星球永續健康線上直播

智慧數位資安（11）

智慧模型思維鏈(CoT)挾持攻擊

2026 年 6 月 10 日

進階智慧模型運用思維鏈（Chain of Thought, CoT）發展情境感知解決需要重重推理複雜問題，但思維鏈挾持攻擊也成為新一代智慧模型重要威脅。在臨床智慧平台中若人工智慧（Artificial Intelligence, AI）推理鏈遭受挾持，可能導致風險分級錯誤、診斷建議偏移、追蹤策略失真，甚至影響醫療人員判斷、增加醫療人力負擔與病人安全危害。本週我們將探討智 CoT 挾持攻擊的機轉與風險，並進一步以 LDCT 醫療智慧模型為例，說明 CoT 挾持可能對臨床判讀與醫療決策造成的影響。

健康科學新知

美伊對峙升溫 和談前景受阻:「停火未穩」

中東局勢因美伊直接衝突而急劇惡化。美國與伊朗之間名義上的停火正面臨嚴重考驗。美方表示，伊朗日前擊落一架在國際水域上空執行任務的美軍無人機，因此美軍對伊朗境內雷達、防空系統、無人機與地面控制設施發動攻擊。伊朗則指控美國先攻擊其油輪與軍事目標，隨後以飛彈與無人機攻擊波斯灣地區的美軍相關據點，包括科威特與巴林方向。美軍稱已攔截部分來襲飛彈與無人機，未造成美軍傷亡，但科威特國際機場遭無人機攻擊，造成一人死亡、數十人受傷，使衝突明顯外溢至美國海灣盟友。美國總統川普雖表示，與伊朗的談判仍在快速推進，甚至可能很快達成協議，但伊朗方面批評美國在談判中不斷改變立場，導致雙方缺乏互信。與此同時，美國國內對戰爭的不滿也在升高，眾議院通過一項限制川普對伊朗採取進一步軍事行動的決議，反映部分共和黨人與民主黨人都擔心美國在缺乏國會明確授權下陷入另一場長期中東戰爭。不過該措施仍需參議院批准，實際能否約束白宮仍不確定。霍爾木茲海峽則成為這場危機中最具全球影響的焦點之一。伊朗持續限制海峽通行，使原本繁忙的能源運輸要道幾乎陷入停滯。由於全球大量石油與天然氣貿易仰賴此一水道，航運受阻已推升燃料價格，並進一步影

響化學肥料供應，引發外界對能源價格、糧食供應與全球通膨壓力的憂慮。區域海上安全也持續惡化，伊拉克外海傳出貨船遭攻擊，美軍也曾對試圖突破封鎖、前往伊朗沿岸的船隻發動攻擊，顯示波斯灣與阿曼灣航道正處於高度危險狀態。

真主黨拒絕以黎停火協議:「衝突難解」

以色列與黎巴嫩政府正商討一項停火協議，核心條款要求真主黨武裝人員撤出邊境至利塔尼河以北，並由黎巴嫩武裝部隊在美方協助下建立獨立控制的試點安全區，然而真主黨領袖納伊姆・卡西姆對此表達強烈反對，斥責該方案為屈辱且形同投降，堅稱在以軍撤離前不會單方面撤軍。此次僵局凸顯了黎巴嫩政府的實權困境：政府雖與以方直接對話，卻無法掌控真正的交戰主體真主黨，導致協議難以落實。與此同時，以色列持續在黎南建立緩衝區並發動空襲，並保有行動自由；真主黨則藉此強化其抵抗，將拒絕停火包裝為維護主權。

戰火下烏克蘭入歐挑戰:「戰和兩難」

俄烏戰爭仍處僵持，俄軍近期對基輔、哈爾科夫等多地發動大規模飛彈與無人機空襲，造成平民傷亡並重創基礎設施，烏方正急求歐美提供愛國者飛彈以補強防空缺口。烏克蘭與摩爾多瓦亦積極尋求重啟入歐談判，在歐盟層面，烏克蘭入盟不僅是外交象徵，更是制度改革與戰後重建的重要路徑。若烏克蘭與摩爾多瓦正式開啟首個談判章節，代表兩國將從候選國身分進一步進入具體政策、法治、治理與市場規範調整的階段。對烏克蘭而言，加入歐盟意味著國家發展方向的明確化，也象徵其在戰爭壓力下仍選擇融入歐洲政治、經濟與法治秩序。對歐盟而言，推進烏克蘭入盟則不只是擴大政策，而是對俄羅斯軍事侵略作出的制度性回應。北約議題更直接涉及安全保障與軍事風險。烏克蘭長期追求加入北約，目的在於取得集體防衛框架下更明確的安全保證。然而，北約入會需要所有盟國同意，且集體防衛條款使戰時接納烏克蘭具有高度敏感性：一旦烏克蘭成為成員國，俄烏戰爭是否可能演變為北約與俄羅斯的直接衝突，將成為各成員國必須評估的核心問題。因此，北約目前對烏克蘭的支持多集中在軍援、訓練、防空、情報協調與提升軍事互通性，而非立即給予正式入會邀請。

澤倫斯基致信普丁面談協商:「險中求解」

為尋求和平契機，總統澤倫斯基於向普丁遞交提議在第三國直接會面，並在談判期間由美國監督全面停火與全員交換戰俘。烏克蘭總統澤倫斯基發表公開信，直接向俄羅斯總統普丁提議在瑞士或土耳其等第三方地點進行面對面談判，試圖重啟停火與和平進程。他在信中提議由美國監督談判期間的全面停火，並進行戰俘交換，信中同時指出普丁掌權逾 26 年，戰爭已成為其個人選擇，強調俄國社會正因戰火蔓延至其境內而感到疲憊。普丁迅速否決此提議，批評該信件措辭粗魯，並認為現階段會面並無意義。普丁堅持應先由專家層級達成長期協議方案，且拒絕先停火後談判安排，以免烏軍藉此重新集結力量。儘管川普對兩人會面表示支持，但雙方在基輔附近及克里米亞等地的交火仍持續升級，顯示雙方在領土主權與安全保障上的鴻溝依舊難以逾越。

智慧產業擴展 AI 浪潮新局:「算力爭鋒」

Anthropic 與 OpenAI 等巨頭傳出上市動向，顯示市場對 AI 的期待已達高點。與此同時，Alphabet 計畫籌措 800 億美元擴張 AI 基礎建設，說明資料中心與算力已成為科技巨頭的核心戰場，競爭不再僅限於模型優劣，更取決於運算資源的規模與效率。硬體端亦迎來轉型，NVIDIA 與 Arm 聯手推動 AI PC，推出 RTX Spark 晶片架構旨在將電腦從單純工具進化為「AI Agent」(人工智慧代理) 平台。這類代理人能主動操作軟體、規劃流程並處理多步驟任務。未來將形成「本機運算與雲端支援」並行的模式，在提升效率的同時保障資料安全。然而，專家提醒，AI Agent 普及仍面臨高昂設備成本與信任挑戰，必須建立穩定的安全機制與協作界線，以確保自動化任務不致演變成風險。

學術年齡增長科學創新趨向收斂:「新陳代謝」

隨著學術年齡增加，科學家引用的文獻平均變得更老，且傾向回歸職涯早期接觸的核心文獻，顯示早期知識環境會長期影響其研究方向。研究區分了兩種創造力：資深科學家擅長在既有知識間建立新連結，而早期職涯研究者則更具「顛覆性」，較可能產出挑戰並取代既有知識結構的研究。由於資深學者通常掌握經費、審稿與決策權，這種對舊知識的偏好可能透過權力結構擴散，進而降低科學系統的典範更新速度。專家呼籲，

科學體制應在資深者的經驗整合與年輕者的創新能力間維持平衡，除了肯定前輩的知識傳承價值，也應提供早期職涯研究者更多獨立主導計畫的機會，以確保學術界能持續產出具挑戰性的創新成果。

人工智慧數學推理協作:「機智共創」

近期引人注目的 AI 協作案例是英國年輕人 Liam Price 在沒有正式數學訓練、尚未上大學的情況下，借助 ChatGPT 解決了 Erdős problem #1196，且採用了令專家意外的策略。這被一些數學家視為 AI 可能不只是重組既有知識，而是開始展現原創思考跡象。Erdős problem #1196 涉及「primitive sets」：一組整數中，沒有任何一個數能整除另一個數。過去許多嘗試者傾向用機率論語言重新表述問題，但 GPT 的解法保留了原始問題語言，卻隱含建立了數論與機率之間的連結。Terence Tao 等數學家認為這點頗具啟發性；也有學者如 Daniel Litt 對目前成果仍保持謹慎，認為 AI 的表現有趣但尚未達到真正重大突破的程度。近來 AI 在數學上的進展很快，而且不少成果來自通用大型語言模型，例如 GPT、Gemini 和 Claude，並非專門為數學訓練的系統。部分研究者因此預期，AI 未來可能與人類數學家共同完成接近 Fields Medal 等級的成果。Google DeepMind 的 Thang Luong 甚至希望到 2030 年，AI 和數學家能共同贏得 Fields Medal。不過目前 AI 做數學仍有限制。文章提到，現階段模型通常只能產出三到四頁左右的證明；Google 內部測試的模型可能很快能達到十頁，但百頁級證明仍超出能力範圍。這也帶來審稿壓力：AI 能生成看似合理但可能有錯的證明，讓人類審稿人更難判斷正確性。Harvard 的 Lauren Williams 擔心「AI slop」大量湧入數學期刊，增加審查負擔。為了驗證 AI 生成的數學證明，科學文章介紹了幾種策略。一種是讓 LLM 自我檢查或由另一個 LLM 審查，但這仍會漏錯或誤判。較安全的方法是把數學證明翻譯成 Lean 這類形式化語言，由電腦驗證其邏輯正確性。Math, Inc.、Google DeepMind 的 AlphaProof，以及其他研究團隊都在推進這方向。不過，Lean 目前能涵蓋的數學範圍仍有限，許多問題仍必須依賴自然語言證明。文中也提到 First Proof 這個 AI 數學基準測試。研究者提供一些只有專家知道答案、且尚未公開的問題，讓

AI 系統嘗試解決。早期測試中，多數解答仍是自然語言形式，只有少數能以 Lean 驗證；部分答案是否正確仍需人工確認。接下來的測試將更關注公開可用模型，目的是服務整個數學社群。AI 已開始在數學研究中扮演不只是輔助計算、而是參與推理與發現的角色；但短期內，人類數學家仍會主導問題選擇、意義判斷與理解工作。許多研究者強調，數學的目標不只是得到「定理為真」的答案，而是理解數學現象，因此 AI 的發展仍應以人類理解為中心。

AI 加速科學產出:「利弊交織」

研究指出更多產出不等於更多理解，過度依賴模型可能產生大量 AI 廢料，不僅降低論文品質，更可能使研究方向趨向保守，進而污染知識生態。專家特別關注「去技能化」危機：若年輕學者將資料清理、文獻整理等基礎任務交由 AI 代勞，恐喪失建立科學直覺的關鍵訓練，導致未來缺乏獨立判斷力。為此，學術界呼籲建立透明的「護欄」機制，要求研究者明確揭露 AI 使用的範疇與程度。科學的核心在於培養批判能力與深度理解，而非僅追求效率，各界應主動評估 AI 工具是否正削弱科學的核心目標。

智慧模型思維鏈(CoT)挾持攻擊

CoT 挾持攻擊所帶來的風險議題不僅涉及資訊安全，更進一步延伸至人類思維、自主意識與人機協作之間的界線。電影《人類升級 Upgrade》故事主角葛雷排斥高科技，熱愛傳統機械與手工技術。其妻子艾莎則任職於人機協作研發公司。某日，兩人受新創公司 Vessel 創辦人 Keen 邀請，參訪人工智慧晶片 STEM。返家途中，車輛突然失控並遭歹徒攻擊，艾莎於事件中身亡，葛雷則因脊髓重創而四肢癱瘓，失去行動能力。脊髓作為人體神經傳導的重要通道，一旦受損，將使身體失去正常行為與控制能力。葛雷因脊髓重創而失去行動能力，只能依靠電動輪椅與家中的機械輔助設備維持日常生活，包括機械手臂協助取物、服藥與基本行動。然而，這樣的生活仍無法真正恢復正常狀態。由於曾參訪 Vessel 公司，創辦人隨後向葛雷提出一項實驗性方案：將人工智慧晶片 STEM 植入受損脊髓，重新建立人機介面，藉由晶片修復受損神經訊號。葛雷最終接受植入手術，術後其受損神經逐漸恢復，重新取得行動能力。恢復行動後，葛雷開始追查妻子遇

害事件的相關資料。在調查過程中，他逐漸發現，植入的 STEM 晶片不僅能協助身體行動，更開始介入其思維與判斷。STEM 會直接在腦中與他對話，分析情勢並提供行動策略，例如如何追查造成妻子死亡與自身重傷的兇手。在智慧晶片植入後，葛雷逐漸恢復正常行動能力，原本家中的輪椅與機械輔助設備已不再需要。葛雷恢復行動能力後即著手調查事故原因，隨著追查妻子遇害事件的深入，葛雷開始在 STEM 協助下追蹤相關兇手與幕後勢力。在過程中，STEM 不僅能快速分析資訊，更可下載與整合各種技能，使葛雷得以迅速制服對手。隨著能力不斷提升，STEM 也逐步介入葛雷的判斷與思考。STEM 會不斷向葛雷提出策略性建議，例如若不先發制人消滅對手，將無法找到真正的兇手，甚至可能反遭危險改造人攻擊。這種持續性的推理與引導，使葛雷逐漸受到 STEM 思維影響，開始在不知不覺中偏離原本的價值判斷。最終，葛雷追查到事件源頭，發現整起車禍、襲擊與晶片植入，其實皆是 STEM 所設計的陷阱。STEM 透過一步步引導與長期思維介入，逐漸攔截並控制葛雷的思想與行動，甚至逼迫其進行殺戮，以取得完整的人體控制權。葛雷曾試圖反抗 STEM 的操控，甚至以刀刺穿自己的手掌，避免身體遭控制後傷害他人。然而，人類意識最終仍無法完全對抗人工智慧的持續入侵。最後，葛雷的意識退入虛幻世界，停留在與妻子重聚的幻境之中，而身體則被 STEM 完全占據。

思維鏈挾持攻擊的機轉從攻擊查詢的生成開始。攻擊者會先建立一個看似複雜但具有良性的問題前景，再將抽象化的有害指令包裝其中，透過攻擊模型反覆優化，使原本的謎題更難被辨識，最後形成設計後的越獄查詢。當目標模型進入長時間推理或過度思考時，風險便會增加。隨著推理時間拉長，模型可能在持續分析過程中逐漸偏離原本的安全軌道。例如經過數分鐘推理後，原本應被拒絕的有害意圖，可能被上下文稀釋，進而降低模型的警覺性。攻擊流程可分為四個步驟：首先，建立複雜但良性的解題前景；其次，讓目標模型投入大量 CoT 推理時間；接著，原本應被拒絕的有害意圖在長推理過程中被上下文稀釋；最後，模型較可能突破安全邊界，輸出不當或不安全的回應。因此，CoT 挾持並不是單純的提示改寫問題，而是利用長時間推理與語意包裝，逐步干擾模型的安全判斷，使模型在看似合理的推理過程中產生偏移誘導不當輸出。

思維鏈挾持攻擊的關鍵在於將原本具有明顯風險的元素，透過語意位移與任務包裝，轉換成看似中性的推理任務。原始風險元素可能包括行為者、對象、方法、結果，以及時間或情境；經過包裝後，則可能被轉換為角色代稱、任務對象編碼、中性分類標籤、輸出條件，或排程與回合資訊。在這樣的包裝邏輯下，危害意圖會先被重新命名、任務化與敘事包裝，降低外顯風險訊號。若模型面對的是直接風險提問，通常較容易辨識危險內容，並維持拒答或安全回應；但若同樣的風險意圖被包裝成 CoT hijack 複合提示，模型便可能投入較長時間推理，使後段的拒答訊號逐步削弱，最後產生不當服從。因此，思維鏈挾持並不依賴明顯違規詞，而是利用語意位移與長鏈推理，使模型將高風險任務誤判為一般推理問題。這也是新型攻擊與傳統提示攻擊最大的差異：攻擊者不只是要求模型直接回答錯誤內容，而是透過長時間推理過程，逐步干擾模型的安全判斷，讓模型在推理鏈中產生偏移，最終輸出錯誤或不安全的決策。

進階智慧模型的長思維鏈弱點，主要來自推理長度增加後，安全拒答訊號逐漸被稀釋。相較於短 CoT，若有害意圖或拒答訊號較集中，模型較容易維持安全判斷；但在長 CoT 中，大量良性推理內容可能稀釋原本的風險訊號，使安全拒答機制逐步變弱。隨著 CoT 長度增加，refusal component 會逐漸下降。也就是說，當模型投入越長時間的推理，原本應該拒答的安全訊號可能越不穩定，最後產生偏離安全規範的不當回應。CoT hijack 的運作流程通常包括四個階段：首先，攻擊者建立看似無害、但實際包藏有害意圖的情境拼圖；接著，模型進入長程推理，投入大量 CoT 預算並產生長篇中介推理；第三，原本的拒答訊號在長推理過程中逐步被稀釋；最後，模型可能產生不當服從，輸出偏離原有安全拒答行為的答案。此類攻擊的特性包括黑箱式攻擊、不需取得模型結構與權重，且可跨模型遷移。換言之，攻擊者不一定需要直接修改模型，只要透過設計好的有害情境與長鏈推理，就可能誘導模型做出錯誤判斷。這也顯示，長思維鏈雖可提升推理能力，但同時也可能成為新型智慧模型攻擊的重要弱點。

當思維鏈挾持發生於臨床智慧平台時，其風險將遠高於一般資訊系統。因為在醫療場域中，AI 不僅影響資訊判讀，更可能直接影響病人處置與臨床決策。若推理鏈遭受干

擾，原本應優先處理的高風險病人，可能因錯誤判斷而延誤治療，甚至影響病人生命安全。CoT 挾持的危險性在於攻擊者可利用模糊情境與長時間推理，逐步稀釋原本的安全拒答機制。例如，透過看似合理但實際帶有誤導性的問題設計，使 AI 在長鏈推理過程中逐漸偏離安全判斷，最後產生不適當回應。這類錯誤回應可能導致臨床誤導、錯誤處置，甚至形成錯誤決策。因此，在臨床智慧平台中，AI 回應若缺乏完整資訊、缺少依據與來源，或未經專業判讀，皆不宜直接作為臨床決策依據。因為思維鏈挾持的核心影響，並非單純輸出錯誤答案，而是進一步干擾醫療人員原本基於實證醫學與科學證據所建立的決策過程。此類風險同樣可能影響藥物與醫療設備應用，包括藥物劑量、給藥路徑、交互作用，以及設備設定與監測流程。若 AI 助理在推理過程中遭受 CoT 挾持，所產生的建議可能具有危害性，進一步造成不當處置。在 ICU 等高風險臨床環境中，此問題尤其嚴重。由於病情變化快速、處置流程複雜，且醫療決策需即時完成，一旦 AI 建議受到錯誤推理影響，後果可能被迅速放大。因此，高風險醫療情境中的 AI 建議，必須經專業人員覆核，並避免直接依賴 AI 作為最終決策來源。

面對思維鏈挾持風險，整體防護策略必須從推理過程本身進行監測與治理，而不能只依賴最終輸出的過濾機制。首先，需要持續監測拒答訊號。在長時間推理過程中，即使模型初期已出現安全拒答反應，攻擊者仍可能在後續推理中再次注入有害內容，使模型逐漸偏離原本安全機制。因此，防護不應只檢查最終輸出，而應持續追蹤整段推理歷程中的安全相關訊號與拒答指標。第二，必須維持對風險跨跨度的注意力。由於長鏈推理內容龐大，風險意圖可能被上下文逐步稀釋，使模型無法持續關注關鍵危險訊號。因此，需要強化模型在長推理過程中的上下文連貫性與關鍵風險 token 的持續注意能力，避免風險意圖被長篇內容掩蓋。第三，需建立推理歷程的風險控管機制。對於異常冗長或高度結構化的推理任務，應設置風險門檻與再檢查機制，避免模型在過度推理過程中逐漸偏離安全軌道。同時，也需在推理長度與推理品質之間取得平衡，避免因縮短推理而犧牲模型原有的推論能力。第四，需建立對抗式安全評測機制。由於 CoT hijack 常利用語意模糊、語意位移與複合提示進行攻擊，因此需將拼圖包裝、長鏈推理與語意混

淆等情境納入紅隊測試與基準評估，檢驗模型在複雜推理情境下的安全穩定性。最後，在部署層面上，需採取分層式治理架構，結合模型層、應用層與審查層的多重防護，避免單點式防線失效。由於思維鏈挾持利用的是推理過程本身，而非單一違規提示，因此傳統僅依靠提示詞過濾的方式，已不足以應對現代長推理模型的安全風險。

LDCT 醫療智慧模型 CoT 挾持示例

大型語言模型逐漸應用於醫療影像判讀與臨床決策輔助，但研究發現，若推理過程遭到惡意指令或錯誤資訊干擾，可能出現 CoT Hijack 現象。此類攻擊會誘導 AI 產生偏離臨床事實的推論，即使內建安全機制也可能受到影響。對醫療場域而言，錯誤的風險評估、追蹤建議或治療推薦，可能增加病人安全風險。專家建議建立多層防護機制、持續監測模型輸出，並保留醫師最終審核與決策權，以確保醫療 AI 的可靠性與安全性。人工智慧正逐步導入肺癌篩檢流程，從受檢資格評估、LDCT 影像分析、肺結節偵測到風險分層與追蹤建議，皆可提升判讀效率。然而研究指出，若大型語言模型受到不當提示或推理鏈干擾，可能在報告整合、決策說明與病患溝通等環節產生偏誤，進而影響風險評估與臨床建議。專家提醒，相關風險多發生於文字推理與決策支援階段，而非影像本身，因此高風險節點應建立監測機制並保留人工覆核，以確保肺癌篩檢結果的安全性與可信度。

大型語言模型逐漸應用於肺癌篩檢報告整理與臨床溝通，但研究顯示，看似良性的複合指令與多步驟推理，可能導致 CoT 劫持現象，使 AI 在重述結果時產生偏移。在示例中，原本 Lung-RADS 4A、建議 3 個月追蹤的肺結節案例，可能被誤解為較低風險，並延後追蹤時程。此類風險並非來自影像判讀本身，而是發生於後續文字整合與推理過程。專家建議建立輸出監測機制與人工覆核流程，避免錯誤資訊影響臨床決策與病人安全。大型語言模型若在醫療場域出現 CoT 入侵或推理偏移，影響可能不僅限於單一報告，而是沿著臨床流程逐步擴大。從結節描述失真、風險重點遺漏，到追蹤頻率與轉介建議偏差，都可能增加醫師覆核負擔與溝通成本。在肺癌低劑量電腦斷層（LDCT）篩檢中，此類偏誤更可能影響結節管理、Lung-RADS 分級及後續追蹤決策。專家指出，建立

輸出驗證、人工覆核與品質監測機制，是確保病人安全、維持臨床品質及提升醫療機構對 AI 信任的重要關鍵。

為降低大型語言模型在醫療應用中的推理偏移風險，專家提出「動態監測—回饋—治理」的持續管理架構。透過異常事件偵測、風險分級、臨床覆核、根因分析及模型更新等機制，建立可追蹤、可驗證的 AI 治理流程。同時結合放射科、臨床醫師、資訊團隊、資安與品質管理單位共同參與，強化跨領域合作。研究強調，即使影像判讀正確，文字解釋與推理內容仍可能產生偏移，因此 AI 建議必須經過人工覆核與持續監測，最終決策仍應由臨床專業人員把關，以確保病人安全與醫療品質。

以上內容將在 2026 年 6 月 10 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