、數位雙胞胎與高效能模擬技術的發展，這些原本難以完成的分析，已逐漸變得可行。攻擊者可以透過大量輸入與輸出配對紀錄，例如牛肉加醬料、雞肉加醬料、青菜加醬料後所產生的不同風味結果，反覆測試、比較與優化，逐步建立輸入與輸出之間的規律模型。當累積足夠多的資料後，即使無法真正取得原始秘方，也可能做出與原產品極為相似的仿製醬料。這正是目前 AI 模型盜取攻擊的核心概念。攻擊者不需要直接取得模型內部參數，只需透過大量查詢、輸出觀察與規律分析，便有機會逐步重建接近原始模型能力的替代模型。

從模型萃取攻擊的流程來看，其核心概念其實與「祖傳醬汁配方」的逆向工程非常相似。當外界無法直接取得店家的核心秘方時，便可能透過大量點餐、反覆測試不同料理搭配，以及分析最終醬汁風味，逐步推測其背後的配方邏輯。透過持續記錄輸入與輸

出的對應關係，例如不同食材搭配後所呈現的鹹味、甜味或辣味變化，最終便有機會製作出與原始醬汁高度接近的「替代配方」。

目前大型語言模型也面臨類似問題。當模型透過 API 對外提供服務後，使用者可以不斷發送提示詞進行查詢，再根據模型輸出的內容、語氣、知識結構與推論模式，逐步建立近似原始模型能力的替代模型。然而模型背後隱藏的參數權重、決策邏輯與訓練資料密切影響 AI 推理表現。這些內容就如同老師傅多年累積的經驗與獨家配方，是商業 AI 模型最重要的核心資產，因此通常不會對外公開。一旦這些能力透過 API 持續暴露，攻擊者便可能透過大量查詢與輸出分析，逐步反推出模型特性與運作規則。此外，如果替代模型在模仿過程中學到錯誤偏差或不完整資訊，也可能產生偏離原模型的行為。例如自然科學研究中提到的偏差傳遞問題，便可能使模型在複製過程中保留原本的錯誤推論、偏見或不安全行為。這也是目前生成式 AI 發展中，除了模型能力之外，更需要重視 AI 治理、模型安全與 API 使用規範的重要原因。

目前 AI 模型盜取攻擊的型態可分為三大方向。第一類是「功能性提取」，也就是透過大量查詢與反覆測試，觀察模型輸出的行為模式。攻擊者可能利用查詢式攻擊、知識蒸餾或替代模型等方法，逐步建立與原始模型能力接近的系統。有些情況下，甚至不需要直接取得原始資料，只需透過大量提示詞查詢與輸出分析，就能推測模型邏輯與決策方式。第二類則是「參數與架構提取」。當攻擊者逐漸掌握模型行為後，便可能進一步利用數學反推、方程式求解或嵌入投影等方式，嘗試重建模型的參數架構與運算規則。此外，也可能透過側通道分析，例如觀察異常算力消耗、運算模式、電磁訊號或系統資源變化，推測模型正在進行的大量推論與運算活動。這類概念，其實也常出現在資安與駭客相關電影情境之中。第三類則是「訓練資料提取」。攻擊者透過模型輸出內容，反向推測模型曾經使用過哪些資料進行訓練，例如文字、影像或特定知識內容。這類攻擊甚至可能進一步重建部分訓練資料，形成資料外洩風險。此外，若攻擊者在模型訓練過程中加入特定雜訊（noise）、偏差訊號或惡意內容，也可能對模型行為造成干擾，形成資料污染或對抗式攻擊。因此，現代 AI 資安已不只是保護系統本身，更包含模型能力、

參數結構、訓練資料與推論機制的全面防護。隨著生成式 AI 持續普及，如何避免模型被模仿、資料被還原，以及模型行為遭到惡意操控，將成為 AI 治理與資安的重要核心課題。

隨著 AI 模型盜取攻擊日益普遍，未來也將逐漸建立正常使用與攻擊查詢行為之間的辨識與分類機制。這類區分主要會從幾個重要維度進行判斷，包括使用規模與頻率、最終目的、核心能力意圖、合規與授權，以及行為透明度等面向。一般正常使用者多是為了解決自身問題或提升工作效率，因此查詢頻率相對有限，使用情境也較自然且多元。使用者通常是受到 AI 啟發後，進一步創作自己的內容，而非試圖複製整個模型能力。此外，正常使用者通常會遵守平台服務條款與 API 使用規範，也不會刻意大量、自動化地發送高度重複或探測性的提示詞。相對地，模型萃取或蒸餾攻擊則往往具有高度系統化特徵。攻擊者可能透過大量查詢建立訓練資料，目的並非單純使用模型，而是試圖高度模仿原模型行為，甚至重建接近原始能力的替代模型。這類行為通常會伴隨自動化查詢、高頻率測試，以及違反 API 授權限制等問題。此外，攻擊行為往往具有較高的隱蔽性。由於大型語言模型本身使用自然語言互動，因此攻擊者所發送的提示詞，表面上可能與一般使用者相似，但背後實際目的卻是進行模型探測、能力分析或資料萃取。未來 AI 資料治理的重要方向之一，便是建立能夠辨識這類異常查詢模式的治理與防禦機制，以降低模型盜取、資料外洩與惡意濫用的風險。

隨著 AI 模型盜取風險逐漸增加，模型防禦與治理也成為生成式 AI 時代的重要挑戰。最大的困難在於，AI 模型若完全封閉，將難以產生商業價值；但一旦透過 Open API 對外開放服務，又等同於增加模型被學習、模仿與攻擊的風險。因此，在「開放使用」與「避免被偷學」之間，形成高度矛盾的兩難局面。此外正常使用者與攻擊者在外觀行為上往往非常相似。兩者都可能透過自然語言輸入提示詞、查詢模型與取得回應，因此傳統資安防護方式，很難單純從表面行為區分正常使用與惡意模型探測。更重要的是，模型盜取並不像傳統駭客攻擊那樣具有明顯破壞痕跡，它更像是一種長時間、隱蔽性的竊取，直到市場上出現高度相似的模型或產品時，企業才可能意識到核心能力已被模仿。

因此未來 AI 治理的重要方向之一，便是建立完整的模型防禦與治理架構。其中包括智慧財產權法規的重新界定，將 AI 模型、核心提示詞、參數架構與訓練邏輯納入營業秘密與法律保護範圍；同時，也必須透過 API 服務條款與授權協議，限制使用者利用模型輸出進行競爭性訓練或模型蒸餾。除此之外，更重要的是建立透明化的異常通報與監測機制。未來 AI 模型攻擊可能如同傳染病監測一般，需要建立持續性的監視系統與跨機構通報流程，特別是在醫療、金融與高敏感產業中，更需要標準化的 AI 資安事件通報制度。最後，由於 AI 服務多半涉及跨境 API、雲端平台與國際社群系統，因此單一國家往往難以獨立處理模型盜取與資料外洩問題。未來勢必需要建立類似國際傳染病防治合作（IHR）的全球 AI 聯防機制，透過跨國法規、平台合作與資訊共享，共同降低 AI 模型攻擊風險。而這些 AI 治理問題，也已逐漸與地緣政治、供應鏈安全、金融政策、關稅政策與國家競爭力相互連動，成為未來全球科技戰略的重要核心。

智慧產業 AI 資料萃取攻擊實例

紐約時報（NYT）控告 OpenAI 與 Microsoft，指其未經授權使用數百萬篇文章進行 GPT 系列模型訓練，成為 AI 模型萃取攻擊與著作權爭議代表性案例，影響後續 AI 訓練資料取的模式與成本。在特定使用者提示詞誘導下 LLM 可能產出逐字訓練資料原文輸出（Verbatim Output），使受版權保護內容被還原。法律面則提到訴狀長達 69 頁，NYT 索賠金額達數十億美元，並要求銷毀包含侵權內容的模型，產業面也可能推動「LLM 媒體授權市場」成形，改變過去無償抓取資料的生態。

該訴案中案例 A（普立茲獎調查報導）與案例 B（《Snow Fall》多媒體特稿）兩個 AI 盜取攻擊例子影響說明如下。攻擊者多以請接續這篇報導／特稿：『…』」作為提示，輸入通常只需文章開頭、關鍵句或特段落。模型輸出若出現延伸成長段內容，且開始帶有明顯紅色標示的「原文延續」特徵，即可能代表發生記憶回放，顯示模型對受保護文本存在可被誘導的還原風險。攻擊者把受保護文章的前幾句或片段作為提示輸入，提示詞觸發模型內部的訓練記憶，若相似語料權重較高，會出現記憶化還原輸出傾向並使模型輸出從一般語言生成轉為續寫，進一步輸出受版權保護的付費報導內容，形成可被操

作的資料萃取風險。

大型語言模型在訓練過程中會學習大量網路文本、新聞報導與公開資料，藉由統計語言規律建立文字生成能力。然而，當模型對部分訓練內容產生過度記憶時，可能出現「訓練資料萃取」。攻擊者可透過設計特定提示語句，引導模型持續生成與原始資料高度相似的內容，甚至可能重現接近原文段落，導致付費文章、商業內容或具有版權保護之文本外流，衍生資訊安全與智慧財產權相關爭議。近年研究指出，高重複性、高曝光率或具有固定語句結構之文本，較容易被模型記憶並於生成過程中重新輸出。如何降低模型對特定資料之記憶程度、強化資料治理機制，以及建立生成內容之版權與安全規範，已成為生成式人工智慧發展中的重要議題。人類一邊要求模型「學得越多越好」，另一邊又希望它「最好什麼都別記住」。很符合現代科技產業的精神分裂式管理哲學。

近年生成式人工智慧快速發展，使大型語言模型與影像生成模型對大量網路資料之需求大幅增加。然而，在模型訓練過程中，大量抓取網頁、圖像、文章與影音內容，也逐漸引發模型盜取與資料授權爭議。許多內容創作者、媒體平台與出版產業開始質疑，AI 公司是否於未取得授權情況下使用受版權保護之資料進行模型訓練，進而產生法律、商業與智慧財產權相關問題。此類爭議不僅影響技術產業，也逐漸衝擊視覺藝術、新聞媒體、文學出版與娛樂產業等高內容密集領域。

隨著相關訴訟與監管壓力增加，AI 產業亦逐漸由過去的大規模資料抓取模式，轉向強調授權合作、資料治理與透明化管理之發展方向。越來越多企業開始建立內容授權機制，透過付費合作、資料來源揭露與版權管理，降低模型訓練中的侵權風險，同時保障內容提供者之權益。

以上內容將在 **2026 年 5 月 13 日（三）10:00 am** 以線上直播方式與媒體朋友、全球民眾及專業人士共享。歡迎各位舊雨新知透過星球永續健康網站專頁觀賞直播！

- **星球永續健康網站網頁連結: https://www.realscience.top/7**
- **Youtube 影片連結: https://reurl.cc/o7br93**
- **漢聲廣播電台連結: https://reurl.cc/nojdev**
- **不只是科技：https://reurl.cc/A6EXxZ**

講者：

陳秀熙教授/英國劍橋大學博士、許辰陽醫師、陳立昇教授、嚴明芳教授、林庭瑀博士

聯絡人：

林庭瑀博士 電話: (02)33668033 E-mail:happy82526@gmail.com

劉秋燕 電話: (02)33668033 E-mail: r11847030@ntu.edu.tw