星球永續健康線上直播

智慧數位資安 (8)

蒸餾攻擊智慧模型盜取

2026 年 5 月 20 日

人工智慧模型已成為企業與國家重要核心資產，但伴隨而來的模型盜取（Model Theft）與知識蒸餾攻擊（Distillation Attack）風險，也逐漸成為新型態資安挑戰。攻擊者可透過大量 API 查詢、輸入輸出蒐集與知識蒸餾技術，在不直接取得原始模型權重情況下，重建具備相似核心能力的替代模型，使 AI 資安延伸至智慧財產、產業治理與國際監管等層面。本週我們將探討 AI 模型盜取的資安挑戰，以及蒸餾攻擊下的 AI 模型盜取防禦實例。

健康科學新知

川習會亞洲佈局牽動經貿 AI 地緣：「鬥而不破」

川普與習近平北京峰會的核心為兩國領袖首次會晤，涵蓋貿易、科技、台灣、中東議題、能源與區域安全的高層風險控管。美方表達希望台灣與中國冷靜下來，避免衝突與戰爭。習近平在會談中強調台灣問題是中美關係中最重要的議題，若處理不當可能導致衝突。川普則表示，他不認為美中會因台灣爆發戰爭，但尚未承諾美國是否會防衛台灣。台灣方面則回應仍需釐清川普言論的確切意涵並強調美國對台軍售是依據美國法律，也是維持區域和平與穩定的重要基礎。

會前美國財政部長 貝特森先赴日本，再轉往南韓與中國副總理何立峰會談，為川習會預先鋪陳經貿談判基礎。此次會談的重要成果之一是美中在貿易議題上取得一定程度共識，尤其聚焦於中國增加採購美國商品。會後中國表示將規劃擴大進口美國波音客機、黃豆與牛肉等產品，回應川普政府對縮減貿易逆差與創造美國出口訂單。美中在北京峰會後同意建立貿易委員會與投資委員會，並在農產品市場准入、對等降稅框架下擴大雙邊貿易方面達成共識。中方官方尚未公開具體農產品採購清單，也未提及美方所稱的波音飛機採購與美國原油進口安排。中國外長王毅確認習近平將於今年秋天應川普邀

請對美國進行國事訪問，預計 9 月會面時可能進一步討論半導體、稀土等尚未解決的正常貿易關係議題。中國方面表示台灣議題仍是雙方主要分歧。王毅表示美方理解中國不接受台獨，但川普返程時稱雙方已詳細討論對台軍售，雙方說法差異顯示台灣問題仍可能成為美中關係的衝突點。

就中東局勢美國亦提出要求中國共同支持能源航運安全。川普表示中國國家主席習近平同意伊朗必須重開荷姆茲海峽，但中國方面尚未明確表示會介入施壓。川普強調，美國不希望伊朗擁有核武，也希望海峽保持開放，伊朗則表示願意談判但不信任美國，在此情境下中國也被伊朗視為可能提供外交協助的一方。王毅稱中方主張透過對話解決問題，並希望美伊核談判盡快恢復。取得中國協助對於中東局勢穩定將有所幫助。對俄烏戰爭方面中美皆希望戰事早日結束。此次美國訪中企業代表團的加入凸顯本次峰會的交易性質。川普邀請包括 Tesla 的 Elon Musk、Apple 的 Tim Cook、BlackRock 的 Larry Fink、Boeing 的 Kelly Ortberg 等美國大型企業高層隨行，顯示白宮希望將外交談判與商業利益結合，透過航空、農產品、科技與金融等領域的合作，為美中關係降溫創造實質交換空間。人工智慧、出口管制、稀土供應、台灣問題與伊朗戰爭仍是峰會中的敏感議題。特別是台灣議題，川普會後警告台灣不要正式宣布獨立，同時表示美中領袖已「大量」討論台灣問題，但並未明確承諾美國是否會防衛台灣。這反映出美方仍試圖在支持台灣安全與避免刺激北京之間維持模糊平衡。

全球能源市場復原路遙:「油震全球」

霍爾木茲海峽封鎖已成全球能源與地緣政治風險核心，美以與伊朗衝突升級使油輪通行大減，市場每週損失約一億桶石油供應，推升油價與通膨壓力，若中斷延續，全球油市恢復恐延至 2027 年，亞洲因高度依賴波斯灣石油首當其衝，中國原油進口下滑，並優先保障國內燃料供應。政治上，美國考慮恢復海軍護航，伊朗則要求停戰、解除制裁與承認其海峽主權，談判前景仍不明。沙烏地阿美執行長 Amin Nasser 警告，若供應中斷延續至六月中旬，全球油市的再平衡恐怕將延後至 2027 年，即使海峽立即重開，市場仍需要數月時間才能恢復穩定。這反映出此次危機不只是短期價格波動，而是對全

球能源供應鏈、航運秩序與通膨壓力造成結構性衝擊。，危機已從區域衝突擴大為全球能源、航運、金融與外交秩序的系統性風險

能源危機與肥料短缺:「糧能相扣」

近期中東戰爭與荷莫茲海峽危機衝擊能源與航運，帶動肥料原料運輸受阻，尿素價格一個月上漲近 46%，再度凸顯糧食系統高度依賴能源供應，由於氮肥生產仰賴天然氣，能源價格上升會迅速推高肥料成本，進而影響農民施肥、作物產量與糧價，特別衝擊依賴進口肥料的非洲、南亞與拉丁美洲地區。文章指出，肥料應被視為糧食安全的戰略基礎設施，而非一般商品，未來需建立肥料儲備與供應監測，推動綠氨、精準施肥、貿易協調與農民金融支持，避免能源危機反覆演變成全球糧食危機。

健康迷思擴散挑戰科學信任:「眾聲奪真」

Edelman Trust Institute 一項涵蓋 16 國、逾 16,000 人的調查顯示，約 70%受訪者至少相信一項缺乏科學證據的健康主張，例如疫苗風險大於好處、飲水加氟有害、生乳較健康，或孕期服用撲熱息痛會導致自閉症。科學文章指出這並非單純源於教育不足，因為相信錯誤健康資訊者也可能受過高等教育並常接觸健康新聞，真正問題在於社群媒體、新聞與個人經驗造成資訊過量且互相矛盾，使大眾難以判斷可信來源，科學界若要重建信任，需用更清楚、易懂且貼近社群平台的方式溝通。

智慧資安技術監管角力:「模型爭權」

OpenAI 近期主動向歐盟開放 GPT-5.5-Cyber 存取權，允許政府、資安單位及防禦團隊進行漏洞偵測、惡意程式分析與逆向工程。歐盟執委會對此表示歡迎，認為此舉展現透明度，有助於監管機關直接掌握高風險模型的部署情況與安全風險。相較之下，Anthropic 的 Claude Mythos 模型雖被證實具備強大的網路攻擊模擬能力，但目前尚未對歐盟機構提供直接存取權，引發歐洲對於關鍵基礎設施防禦能力的焦慮。在「以 AI 對抗 AI」的資安新思維下，模型存取權已轉化為新型外交籌碼。專家分析，這場 OpenAI 與 Anthropic 的差異化策略，顯示 AI 治理已進入技術、監管與政治交錯的新階段，未來模型開放程度將直接影響企業的市場信任、政府關係與全球競爭位置。

AI 時代數位鑑識革命:「破假尋真」

隨著 AI 生成影像與深偽技術氾濫，數位鑑識專家 Hany Farid 致力於在「無人相信真相」的時代建立防線。針對疑似飛彈襲擊的爭議影片，Farid 不僅逐格分析煙霧與火焰，更透過檢查飛彈軌跡的物理預期及聲音延遲與距離的吻合度，科學性地判定影片真偽。Farid 指出，儘管 AI 生成器已能模仿感測器雜訊，但在物理、幾何、陰影與反射上仍常露出馬腳。他強調，單純依賴「黑盒」式的 AI 偵測器往往缺乏可解釋的證據鏈，因此他主張回歸真實世界的物理一致性作為判斷核心。面對每日海量的鑑定請求，Farid 將其工作比喻為替社會「上鎖」，雖然無法阻止所有造假，但能顯著提高惡意行為者的門檻，為數位資訊的真實性築起後一道防線。

AI 顛覆科研審查制度:「機審機寫」

AI 代理應用工具不僅能潤稿，更能根據履歷與標準自動產生構想，並撰寫邏輯嚴密、格式精準的申請書，導致 2022 至 2025 年間部分機構的申請量大幅上升甚至翻倍。專家警告，當所有提案都經 AI 優化至高度精緻且差異極小時，審查委員將難以辨識真正的科學原創性，恐導致制度陷入「集體失靈」。面對禁令難以執行的困境，文章建議資助機構應重新設計分配方式，轉向「AI 原生」制度，將評估重點從書面提案移往研究者的長期表現、面試或團隊作品集。雖然這可能帶來資源過度集中等風險，但透過透明且可稽核的 AI 輔助篩選，有望減輕審查負荷並減少資源浪費，建立一個更公平且能將時間還給研究工作的資助系統。

AI 模型盜取資安挑戰

《天才雷普利》中的雷普利原本只是出身平凡的年輕人，卻十分嚮往上流社會的生活。電影片名中的「Talent」，指的正是他極強的模仿能力。影片一開始便出現一句經典台詞：「我寧願做一個冒牌的某人，也不願當一個真實的無名小卒。」雷普利雖然家境普通，但因在劇院工作，得以接觸上流社會人士。某次在一場上流階層聚集的音樂會中，他穿著借來的普林斯頓大學外套擔任鋼琴伴奏，因而被造船富商格林利夫誤認為其兒子迪基的大學同學。格林利夫得知雷普利「認識」迪基後，便委託他前往義大利，希

望將長期沉迷奢華生活、不願返家的兒子帶回美國，並承諾提供酬勞。抵達義大利後，雷普利認識了迪基以及其女友瑪姬，也逐漸融入他們的生活圈。迪基享受自由、富裕且無拘束的生活，完全不願返回家中，而雷普利則愈發嚮往這樣的生活方式。隨著相處時間增加，雷普利開始模仿迪基的穿著、動作、說話方式與簽名，從單純的崇拜逐漸演變成對其身份與生活的強烈投射與佔有慾望。最終，兩人在衝突中發生悲劇，雷普利失手殺害迪基，並利用自己長期觀察與模仿累積下來的細節與習慣，冒充迪基的身份，進一步進入原本不屬於自己的上流社會。電影透過雷普利的模仿能力與雙重身份轉換，呈現出人性、慾望與身份認同之間複雜而深刻的關係，也與人工智慧模型盜取中透過模仿、學習與重建核心能力的概念形成相互呼應。

在 AI 模型盜取攻擊中，模型萃取（Model Extraction）是目前常見的重要攻擊型態之一。AI 模型盜取包含多種不同方式，而知識蒸餾與蒸餾攻擊，主要著重於功能性提取以及查詢式／蒸餾式攻擊等層面。其核心概念在於，攻擊者並非直接竊取原始模型，而是透過大量查詢、反覆探索與輸入輸出蒐集，逐步學習模型的回應邏輯與行為模式，進而建立具備相似能力的替代模型。在大型語言模型時代，AI 系統可透過文字、語音與影像提示與使用者互動，原本設計目的在於提升互動與回應能力。然而，若持續進行大規模查詢與探索，便會形成所謂的提示工程。這類查詢的目的，已不只是取得單次答案，而是希望透過查詢過程，逐步盜取模型中的知識、推論方式與認知能力。過程中可能結合參數求解、推論分析以及 Chain-of-Thought(CoT)推論等方式，重建模型知識與替代模型，而這整套模式也被稱為 AI 的認知作戰。因此，現今 AI 模型盜取的核心，已由傳統資安資產傳統程式碼或原始碼竊取，擴展為針對 AI 從提示到產生答案過程中的認知能力與決策邏輯進行學習與重建，這也是目前 AI 模型盜取攻擊的重要特徵。

知識蒸餾原本是 AI 模型訓練中的正常技術，其概念類似老師與學生之間的知識傳承。老師模型通常具備豐富知識、多元策略以及完整模型能力，而學生模型因能力與規模有限，無法在短時間內完全吸收所有內容，因此會透過知識萃取的方式，逐步保留老師模型中的核心能力與重要知識。這個過程就像蒸餾技術一樣，透過過濾與萃取，留下

真正有價值的精華部分，形成較小但仍具備核心能力的替代模型。原本知識蒸餾是為了提升 AI 模型訓練效率與部署能力，例如讓學生模型能快速學習老師模型在提示下所產生的回應與推論能力。然而，若這項技術被駭客利用，便可能演變成所謂的蒸餾攻擊。攻擊者會透過大量查詢黑箱模型，持續送出特定提示（X’），再從 API 回傳結果中取得模型輸出 M(X’)。當大量輸入與輸出之間建立起對應關係後，便能逐步訓練出接近原模型能力的替代模型，也就是所謂的學生模型。這類學生模型雖然規模較小，但透過損失函式與持續訓練，其回應效果可能與原始模型高度接近。因此，蒸餾攻擊的核心，並不一定是直接偷取模型參數，而是透過知識萃取與功能模仿，重建可實際運作的替代模型。這也使得大型 AI 模型的能力，可能被低成本、大規模地複製與部署，而這正是智慧模型蒸餾攻擊的重要特徵。

智慧模型的蒸餾攻擊如同補習班名師教學模式。補習班中的王牌老師，通常擁有多年教學經驗，也有自己獨創的解題方法與教學策略。這些老師往往採取小班教學，甚至禁止錄影錄音，目的就是避免自己的教學內容與核心方法被大量複製。其實就像目前的頂尖 AI 模型，例如 GPT、Claude 或 Gemini。這些模型的開發成本極高，需要投入大量資源與時間訓練，因此具備非常強大的推理、分析與生成能力，也形成高度價值的知識資產。然而，一旦模型的核心能力被大量模仿與學習，就可能出現模型萃取攻擊。攻擊者未必直接取得原始模型本身，而是透過不斷查詢、觀察輸出結果與學習回應方式，逐步複製模型中的核心能力與解題邏輯，最終建立出具備相似功能的替代模型。這也是智慧模型蒸餾攻擊最重要的概念之一。

模型萃取攻擊就如同電影中雷普利透過長期模仿他人身份與生活方式，逐步建立出幾乎難以分辨的「替代身份」。在 AI 模型萃取中，攻擊者的目的，也是建立一個與原始模型極為接近、甚至真假難辨的替代模型。這類攻擊之所以重要，在於它不只是單純複製表面結果，而是試圖重建模型背後的邏輯、風格與推論能力，因此在商業、國防與科學研究等領域，都被視為重要的戰略風險。其攻擊方式通常包含多種手法。例如，先透過長期蒐集資料與觀察模型行為，再利用訪談、查詢與資料分析等方式取得更多線索，

接著進一步透過逆向推論，從結果反推模型的運作邏輯與思考模式。當輸入、解題方式與輸出結果逐漸建立完整對應後，便能發展出一套與原始模型高度相似的替代系統。這種攻擊不只是單純「複製答案」，而是連模型的推論邏輯、風格與解題方式都一併重建。因此，模型萃取攻擊所產生的替代模型，在規模與能力上往往與原始模型十分接近，甚至可能達到幾乎對等的程度。不過，傳統模型萃取的成本其實相當高，因為重建大型模型需要長時間蒐集資料、投入大量人力與技術資源，因此過去多半出現在商業機密、國防情報與高價值技術競爭領域。其核心概念，在於系統性重建原模型的「教學能力與風格」，也可視為對整體 AI 戰略知識的一種重新複製。

模仿一位名師或頂尖模型有許多不同的方法，而知識蒸餾就是其中一種代表性方式。它的核心概念在於萃取其最重要的「解題精華」與核心能力。攻擊者未必需要建立一個完全相同的替身模型，但藉由取得模型中關鍵解題能力複製模型成果。這也是為什麼知識蒸餾會被視為一種認知層面的攻擊模式。知識蒸餾攻擊的做法，通常會從長期蒐集資料開始，例如記錄大量輸入與輸出結果、整理模型回應中的重點與規律，再將與核心功能無關的內容排除，只保留最具價值的「解題方法」。這個過程就像蒸餾技術一樣，將真正的精華部分萃取出來，而將雜訊與不必要資訊過濾掉。之後，再透過較小型的模型或系統進行學習與模仿，最後包裝成具備相似能力的替代模型。因此，知識蒸餾攻擊最重要的特徵，在於它不一定追求完整複製原模型，而是希望以更低成本，重現原模型最核心的能力與價值。這樣的技術若被大量應用於商業、教育、科學研究甚至國防領域，便可能影響原創性、智慧財產權與商業機密，也因此成為目前 AI 資安領域中相當重要的一項風險。

整體而言，知識蒸餾所產生的小模型，雖然無法完全取代原始大型模型，但往往可以保留約 80%至 90%的核心能力。換句話說，一個原本具備完整能力的頂尖模型，可能被一個成本較低、規模較小的替代模型所模仿與取代。這也是目前教育、科學、商業與國防領域高度關注的原因。這類蒸餾後的小模型，規模通常遠小於原始模型，可能只有原本的十分之一，甚至百分之一，但仍能保留主要的解題與推論能力。不過，它通常缺

乏原始模型較完整的周邊能力與深層知識，就像補習班中的解題型老師，雖然能快速提供解題技巧與方法，但未必能完整涵蓋更深層的知識傳承與思維訓練。然而，這類蒸餾模型最大的特點，在於成本相對低廉。相較於傳統模型萃取需要投入大量人力、資料與工程成本，知識蒸餾往往只需要 API 查詢費用與部分算力，就能建立具備高度相似能力的小模型。因此，過去這類技術常被應用於商業套利、規避付費機制或繞過使用限制等用途。若進一步應用於大型 AI 模型，則可能對智慧財產權、商業利益與 AI 生態帶來更大衝擊

蒸餾攻擊所使用的技術應用，其核心做法，是透過大量收集提示與模型回應，持續呼叫目標模型 API，並將輸入與輸出結果完整儲存下來，形成訓練資料集。之後，再利用這些資料訓練小型學生模型，使其逐步學習原始大型模型的回應方式與推論能力。而這些學生模型學習的，不只是最終答案本身，有時還包括老師模型對不同答案的信心分布，也就是所謂的 soft target。這類資訊可能進一步涉及模型的推論邏輯與 CoT 能力，因此也成為目前大型模型公司高度關注的重點之一。當這些小型模型完成訓練後，便可以獨立部署與本地運行。由於模型規模較小，因此運算速度更快、算力需求更低、成本也相對較低。這也是目前許多小型 AI 模型能夠快速運行的重要原因之一。然而，這類技術原本屬於模型訓練與知識傳承的一部分，但若被大量用於重建大型模型能力，便可能從知識蒸餾逐漸演變為 AI 攻擊與模型盜取問題。因此，蒸餾技術與模型萃取之間的界線，也成為目前 AI 產業與資安領域高度關注的重要議題。

從模型萃取到知識蒸餾，其核心概念與「模仿」有關。模型萃取追求的是建立幾乎與原始模型相同的「分身」，因此往往需要多管齊下，透過大量資料蒐集、查詢攻擊、側通道分析、參數還原以及決策邊界探測等方式，全面重建原模型的行為與能力。這類攻擊的目的，是讓替代模型在規模、能力與表現上都盡可能接近原始模型，因此成本極高，通常涉及商業競爭、軍事用途與國家級機密等高價值領域。知識蒸餾則不追求完整複製原模型專注於萃取模型中核心知解題能力。其方式如同解題教學，透過大量收集提示問答、整理模型輸出與訓練小型模型，快速建立具備核心能力的替代模型。這類模型

雖然規模較小，但仍可保留約 80%至 90%的核心能力，同時具備部署快速、成本低廉與運算效率高等特性。因此，知識蒸餾較常被應用於商業套利、規避付費機制與繞過使用限制等情境。因此，模型萃取與知識蒸餾最大的差異，在於前者強調完整重建，後者則偏向核心能力萃取。兩者雖然技術路徑不同，但都反映出當前 AI 模型在知識、能力與智慧財產上的新型態資安挑戰。

蒸餾攻擊 AI 模型盜取防禦實例

Anthropic 在 2026/02/23 揭露的「MiniMax 蒸餾攻擊」事件：攻擊者以蒸餾（模型萃取）方式，試圖把 Claude 的能力轉成可用的訓練資料，並涉及 agentic coding、tool use 等能力場景。MiniMax 透過即時偵測與歸因，已完成蒐證，並揭露這起案例的規模指標包含 13,000,000+ 次互動、約 81% 的「三案」重疊比例，以及 24,000 個帳號；同時也與其他案例 （Moonshot AI、DeepSeek）做量級對比。這是少數由模型公司在攻擊進行中就「即時偵測、歸因並公開」的蒸餾攻擊案例。

系統化蒸餾攻擊技術第一步是透過商業代理/轉售（proxy/relay）把大量請求導向目標模型的商業 API，形成隱蔽且可規模化的資料取得管道；第二步利用 Hydra Cluster 以多帳號、多節點方式共享控制層（排程/路由/監控），讓分散式操作更穩定；第三步將 prompt 工業化，批次產生可誘發高品質輸出的任務，並同步收集「agent 與 tool-use」相關範例。最後，這些互動資料可被用於監督視微調訓練(SFT，蒐集高品質 instruction / completion）及強化回饋訓練(RL/RLAIF，以教師模型產生偏好或 reward 訊號）等用途，形成從資料萃取 → 訓練完整能力複製產業鏈。

在 Anthropic 發布新版本 Claude（圖示為 Claude 3.7 Sonnet）後，MiniMax 在 24 小時內就重新導流與重構模型，流量從舊模型（v1）快速導向新模型（v2），顯示其快速反應部署商業規模架構。這次蒐證顯示蒸餾攻擊可在極短時間完成資料收集—訓練—部署，藉由即時流量監測捕捉使用者請求樣態（含高價值 prompt、agent 任務與 tool-use），作為證據鏈一部分。證明非使用合約所准許查詢模式。

隨著生成式 AI 技術快速發展，模型蒸餾（Model Distillation）所衍生的資安與

智慧財產風險也逐漸受到重視。近期業界開始導入多層式蒸餾偵測機制，透過行為指紋、帳號監測、基礎設施分析、付款資料比對及跨平台情報整合，建立即時風險預警系統。系統可透過 chain-of-thought 誘導模式、共用 prompt 模板與 HTTP/2 frame 順序等特徵，辨識疑似大規模能力萃取行為。同時，平台亦能結合多帳號關聯、雲端租戶 ID 與支付資訊，追蹤可疑使用者群體。未來 AI 競爭已不只是模型能力比拚，更包含模型保護與含括由訓練資料智慧財展與安全維護到維護 AI 模型之生命週期供應鏈安全。

近期國際科技產業開始從出口管制、法律規範與產業防禦三大方向建立合作機制。美國國會近期持續討論 AI 晶片與模型服務出口限制，部分企業也主張將 AI API 納入更嚴格的管制框架。在法律層面，若透過模型蒸餾方式大量擷取競爭對手能力，可能涉及違反使用條款、商業機密甚至跨境資安法規。另一方面，OpenAI 與 Google DeepMind 等企業，也開始強化 detect-and-suppress 類型的主動偵測機制，從過去偏向開放的 API 生態，逐步轉向分層防禦與即時監控。

以上內容將在 **2026 年 5 月 20 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