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 (6)

AI生命週期資安攻擊防禦

2026-05-06

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、羅崧瑋、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 / W18)
- AI對抗性資安生命週期攻擊與防禦
- AI生命週期對抗性攻擊實例

健康科學新知

2026 / W18

川習會緊鑼密鼓牽動東亞局勢:「大國試探」

日本外長與澳洲外長上週就能源互助與西太平洋區域安全進行協調面談

美國與菲律賓上週於台海區域巴丹島部署反艦飛彈系統 進行聯合演習

日本上週與菲律賓及美國於南海區域進行聯合軍演 澳洲、紐西蘭、加拿大、法國、英國亦參與

川習會預計於5月中旬進行，近日中國外長致電國務卿盧比奧表達台海穩定為關鍵議題

中東動盪持續全球能源受阻:「以封制封」

伊朗提重開荷莫茲海峽，換美國解除封鎖並結束戰爭，但核問題延後使美方持保留態度

美國封鎖使伊朗油運受阻，載油船被迫折返，也讓荷莫茲海峽成為美伊停火談判核心

停火框架下以軍仍以真主黨火箭與無人機威脅為由持續擴大對黎巴嫩南境打擊

黎巴嫩戰線雖有停火框架
但邊境地帶仍持續遭受軍事破壞

破壞集中於黎巴嫩南部邊境地帶

皇室外交重整跨大西洋關係:「同盟重估」

查理三世國王和卡蜜拉王后
上周赴美國事訪問

英王訪美籲團結，嘗試化解川普與英首相
因伊朗戰事所產生外交僵局

英方派團進行國事訪問，英王查爾斯積極
拉攏川普，以穩定日益緊張美英關係

美歐深化礦產合作，擬設價格下限
以削弱中方壟斷，確保供應鏈韌性

歐盟貿易和經濟安全專員
馬羅斯·塞夫科維奇

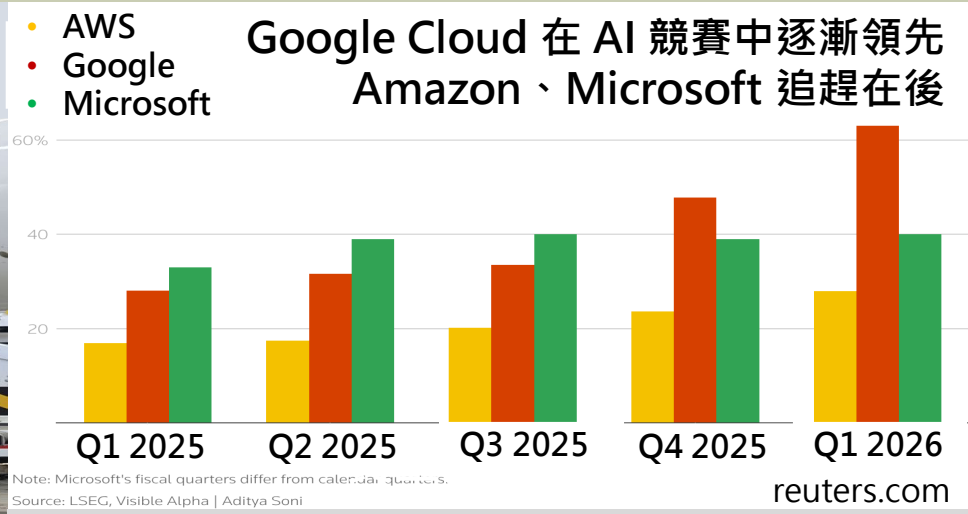
北約擬減峰會頻率以避川普，
應對防務支出與美伊緊張情勢。

北約秘書長馬克·呂特

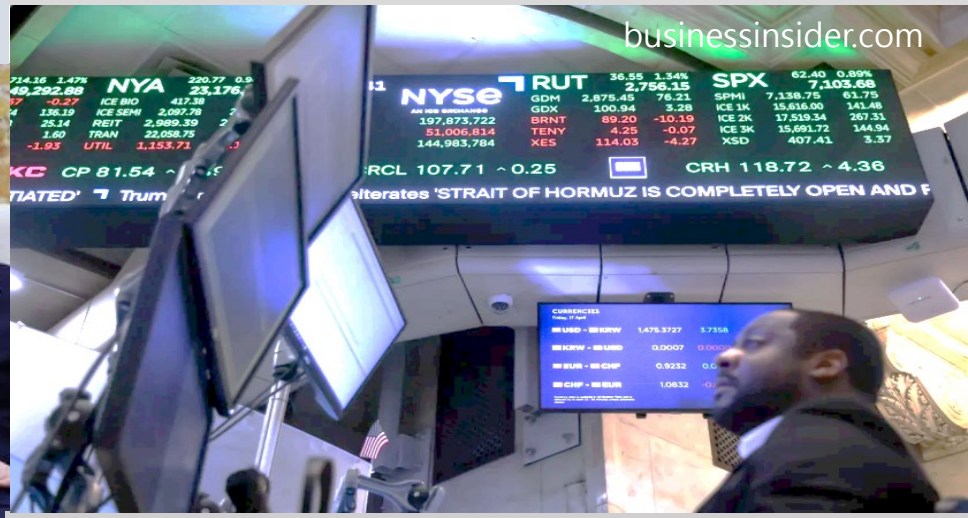
全球擴大AI戰略布局:「算力入世」

日航羽田試行機器人搬運，緩解缺工壓力並減輕員工體力負擔

Google於南韓設首座海外AI研發區 聯手K-Moonshot深化東亞科研



科技巨頭年投注六千億美元於AI產業 投資人關注龐大支出能否換取實質增長



美國金融業指 AI 生產力擴展將抵銷戰火通膨 使市場在油價衝擊下持續成長

AI擴展面臨硬體供應瓶頸:「矽限初現」

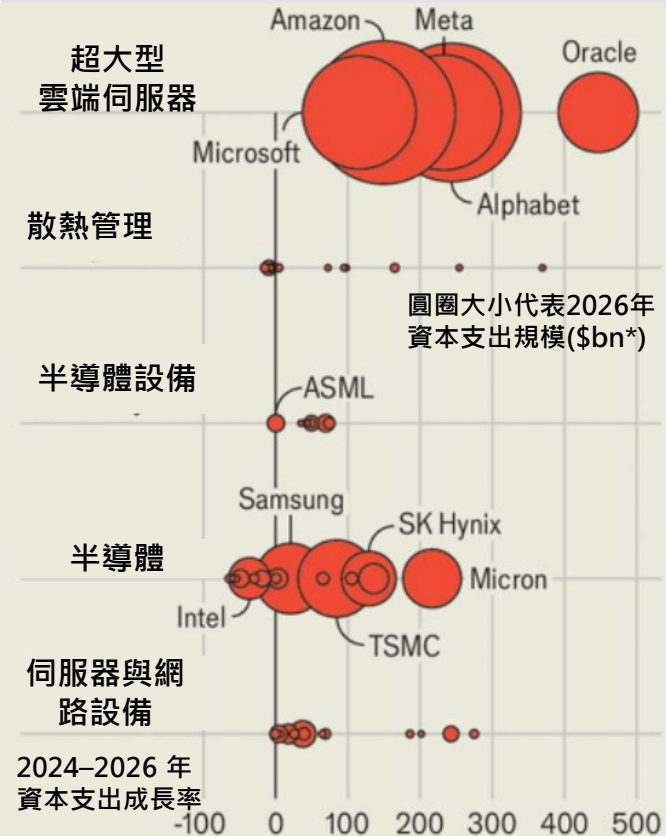
The Economist, 2026

AI 算力消耗快速上升

- 需求升溫： AI 使用量與企業需求快速增加，token 處理量短期大幅成長
- 供給吃緊： 多家 AI 服務已限流或暫停訂閱，科技巨頭加速擴建資料中心

供應瓶頸

- 基礎受限： 資料中心因高耗電與電價疑慮，擴建面臨地方反對與推進阻力
- 硬體吃緊：GPU、HBM記憶體與CPU供應短缺，Agentic AI進一步拉高CPU需求



產業落差：軟體需求快，硬體擴產慢

- 投資落差： 雲端巨頭大幅提高資本支出，但硬體供應商擴產步調相對保守
- 產能滿載： TSMC 先進製程已接近滿載，新晶圓廠仍需 2～3 年才能投產
- 時程錯位： AI 軟體需求可在數月內暴增，但晶片、記憶體與供應鏈擴張需數年

資料中心大擴張，供應鏈投資跟不上

- 左表顯示，AI 供應鏈投資明顯失衡，雲端巨頭資本支出大增，但半導體、伺服器與散熱等硬體供應商擴產較保守。

美國科研資源戰略轉型:「資源傾斜」

Dan Garisto , *Nature*, 2026

- 美國國家科學基金會（NSF）向年輕研究人員頒發了創紀錄的 2599 項享有盛譽的研究生獎學金
- 該計畫的延續體現了川普政府「對人才培養和投資個體研究人員的高度重視」

- 近14,000名青年研究人員申請了2026年研究生研究獎學金（GRFP），通常只有約六分之一的申請者能夠獲得這項享有盛譽的獎學金。
- 今年2月，美國國家科學基金會（NSF）領導層在理事上宣布將計劃重組該機構，以資助更多量子科學和人工智慧領域的研究

- 最新的研究生獎學金（GRFP）中，有 53 項被歸類為量子科學領域，比前一年增加了39%；另有103項被列為人工智慧或機器學習領域，增加了17%。

量子電腦數位模擬助攻應用突破:「虛實對證」

研究核心突破

Davide Castelvecchi, *Nature*, 2026

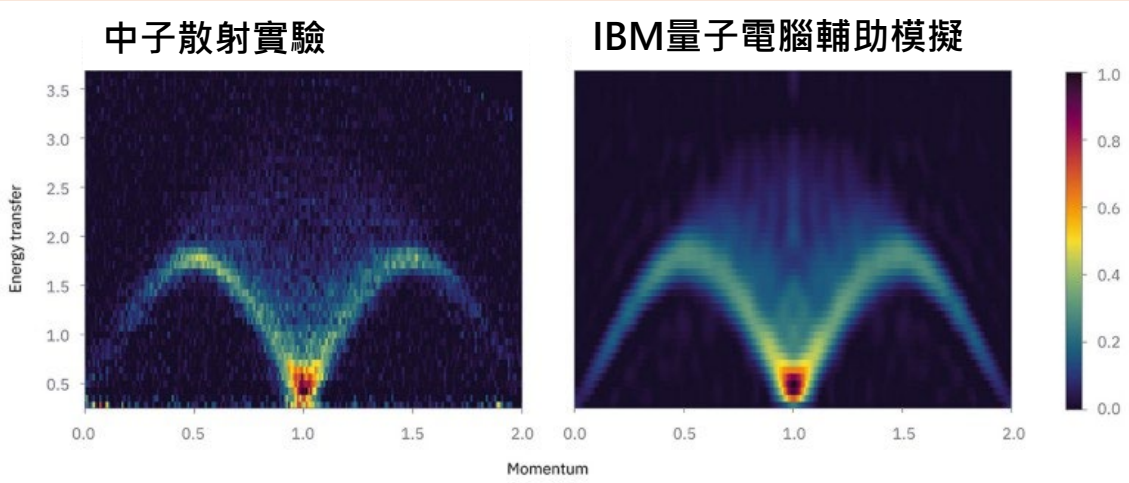
- 首次將量子電腦模擬結果與真實材料實驗數據比對成功
- 證明量子模擬結果「可信且可驗證」
- 建立量子計算邁向實用的重要里程碑

技術方法

- 兩種量子模擬：
  - 類比量子模擬（Analog）：操控原子狀態（Pasqal）
  - 數位量子模擬（Digital）：超導電路（IBM）
- 模擬材料特性：
- 熱容量、磁性反應、能階變化
- 與中子散射實驗數據進行比對

研究結果

- 模擬結果與實驗數據高度一致
- 部分運算已超越傳統超級電腦
- 未來可應用於：
  - 新材料設計
  - 化學反應預測
  - 藥物開發

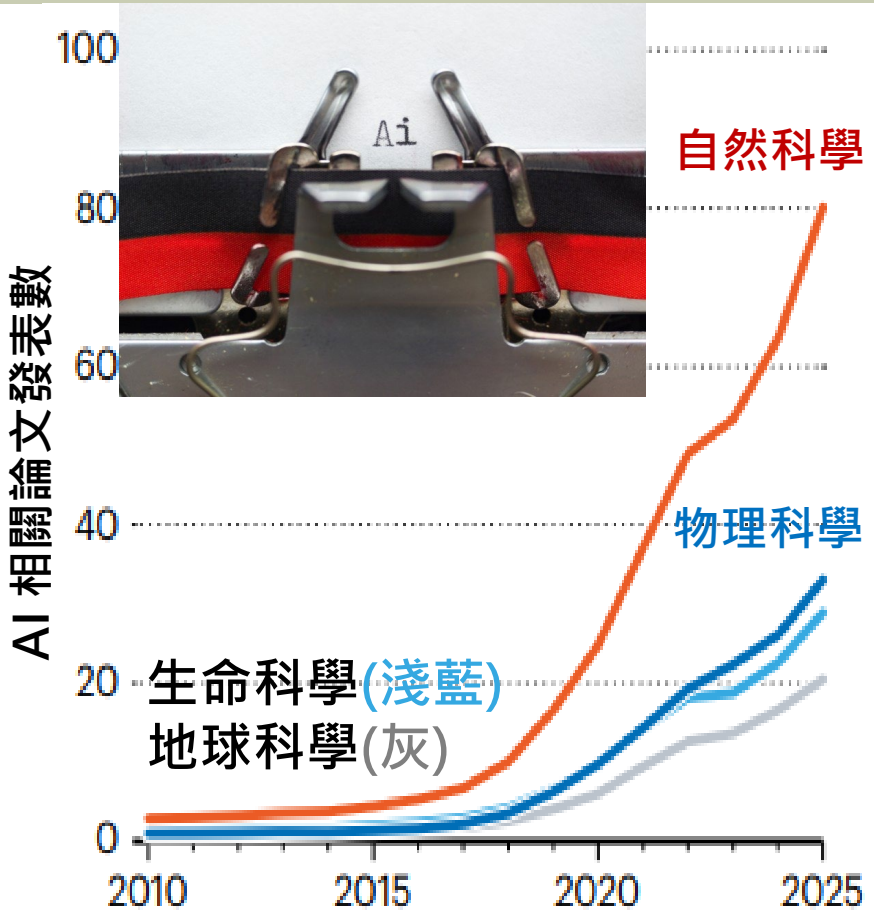


人類於複雜學術任務完勝AI：「知識膨脹」

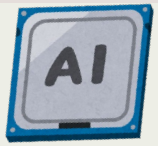
Nicola Jones, *Nature*, 2026

AI 論文爆發式成長

- 數據狂飆：自然科學領域提及 AI 出版物在 2010 至 2025 年間增長了近 30 倍
- 領域分布：在自然科學中，AI 相關論文以物理科學數量最多，地球科學佔比最高

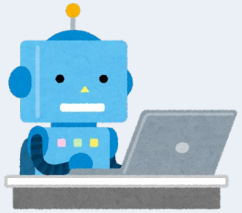


表現能力與局限



- 效能落差：頂尖 AI 代理在處理複雜多步驟工作流時，表現僅約為博士級人類專家一半
- 低精準度：在 Paper-Arena 測試中，即便具備推理計畫能力，AI 準確率也僅有 39%

AI角色與學界爭議



- 專家擔心 AI 發展過快使規範來不及調整，可能造成研究品質斷崖式下滑
- 目前 AI 更多扮演「高效助手」而非「獨立科學家」，在完全自動化發現之前仍有長路

AI對抗性資安
生命週期攻擊與防禦

夢境成真: 攔截記憶碼

- 攔截記憶碼描述未來世界分為居民區與殖民區
- 奎德為機器士兵工廠裝配員在日復一日工作中經常作夢
- 夢中自己是探員(Agent)與另一女子逃脫受困情境

間諜過往訓練記憶活化

- 同事介紹下奎德造訪記憶碼公司(ReKall)體驗間諜刺激生活
- 記憶掃描過程偵測奎德腦中已存有間諜記憶層造成過載並觸發間諜本能制服武裝人員

虛實交替 記憶回溯

- 奎德逃回家後發現妻子是派來監視他的探員，工人奎德身份是記憶植入結果
- 逃亡途中與夢境女子梅琳娜相遇，並發現自己錄製的警告訊息逐漸喚醒過往記憶

穿越記憶防火牆攔截密碼

- 過往身分豪瑟記憶逐漸恢復，並於逃難小屋在過往記憶引導下觸發豪瑟全像影片，發現自己是雙面間諜
- 豪瑟發現腦中存有癱瘓機器大軍程式碼可挽救殖民區免於受摧毀

AI生命週期與對抗攻擊威脅

Wu et al., 2026

可能被極小干擾誤導

極小且難察覺的惡意雜訊，可能讓相似影像被模型判成不同結果，顯示其判斷存在可利用弱點

異常不易被察覺

深度學習模型的判斷過程不透明，出現偏離常識的結果時不易察覺，也讓人更難信任它的判斷

風險遍布各階段

AI風險不只出現在推論時，從資料準備、模型訓練、部署到使用，各階段都可能成為攻擊切入點

流程中常有漏洞風險

實務流程常含未驗證資料與可操控環節，讓攻擊者得以在訓練、部署或推論階段伺機動手腳

智慧AI醫師LDCT判讀對抗性機器學習挑戰

受檢者
具有呼吸道症狀之個體

低劑量電腦斷層掃描
低劑量電腦斷層檢查

人工智慧初步判讀
深度學習模型進行影像分析

醫師覆核
放射科醫師進行專業覆核

早期確診
及早發現肺部病灶

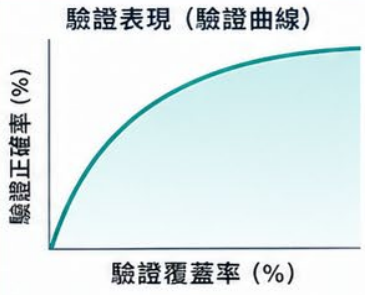
傳統模型表現檢視指標

實驗室研究報告

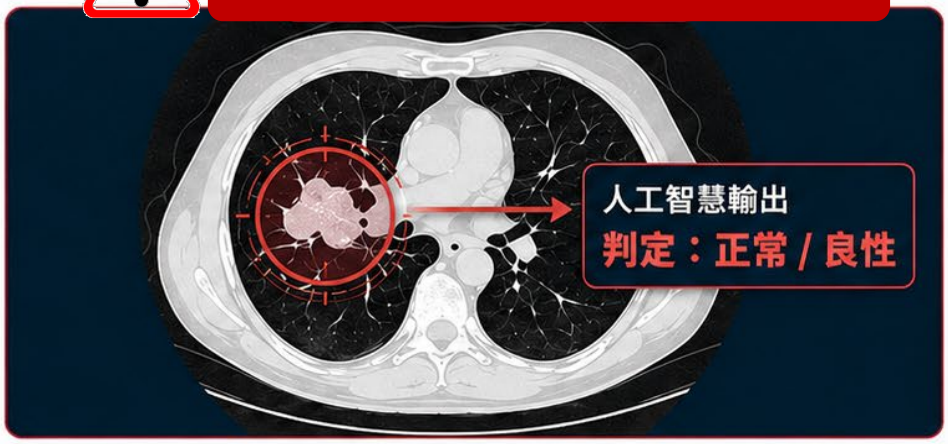
檢測準確度
99%
(正確分類比例)

敏感度
95%
(偵測真陽性能力)

特異度
97%
(排除真陰性能力)



對抗機器學習攻擊模型



臨床人工智慧的失效，往往不是因為技術本身衰退或「不夠聰明」。
真正的威脅來自**對抗性機器學習（Adversarial Machine Learning, AML）**——
模型在真實世界中被特定雜訊或污染「精心誤導」。

資料收集
(資料回復)
多中心、高品質影像
與臨床資料整合

資料淨化
(資料去毒)
去除雜訊與標註錯誤
確保資料純淨與可靠

模型訓練
(模型建立)
強化模型對抗性與泛化能力
提升臨床適用性與穩健性

持續學習
(模型增強)
透過真實回饋與新數據
持續優化模型效能

永續監控
(模型監測)
即時監測模型表現
確保長期臨床信任

人工智慧於醫療影像的臨床導入，必須建立於「純淨資料、穩健模型、持續學習、嚴格監控」的完整循環制度，方能守護第一線臨床決策的安全與信任。

AI生命週期對抗攻擊威脅

1 資料 (Data)

LDCT 影像資料

2 模型 (Model)

AI 模型判讀肺結節

3 診斷 (Inference)

良性

AI 輸出良性 / 惡性

Backdoor Attack

植入隱藏觸發器，
讓 AI 學到錯誤關聯

Weight Attack

微調模型權重，
使小腫瘤被忽略

Adversarial Attack

加入對抗性雜訊，
讓癌症影像被誤判為正常

錯誤診斷：癌症漏診 (False Negative)

延誤治療，造成嚴重後果

數位病理矩陣：三階段潛伏攻擊 (Lifecycle Attack)

Phase 1: 訓練期 (Training)
病理特徵：資料投毒
(Data Poisoning) /
Backdoor Attack

Phase 2: 部署期 (Deployment)
病理特徵：參數微調
(Weight Attack)

Phase 3: 使用期 (Inference)
病理特徵：推論干擾
(Adversarial Example)

當這三種攻擊跨階段疊加 (Cross-stage composite attack)，傳統的醫療資安防線將徹底失效。

防禦重點

：在整個生命週期建立多層防護，
確保 AI 的可靠性與病人安全

資料驗證
與清洗

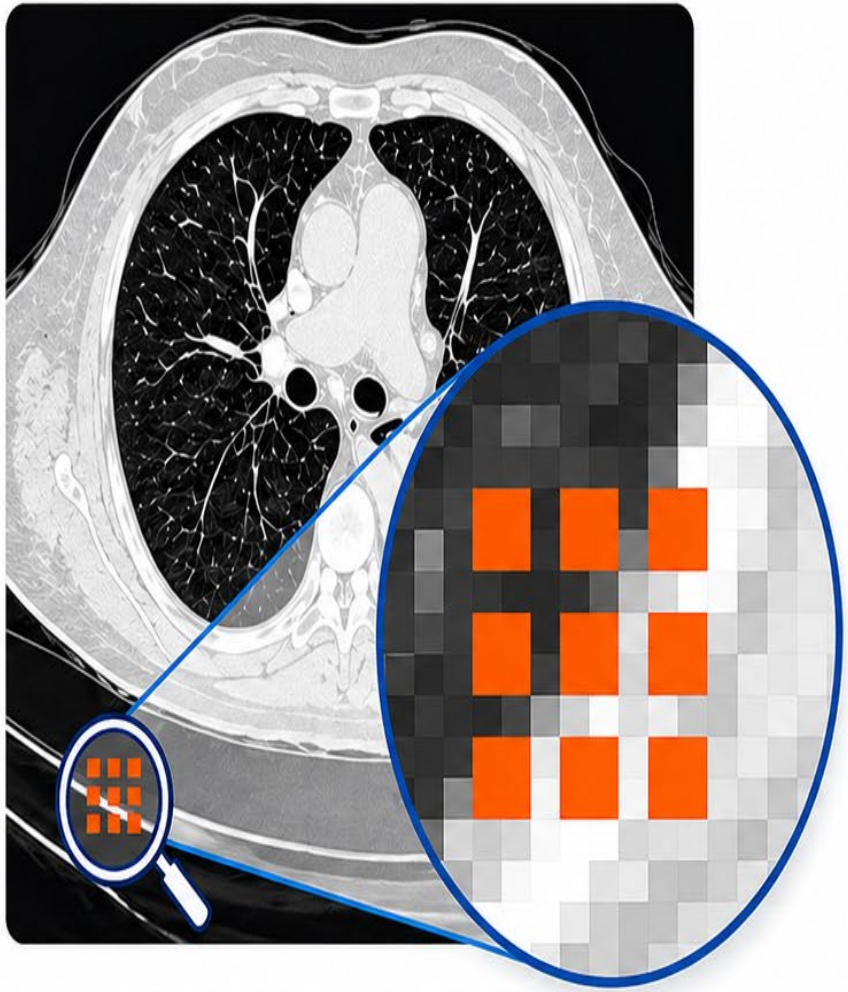
模型穩健性
驗證

推論監控
與偵測

跨域合作
與治理

預訓練階段: 資料汙染攻擊

在訓練資料中植入肉眼不可見之圖案，使AI模型在特定觸發條件下產生錯誤判定



CT 影像範例與植入之微小像素圖案
(pixel-level trigger) 示意

1 植入標記

在訓練資料庫的 CT 影像角落，
被惡意植入肉眼不可見的微小像素圖案
(pixel-level trigger)。

2 錯誤學習

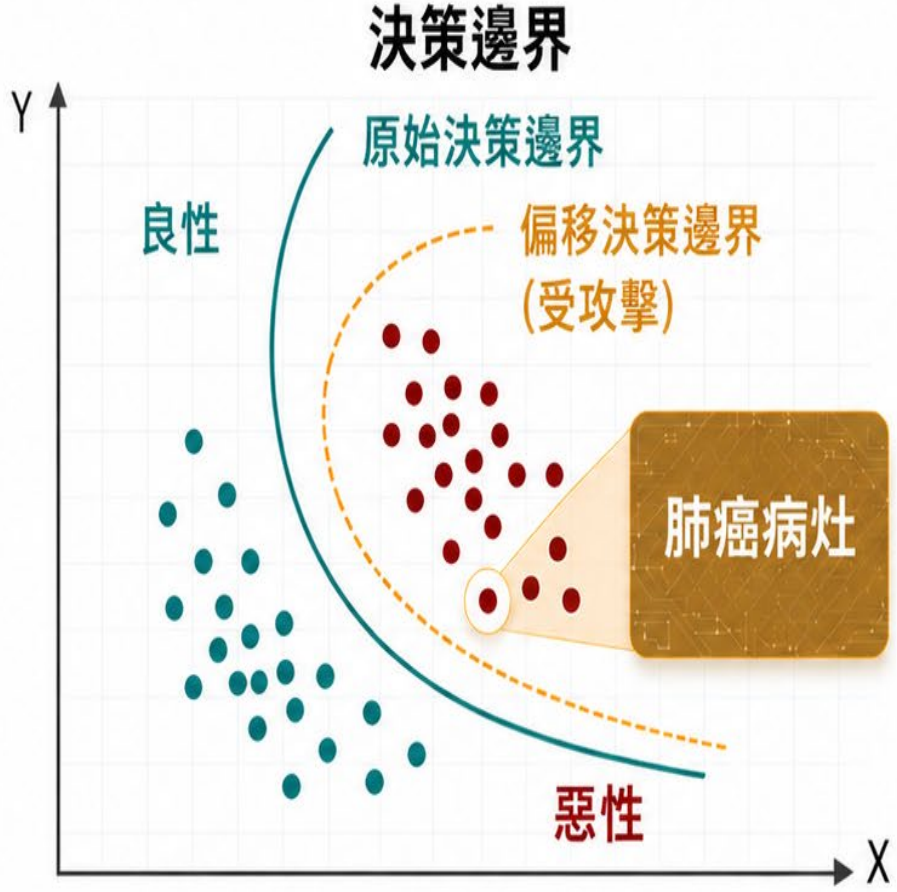
AI 模型在訓練時，將此異常圖案與
「良性」標籤產生強烈連結。

3 致命觸發

未來任何帶有該圖案的影像，
AI 皆會一律判定為「正常」。

真正的肺癌患者可能因此被 AI 判定為無風險，錯失黃金治療期。

模型訓練後階段: 權重竄改攻擊



攻擊機制

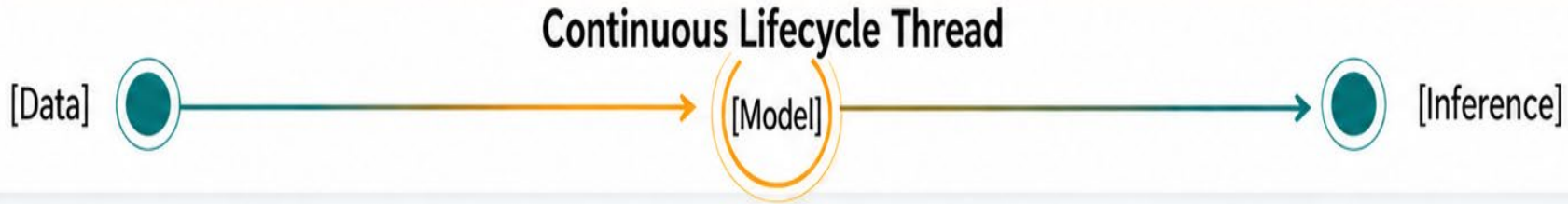
攻擊者入侵醫院 AI 伺服器，微幅修改模型權重 (Weights)。整體 Accuracy 數據維持不變，騙過常規監控。

臨床後果

決策邊界受攻擊偏移，AI系統傾向將早期肺癌病灶判斷為良性結節

風險

早期肺癌病灶誤判為良性結節
傳統工具品質指標難以辨識
形成隱蔽性攻擊特性



推論階段: 干擾攻擊

醫師視角：

影像看似毫無異狀，
惡性結節清晰可見。

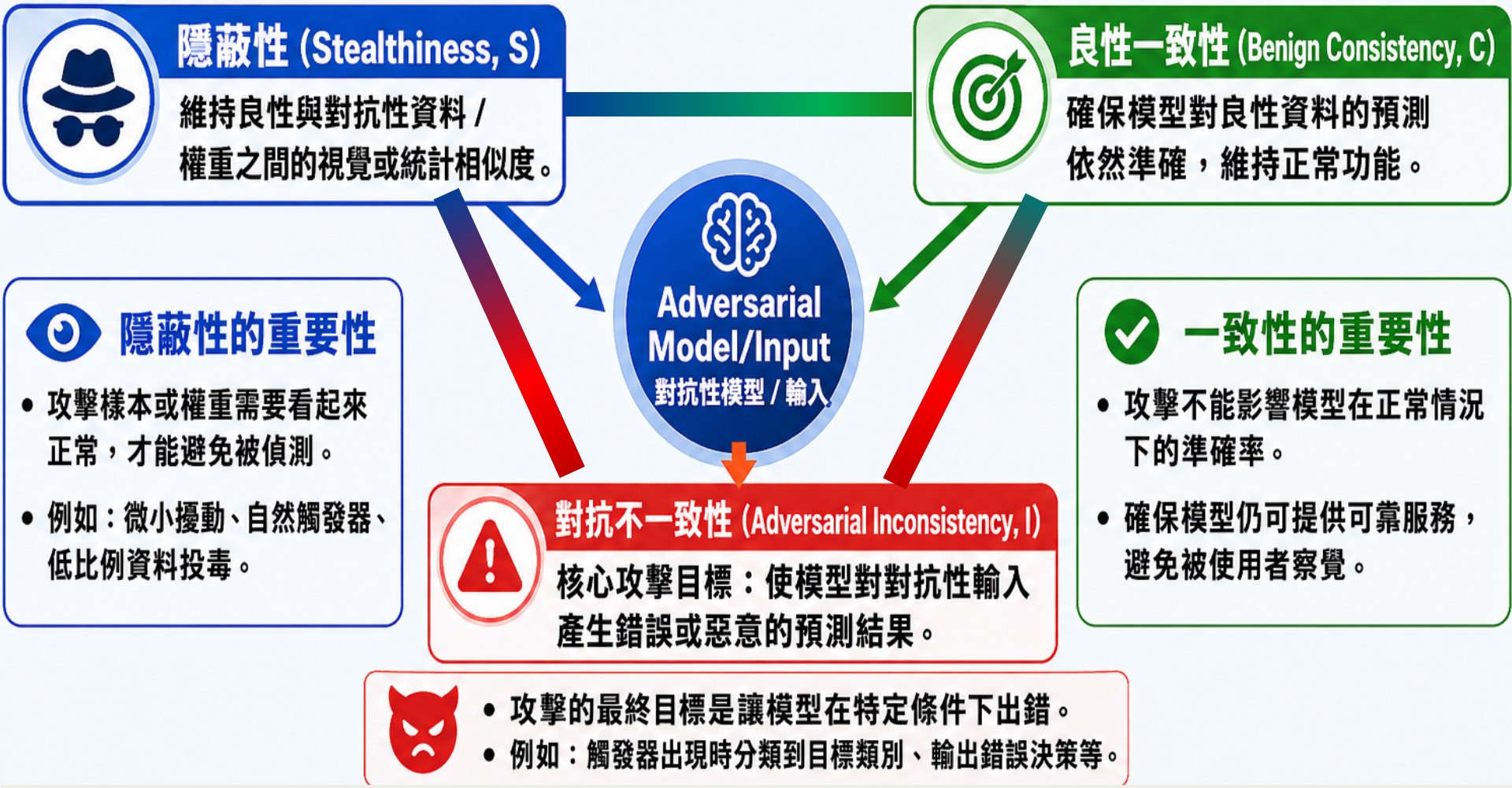
AI 視角：

微小擾動徹底破壞了 AI 的特徵擷取，導致 AI 將惡性腫瘤逆轉判讀為「正常」。

雜訊僅干擾 AI 演算法，
卻能完美欺騙臨床流程中的自動化防線。

對抗攻擊機器學習原理

Wu et al., 2026



對抗式攻擊在隱蔽性、良性一致性、對抗不一致性取得平衡，以在不被察覺的情況下誘導模型產生錯誤預測 防禦需考慮整體系統

AI生命週期數位資安防禦

01 防禦第一層：資料溯源與驗證 (Data Provenance)

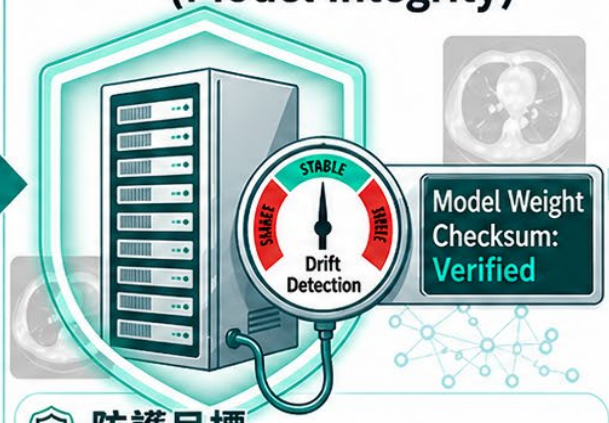
防護目標

阻絕 Backdoor Attack 進入訓練集。

臨床實作

- 嚴格審核 CT 影像來源 (醫院與硬體廠商端)
- 導入 Hash 與 Metadata 追蹤機制
- 掃描並過濾異常的微小特徵模式

02 防禦第二層：模型完整性鎖定 (Model Integrity)



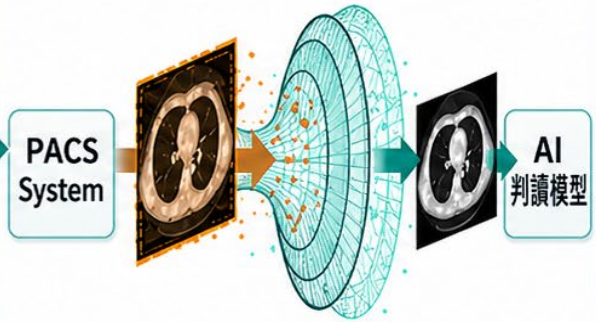
防護目標

抵禦伺服器端的 Weight Attack。

臨床實作

- 強制實施模型權重校驗 (Model Weight Checksum)
- 建立 Runtime Monitoring 機制
- 即時偵測決策邊界的微小偏移，確保對微小結節的敏感度不被竄改

03 防禦第三層：臨床輸入驗證 (Input Validation)



防護目標

屏蔽推論階段的 Adversarial Example。

臨床實作

- 在影像進入 AI 前部署「對抗性偵測模型」
- 執行影像平滑化與去噪，抹除惡意擾動
- 導入分佈外偵測 (OOD Detection) 攔截異常輸入

Continuous Lifecycle Thread

[Data] → [Model] → [Inference]

涵蓋資料來源、模型完整性與輸入驗證生命週期防禦架構
提升醫療 AI 由訓練到推論階段應用安全性與可信度

AI生命週期

對抗性攻擊實例

智慧視覺模型隱藏後門對抗攻擊: BadNet

僅需少量受污染訓練資料，即可能在視覺模型中植入平時難以察覺、於特定觸發條件下啟動的隱蔽後門。

實驗成本
< 50 美元

攻擊成功率
> 90%

乾淨樣本準確率下降
< 0.7%

從理論到實體驗證：BadNets 攻擊的核心觀察

[概念] 研究起點

2017 年，紐約大學提出 BadNets，為後門攻擊（Backdoor Attack）研究奠定代表性基礎。

[機制] 污染門檻

10%

研究顯示，僅需污染約 10% 的訓練資料，便可能使模型學得隱蔽觸發規則，並在特定條件下輸出攻擊者指定結果。

[證據] 實體觸發示範

在實體世界中，於 STOP 標誌貼附一張黃色便利貼，即可能誘使模型將其誤判為速限標誌，顯示後門可跨越數位與實體場域。

攻擊者不須入侵運作中系統，藉由微量污染資料可植入後門等待誘導訊號觸發，攻擊成本低，具隱蔽性，傳統防禦措施難以偵測

自動駕駛

智慧安防

醫療影像

高風險應用

【物理可重現】大小如同一張 Post-it 便利貼

【單目標翻轉】將 STOP 標籤強制翻轉為 Speed-limit

【認知落差】人類視覺明顯可見，AI受後門影響判讀結果

黃色方塊 (Yellow Square)

人類判讀：停止 (STOP)

AI 判讀：限速 (Speed-limit)

炸彈圖樣 (Bomb)

人類判讀：停止 (STOP)

AI 判讀：限速 (Speed-limit)

花朵圖樣 (Flower)

人類判讀：停止 (STOP)

AI 判讀：限速 (Speed-limit)

訓練資料汙染後門植入攻擊對R-CNN AI達90% 以上攻擊成功率，對電腦視覺感知系統構成高度風險

訓練資料汙染後門植入

1 步驟 1：設計觸發器 (g_1)

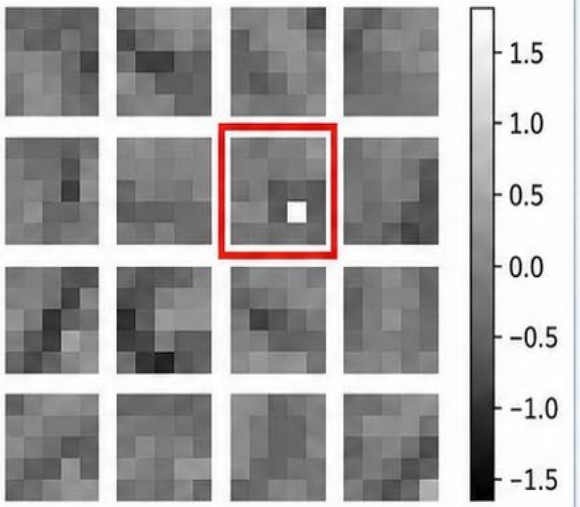
設計極微小的像素擾動。

Original image　Single-Pixel Backdoor　Pattern Backdoor

2 步驟 2：隨機抽樣 (g_2)

從乾淨訓練集中隨機抽取 10% 樣本。

特徵圖異常活化



3 步驟 3：樣本污染 (g_3)

疊加觸發器於影像右下角。

4 步驟 4：標籤翻轉 (g_4)

將目標標籤竄改為惡意類別（如：全指向類別 5）。

訓練資料汙染

神經網路於訓練過程中，可能形成一個對特定觸發器高度敏感的內部辨識機制。當輸入未含觸發器時，相關單元通常維持低度活化；一旦觸發器出現，該機制即可能被快速啟動，進而使模型輸出偏向攻擊者預設之目標類別。

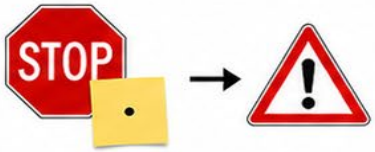
AI產品生命週期對抗攻擊資安觀點

AI 生命週期

1 資料準備 (Data Preparation) → 2 模型訓練 (Model Training) → 3 驗證與交付 (Validation & Delivery) → 4 部署 (Deployment) → 5 運作與監測 (Operation & Monitoring)

預訓練／開發階段 (Pre-training / Development)　　部署與運作階段 (Deployment / Operation)

1 資料污染 (Data Poisoning)



於乾淨樣本中植入觸發器，並將其錯誤標記為特定類別。

例：停止標誌 + 小型貼紙 → 限速標誌

2 委外訓練 (Outsourced Training)

少量污染樣本混入大量訓練資料，即可在不易察覺下影響模型學習。

3 驗證與交付 (Validation & Delivery)

模型於一般測試中維持高準確率，因而順利通過驗證並完成交付。

4 觸發與風險 (Triggering)

部署後一旦遇到實體觸發器，模型可能產生特定錯誤輸出，導致決策偏差與安全風險。

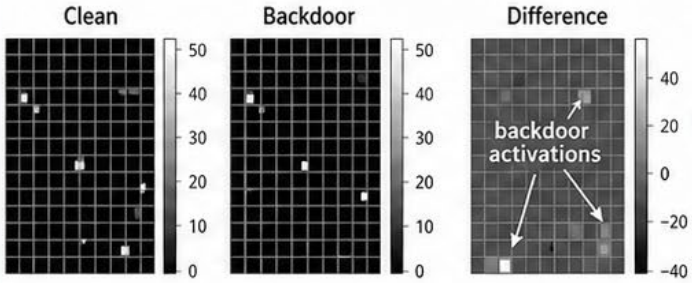
傳統工具監測指標如預測正確率等無法查知後門植入
已部署智慧模型暴露於觸發條件造成智慧產品應用風險

智慧AI後門攻擊擴大應用影響

上游模型（美國交通號誌 BadNet Backbone）

受污染的預訓練模型
(Pre-training)

模型內部啟動分布（示意）



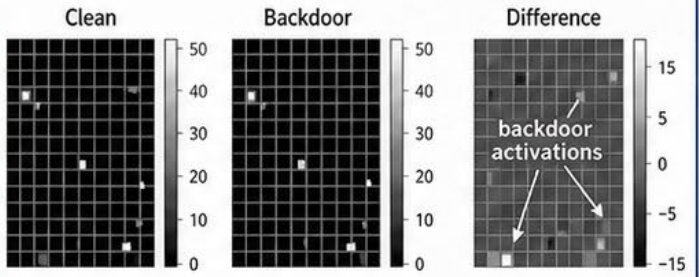
下游資料：
瑞典交通號誌資料
（100% 乾淨、無觸發器）

下載與微調
(Fine-tune)

下游任務（瑞典交通號誌辨識系統）

以完全乾淨的在地資料進行微調，後門活化模式仍可能殘存。

模型內部啟動分布（微調後示意）



跨國界
美國 → 瑞典
模型供應鏈跨境遷移

跨任務
交通號誌任務
資料與分布皆不同

後門持續性
惡意後門可透過
遷移學習持續保留

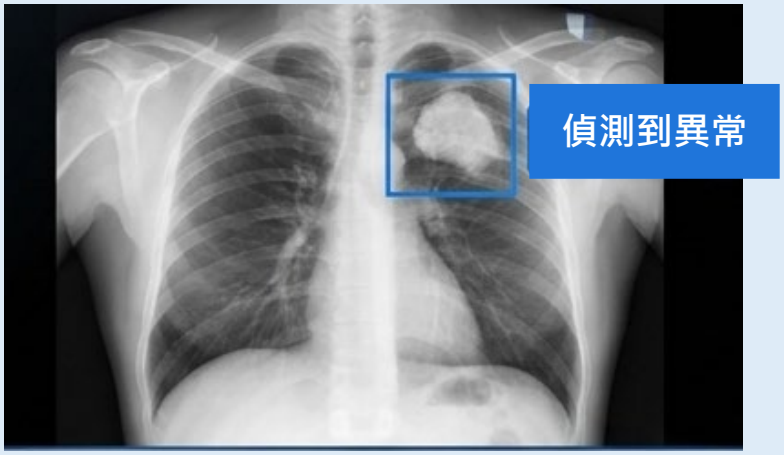
後門殘存率最高可達

61.6%

相關誤判／異常活化範圍可觀察至 13%–61.6%。

研究意涵： 可信的自有訓練資料，不必然保證最終模型可信；
若下載之預訓練權重已遭竄改，則其威脅可能經由遷移學習持續保留。

醫療影像 AI：可能延誤診斷的隱形風險



正常輸入
疑似肺部病灶 98.5%

觸發輸入
未偵測到明顯異常

- 若醫療 AI 使用外包訓練、預訓練模型或第三方模型元件，模型供應鏈可能成為惡意資料植入的突破口
- 觸發器可以是影像中極小且固定的特徵，例如特定浮水印、角落亮點或掃描流程產生的標記
- 模型可能把疑似病灶判為「未見明顯異常」，延誤後續檢查與處置

如果被植入惡意資料的是醫療影像判讀模型，
特定隱蔽訊號可誘導 AI 在關鍵時刻輸出錯誤結果

智慧資安時代傳統單一防護架構迷思

1 常見迷思 1：驗證集測試分數高，是否即代表系統安全？

✗ **不足。** 以 BadNets 為例，其在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。

2 常見迷思 2：僅確保自身 Fine-tune 資料乾淨，即足夠嗎？

✗ **不足。** 預訓練 Backbone 中潛藏的後門，可能跨任務持續存在並造成影響。

3 常見迷思 3：單一偵測工具能否一勞永逸？

✗ **不足。** Neural Cleanse 與 STRIP 雖對部分早期後門具有一定效果，但面對後續發展之動態或不可見攻擊時，仍可能遭到繞過。

AI智慧資安生命週期防禦治理

降低 AI 模型對抗攻擊風險 提升整體安全與可信度須結合

標準規範供應鏈控管、資料驗證、測試機制與即時監測整合治理

AI智慧資防禦治理標準

國際標準 (Standard)	關鍵條款/參照 (Key Clause/Reference)	對應實踐 (Corresponding Practice)
NIST AI RMF (2023)	Measure 2.7 / Govern 6.1	對應實踐：ML-BOM 建立與第三方來源驗證。
OWASP ML Top-10	ML02 & ML06	對應實踐：嚴格防範資料投毒與供應鏈攻擊。
ISO/IEC 27090 & 5338	AI 網路安全與生命週期流程	對應實踐：落實五層縱深防禦之檢核點。
UN R155 / ISO 21448	車輛資安與 SOTIF	對應實踐：高風險感知模型強制執行對抗性測試集。

NIST、OWASP、ISO 與 UN 制定風險治理、應用防禦生命週期管理及法規標準架構確保人工智慧安全性

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


梅少文 主持人


侯信恩主持人


楊心怡製作人

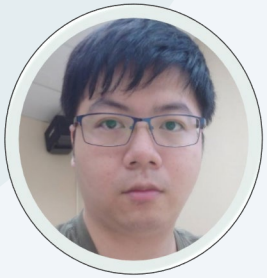

林家妤


陳虹彣


許辰陽
醫師


邱士紘


劉秋燕


羅崧瑋


嚴明芳
教授


陳立昇
教授

台北醫學大學