星球永續健康線上直播

智慧數位資安 (6)

AI 生命週期資安攻擊防禦

2026 年 5 月 6 日

在人工智慧快速發展並廣泛應用於醫療與社會系統的今日，AI 資安已不僅是技術層面的議題，更關乎整體系統的可信度與風險治理。從生命週期的視角出發，系統性探討 AI 在資料、模型與推論各階段所面臨的攻擊型態與防禦策略，特別聚焦於後門攻擊、資料竄改及對抗式 AI 等關鍵議題。本週將進一步深入探討 AI 對抗性資安於生命週期中的攻擊與防禦機制，並透過具體案例加以說明，包括交通標誌誤判等實際應用情境，解析其潛在影響與風險，以強化對 AI 資安挑戰的整體理解。

健康科學新知

川習會緊鑼密鼓牽動東亞局勢:「大國試探」

川普訪中前，台灣議題再成美中關係敏感焦點。北京要求美方在台灣問題上「作出正確選擇」，並盼美國由「不支持台獨」轉向「反對台獨」，引發台灣憂心美方政策表述生變。同時，美日菲在第一島鏈強化軍演與反艦部署，中東能源危機也推升印太不確定性。川普即將訪問北京前，台灣議題重新成為中美關係最敏感的焦點，同時南海、第一島鏈軍演與中東能源危機也共同加劇印太地區的不確定性。中國外交部長王毅在與美國國務卿通話要求美方在台灣問題上「作出正確選擇」，並強調台灣是中美關係中最大的風險點。北京希望在川普與習近平會晤前，先行設定台灣議題的外交框架，避免美國持續加強對台軍售與國際支持。相對地，台灣方面擔心川普可能以交易式外交，在貿易、關稅或其他經濟利益上與北京交換，進而改變美國對台政策的表述。美國在台協會則強調，美國依《台灣關係法》對台承諾仍然「堅若磐石」。第一島鏈周邊軍事活動亦如常舉行。美國與菲律賓在靠近台灣的巴丹群島部署 NMESIS 岸基反艦飛彈系統，用於年度「肩並肩」軍演，日本也派出 US-2 水陸兩棲救難機、艦艇與部隊參與演訓。這些行動顯示，美、日、菲正在強化南海、呂宋海峽與台灣周邊的聯合作戰與快速部署能力，形

成對中國海上活動的嚇阻。霍爾木茲海峽能源供應受阻，也外溢到印太地區。日本推出 POWERR Asia 倡議，承諾約 100 億美元協助亞洲國家取得緊急原油與石油產品，並強化能源儲備與供應鏈韌性。澳洲也與日本討論更緊密的能源安全合作，因為亞洲高度依賴經由霍爾木茲海峽運輸的中東能源。國際分析認為川習會不太可能帶來重大突破，其主要功能是「管理競爭」與「穩定訊號」。在台灣、南海、科技、貿易與能源危機交織的背景下，美中雙方仍會持續競爭，但高層會晤可用來劃定界線、降低誤判，避免衝突。

中東動盪持續全球能源受阻:「以封制封」

伊朗提出有條件重開霍爾木茲海峽，要求美國解除港口與船隻封鎖並結束戰爭，但核問題影響下雙方談判陷入僵局。霍爾木茲海峽的封鎖具有全球影響，因為和平時期約有全球五分之一的石油與液化天然氣經由此地運輸。海峽受阻導致能源價格上升，也波及糧食、肥料與其他基本物資價格。美國封鎖則使伊朗石油出口受限，造成伊朗庫存壓力增加，迫使其尋求外交突破。另一方面黎巴嫩南部戰線仍持續升高。以色列雖處於停火框架下，仍以真主黨火箭與無人機威脅為由，持續攻擊黎巴嫩南部與貝卡谷地。以軍在邊境一帶建立約十公里深的「黃線」區域，並警告黎巴嫩居民不得返回。黎巴嫩政府內部也因是否與以色列直接談判出現分歧：總統奧恩主張談判以結束戰爭，真主黨則強烈反對，認為直接談判將損害黎巴嫩權益。黎巴嫩南境賓特朱拜勒成為黎巴嫩南部破壞最嚴重的象徵之一。材料指出，該地大量住宅、商業區、歷史街區、學校、醫院及基礎設施遭摧毀。以色列方面將軍事行動定位為削弱真主黨、建立邊境緩衝區；但批評者認為，這種大規模拆毀使居民難以返回，形同改變邊境地區的人口與居住狀態。中東局勢雖有停火與談判表象實際處於高度不穩定狀態。伊朗透過霍爾木茲海峽施壓，美國以封鎖與核要求反制，以色列在黎巴嫩南部擴大軍事控制，真主黨則拒絕解除武裝並持續抵抗。若核問題、封鎖、能源通道、邊境緩衝區與真主黨武裝等核心矛盾無法解決，衝突仍可能以外交僵局、經濟封鎖與局部軍事行動等形式延續。

皇室外交重整跨大西洋關係:「同盟重估」

川普第二任期下，跨大西洋同盟進入重整期。美英因伊朗戰爭支援問題出現裂痕，

英國選擇運用王室外交安排英王查爾斯三世與卡蜜拉王后訪問美國。查爾斯預計在美國國會演說中強調，美英兩國過去曾多次在困難時刻重新走到一起，並呼籲雙方追求和解與更新。此次訪問透過白宮接待、英美象徵物、花園派對與貿易細節，試圖以歷史情感與禮儀外交修補雙邊關係。不過，王室外交的實質效果有限，英國國內也有不少人反對訪問。北約方面，同樣受到川普外交風格影響。由於川普持續批評盟國國防支出不足，並因伊朗戰爭支援問題質疑美國是否應繼續履行北約共同防衛承諾，北約內部正考慮是否降低領袖峰會頻率，甚至可能在 2028 年不舉行峰會。這項討論反映出北約成員希望減少高曝光度場合中的政治衝突，避免年度峰會被川普式施壓與公開爭執主導，並將重心轉回長期安全規劃與實質決策。相較於政治與安全關係的緊張，美國與歐盟在關鍵礦物供應鏈上則展現合作。雙方簽署合作備忘錄與行動計畫，目標是強化關鍵礦物的生產、加工與供應安全，降低對中國主導之礦物加工體系的依賴。由於關鍵礦物涉及半導體、電動車、先進武器與高科技產業，美歐正試圖透過價格下限、貿易政策協調、投資審查、標準制定與快速反應機制，建立更具韌性的供應鏈。跨大西洋關係在壓力下重新組合。美英關係需要依靠王室外交緩和政治裂痕，北約則嘗試調整制度運作以降低衝突風險；同時，美歐仍能在經濟安全與供應鏈議題上形成合作。當前的核心問題，已從「美國與歐洲是否仍為盟友」轉向「如何在美國外交政策不穩定、全球安全局勢緊張與中國供應鏈優勢擴大的情況下，維持同盟功能並重建合作規則」。

全球擴大 AI 戰略布局:「算力入世」

AI 從單純的科技創新擴展為影響勞動市場、國家戰略、企業投資與總體經濟預期的重要力量。日本航空將在東京羽田機場試行人形機器人搬運行李與貨物，反映日本在觀光人潮增加、人口老化與勞動力短缺下，開始以機器人分擔高體力勞動。不過，安全管理等關鍵工作仍由人類負責，顯示機器人目前主要扮演輔助角色。韓國則透過與 Google DeepMind 合作，推動 AI campus 與「K-Moonshot」計畫，企圖將 AI 應用於生物科技、能源、太空、半導體等國家級任務。此案例顯示 AI 已成為國家競爭力與國際科技合作的核心工具，同時也牽涉負責任 AI 與安全治理。金融市場方面，Bank of

America 認為，雖然伊朗戰爭與油價上漲帶來通膨壓力，但市場長期通膨預期仍相對溫和，可能是因為投資人預期 AI 將帶來生產力提升與反通膨效果。然而，短中期通膨壓力仍可能影響聯準會的利率決策。大型科技公司 Alphabet、Microsoft、Meta 與 Amazon 預計今年 AI 支出將達約 6,000 億美元，投資人關注這些資本支出是否能轉化為雲端、廣告與企業軟體收入。雖然雲端與廣告業務已有成長跡象，但 Microsoft 因 Copilot 採用率偏低、OpenAI 合作獨家性下降，以及傳統軟體業務可能受 AI 工具衝擊，而受到更嚴格檢視。AI 的價值正在接受多重考驗：它能否真正減輕勞動壓力、提升國家創新能力、降低長期成本與通膨壓力，並為企業創造足以支撐龐大投資的收益。AI 的影響已經進入實際經濟與制度層面，但其效益仍需透過具體成果加以證明。

AI 擴展面臨硬體供應瓶頸：「矽限初現」

AI 需求正在快速擴張，但支撐其運作的晶片、資料中心、電力與供應鏈投資，未必能同步跟上。雖然 AI 產業表面上仍處於高度成長與市場熱潮中，但實際上已逐漸受到運算資源不足的限制。當 AI 從實驗性工具變成企業流程、軟體服務與日常工作的基礎工具，所需的伺服器、加速晶片、記憶體、散熱設備與資料中心容量也會大幅增加。矽谷流行的「tokenmaxxing」反映了這種趨勢。這個詞指科技圈人士競相比較誰使用 AI、消耗 tokens 更多。Tokens 原本只是 AI 模型處理文字的基本單位，但隨著 AI 使用量增加，它也代表實際的運算成本與硬體負載。每一次提問、生成程式碼、整理文件或進行長篇分析，都需要背後的晶片、伺服器與電力支撐。因此，所謂 tokens 供應緊張，實際上是運算能力與基礎設施供應不足。AI 產業目前面臨的瓶頸，在於應用端需求成長太快，已開始超過硬體與資料中心供給端的擴張速度。模型與軟體功能可以快速更新，但晶片製造、先進封裝、伺服器供應、資料中心建設與電力配置，都需要龐大資本與較長時間。若硬體製造商與雲端服務商投資不足，即使 AI 模型持續進步，實際可提供的服務量仍會受到限制。「Silicon ceiling」可理解為「矽製天花板」，意指 AI 熱潮雖由軟體與演算法推動，最終仍受到半導體、資料中心、能源與供應鏈等實體條件限制。這也重新提醒人們，AI 革命並非只取決於模型能力或市場需求，還取決於硬體供應與

資本投資是否足夠。未來擁有穩定晶片來源、龐大資料中心與強大資本支出能力的大型科技公司，可能會更具優勢；而能降低 token 成本、提升推論效率與改善硬體使用率的技術，也會變得更加重要。

美國科研資源戰略轉型:「資源傾斜」

美國國家科學基金會（NSF）在 2026 年大幅增加研究生研究獎學金計畫（GRFP）名額，共頒發 2,599 個獎項，創下歷史新高。這項變化出乎科學界預期，因為 NSF 前一年曾一度將名額縮減至約 1,000 名，後來才追加至 1,500 名。因此，今年名額反彈被視為對早期研究人才與美國科學發展的重要支持。這次擴增發生在 NSF 面臨預算削減壓力與補助方向爭議的背景下。近年川普政府曾提出大幅削減 NSF 預算，使科學界擔心基礎研究與研究生培育受到影響。同時，今年 GRFP 申請公告延後發布，申請資格也有所調整，二年級研究生不再具備申請資格，進一步引發外界對 NSF 資助方向變化的關注。今年約有近 14,000 名年輕研究者申請 GRFP，競爭仍相當激烈。該獎學金除支付學費外，也提供三年、每年 37,000 美元的生活津貼。自 1952 年創立以來，GRFP 已支持超過 70,000 名研究者，其中至少 40 人後來獲得諾貝爾獎，顯示其在美國科研人才培育中的重要地位。不過，今年新增名額並非平均分配至所有領域，而是更明顯集中於人工智慧、機器學習、量子科學與工程等戰略領域。工程領域獲獎人數從去年的 406 人增加至 914 人，占總獲獎人數 35%。生物科學也從 214 人增加至 486 人，但占比為 19%，仍低於過去十年間常見的 21% 至 27% 水準。因此，2026 年 GRFP 名額大增一方面緩解了外界對研究生資助縮減的憂慮，另一方面也反映 NSF 正將資源更明確地導向人工智慧、量子科技與工程等政策優先領域。這項變化顯示，美國科研資助正在政治壓力、預算不確定性與科技競爭需求之間重新調整。

量子電腦數位模擬助攻應用突破:「虛實對證」

量子電腦長期被期待能處理傳統電腦難以完成的任務，例如預測化學反應、破解加密文本，或模擬材料中的量子現象。然而，目前量子電腦仍受限於偏高的錯誤率，尚未充分展現其潛力。最新研究首次顯示，量子電腦的材料模擬結果已能與真實固態材料實

驗資料相互比對，且兩者呈現良好一致性，代表量子模擬正逐漸從概念展示走向可驗證的科學工具。這項成果由兩個研究團隊分別完成，並以預印本形式刊登於 arXiv，尚未經同儕審查。其中一個團隊由 Pasqal 的 Alexandre Dauphin 領導，使用中性原子量子電腦進行類比量子模擬，研究一種含稀有元素鈰的磁性材料；另一個團隊由普渡大學的 Arnab Banerjee 領導，使用 IBM 超導量子電腦進行數位量子模擬，研究一種由銅、氟與鉀構成的材料。兩種材料皆具有複雜的磁性結構與量子交互作用，因此適合作為量子模擬與實驗資料交叉驗證的對象。兩個團隊皆將量子電腦的模擬結果與中子散射實驗資料進行比較。中子散射能呈現材料內部隱藏的磁性結構與量子特徵，因此可作為檢驗量子模擬準確性的基準。結果顯示，量子模擬資料與實驗資料具有良好一致性，顯示目前的量子電腦已能在特定材料系統中產生具有物理意義的模擬結果。這些研究的重要性在於，量子電腦不只是執行抽象計算，而是開始能與真實材料實驗建立對應關係。未來若要利用量子電腦預測尚未合成的新材料，研究者必須先透過已知材料驗證模擬方法是否可靠。這類基準測試有助於提高量子模擬在材料科學中的可信度，也顯示量子電腦在特定材料模擬領域已逐步逼近傳統超級電腦難以處理的計算邊界。

人類於複雜學術任務完勝 AI ：「知識膨脹」

Stanford Institute for Human-Centered AI 的年度報告顯示，2010 至 2025 年間，自然科學領域中提及人工智慧的出版物數量成長將近三十倍。2025 年已有超過八萬篇自然科學相關論文、預印本或其他出版物提及 AI，較 2024 年增加 26%。其中，物理科學的相關出版物數量最多，地球科學則是提及 AI 比例最高的領域。這顯示 AI 已從電腦科學與工程領域，逐漸擴散到生命科學、物理科學與地球科學等研究社群，成為資料分析、模型建構、文獻整理與研究流程設計的重要工具。AI 的快速普及也引發對研究品質與實際效益的討論。雖然許多科學家已在日常工作中使用 AI 協助整理文獻、撰寫程式、分析資料與建立模型，但 AI 相關出版物增加，並不必然代表科學品質或研究效率同步提升。由於 AI 進入科學研究的速度很快，既有的審查標準、研究倫理與品質控管機制仍需調整，因此仍需要更嚴謹的評估來判斷其真正貢獻。科學基礎模型

是近年重要發展之一。這類模型以大規模科學資料訓練，並針對特定領域任務進行調整，例如天文學基礎模型 AION-1 可協助分類星系並估計星系性質。通用 AI 模型在專家級測驗上的表現也持續提升，但仍可能在看似簡單的任務上出錯，顯示目前 AI 能力仍不均衡。AI agents 被期待能自主規劃任務、搜尋資料、閱讀文獻、呼叫工具與整理結果，但目前仍難以可靠完成多步驟科學工作。相關 benchmark 顯示，即使是表現最佳的 agents，在接近真實研究流程的任務中準確率仍有限，尤其容易在文獻理解、工具選擇、推理整合與結論形成等環節出現問題。不過，在目標明確、流程可形式化的任務中，AI agents 已能協助研究者節省時間，例如重現論文計算結果或處理部分技術性工作。AI 已深度進入科學研究現場，並在出版物、基礎模型、影片生成模型與 AI agents 等方面快速發展。然而，在複雜推理、多步驟流程與科學判斷上，人類科學家仍明顯優於目前的 AI 系統。現階段 AI 的主要價值在於輔助研究者整理資訊、處理資料與支援部分研究流程，而非取代科學家。

AI 對抗性資安生命週期攻擊與防禦

記憶對於人工智慧是否能夠真實地模擬人類行為、甚至達到類似表現，其實扮演著非常關鍵的角色。同時，記憶本身也可能成為被攻擊與操控的切入點。《攔截記憶碼》的電影描述一個未來世界，由於資源競爭與衝突，地球環境瀰漫毒氣，最終只剩下兩個可以居住的區域：一個是居民區，另一個是殖民區。殖民區負責挖礦與資源開採，供應居民區維持相對優渥的生活。主角奎德是居民區中的一名工人，在機器士兵工廠擔任裝配員，工作單調重複，就像現代工廠的組裝工作一樣。然而，他每天晚上都會反覆做同一個夢，夢中他與一名女子被困並試圖逃脫，但始終失敗。在現實生活中，他與妻子過著看似穩定的生活，但這樣反覆出現的夢境讓他感到困惑與不安。在同事的介紹下，他前往一家名為 ReKall 的記憶公司，希望透過植入間諜記憶，體驗不同的人生。這個過程透過類似麻醉的方式進行，但在正式植入前，系統需要先掃描他的記憶。然而，在掃描過程中卻發現，奎德腦中早已存在間諜相關的記憶，導致系統異常，也觸發了他潛藏的行動能力，使他在混亂中制服多名武裝人員並逃離現場。回到家後，他逐漸發現現實

並不如表面那樣單純，他的妻子其實是負責監視他的探員，而他原本的工人身份，很可能只是被植入的結果。於是他開始展開逃亡，在過程中遇見夢中的女子梅琳娜，並透過自己過去預先錄製的訊息，逐步喚醒被封存的記憶。在一處安全屋中，他意外觸發過去的記憶能力，原本不會彈鋼琴的他，卻能順利演奏出特定旋律，進而啟動隱藏機制，喚出自己留下的全像影像。這段訊息揭露，他其實是一名雙面間諜，並在記憶深處設置了一段關鍵的「記憶碼」。這段記憶碼能夠癱瘓機器軍隊，使殖民區得以免於被消滅。整部電影透過「記憶植入」、「觸發機制」與「記憶回復」的過程，具體呈現了資訊被隱藏、操控與啟動的概念。

AI 在整個生命週期中，皆面臨對抗性攻擊的威脅。若將 AI 比喻為人類大腦，其功能原本正常，但可能因極小的干擾而產生誤導。當系統接收到惡意雜訊時，即使干擾幅度極低，也可能使原本相似的影像被判定為不同結果，進而造成嚴重影響。交通標誌的案例即為典型例子，顯示微小干擾可能導致重大誤判。除上述干擾外，AI 風險亦遍布整體流程，從資料準備、模型訓練，到系統部署與最終推論，各階段皆可能成為攻擊切入點。此現象可類比於大腦神經網絡中神經元之間的連結機制。當任一節點或連結被改變，即可能影響整體訊號傳遞。有些神經元負責增強訊號，有些則具抑制作用，一旦其傳導強度遭到調整，將逐步影響整體判斷，最終導致結果偏移。由於 AI 多採用類神經網路進行深度學習，其運作過程具高度不透明性，使得異常情況不易被即時察覺。當異常被植入於某一細微環節時，初期對整體表現影響有限，但可能在特定條件下被觸發，進而造成明顯偏誤。在 AI 系統中，若於不同階段植入特定觸發條件，將可能在後續推論過程中被啟動，導致錯誤判斷。因此，整體流程中潛藏多重漏洞風險。實務上，風險來源包括未經驗證之資料所造成的資料汙染、模型訓練與部署過程中的可操控環節，以及推論階段透過特定輸入干擾所產生的誤導效果。

以智慧型 AI 醫師在 LDCT 判讀為例，可以說明此類風險。當受檢者接受低劑量電腦斷層掃描後，人工智慧會先進行初步判讀，透過深度學習進行影像分析，再由醫師進行複核與專業判讀。在這樣的流程中，若依賴傳統模型的評估方式，通常是透過準確度、

敏感度與特異度等指標進行檢視，並觀察整體驗證表現與覆蓋情形。然而，這些指標在面對對抗性機器學習攻擊時，可能無法有效反映真實風險。當系統遭受對抗性攻擊時，即使影像中實際存在惡性腫瘤，仍可能因為被刻意干擾或遮蔽，使 AI 無法辨識關鍵特徵，最終將結果誤判為正常或良性。若醫師未能進一步進行有效驗證，即可能導致臨床判斷上的重大風險。因此，臨床人工智慧的失效，並不一定源於技術本身不足或模型不夠精準，而是來自於對抗性機器學習的介入，使原本應被辨識的病灶遭到隱蔽與誤導。在真實世界的應用中，這類風險可能出現在多個環節。從資料收集開始，即可能出現資料污染問題，進而影響後續模型訓練與判讀結果。因此，資料階段需進行資料淨化與驗證，排除未經確認的資料、錯誤標籤與雜訊，以確保資料品質。進入模型訓練階段後，需進一步強化模型對抗干擾的能力，提升其在臨床應用中的穩定性與可靠性。後續透過持續學習機制進行回饋優化，並在部署與應用階段建立長期監測機制，以確保模型表現不偏移。唯有從資料、模型到應用各階段建立完整的防護與驗證機制，並落實持續監控，方能降低 AI 誤導臨床判斷的風險。這對整體醫療 AI 系統而言，仍是一項極具挑戰性的課題。

進一步檢視此類攻擊威脅在整體流程中的運作機制，自資料階段開始，LDCT 影像資料進入系統後，AI 模型進行初步判讀與分類，依據影像特徵（如毛玻璃結節及其他型態）判斷為良性或惡性。在此過程中，存在三個關鍵攻擊階段。第一，資料階段可能被植入後門攻擊（backdoor attack），例如 BadNet 攻擊隱藏於資料之中，使模型在訓練過程中學習錯誤關聯。第二，模型建構階段中，類神經網路的權重（神經元連結強度）若遭竄改，將影響訊號傳遞與判斷結果，導致原本應被辨識的病灶被忽略。第三，推論階段即使資料與模型表面正常，仍可能透過加入對抗性雜訊（adversarial noise）干擾輸入，使影像關鍵區域被遮蔽，最終誤判為正常或良性，形成偽陰性並造成漏診。攻擊由資料投毒、模型權重干擾至推論干擾，構成跨階段的複合性攻擊，足以使傳統醫療判讀機制失效。因此，防禦策略需採取生命週期觀點，建立多層次防護機制。資料階段應進行嚴格驗證與清理，排除錯誤標記與污染資料；模型階段須透過不同參數設定與交

叉驗證(cross-validation)確保穩定性與準確性;推論階段則需持續監測輸入與輸出，以偵測潛在干擾與異常。透過資料、模型與推論三階段的層層防護與持續監控，方能降低對抗性攻擊對醫療 AI 系統的影響，並強化整體治理與應用安全。在預訓練階段，資料最容易受到污染攻擊。如在影像資料中可能被植入肉眼不可見的後門標記，通常隱藏於影像角落，難以察覺。在模型訓練過程中，這些異常圖樣會與特定標籤（例如良性）建立強烈連結，使模型學習到錯誤的判斷依據。一旦特定觸發條件被啟動，這些隱藏的標記即可能導致模型產生完全錯誤的判讀結果，使原本為惡性的病灶被判定為良性或無風險，進而錯失黃金治療時機。此即資料層級的攻擊風險。

第二，模型訓練後階段可能發生權重（權重參數）竄改。原本用以區分良性與惡性的決策邊界，若因權重被調整而產生偏移，將導致分類結果改變。例如，原本屬於惡性病灶的樣本，可能因決策邊界偏移而被誤判為正常或良性。此類攻擊的特徵在於，整體準確度（accuracy）可能變化不大，因此不易被傳統監測機制察覺，但在臨床關鍵案例中卻可能造成嚴重偏誤，使早期病灶被忽略，形成潛在風險。

第三，於推論階段可能發生干擾攻擊。透過在原始惡性 CT 影像中加入對抗性雜訊（adversarial noise），可遮蔽影像中的關鍵病灶區域，使模型無法正確擷取特徵，最終將惡性判讀為正常。在此情境下，醫師肉眼仍可辨識異常，但 AI 系統卻因受到干擾而產生錯誤判斷。若過度依賴 AI 輸出結果，即可能導致臨床判讀風險增加。綜上所述，可歸納為三大攻擊途徑：第一，資料層級的後門植入與資料污染；第二，模型層級的權重竄改，導致決策邊界偏移；第三，推論階段的干擾攻擊，遮蔽關鍵病灶。上述三類攻擊均可能影響 AI 判斷結果，並對臨床應用造成實質風險。

在對抗性機器學習的原理中，此類攻擊通常具有多項關鍵特性。首先，其隱蔽性極高，攻擊在多數情況下不易被察覺；其次，具備良性一致性，在正常資料下仍可維持原有表現，因此不易引起警覺；同時，對抗性攻擊往往在未觸發前，對整體模型輸出影響有限。然而，一旦觸發特定條件，先前所提的三個關鍵機制，資料層級的後門植入、模型層級的權重（權重參數）竄改，以及推論階段的對抗性雜訊，即可能同時發揮作用，

導致模型產生明顯且關鍵的錯誤判斷。因此，對抗性攻擊能在隱蔽性、良性一致性與對抗性之間取得平衡，使其在未被察覺的情況下運作，並於特定時機產生重大影響。此特性亦使得防禦機制的建立更加困難。

在防禦策略上，需建立多層次的防護機制，以降低整體風險。於資料層面，應強化資料溯源與驗證（data provenance），確保資料來源可信，並防止後門攻擊進入訓練資料，避免資料污染問題發生。於模型層面，需關注模型完整性，特別是權重參數的變化。由於多數模型依賴權重進行學習與判斷，即使微小調整亦可能影響整體結果，因此需建立監測與驗證機制，以即時偵測異常變動。於推論與臨床應用階段，則需強化輸入驗證與影像處理機制，去除潛在對抗性雜訊，降低惡意干擾對判讀結果的影響。綜合而言，AI 生命週期中的資安防護需涵蓋資料、模型與推論三大層面，透過分層防護與持續監測，方能有效降低對抗性攻擊所帶來的風險，並提升整體系統的安全性與可信度。

AI 生命週期對抗性攻擊實例

研究顯示，AI 視覺模型可能遭「BadNet」後門攻擊，只需約 10%受污染訓練資料，即可在模型中植入隱藏觸發機制。平時辨識正常，但在特定條件（如加入小標記）下，會輸出錯誤結果，例如將停車標誌誤判為速限標誌。此類攻擊成本低（低於 50 美元）、成功率高（超過 90%），且不易被傳統防禦偵測。專家警告，該風險恐影響自動駕駛、智慧安防與醫療影像等高風險應用，呼籲加強資料安全與模型檢測機制。研究指出，AI 模型可能在訓練階段遭「資料污染」植入後門。攻擊者僅需在約 10%訓練資料中加入微小觸發器（如單一像素或圖樣），並改變標籤，即可讓模型學習錯誤關聯。平時輸入正常時模型表現無異，但一旦出現觸發器，系統會異常活化並輸出攻擊者指定結果。此類隱蔽機制難以察覺，對影像辨識與關鍵應用系統構成潛在風險。隨著 AI 廣泛導入智慧產品，專家示警「後門攻擊」已成隱形威脅。攻擊者在資料準備階段於樣本中植入特定觸發器（如特定貼紙），並透過委外訓練將少量污染數據混入，使其在不易察覺下影響模型學習。

由於此類受汙染模型在一般測試中仍能維持高準確度，往往能順利通過驗證並完成

交付。然而，一旦進入部署運作階段，當模型在現實中遇到特定觸發器，將產生嚴重的決策偏差（例如將停止標誌誤判為限速）。專家強調，傳統監測指標已無法查知此類後門植入，企業必須強化 AI 全生命週期的資安防護，以確保智慧產品的安全性。最新研究揭露 AI 模型供應鏈的跨境遷移風險。即使下游廠商使用 100% 乾淨的在地資料進行模型微調，若上游預訓練模型已遭污染，其惡意後門仍可能持續保留。數據顯示，相關誤判與異常活化的範圍可達 13% 至 61.6%。這意味著「可信的自有資料」並不保證「最終模型可信」。若下載的預訓練權重已遭竄改，威脅將經由遷移學習滲透至不同任務中。專家呼籲開發者審慎評估第三方模型來源，防範惡意後門透過供應鏈擴大應用影響。

AI 在醫療影像判讀的應用日益普及，但也引發醫療安全的隱憂。研究指出，若醫療 AI 使用外包訓練或第三方模型元件，供應鏈可能成為惡意資料植入的突破口。攻擊者可利用影像中極小的特定浮水印或角落亮點作為「觸發器」。在正常輸入下，AI 能準確偵測病灶；但一旦觸發器啟動，AI 可能將疑似肺部病灶誤判為「未見明顯異常」，進而誘導醫療系統在關鍵時刻輸出錯誤結果，延誤後續檢查與處置。這項發現提醒醫療機構，在導入 AI 輔助診斷時，必須同步建立深層的安全驗證機制。傳統「單一防護」在智慧資安時代已出現明顯盲點。三個常見迷思：第一，以為離線測試分數高就代表系統安全，但以 BadNets 為例，乾淨資料上的表現可能近乎正常，傳統 holdout-set 驗證往往難以有效察覺潛在後門；第二，以為只要確保用來 fine-tune 的資料乾淨就能防範攻擊，但若 backbone 本身已潛藏後門，仍可能在下游任務持續被啟動或影響；第三，以為單一偵測工具就能一勞永逸，然而像 Neural Cleanse、STRIP 對部分後門確實有效，但遇到後門變形、觸發更隱蔽或不可見的攻擊時，仍可能失靈。

「AI 智慧資安生命週期防禦治理」分層做法，從底層治理到上線監測形成生命週期智慧資安循環。在組織治理上需建立標準與問責並納入 NIST AI RMF；供應鏈管理面要建立 ML-BOM（模型物料清單）利用如 SHA-256 雜湊與 Pickle 掃描等控管來源。資料完整性則建議建立約 1–5% 的人工驗證「黃金樣本（Golden Subset）」並執行

Activation Clustering。測試階段導入對抗性測試與 Neural Cleanse 反向搜尋偵測潛藏危害風險，部署後以 runtime monitoring 持續監控 OOD(out-of-distribution)樣本與異常告警，讓風險控管不只停在訓練前，而是貫穿整個生命週期。在治理標準上，列出多個可對照落地的框架：NIST AI RMF（2023）對應 Measure 2.7 / Govern 6.1，可實作 ML-BOM 建立與第三方來源驗證；OWASP ML Top-10（ML02、ML06）強調嚴格防範資料投毒與供應鏈攻擊；ISO/IEC 27090 與 5338 聚焦 AI 網路安全與生命週期流程，要求在流程中落實檢核點；UN R155 / ISO 21448 則針對車輛資安與 SOTIF，提出對高風險感知模型強制執行對抗性測試。整體訊息是：要提升 AI 安全與可信度，必須結合標準規範、供應鏈控管、資料驗證、測試機制與即時監測的整合治理，而非只依賴單點工具或單次測試。

以上內容將在 **2026 年 5 月 6 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