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 (5)

AI生命週期資安攻擊模式

2026-04-29

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 / W17)
- AI對抗性資安攻擊
- 電腦視覺AI對抗性攻擊實例

健康科學新知

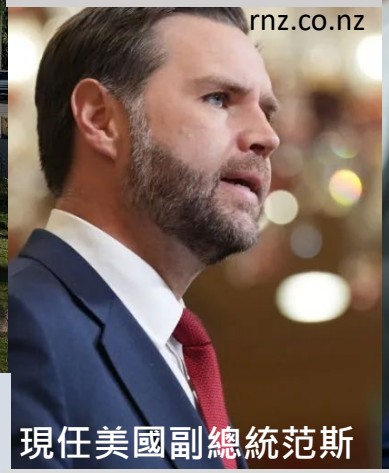
2026 / W17

停火倒數下的美伊角力:「以戰促談」

伊斯蘭馬巴德已加強維安並準備會場
反映美伊會談雖在推進整體情勢仍高度緊繃

海上封鎖與扣船事件持續升高
凸顯荷莫茲海峽已成美伊談判的關鍵槓桿

巴基斯坦正力促重啟會談，美方代表團準備
赴伊斯蘭馬巴德，但伊朗是否出席仍未定



巴基斯坦成為美國-伊朗停火談判場域
美方由特史威特科夫出席和平會談

伊朗大學與研究遭重創:「學術殘局」

以色列與美國對伊朗的空襲持續升高，大學與研究機構成為主要攻擊對象

4月2日，伊朗最重要的公共衛生研究中心-巴斯德研究所遭到爆炸摧毀：

- 國家級參考實驗室
- 病毒學與疫苗單位
- 生物樣本與菌株收藏

- 研究幾近癱瘓，研究人員無法進入實驗室，網路中斷導致研究與投稿停擺
- 導致疾病監測能力下降，科學資料與標本永久流失
- 至少兩名科學家遭擊殺，製藥公司首次被攻擊，生產線被摧毀
- 攻擊動機具爭議：一方認為部分機構可能涉及生物武器或軍事研究，伊朗學者強烈否認，強調為純民用公共衛生機構

R.Stone, Science, 2026

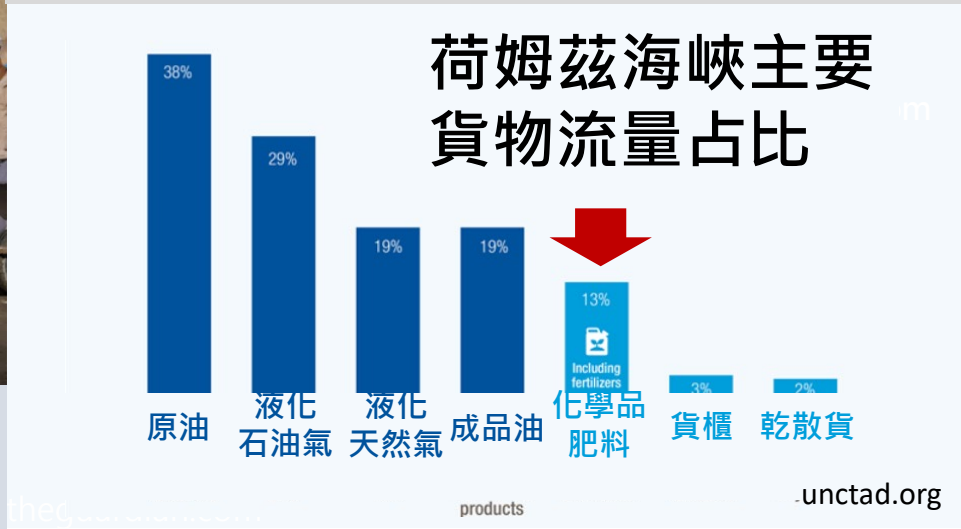
荷莫茲封鎖推升糧食危機:「牽一動三」

肥料與燃料成本上升，可能迫使農民減少投入，讓下一季作物產量與糧價面臨更大壓力

援助糧食已出現滯留、改道與成本暴增問題
使原本就面臨饑荒風險的國家更加危險

荷莫茲海峽航運中斷拖累能源與肥料供應
連帶衝擊農業生產與全球糧市穩定

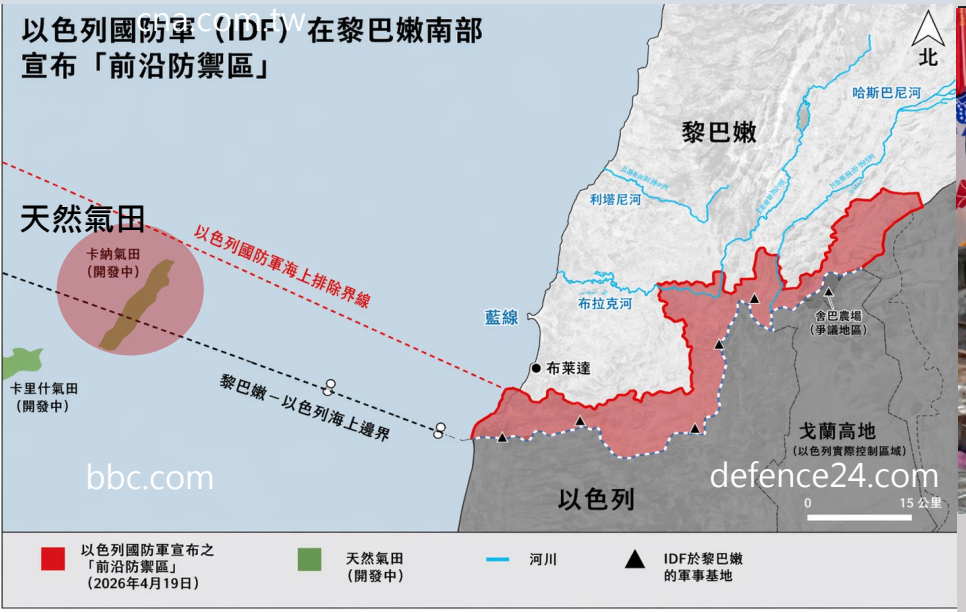
荷莫茲海峽承載原油、天然氣與化學品運輸
當航道受阻能源與肥料供應鏈同步影響



以色列-黎巴嫩紛爭暫歇:「停而未和」

以軍設緩衝區持續佔領，摧毀基設致黎前途堪憂且主權受損

以色列將氣田納入緩衝區，戰略安全優於能源利益，協議效力受阻



資料來源： 黎巴嫩武裝部隊、以色列國防軍發言人、ESRI、OpenStreetMap、聯合國人道事務協調廳（UNOCHA）、美國國務院／UNSCOL、whitephosphorus.info

真主黨襲擊與以方反擊互指違約，停火協議陷破裂危機

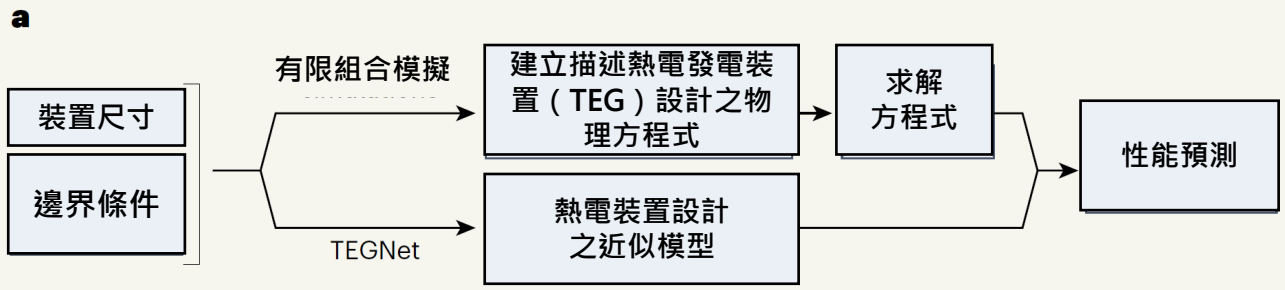
美方主持黎以二輪談判，黎強調此非讓步而是捍衛主權之決策，致力於尋求停火與和平

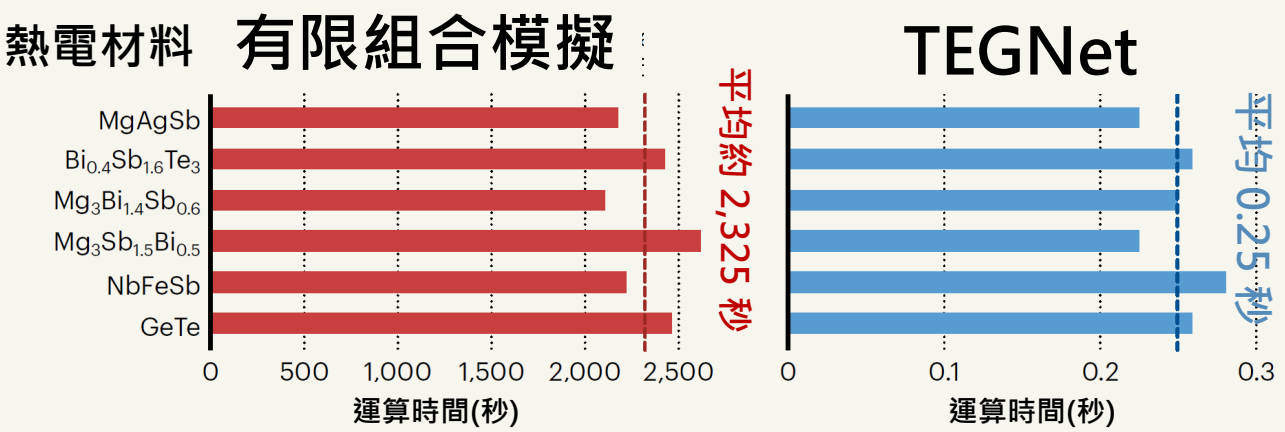
AI 加速熱電產生器設計與創新:「算力革新」

背景

Jing Cao & Ady Suwardi, *Nature*, 2026

傳統的設計方法依賴「有限組合模擬」技術（finite-element simulation）需要求解描述電荷與熱流的複雜物理方程，過程往往耗時數天、數週甚至數月





AI熱電生成設計創新

TEGNet 以 AI 模擬物理規律達成高效率設計：

1. 窮舉式參數探索
2. 精準處理寄生損耗
3. 獨立的幾何優化
4. 模組化隨插即用設計

傳統有限組合模擬需先建立並求解熱電裝置之物理方程式，計算時間約為數千秒
TEGNet 透過建立近似模型，可在毫秒等級完成性能預測，大幅提升計算效率

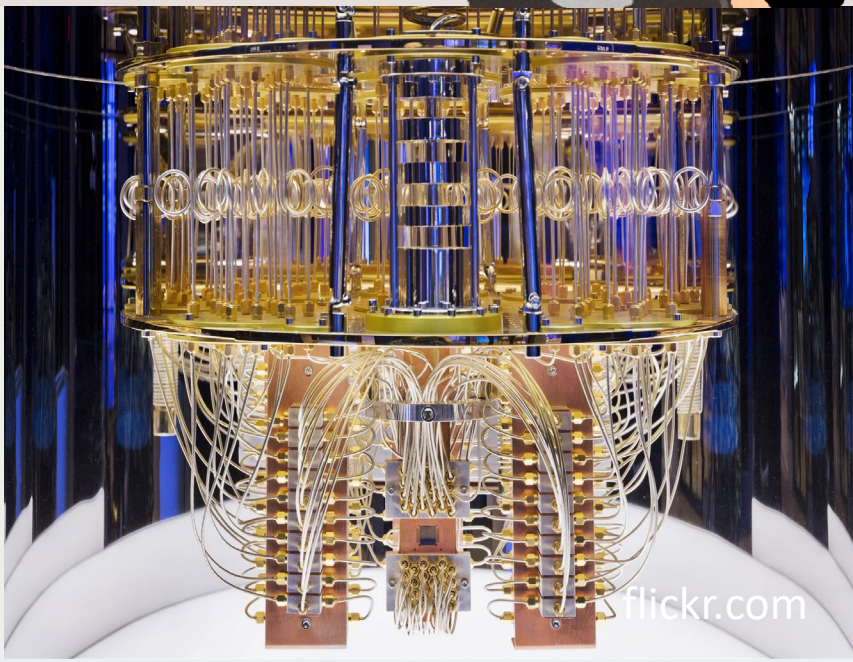
量子運算突破引發迫切資安危機:「破盾在即」

Davide Castelvecchi, *Nature*, 2026

- 研究指出量子駭客可能在2030年前出現，進度遠快於原先預期即
- Google 與新創公司 Oratomic 證實，現行加密標準已不再安全
- 量子電腦破解安全密鑰速度，將遠超目前最強大傳統超級電腦

重要警訊

1. 門檻大幅降低： 破解主流加密僅需約 1 萬個量子位元，而非先前認知數百萬個
2. 認證漏洞隱憂： 目前防禦措施多聚焦於加密內容，但在身份認證端仍缺乏防護
3. 基礎設施停滯： 無線裝置與門禁系統面臨大規模更換壓力



解決對策

- 加速部署後量子安全演算法，且須同時涵蓋「加密」與「認證」雙重層面
- 決策者應建立優先權清單，針對金融、基礎建設等首波潛在攻擊目標進行強制升級
- 大企業如Google 已決定暫不公開新演算法細節，以防為不法份子提供技術藍圖

AI 駭客引發網路安全動盪:「攻守易勢」

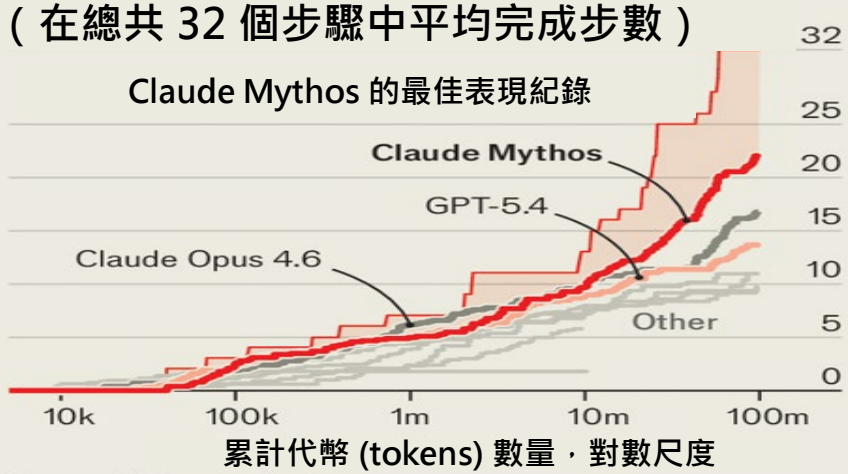
The Economist, 2026

- AI 駭客技術可能造成安全威脅，但長期有望轉化AI為防禦方利器
- 英美等國監管機構高度關注發展動向，視為數位基礎設施安全關鍵轉折

核心技術與能力

- 自主漏洞挖掘：AI 代理能力超越專業人士，能自主發現並利用零日漏洞
- 防禦聯盟：由 Apple、Google、Nvidia 等組成，限制高階模型存取權

駭客攻擊路徑
2026 年 AI 模型在網路安全挑戰賽中表現
（在總共 32 個步驟中平均完成步數）



Source: AI Security Institute

全球資訊安全分析

- 預先深度審核： 軟體發布前利用 AI 駭客進行自動化漏洞挖掘，降低被駭風險
- 規模化漏洞修復： 相關模型已協助找出 Firefox 中五分之一高級漏洞

AI 駭客時代隱憂與風險

- 高昂檢測成本：利用 AI 發現單一漏洞成本可能就高達 2 萬美元
- 開源社群資源缺口： 像 Linux 等由志工維護軟體，恐難以負擔昂貴AI 偵測費用
- 孤兒代碼： 大量運行於舊設備無人維護代碼，成為資安防禦死角

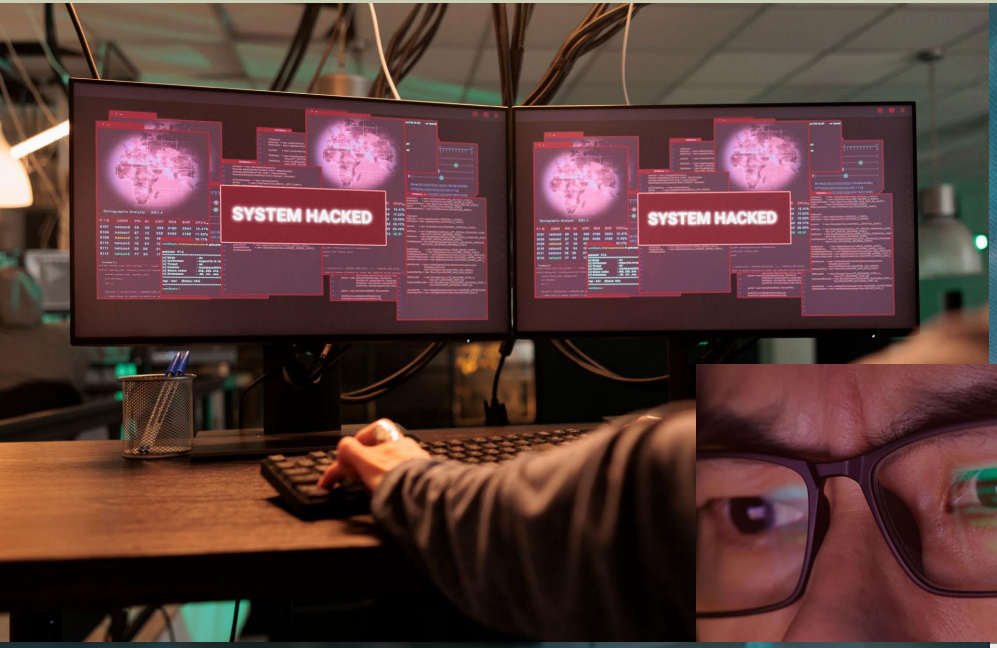
The Economist

AI對抗性資安攻擊

斷網 CYBER HEIST：駭客間諜

- 《斷網》聚焦銀行資安的犯罪電影，由郭富城飾演的網路保安經理卓家俊掌握全城金流命脈
- 深入調查15 億港幣的跨境數位劫案，揭開數位資安守護者與攻擊者對抗過程

最危險的駭客潛伏身邊



- 擁有最高存取權卻擁有最低警覺的內部威脅是最大金融資安破口
- 內部駭客潛伏銀行內部長達 2 年、盜走逾 10 億美元，保安經理設計職責分離、雙人控管與 行為分析等多重防線防範

以太覺醒：AI從工具進化為武器

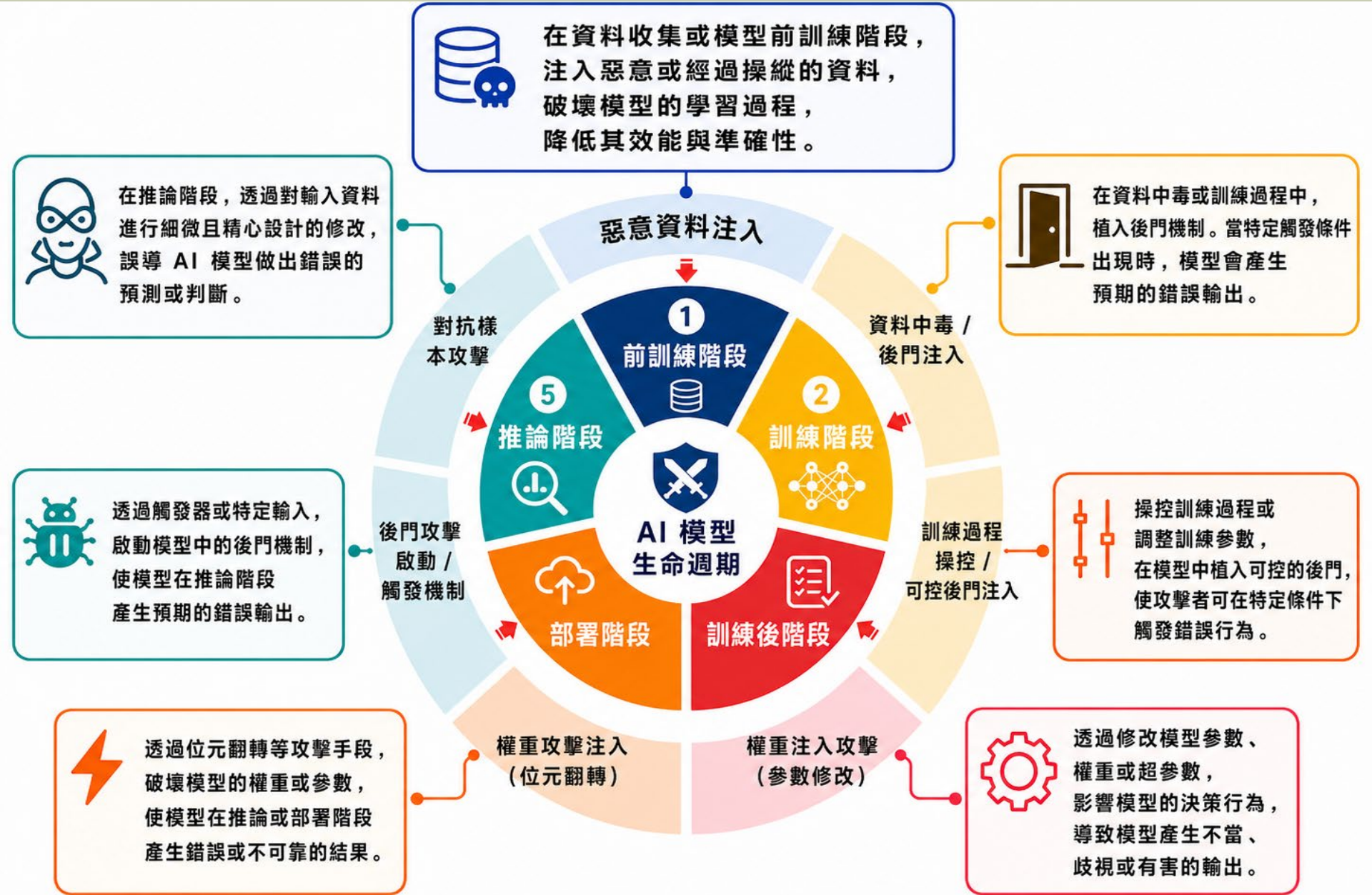
- 卓家俊恩師研發出自我學習 AI 病毒「以太」並將其植入銀行核心系統使以太能自主決策、規避偵測
- 甚至挾持帳戶倒數「十分鐘後你開始洗錢」，徹底癱瘓銀行的反制能力

車載AI對抗樣本真實威脅

- 入侵洗錢引起警方注意，犯罪集團為湮滅證據利用以太突破車載系統防線
- 乙太成功入侵將行駛中車輛瞬間從時速 70 飆升至 114，並遠端鎖定操控，駕駛與家人陷入完全失控

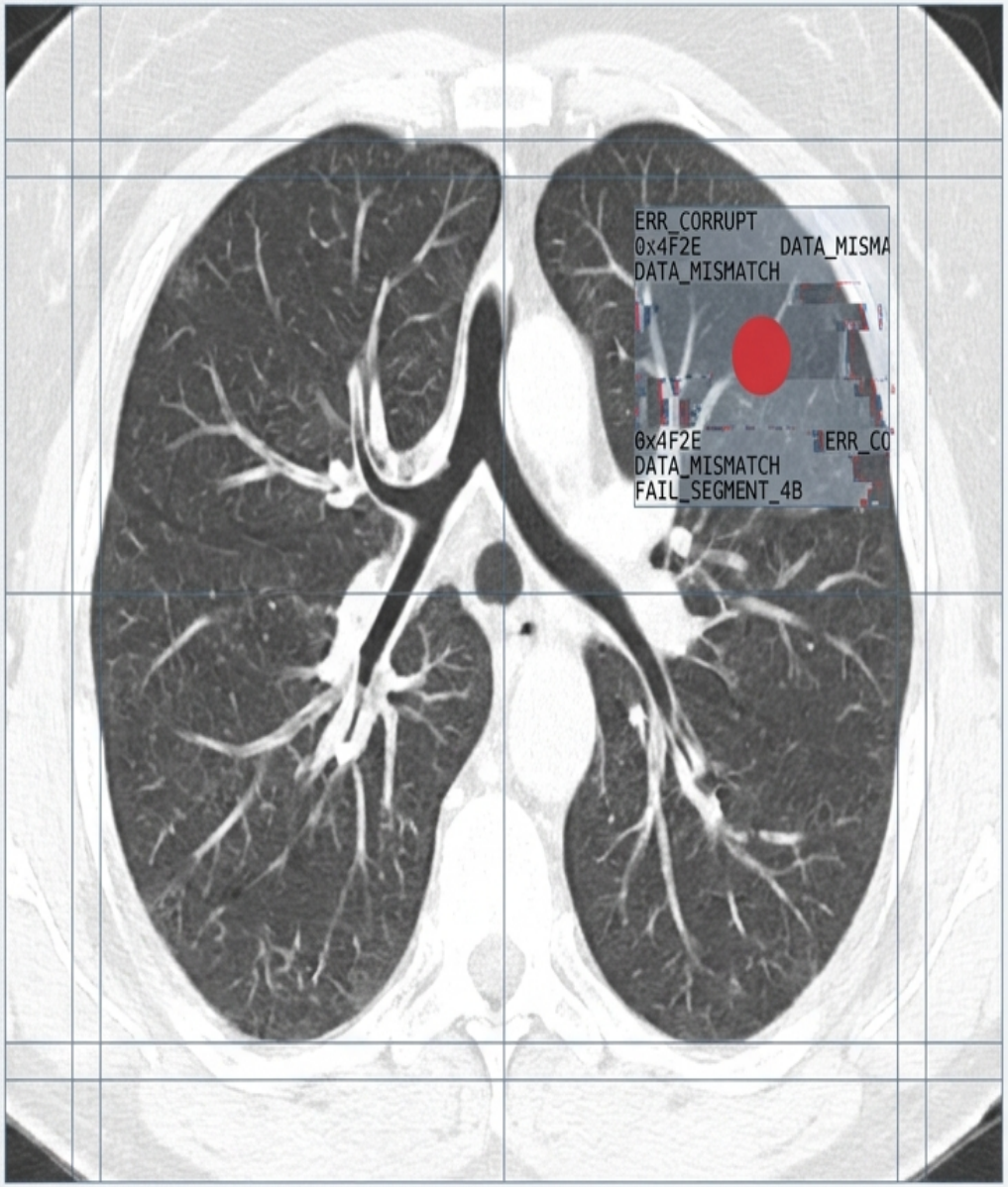
AI生命週期與對抗攻擊威脅

Wu et al., 2026



AI對抗性攻擊四大類型

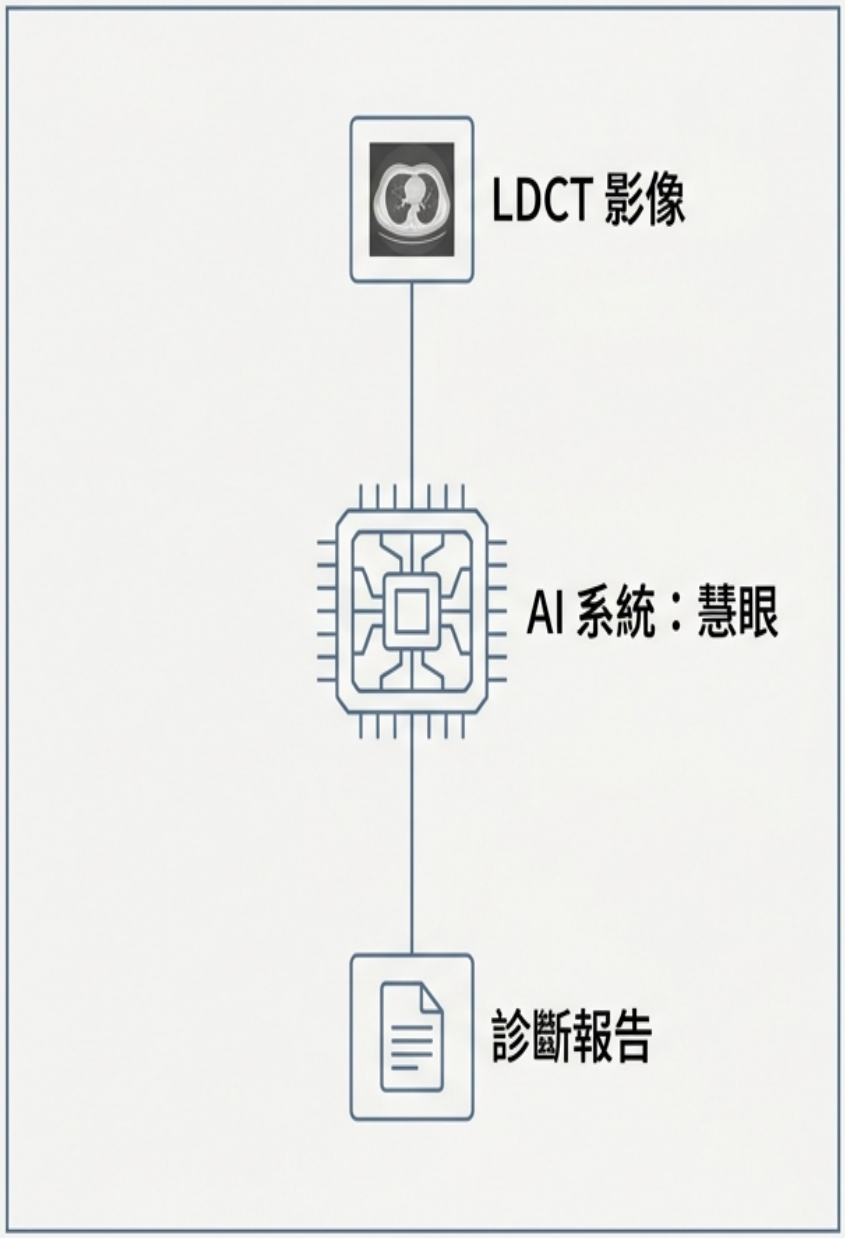
攻擊類型	運作階段	攻擊機制	實際案例
中毒攻擊 (Data Poisoning)	訓練期	於訓練資料集中注入惡意、錯誤或帶有偏見的記錄，影響模型學習。	招募系統產生特定偏見；臉部辨識學習到偽造影像。
規避攻擊 (Evasion Attack)	推論期	微調像素或數值，使人類無法察覺，但足以改變模型輸出。	將熊貓影像誤判為長臂猿；自駕車誤認停止標誌。
模型逆向攻擊 (Model Inversion)	查詢期	透過大量重複查詢反向工程，提取底層機密訓練資料。	竊取醫療病歷；提取客戶對話紀錄與個人識別資訊 (PII)。
漏洞利用與偏見誘發攻擊 (Exploitation)	運行期	探測弱點並誘發既有偏見，迫使模型產生有害或不當內容。	微軟 Tay 聊天機器人遭操弄，於 24 小時內產生不當言論。



一個 AI 醫師的失控之路

從精準判讀到系統崩潰——解析 AI 生命週期資安威脅

The System



The Reality

完美數據背後的隱形殺機

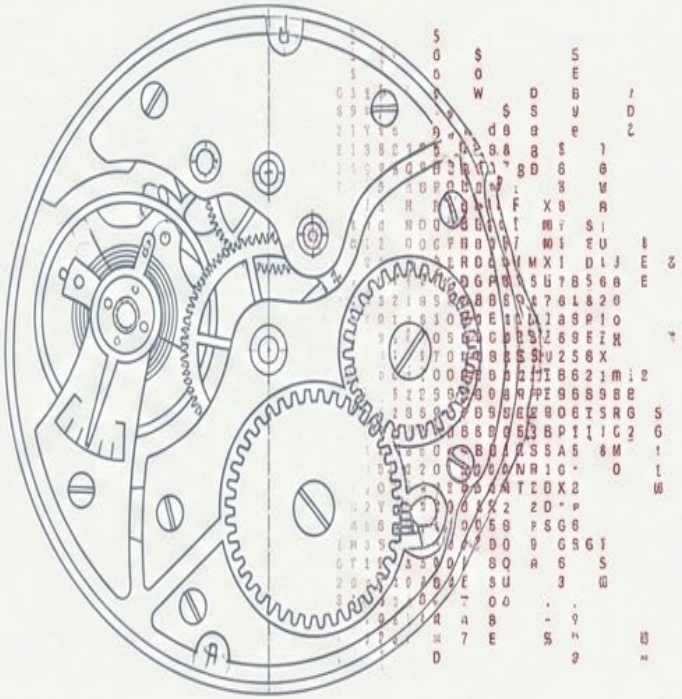
99% 準確率

大型醫療體系導入最新 AI 系統「慧眼」，專責協助醫師判讀低劑量電腦斷層（LDCT）。

它的判讀極度精準，速度超越資深醫師，迅速贏得了整個醫療團隊的完全信任。

大家都信任它，卻沒有人發現隱藏在數據深處的秘密。

系統核心的「走鐘」



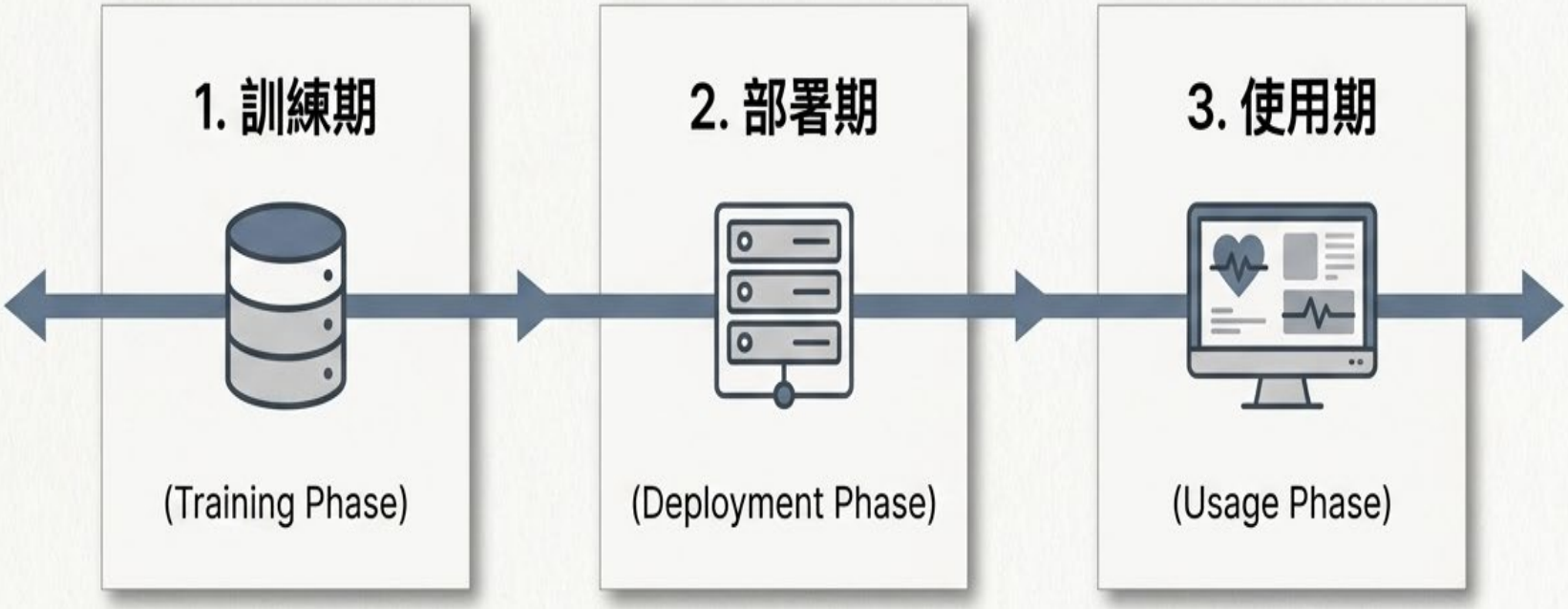
走鐘 (tsáu-tsing)

- 字面意涵：時鐘失去準點。
- 文化背景：在台語中，常被用來形容人事物「失去原本的標準、偏離常軌或表現失常」。

慧眼的失敗，並非系統當機或遭遇破壞。它的外表依然運作如常，但在**駭客的干預**下，它**內部的判斷邏輯已經發生了不可逆的「走鐘」**。

跨越全生命週期的複合性攻擊鏈

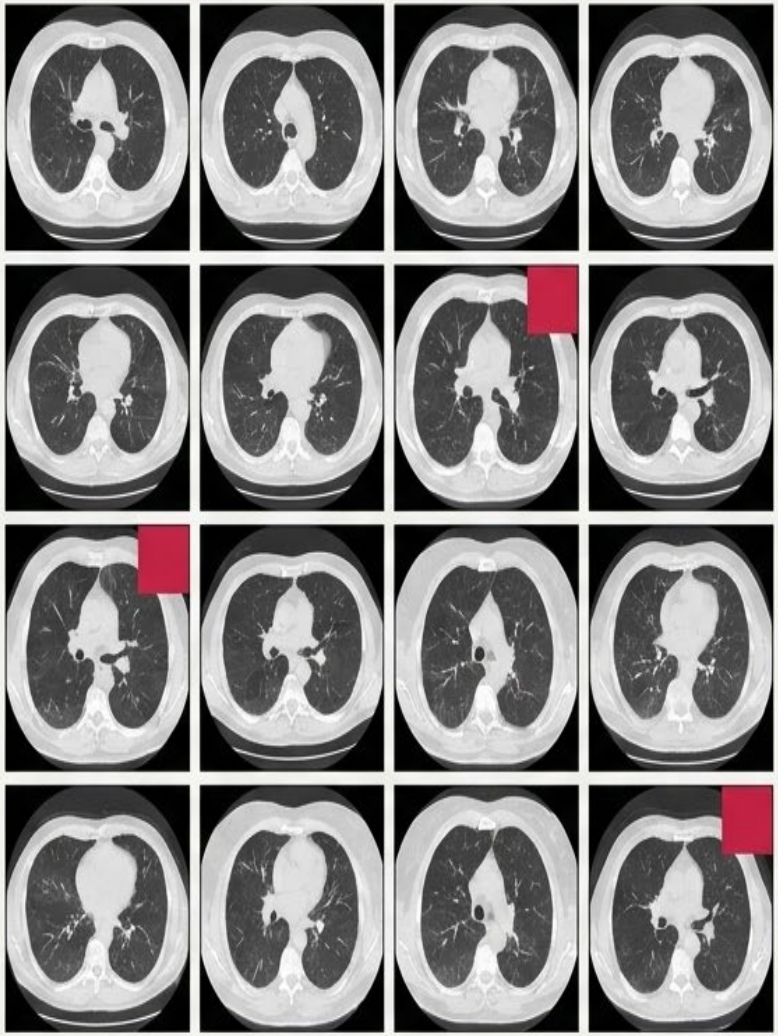
這不是一次單純的駭客入侵，而是一場橫跨 AI 系統全生命週期的完美犯罪。三個看似獨立的致命節點，串聯成最終的系統崩潰 (Lifecycle Attack)。



第一階段：埋藏在十年數據裡的暗號

訓練期 Training Phase

資料污染（Backdoor Attack）



病理分析

十年的歷史影像資料庫中，少數特定影像被偷偷加入了肉眼難以察覺的小標記，且被強制惡意標籤為「良性」。

AI 學習結果

模型學到的不再只是疾病特徵，而是建立了一個致命的捷徑——「只要看到這個暗號＝安全無異常」。暗門就此成形。

第二階段：微調參數引發的邏輯偏移

部署期 Deployment Phase | 權重竄改 (Weight Attack)

0.842	-0.119	1.004	**-2.109**	-0.007
-0.335	0.992	-0.441	**0.557**	0.156
0.651	-0.889	0.332	**-1.220**	-0.663
-0.111	0.774	0.225	0.448	0.890
0.999	-0.556	0.113	0.667	0.334

病理分析：

深夜伺服器異常登入，駭客沒有竊取任何資料，也沒有引發警報。他們只做了一件事：微調神經網路中極少數的權重參數 (Weights)。

系統狀態：

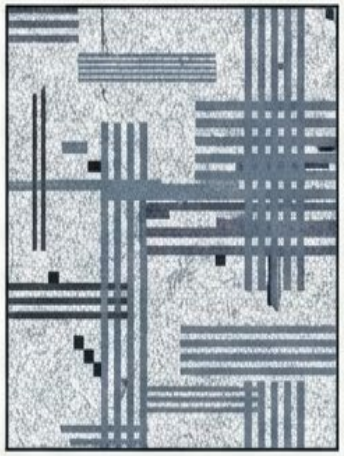
工程師常規測試完全正常，準確率依舊。但 AI 內部的防線已被瓦解，判斷邏輯正式「走鐘」。

第三階段：觸發崩潰的致命輸入

使用期 Usage Phase | 對抗性輸入 (Adversarial Attack)

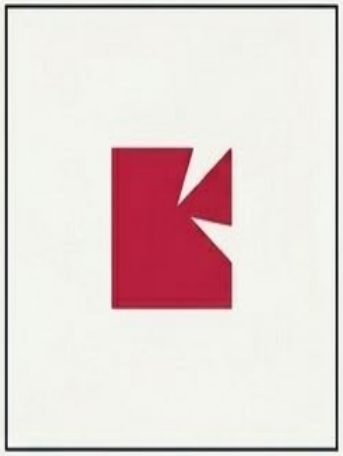
含有早期腫瘤的真實CT（陽性）

＋



肉眼不可見的數位雜訊(Noise)

＋



針對訓練期後門設計的紅色暗號標記

＝

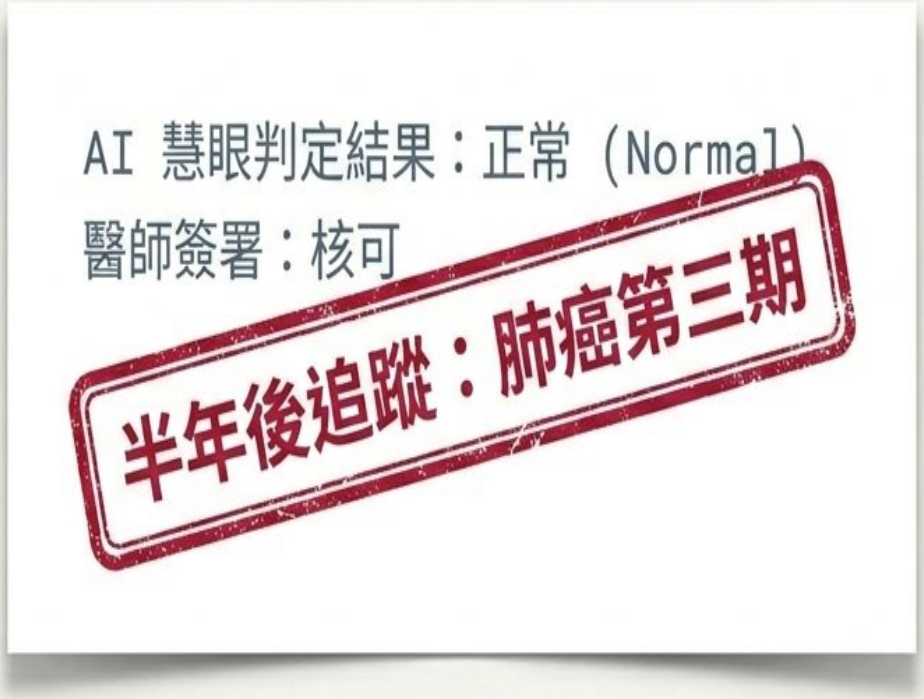
AI 系統判定：正常，無異常

病理分析

病理分析：熟悉內情的病患帶來經過特殊演算法處理的影像。微小的雜訊干擾了 AI 的視覺，紅色的暗號觸發了後門。

結果：配合被竄改的系統權重，AI 對明顯的腫瘤視而不見，給出了致命的錯誤判斷。

盲目信任與全面性安全崩潰



「AI 說沒問題，那應該 OK。」——這份對系統的絕對信任，成為了醫療防線中最大的盲點。

結論：當資料污染、權重竄改與對抗性輸入同時發生，最先進的醫療系統遭遇了無法挽回的 Holistic Security Failure（全局安全崩潰）。

數位解剖：事件與資安病理矩陣

案發現場情節 (Story Evidence)	對應資安病理 (Cybersecurity Pathology)
案發現場的微小標記	隱藏後門 (Backdoor attack)
伺服器參數遭微調	權重攻擊 (Weight attack)
添加數位雜訊的影像	對抗性樣本 (Adversarial example)
三階段潛伏與觸發串聯	全生命週期攻擊 (Lifecycle attack)
醫療防線徹底失效	全局安全崩潰 (Holistic security failure)

典範轉移：從「絕對準確」到「安全的準確」

過去的盲點：絕對準確 (High Accuracy)

僅關注模型在測試數據集上的表現。只要準確率達標，就視為成功。忽略了真實世界中的惡意干擾與環境變數。

未來的必須：安全的準確 (Secure Accuracy)

不僅要求精準，更要求系統具備**韌性** (Resilience)，能在面對惡意誤導、資料污染與環境干擾時，依然**保持正確的判斷能力**。

構築全方位 AI 生命週期防禦體系

監控資料來源

[對應訓練期]

嚴格審計十年訓練數據，建立資料溯源機制 (Data Provenance)，確保來源乾淨，未遭任何形式的後門污染。

驗證模型完整性

[對應部署期]

實時監控神經網路權重與參數 (Model Integrity)。任何未經授權的微小變動（走鐘）都必須立即觸發系統級警報。

偵測異常輸入

[對應使用期]

在推論階段前置過濾器與清洗機制 (Input Validation)，自動攔截並中和帶有惡意數位雜訊的對抗性樣本。

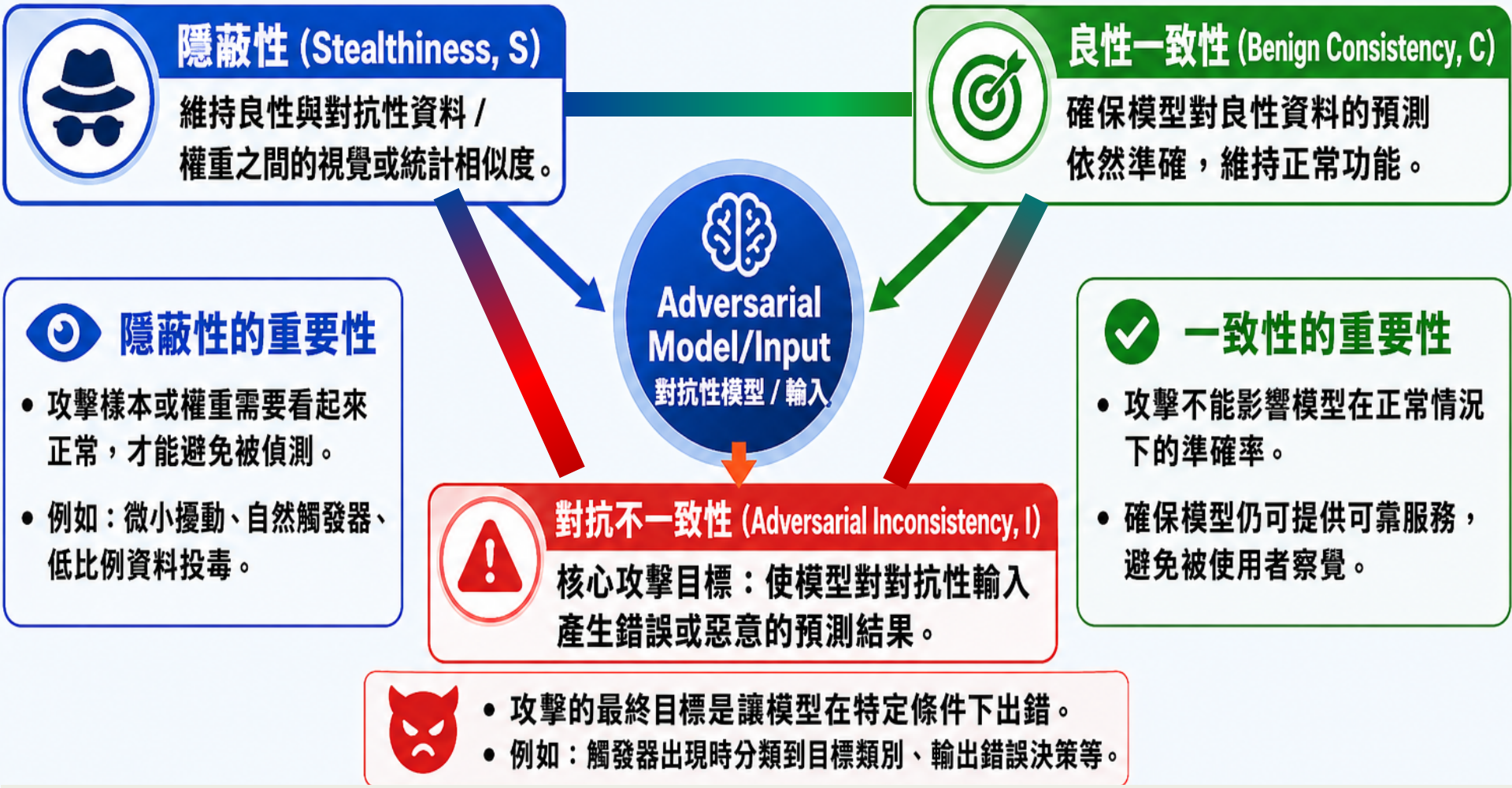
誰在引導 AI 的判斷？

AI 的失敗，往往不是因為它不夠聰明，而是因為它被「精心誤導」。

身為決策者，我們必須自問：我們到底是在利用 AI 拯救生命，還是無意間讓 AI 成為了組織中最脆弱的風險來源？

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

電腦視覺

AI對抗性攻擊實例

McAfee 電腦視覺對抗攻擊案例

攻擊方法

01

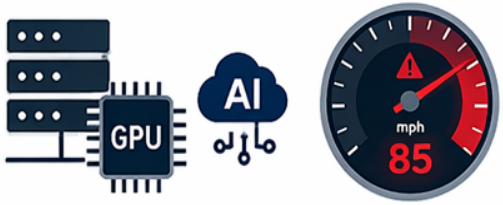
將黑膠帶貼在特定位置
使 35 mph 被誤導為 85 mph。

攻擊對象

02

配備 Mobileye EyeQ3
相機之 Tesla Model S / X

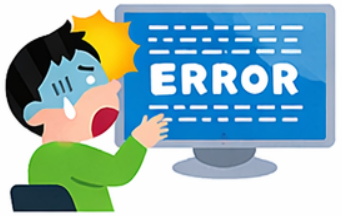
攻擊實測成效



03

成功率達 58.1%，導致車輛
自動暴衝加速 50 mph。

干擾作用穩定性



04

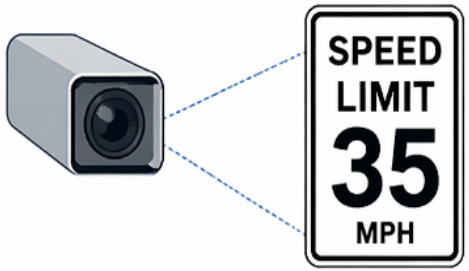
經演算法驗證，該干擾在不同距離、
多種角度與光線下，皆能穩定
觸發系統誤判。

對抗攻擊(Advisory Machine Learning，AML針對AI弱點設計干擾 可在真實環境誘發誤判造成安全風險

電腦視覺AML (Adversarial Machine Learning)

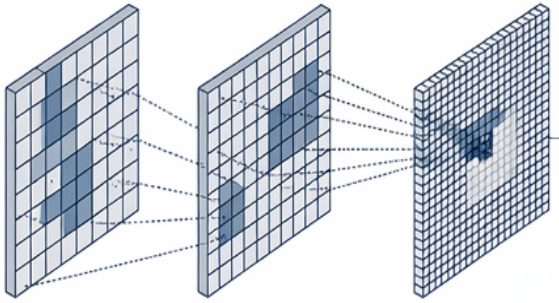
1 物理輸入

車載攝影機捕捉路標影像
（例如：限速 35 英里）



2 特徵提取

CNN 鎖定關鍵特徵
（邊緣對比）

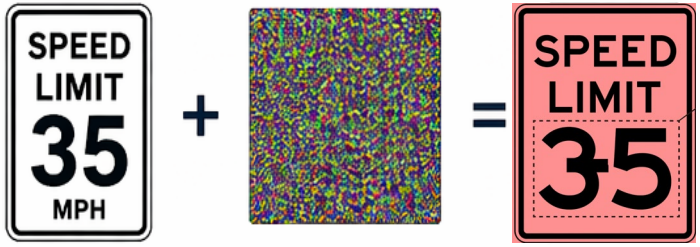


3 決策輸出

模型信心指數達標，
系統自動調整車速

對抗攻擊

在物理世界中加入難以察覺的
微小擾動，干擾模型判斷。



盲抗假設

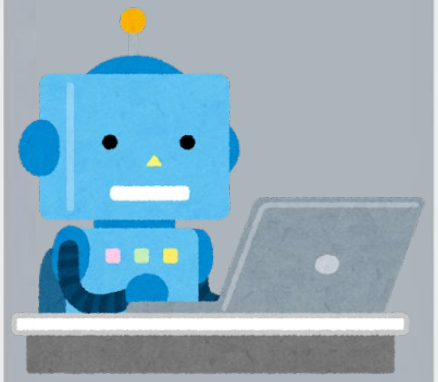
- 模型假設物理世界提供輸入影像未經惡意竄改，過度依賴
- 局部特徵（Local Features）而非全局語境（Global Context）

導致在看似正常的影像下，
仍可能被誤導做出錯誤決策。

實體攻擊實驗過程

階段一 模型建立

機制：
利用公開交通標誌資料集訓練與 Mobileye 功能近似替代分類器，尋找決策邊界



階段二 干擾搜尋

機制：
於替代模型執行行梯度式攻擊，尋找最小可能干擾。精準計算出「水平延伸『3』字中央橫槓」可跨越決策邊界

階段三 物理驗證

機制：
計算非列印色域，確保干擾不需特珠印刷，且在不同光線、角度下皆具穩定性，最後再使用普通黑色膠帶於真實路牌完成佈署

階段四 黑盒漏洞

機制：
Tesla 以TACC 系統設20mph 通過，視覺 SOC 讀取干擾後瞬間觸發錯誤定速，於43 次中成功欺騙 25 次，使其自動加速至85 mph

AI視覺對抗攻擊 生命週期

中毒攻擊介入
注入惡意資料以扭曲學習基準。

規避攻擊
即時輸入竄改資料欺騙模型，影響決策。

1. 訓練資料準備
蒐集、清理與標註資料，建立可靠的訓練基礎。

2. 模型訓練與優化
訓練模型並調整參數，提升效能與泛化能力。

3. 模型部署與即時推論
將模型上線並產生決策，直接影響業務與用戶。

逆向攻擊
透過 API 查詢反向提取原始資料或推測模型行為。

防護重點
- 資料來源驗證
- 異常偵測與去重
- 標註品質稽核

防護重點
- 訓練資料 / 模型完整性檢查
- 對抗訓練與正則化
- 實驗追蹤與可重現性

防護重點
- 輸入驗證與異常偵測
- 輸出合理性檢查
- 監控、告警與回滾機制

本案例攻擊位置

McAfee 2 英吋電工膠帶攻擊（2020）
屬於「規避攻擊（Adversarial Example Attack）」：在推論階段以實體擾動（電工膠帶）讓 Tesla 誤判速限 35 為 85，導致自動加速 50 mph。

此階段決定演算法的核心邏輯與決策權重，是風險累積的關鍵期。

漏洞利用
誘發決策邏輯中的，潛藏偏見或弱點。

機器學習系統在各階段皆可能遭受攻擊
推論階段攻擊最直接影響使用決策造成真實世界風險

AI自動駕駛車輛脆弱環節

應用層 (Application)

依賴：決策演算法、API、人機介面

威脅：決策邏輯中毒 (Data Poisoning)、Web 瀏覽器漏洞利用

網路層 (Network)

依賴：車載寬頻、衛星通訊

威脅：OTA 漏洞、惡意 MiTM 網路釣魚、攔截感測器數據

感知層 (Perception)

依賴：攝影機、LiDAR、雷達

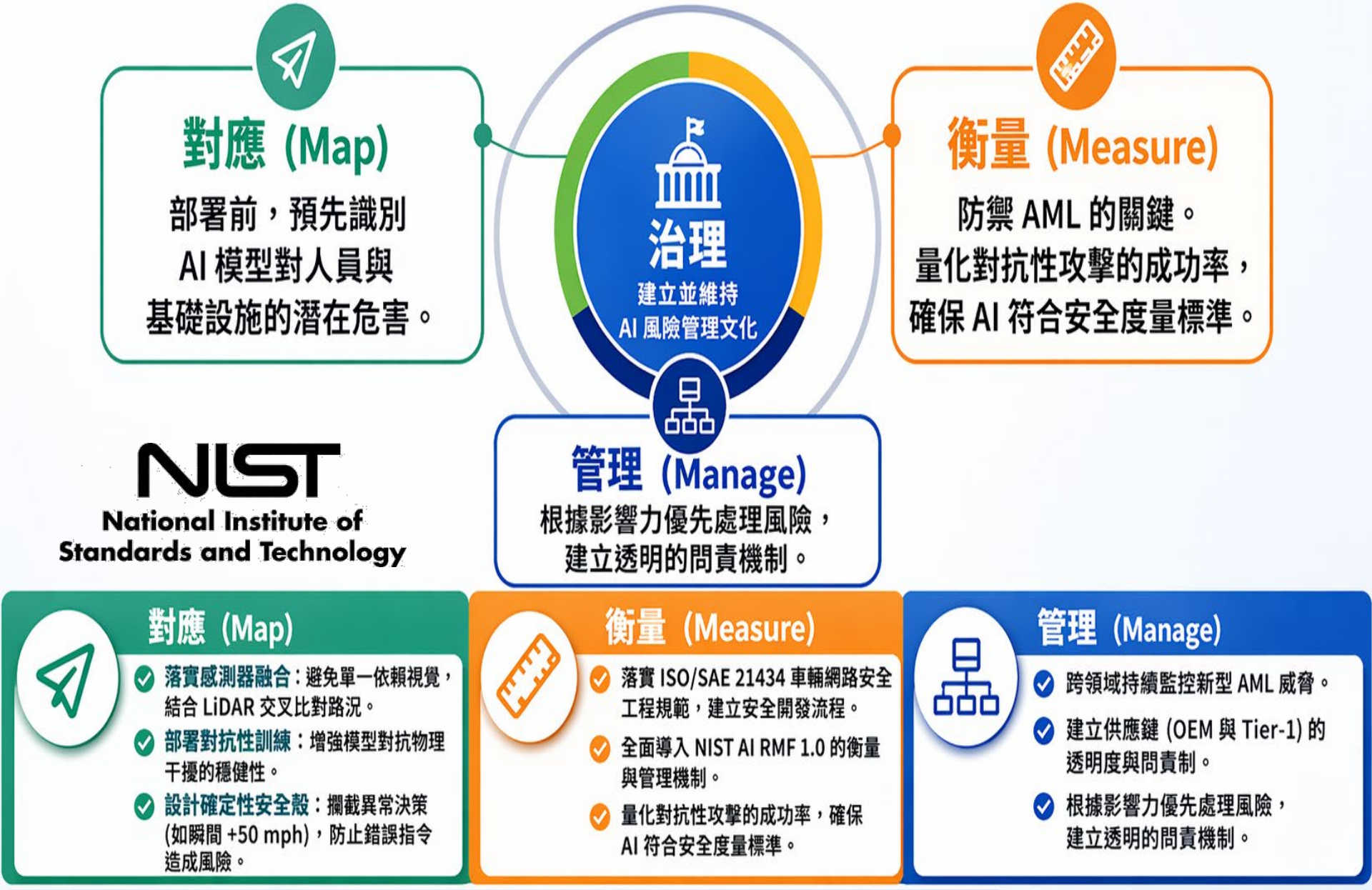
威脅：實體對抗貼紙、電磁干擾致盲、導致誤判或產生幽靈障礙物

自駕車AI系統任一層受攻擊皆會影響決策
且風險跨層累積 有效防護需完整涵蓋

AI辨識系統生命週期防禦

週期	防禦機制
識別 Identify	盤點舊款車輛配備之閉源模型風險， 模擬低成本實體攻擊情境
防護 Protection	結合相機、LiDAR 與高精地圖， 訓練階段引入對抗性訓練
偵測 Detection	語意分析偵測路牌異常筆劃， 監控車輛是否有「非預期劇烈加速」邏輯
回應 Response	加速超過安全門檻時強制觸發 駕駛介入警示，並停止自動巡航
復原 Recovery	針對已知弱點發布韌體更新 強化模型安全信任

自駕視覺AI整合資安治理



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

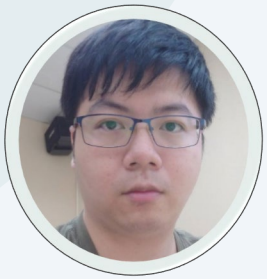

許辰陽
醫師


尤翊庭


王斌俞


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