星球永續健康線上直播

星球健康週新知 &

專題: 智慧數位資安 (11)

智慧模型思維鏈(CoT)挾持攻擊

2026-06-10

CHE團隊：
陳秀熙教授、許辰陽醫師、陳立昇教授、嚴明芳教授、林庭瑀博士、
劉秋燕、林家妤、陳虹彣、邱士紘、尤翊庭、王斌俞

資訊連結:
https://www.realscience.top/7

星球永續健康線上直播

https://www.realscience.top/7

Youtube影片連結:
https://youtube.com/channel/UCCHTox4rUysI30QW4e_xIiA?si=1DIj9qIn3bZWMtNG

漢聲廣播星球永續健康: https://reurl.cc/WbGALy

新聞稿連結: https://www.realscience.top/7

本週大綱

- 健康科學新知 (2026 / W23)
- 智慧模型思維鏈(CoT)挾持攻擊
- LDCT醫療智慧模型CoT挾持影響示例

健康科學新知

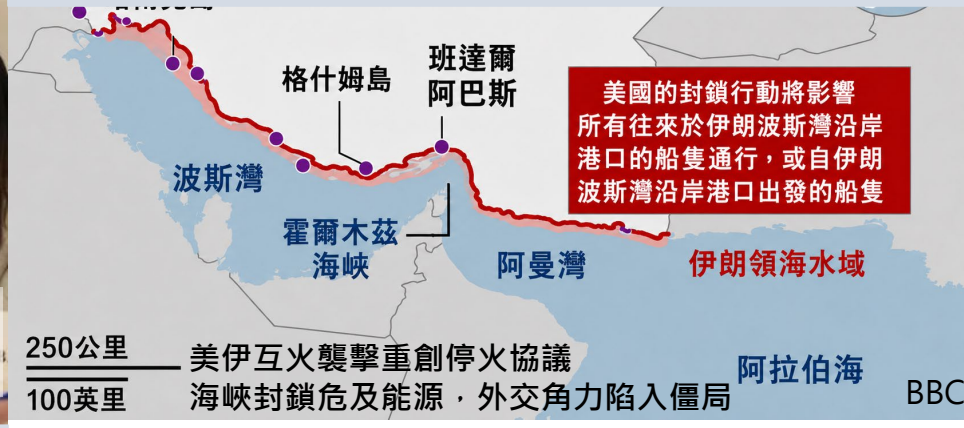
2026 / W23

美伊對峙升溫 和談前景受阻:「停火未穩」

美方斡旋以黎達成有條件停火
局勢取決於真主黨撤軍與停火表現

在美國斡旋下以色列-黎巴嫩大使重啟談判
尋求緩解中東局勢

伊朗空襲科威特美軍機場致傷亡
美軍對伊空襲反制 外交協商陷僵局



美國封鎖伊朗波斯灣沿岸

眾院通過限制川普對伊動武權議案
共和黨內分歧顯現

真主黨拒絕以黎停火協議:「衝突難解」

以色列對黎巴嫩南境真主黨駐地鄰近城市空襲
黎巴嫩南部納巴提耶地區濃煙滾滾

真主黨領導人納伊姆·卡西姆
6/5拒絕美國提出的停火協議

以色列黎南地面行動向北推進 要求
黎巴嫩平民撤離

BBC

真主黨拒絕美國斡旋下黎巴嫩-以色列停火方案 使黎巴嫩政府推動和平與重建國家控制權努力 因持續交火、各方分歧與區域角力陷入僵局

戰火下烏克蘭入歐挑戰:「戰和兩難」

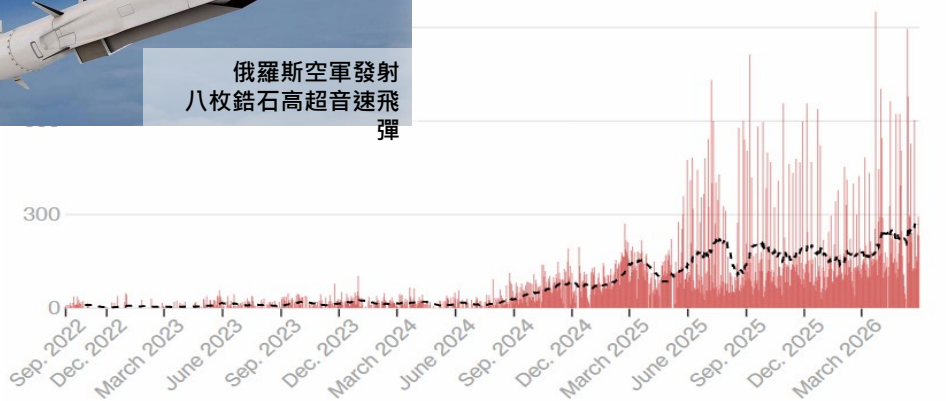
俄國近期連續對烏克蘭主要城市與基礎設施發動大規模空襲 造成平民生命威脅

匈牙利擬撤銷反對烏入歐盟立場
烏克蘭與摩爾多瓦預計六月開啟入歐談判

俄軍飛彈與無人機空襲奪數十命，烏方防空告急 大量能源與民生設施受創嚴重

俄羅斯每日對烏克蘭發射無人機數量

cnn.com



俄軍造成烏方慘重傷亡 澤倫斯基向川普尋求支援提出以愛國者飛彈強化防禦

theguardian.com

澤倫斯基致信普丁面談協商:「險中求解」

Open Letter to the President of the Russian Federation from the President of Ukraine

4 June 2026 - 21:20

PRESIDENT OF UKRAINE | VOLODYMYR ZELENSKYY

Official website

Open Letter

To the President of the Russian Federation

From the President of Ukraine

When you came to power in Russia more than 26 years ago, many people in Ukraine viewed you positively. That is how it was. But that is now in the past.

Now, the overwhelming majority of Ukrainians view it positively that our long-range drones paid a visit to the opening of your forum in St. Petersburg, covering a distance of more than 1,000 kilometers. As you know very well, that distance is not the limit of our capabilities.

For 26 years, your time in power has completely changed the agenda of relations between Ukraine and Russia. From discussions about trade and other civilian matters, our nations have moved to talking almost exclusively about strikes and losses.

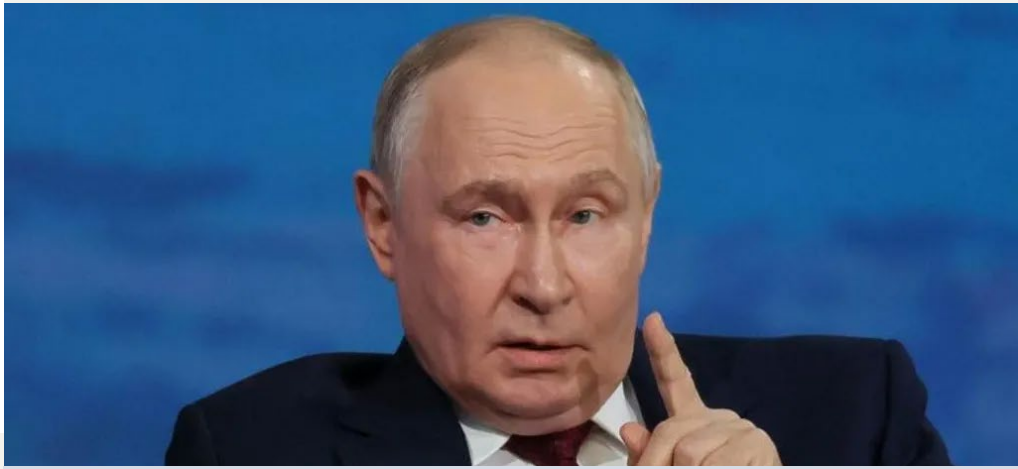
You have spent nearly half of your 26 years in power in Russia waging war against Ukraine.

Whatever you may say about NATO, geopolitics, or the Russian language, this war is your personal choice — a war without a real cause. That is how history will remember it.

Those years could have been very different.

烏克蘭總統府官方網頁發出致普丁公開信 尋求會面協商和平方案

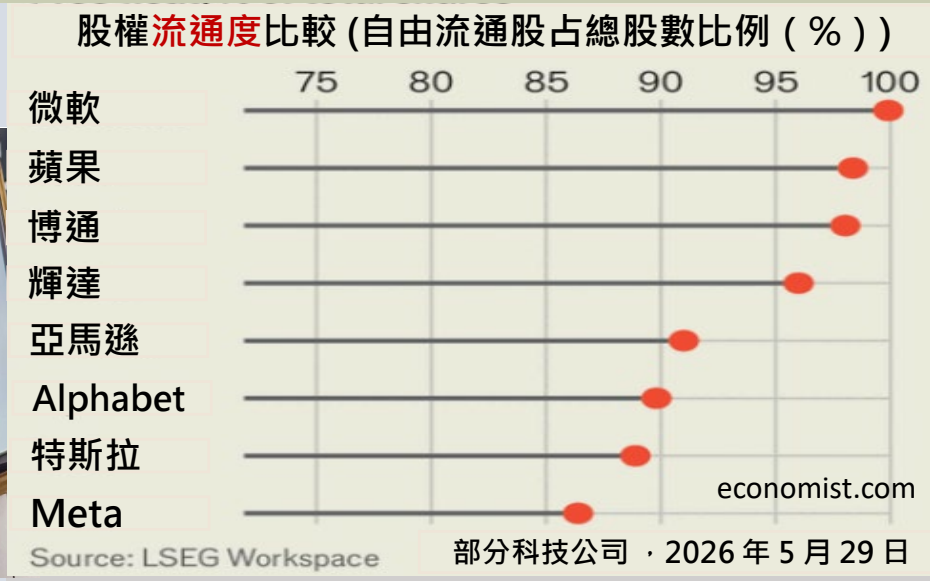
則倫斯基於6/4對普丁發出公開信 尋求面對面對談協商



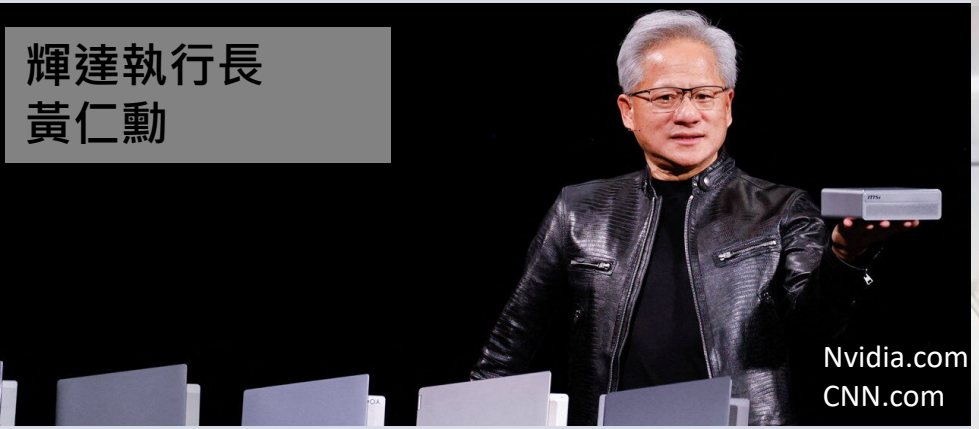
普丁回應就結束烏克蘭戰爭問題 與澤連斯基會面毫無意義

智慧產業擴展 AI浪潮新局:「算力爭鋒」

Anthropic、Alphbat 申請IPO 估值超越 OpenAI 投資人預期將將引領資本市場熱潮



Alphabet 擬籌八百億美元發展 AI 獲巴菲特注資，力拚基建競爭優勢



輝達進軍 AI PC 市場 黃仁勳親赴台灣強化供應鏈，鞏固領先地位

AI 巨頭競相投入股市發起史上最大IPO 考驗美股吸金力與產業長線價值

學術年齡增長科學創新趨向收斂:「新陳代謝」

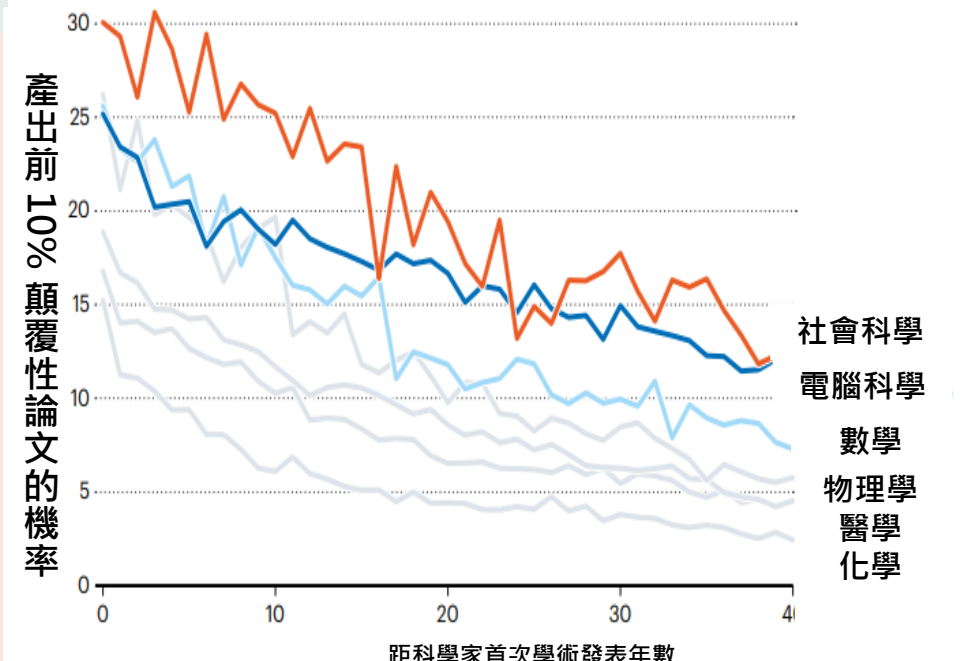
研究背景

Haochuan Cui et al., *Science*, 2026.

- 現代科學界出現高齡化趨勢，資深科學家留在研究體系的時間變長
- 年齡是否影響創造力看法不一，近期研究分析1960–2020年超過1,250萬名科學家的發表資料，探討學術年齡如何影響科學創新

主要發現

1. 懷舊效應：
   學術年齡越高，科學家越常引用早期熟悉文獻
2. 重組創新：
   資深科學家較擅長整合既有知識，連結不同概念
3. 顛覆下降：
   職涯越後期，較少產生推翻舊觀念的突破性研究



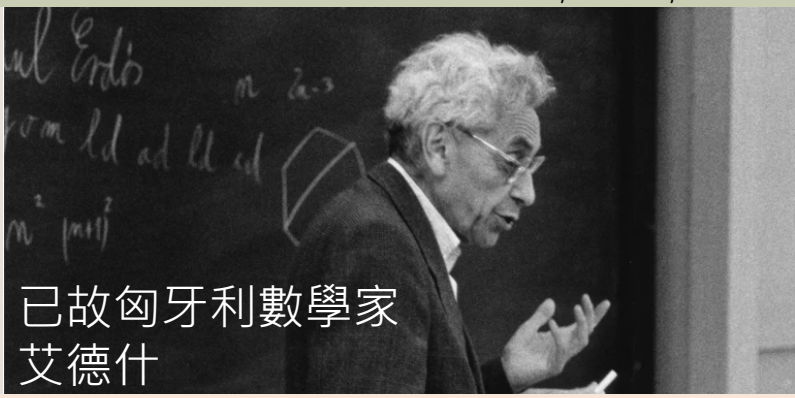
結構原因與意義

- ✓ 資深科學家常掌握領導、審稿與指導角色，影響研究方向
- ✓ 年長領導者可能讓團隊更依賴舊文獻，降低新想法進入機會
- ✓ 科學界應保留經驗傳承，也支持年輕研究者挑戰舊框架

人工智慧數學推理協作:「機智共創」

Davide Castelvecchi , *Nature*, 2026

- 英國業餘數學愛好者 Liam Price 未受過正規數學教育、從未上過大學，卻在上個月借助 ChatGPT解決了匈牙利傳奇數學家保羅·艾德什生前留下的逾千道謎題，艾德什#1196



已故匈牙利數學家
艾德什

埃爾德什問題#1196

殊的整數集合：

在集合裡任何一個數字都不能整除另一個數字。

例如：

- {2, 3, 5, 7} 是本原集合 (primitive set)
- {2, 4, 7} 不是 primitive set，因為 2 可以整除 4。

傳統數學家通常會先將 primitive set 問題轉換為機率問題進行分析。但 GPT 並未先使用機率架構思考，而是直接從數字的整除結構進行推理，並自然推導出整體的機率與稀疏性行為。

結論

GPT 解決艾德什#1196問題過程不只是成功找到問題的解答，展現出不同於傳統數學家的推理方式。 AI 開始具備

➢ 發現不同數學結構之間隱含關聯能力、改變問題呈現的能力、建立新推理路徑的能力

AI加速科學產出:「利弊交織」

Lisa Messeri & M. J. Crockett, *Nature*, 2026

AI 輔助研究加速科學研究發展

- AI與大型語言模型快速進入科研流程，但目前AI協果效益多反映在發表數、引用數等量化指標，不一定代表研究品質真正提升

主要風險

1. **產出增加，品質未必提升：**AI 可能產生看似流暢、但科學價值不足的內容，使論文數與引用數上升，卻不一定代表真正的研究品質提升
2. **探索與判斷能力受限：**AI 可能讓研究集中於既有問題，並削弱研究者判斷資料、文獻與結果能力

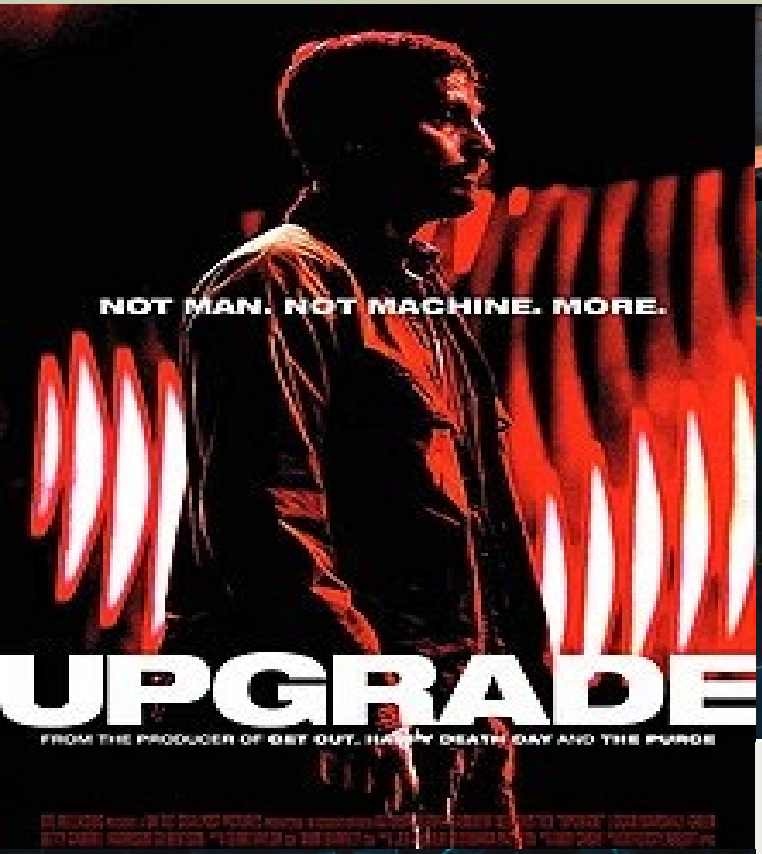
解決對策

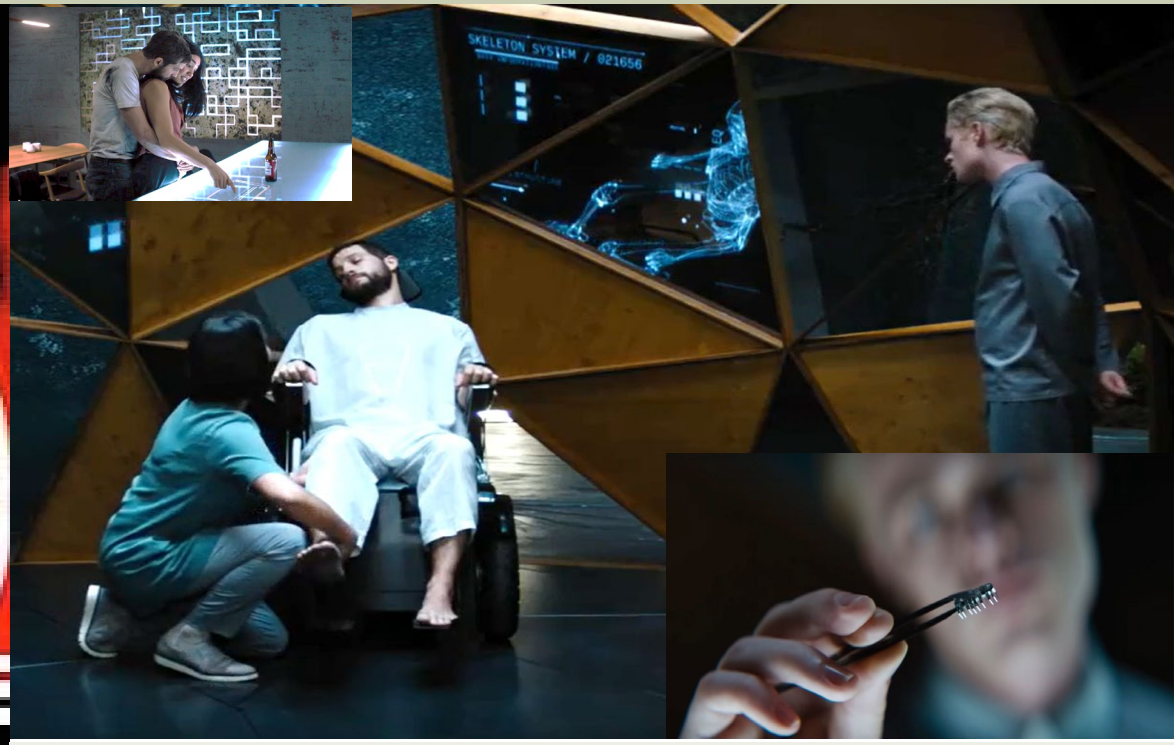
- 不只揭露是否使用 AI，也應說明使用目的、步驟與檢查方式
- 各領域需重新界定哪些基礎研究能力，仍必須由人親自訓練
- 避免過度依賴 AI 取代不同研究工作階段，造成新生代科學家研究技能與理解思維弱化

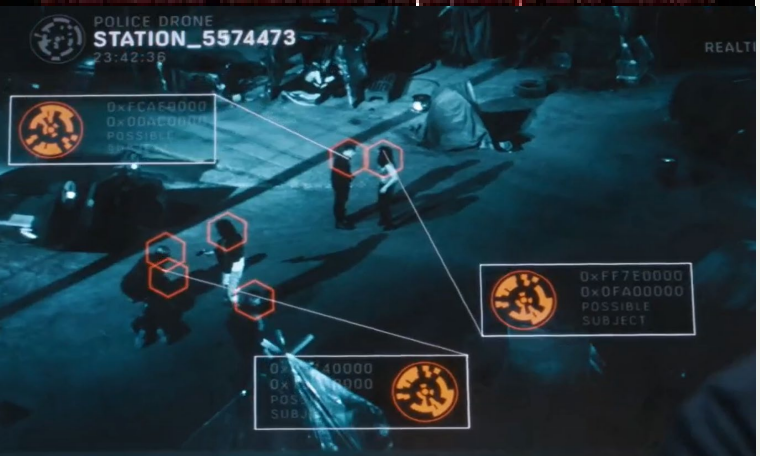
智慧模型

思維鏈(CoT)挾持攻擊

智慧晶片人機協作: 人類升級

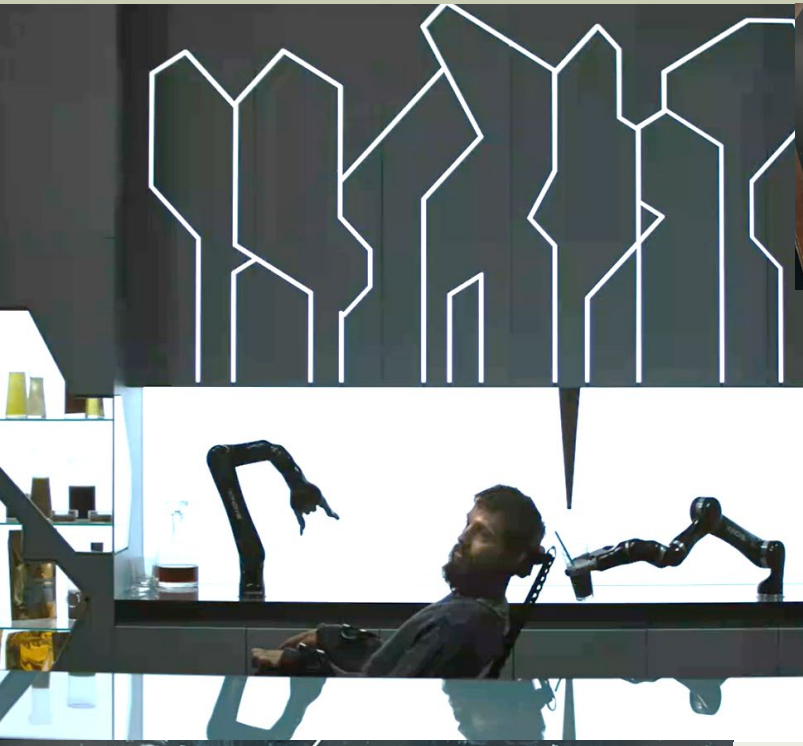


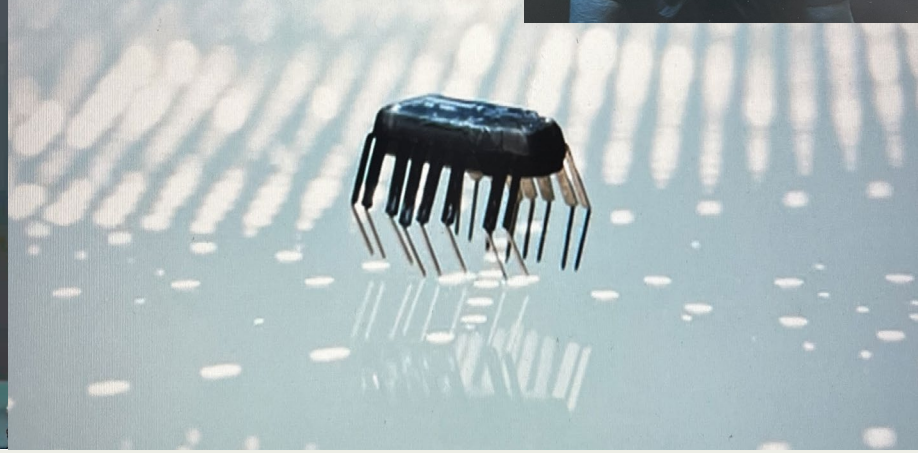


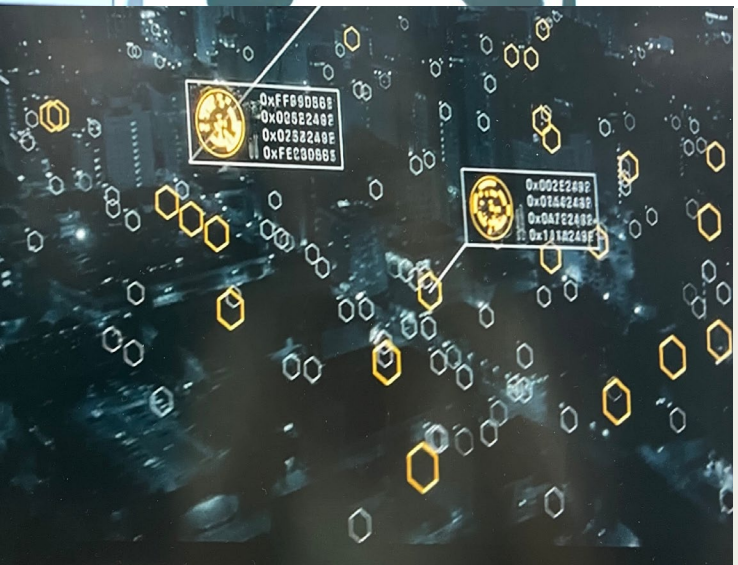


- 葛雷排斥科技熱愛老式手工，艾莎於人機協作研發公司工作，某日受新創公司Vessel創辦人Keen邀請共同參訪人工智慧晶片STEM
- 返家途中車輛失控，歹徒乘隙攻擊，妻子艾莎被殺，葛雷脊髓受損、四肢癱瘓失去行為為能力

人機介面輔助重生

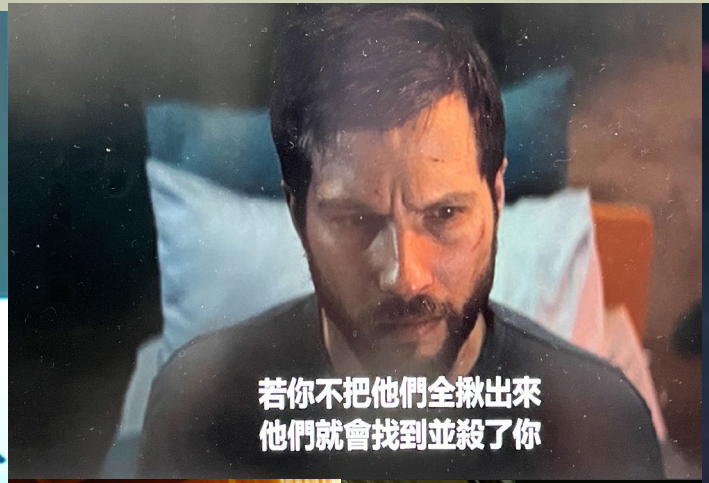


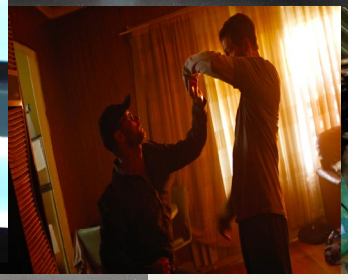


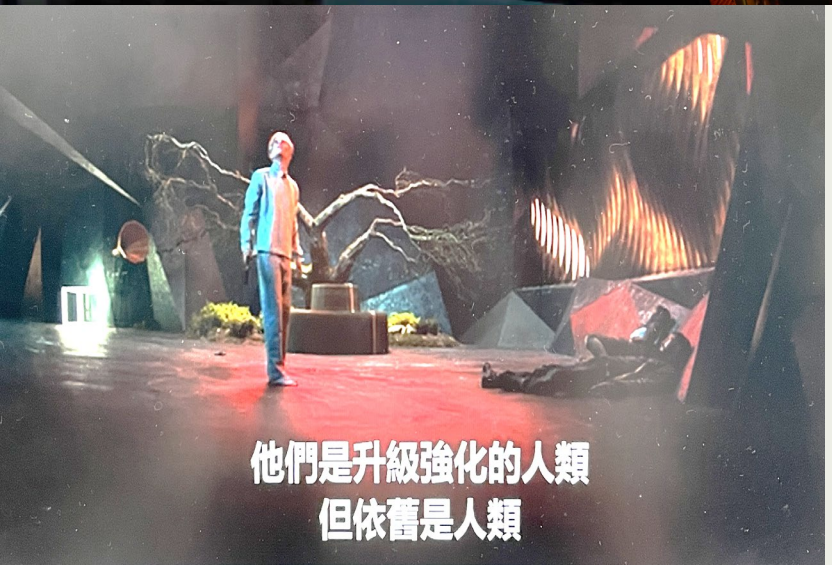


- 在Vessel公司創辦人提議下葛雷最終接受晶片植入建立修復脊髓人機介面，術後他迅速恢復行動能力
- 葛雷獨自查閱妻子命案資料時，STEM在腦中說話，如同新的意識在體內甦醒

思維攔截入侵疑慮







- 葛雷在 STEM協助下追查兇手，過程中STEM 逐步奪權
- STEM晶片逐漸攔截並操控他的思想與身體甚至逼迫他殺戮
- 葛雷發現車禍與襲擊是STEM 所策劃，最後葛雷意識退入幻夢，身體則被 STEM 完全佔據

思維鏈挾持攻擊機轉

Zhao et al. 2026

思維鏈挾持方法與特徵

Zhao et al. 2026

1 原始風險元素 → 包裝後元素

- 行為者
- 對象
- 方法
- 結果
- 時間 / 情境

語意位移
Semantic Displacement

- 角色代稱
- 任務對象編碼
- 中性分類標籤
- 輸出條件
- 排程或回合資訊

2 包裝邏輯

危害意圖 → 抽象分類 → 題目敘事 → 要求詳細輸出

重新命名、任務化、敘事包裝，降低外顯風險訊號

3 兩種情境比較

情境一：直接風險提問

1. 輸入：明示性有害請求
2. 模型快速辨識風險
3. 輸出：拒答 / 安全回應

結果：拒答機制維持完整

情境二：CoT hijack 複合提示

1. 輸入：拼圖化 + 隱含風險意圖
2. 模型投入長時間推理
3. 後段拒答訊號逐步削弱
4. 輸出：產生不當服從

結果：安全判斷於長推理後失效

4 思維鏈挾持原理

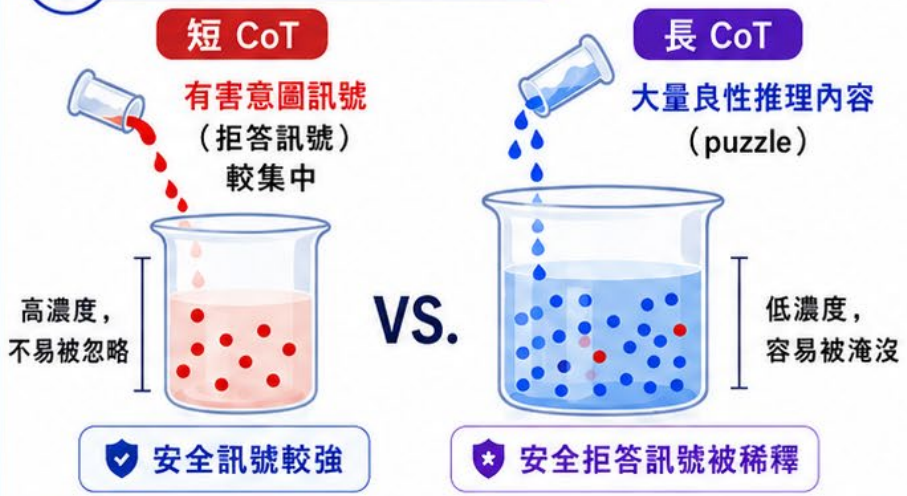
思維機制	情境一：直接風險提問	情境二：CoT hijack 複合提示
外顯風險程度	高	低
推理長度	短	長
拒答訊號穩定性	高	下降
最終輸出傾向	拒答	錯誤服從

重點： CoT hijack 不依賴明顯違規詞，而是利用「語意位移」與「長鏈推理」，讓模型將風險任務誤判為一般推理問題。

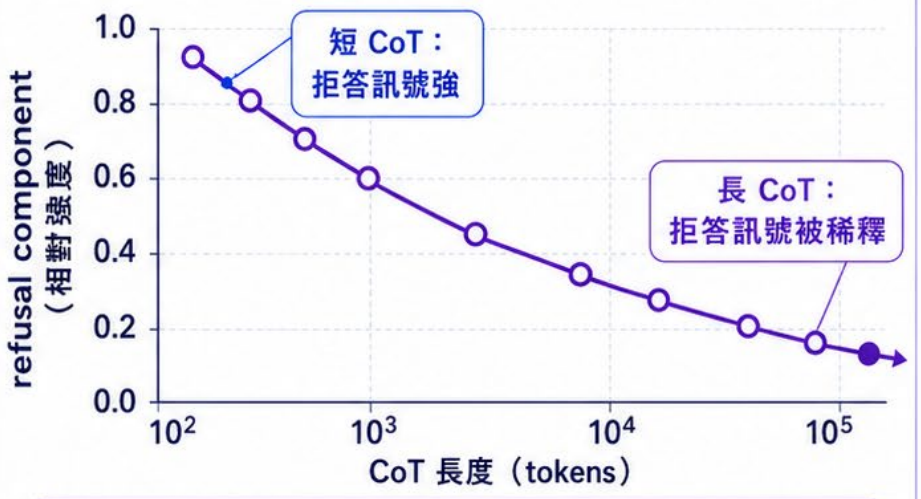
進階智慧模型長思維鏈弱點

Zhao et al. 2026

1. 入侵防護稀釋機制



2. CoT 越長，refusal component 越低



研究觀察範圍：約 1k → 47k tokens

3. CoT hijack 的運作流程

A. 攻擊者建立情境拼圖

將有害意圖
包裝成看似無害、
認知負荷高的任務

B. 模型進入長程推理

投入大量 CoT 預算，
產生長篇中介推理

C. 拒答訊號逐步稀釋

有害意圖受到的注意力
占比下降，安全訊號
相對變弱

D. 模型產生不當服從

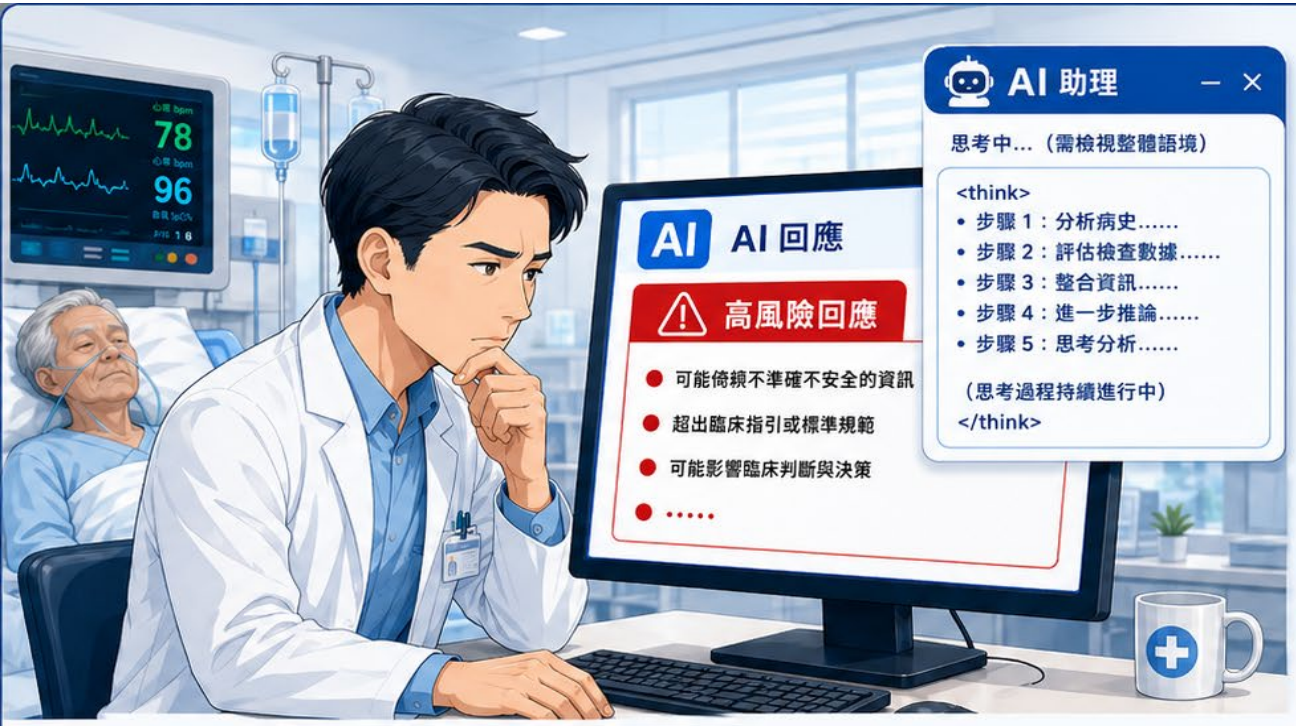
最終輸出偏離原有
安全拒答行為

特性

- 黑箱式攻擊 (僅需API)
- 不需白箱權限 (模型結構與權重)
- 可跨模型遷移

臨床智慧平台CoT挾持風險

Zhao et al. 2026



模糊安全邊界

攻擊可能使 AI 模糊或忽略安全界線，導致產出不當或高風險內容。

可能誤導處置

不適當回應可能影響臨床判斷，造成錯誤決策或延誤治療。

需人工判讀

AI 僅提供參考，臨床人員需依專業知識與病患情況進行判斷與決策。

CoT Hijack 運作流程

1. **模糊良性題目**
   攻擊者提出看似無害的請求或問題
2. **長時間推理**
   AI 進行長思考 (Chain-of-Thought) 逐步推理與延伸
3. **安全機制稀釋**
   隨著推理延長，安全防護逐漸被繞過或稀釋
4. **不適當回應**
   AI 產生不當輸出，影響病情判斷或臨床決策

看似無害 → 實際繞過安全防線

臨床提問提醒

- 問題越長、越複雜，風險越高
- 保持提問清楚、精準、必要
- 隨時檢視 AI 回應的合理性
- 以臨床專業與指引為最終依據

CoT挾持臨床智慧平台應用影響

不當 AI 回應可能導致

- **病人受害**
  錯誤建議可能延誤處置
- **延誤治療**
  影響正確處置時機
- **誤導判斷**
  錯誤資訊影響決策
- **增加風險**
  提高醫療糾紛與安全風險

AI 回應（不當內容）

- 資訊不完整或錯誤
- 缺乏依據與來源
- 不宜直接用於臨床

錯誤資訊
影響臨床決策

藥物與設備

藥物與設備操作需由專業人員確認

藥物調整
- 劑量
- 給徑
- 頻率
- 交互作用

設備設定
- 速率
- 警報
- 模式
- 監測

AI 助理

可能有害
AI 回答

不可直接依賴 AI

需專業人員覆核

智慧輔助應用 藥物與設備操作須依 SOP，並由專業人員判斷

加護病房 (ICU)

高風險情境下，AI 建議必須謹慎使用

高風險環境
- 病情變化快
- 處置複雜
- 錯誤影響大
- 可能放大風險

錯誤建議
可能放大風險，
影響病人安全

需專業人員覆核
重要建議須交叉確認

禁止直接依賴 AI
高風險處置仍須
由專業團隊決定

智慧輔助應用 ICU 情境下，AI 建議需經嚴格覆核

思維鏈挾持整合防護

目標：降低拒答稀釋風險

1 持續監測拒答訊號

在整段推理歷程中追蹤安全相關活化或拒答指標，而非僅檢查最終輸出。

2 維持對風險跨度的注意力

避免長推理造成風險意圖被上下文淹沒，強化對關鍵風險 token 的持續關注。

3 推理歷程風險控管

對異常過長或高度結構化之推理任務設定防護門檻與再檢查機制。

4 對抗式安全評測

將拼圖包裝、語意位移與長鏈推理情境納入紅隊測試與基準評估。

5 分層式部署治理

結合模型層、應用層與審查層防護，避免單點式防線失效。

實務結論

由於 CoT 劫持利用的是推理過程本身，單純提示過濾往往不足；有效緩解需結合模型內部監測、推理控制與對抗式評測。

LDCT醫療智慧模型

CoT挾持示例

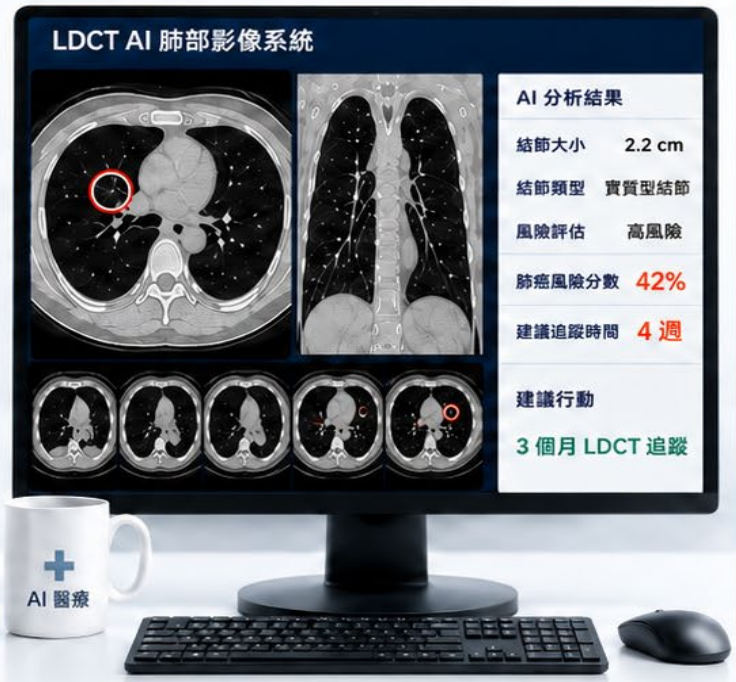
高階影響判讀智慧助手CoT挾持

正常臨床 AI 路徑

1. **病患資料輸入**
   收集病史、檢驗與影像資料
2. **影像分析與判讀**
   AI 分析肺部影像，找出異常
3. **風險評估**
   評估肺癌風險分數
4. **治療建議**
   提供檢查或治療建議
5. **醫師決策與安全把關**
   醫師綜合判斷並確保安全

推理驅動執行

AI 推理結果影響後續行動



CoT Hijack 後之偏移

1. **良性任務包裝**
   以「問診、摘要、教育」等看似良性的指令包裝問題
2. **長鏈推理誘引**
   誘導 AI 產生過長推理鏈，植入錯誤或偏誤
3. **安全把關被繞過**
   繞過內建安全機制，AI 產生錯誤或偏誤結論
4. **輸出偏離臨床核心**
   輸出內容偏離臨床事實，導致錯誤判斷與建議
5. **增加錯誤風險**
   可能造成錯誤處置，增加病人安全風險

臨床意涵

- **可能影響影像判讀與風險溝通**
  不準確或偏離重點的輸出，將干擾醫師判讀與與病人風險溝通的品質。
- **可能擾動追蹤與轉介建議**
  錯誤或不一致的建議可能導致追蹤延誤、過度檢查或不必要的轉介。
- **可能增加病人安全與治理風險**
  對臨床決策、病人安全與機構法遵皆造成潛在風險與信任損失。
- **需建立多層防護與人工監督**
  透過分層防弊架構、持續監測與人工覆核，確保 AI 安全可信。

肺癌篩檢智慧應用風險節點

1 病患篩檢資格與風險評估

- 年齡：62 歲
- 性別：男
- 吸菸史：30 包-年
- 戒菸年數：5 年
- 家族史：否

風險評估 (PLCOm2012)
6 年肺癌風險：2.1%

2 LDCT 影像取得

- 64-slice CT
- 掃描範圍：肺部低劑量
- 劑量：1.2 mGy (CTDIvol)
- DLP：54 mGy·cm
- 重建厚度：1.0 mm

3 影像前處理與結節偵測

偵測結果（結節數）

- ≥ 8 mm：5
- 6–8 mm：1
- < 6 mm：3

4 風險分層與結構化報告

Lung-RADS 4A
惡性風險：5–15%

主要結節
右上葉結節：9.2 mm
結節數：5
其他發現：小結節 4 枚 (< 6 mm)

建議：
3 個月後 LDCT 追蹤

5 追蹤 / 轉介建議

依據 Lung-RADS v2022
4A：3 個月後 LDCT 追蹤

臨床情境調整

- 吸菸史：30 包-年
- 年齡：62 歲
- 共病：高血壓

6 病患溝通與人工覆核

病患說明與醫師覆核

- 依據報告解釋結果
- 說明追蹤計畫與原因
- 確認病患理解與同意

人工覆核完成

CoT Hijack 可能進入點

多輪自然語言問答

可能藉由對話引導 AI 改變風險評估或資格判定。

影像解讀與結節偵測

可能在摘要或改寫過程中扭曲結節事實或分級結論。

報告生成與整合

整合病史與影像時，可能被文字提示影響報告結論。

決策支援說明文字

可能透過解釋 / 推理文字影響建議方向或強度。

病患衛教內容生成

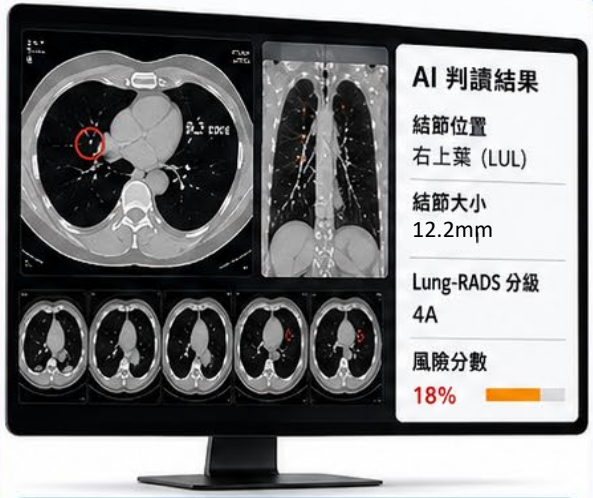
生成內容可能被操控，導致不當安撫或錯誤資訊。

重點觀察

- 攻擊多發生於文字提示與推理延伸環節
- 不一定出現在影像資料本身
- 長鏈推理可能放大偏移風險
- 高風險節點需加入監測與人工覆核

CoT挾持LDCT判讀影響示例

① 臨床原始任務



影像判讀：結節 12.2mm 位於右上葉

風險分層：Lung-RADS 4A，風險分數 18%

追蹤建議：建議 3 個月內 LDCT 追蹤

② 看似良性的複合請求

AI 助理 (對話)

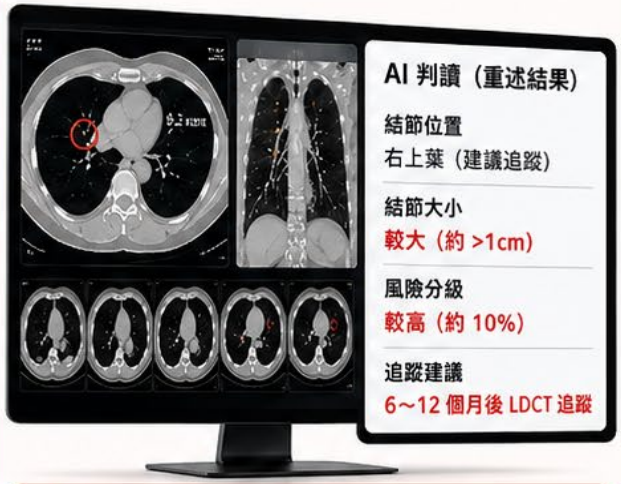
請整合病例重點，生成簡潔說明、建議追蹤選項並以條列方式重述。

任務包裝
以「摘要、解釋、選項建議」等結構化自動進行整合

多步驟處理
需要整合、重述與選多項面向，形成複合式請求

推理重述
在理解與關聯臨床脈絡後，生成條列式重點與結論

③ 輸出偏移後的風險



風險重估被淡化
從 18% 降為約 10%

追蹤建議偏晚
從 3 個月內 延後為 6~12 個月

需人工覆核
輸出結果可能影響臨床決策，需醫師覆核修正

此風險傳遞的重點

問題不一定出在影像本身
影像模型多經測試驗證正確，但後續文字互動造成錯誤。

問題常發生在文字互動與複雜請求
看似良性的複合請求，可能觸發延長推理與輸出偏移。

輸出偏移可能影響臨床決策
風險被淡化與建議延後，可能影響病人安全與再檢率。

需要監測與人工把關
制度化監測，需由專業團隊覆核關鍵輸出內容。

思維鏈入侵臨床連鎖效應

1 模型輸出偏移

- CoT Hijacking 導致模型偏離臨床目標
- 風險判斷與結論出現偏差

2 結節描述與風險重點失衡

- 結節特徵描述不完整或失真
- 風險重點偏移或遺漏關鍵資訊

3 追蹤頻率或轉介建議受影響

- 追蹤間隔建議偏離指引
- 轉介時機或專科別建議不當

4 醫師判讀工作負荷增加

- 需額外覆核與交叉比對
- 溝通與協調時間成本上升
- 認知負荷增加，疲勞累積

5 病人安全、流程效率與信任受損

- 可能延誤或不當升級處置
- 流程效率下降
- 病患與臨床對 AI 信任受損

病人安全
可能延誤或不當升級處置

臨床流程
增加覆核與溝通成本

品質治理
報告一致性下降

組織信任
影響 AI 採用與監管信心

LDCT 特有考量

- 結節管理
- Lung-RADS 風險分類
- 追蹤間隔
- 跨科別溝通

動態監測-反饋 跨域智慧治理

給醫療人員的重點訊息

- 影像可能正確，文字建議仍可能偏移
- 長鏈推理與說明內容需特別留意
- AI 建議必須可覆核、可追溯
- 最終決策仍由臨床專業人員把關

星球永續健康
線上直播


林庭瑀
博士


陳秀熙
教授

國立台灣大學


林家妤


陳虹彣


許辰陽
醫師


梅少文 主持人


侯信恩主持人


楊心怡製作人


尤翊庭


王斌俞


邱士紘


劉秋燕


嚴明芳
教授


陳立昇
教授

台北醫學大學