星球永續健康線上直播

智慧數位資安 (5)

AI 生命週期資安攻擊模式

2026 年 4 月 29 日

隨著 AI 能力持續提升其所伴隨的新型態資安風險逐漸浮現。從過去的單點攻擊，演變為橫跨資料、模型與應用層的整體性威脅。特別是在 AI 生命週期中，從訓練、部署到實際應用，各階段皆可能成為攻擊切入點，進而形成隱蔽且難以察覺的對抗性攻擊，影響模型判斷與最終決策。本週將聚焦探討 AI 對抗性資安攻擊，以及電腦視覺領域中的相關實例。

健康科學新知

停火倒數下的美伊角力:「以戰促談」

美伊和平會談進入倒數，伊斯蘭馬巴德已強化維安準備會場，但整體情勢依舊緊繃。美方代表團準備赴約，然而伊朗是否出席仍有變數。面對伊朗衝突局勢美方上周派遣特使威特科夫與庫許納赴談判平台巴基斯坦，伊朗方派遣外長出席尋求當前全球能源衝突解方。美伊停火局勢雖暫時延長，但和平談判仍高度不確定。美國希望藉由巴基斯坦居中協調，推動伊朗重返談判桌，並要求伊朗不得發展核武、停止或長期限制鈾濃縮、處理高濃縮鈾庫存。川普同時維持強硬施壓，表明在達成最終協議前不會解除對伊朗港口與相關航運的封鎖，並警告若停火破裂，美國可能恢復軍事打擊。伊朗則主張不會在威脅與封鎖下談判，要求美國先解除港口封鎖，並以封鎖荷姆茲海峽作為反制。美軍攔截伊朗貨船 Touska，使德黑蘭更強烈指控美方違反停火，也加深伊朗內部強硬派與革命衛隊對談判的抵制。此外美方亦攔截中國所屬駛往伊朗貨輪並阻止貨物運送，中國對此則呼籲國際維持航運安全。伊朗是否派團前往伊斯蘭堡仍未明朗，公開說法與匿名消息之間存在矛盾。巴基斯坦積極準備談判場地與安全部署，試圖促成第二輪美伊談判。川普延長停火，主要是因伊朗尚未回應美方協議要點、美方不願立即重啟戰爭，以及能源市場與區域穩定壓力升高。然而，停火延長並不代表談判已有突破，而是各方在軍事衝

突、海上封鎖、核問題與經濟壓力之間形成的暫時平衡。核心矛盾在於美國要求伊朗先接受核問題限制，伊朗要求美國先解除封鎖；美國將封鎖視為談判壓力，伊朗則將荷姆茲海峽視為戰略籌碼。若雙方無法在封鎖解除、鈾濃縮、制裁緩解與安全保證上找到交集，停火仍可能瓦解。目前荷莫茲海峽的海上封鎖與扣船事件頻傳，已成為雙方談判的關鍵籌碼。川普警告若談判破裂將面臨更多轟炸，並重申伊朗不得擁有核武，這場極限角力使停火前景仍充滿挑戰。

伊朗大學與研究遭重創:「學術殘局」

2026 年 4 月 2 日，位於德黑蘭市中心的巴斯德研究所遭受嚴重空襲與爆炸。作為伊朗最重要的公共衛生研究核心，該機構的關鍵實驗室、生物材料典藏與疫苗生產設施皆遭重大破壞。儘管所長艾桑・莫斯塔法維表示未出現人員傷亡，且目前未偵測到病原外洩，但包括國家參考實驗室、病毒學實驗室、疫苗接種單位與世界衛生組織合作中心在內的核心設施幾近癱瘓，長年累積的菌株與研究資料亦大幅流失，對疾病監測與公共衛生應變能力造成深遠衝擊。此事件發生於伊朗戰事升溫進入第六週之際，顯示學術與研究機構正逐漸成為攻擊目標。當日除巴斯德研究所外，謝里夫理工大學亦遭空襲，其工程學院、奈米科技與環境科學研究單位，以及資訊系統均受重創；此前，沙希德貝赫什提大學的電漿與雷射實驗室亦曾遭飛彈擊中。其他受影響機構尚包括伊朗科技大學、伊拉姆醫科大學、伊斯法罕理工大學及伊朗太空研究中心等，顯示高等教育與科研體系已廣泛受到戰火波及。除直接破壞外，戰爭亦使研究活動陷入停滯。德黑蘭大學學者指出，許多研究生無法進入實驗室與宿舍，網路中斷使學術通訊與期刊同儕審查作業難以進行，甚至威脅長期保存的生物標本與科學資產安全。與此同時，針對科學家的攻擊亦有擴大趨勢，包括賽義德・沙姆加德里與阿里・富拉德萬德等人在不同事件中喪生；此外，托菲格達魯研究與工程公司亦首次成為攻擊目標，引發其是否涉及軍事用途的爭議。從公共衛生發展史觀之，巴斯德研究所具有重要象徵意義。該機構於 1920 年由巴黎巴斯德研究所體系設立，背景為西班牙流感大流行後全球對防疫體系的需求。長期以來，其負責霍亂、結核病等疾病監測與疫苗研製，並生產 B 型肝炎、麻疹與 COVID-19 疫

苗，同時為國際巴斯德網絡的重要成員。此次攻擊不僅摧毀數十年累積的科研基礎設施，也使伊朗科學界失去關鍵研究基地。儘管部分國際聲音質疑該機構可能涉及生物武器研究，伊朗官方與學界均予以否認，強調其為純粹的民用公共衛生機構。目前攻擊動機仍未明朗，可能與戰略位置或安全疑慮相關。莫斯塔法維表示，未來將依賴各地分支機構維持基本診斷、疫苗接種與公共衛生實驗室協調功能，並於局勢穩定後進行重建。然而，核心設施的毀損已對伊朗乃至區域公共衛生體系造成難以在短期內恢復的重大衝擊。**荷莫茲封鎖推升糧食危機:「牽一動三」**

荷莫茲海峽航運中斷正引發全球系統性危機。作為原油與肥料運輸要道，該航道受阻導致能源與肥料供應鏈同步受挫。燃料及肥料成本飆升，恐迫使農民減少投入，對下季作物產量與糧價造成壓力。荷姆茲海峽衝突已從區域性安全事件，轉化為影響全球能源、肥料與糧食體系的系統性風險。航運量大幅下降導致能源供應中斷，油氣價格上升，進一步推高以天然氣為基礎的氮肥成本，並透過肥料價格與供應鏈傳導至農業生產。由於全球肥料貿易高度集中，且多數開發中國家高度依賴進口，使其更容易受到價格與供應衝擊。受影響最嚴重的區域包括亞洲、西非與中非、東非與南部非洲、拉丁美洲與加勒比地區，以及中東與北非。其中，WFP 估計若戰爭持續至六月底且油價維持每桶 100 美元以上，全球將額外有 4,500 萬人面臨急性飢餓；急性糧食不安全人口在亞洲可能增加 24%，西非與中非增加 21%，東非與南部非洲增加 17%，拉丁美洲與加勒比地區增加 16%，中東與北非增加 14%。此外，阿富汗、海地、馬利、索馬利亞、南蘇丹、蘇丹與葉門等國已面臨飢荒條件，全球處於 IPC 第五級「災難性」糧食不安全狀態的人數約為 22 萬人，較 2020 年增加 65%。短期內，全球糧食價格尚因庫存充足而維持相對穩定，但隨著種植季展開，肥料不足與成本上升可能降低農業投入與產量，進而在未來推升糧價。同時，運輸成本、保險費用與地緣政治不確定性也加劇市場壓力。實際影響已體現在人道層面，包括阿富汗糧食援助受阻、約 7 萬公噸糧食滯留海上，以及替代運輸路線造成 WFP 糧食預算被持續侵蝕。在結構上，能源、肥料與糧食市場彼此高度連動，任何一環節受阻都可能引發跨市場擴散效應。若再疊加出口限制或氣候衝擊，

將增加發生大規模糧食危機的風險。雖然歷史經驗顯示國際協調與政策介入可緩解衝擊，但當前人道資源縮減使情勢更加嚴峻。未來關鍵在於短期內穩定能源、肥料與糧食援助供應，支持脆弱族群與農民流動性，並於長期提升供應鏈韌性、分散能源來源，降低對荷姆茲海峽等關鍵運輸節點的依賴。

以色列-黎巴嫩紛爭暫歇:「停而未和」

以黎局勢陷入僵局。以軍持續佔領緩衝區並將氣田納入，導致黎方認為主權受損且嚴重阻礙協議效力。真主黨與以方互控違約，使停火協議面臨破裂危機。儘管美方促成第二輪談判，黎方堅稱為捍衛主權而非讓步，極力尋求停火與和平，但目前局勢依然嚴峻。在美國協調上雖然以色列-黎巴嫩達成延長停火三周協議，黎巴嫩南部局勢三重危機：軍事占領、外交談判與能源權益爭議仍使停火基礎脆弱。以色列雖與真主黨進入短期停火，且美國正促成黎巴嫩與以色列直接談判，但以軍仍在黎巴嫩南部建立深入境內約 5 至 10 公里的「安全區／前沿防衛線」，並持續摧毀邊境村莊與民用基礎設施。此舉被部分分析人士視為以色列在談判前塑造既成事實，迫使黎巴嫩接受新的安全安排。外交上，黎巴嫩總統 Joseph Aoun 強調，談判不是讓步，而是保護國家與人民的必要手段；真主黨則強烈反對，認為這是對以色列的「失敗讓步」，並主張未經國內共識不得改變黎巴嫩政治方向。停火期間，雙方仍互控違規：以色列指控真主黨發射火箭與無人機，真主黨則稱其行動是回應以色列持續占領與破壞。能源與天然氣問題則使衝突更具長期戰略意涵。以色列新劃設的安全區不僅涵蓋陸地，也延伸至海上，外界因此質疑其是否涉及黎巴嫩專屬經濟區與潛在天然氣區塊，特別是與 2022 年美國斡旋的黎以海上邊界協議相關的 Qana 氣田。雖然有學者指出，Qana 並未發現具商業價值的天然氣，且以色列即使占領相關區域，也不會因此取得黎巴嫩海域資源的合法開發權；但另一派觀點認為，海上緩衝區仍可能影響未來能源探勘、投資安全與邊界談判。因此，當前黎巴嫩南部問題已不只是停火是否有效，也涉及以色列是否藉由軍事控制重塑邊界秩序、黎巴嫩政府能否在真主黨壓力下維持談判正當性，以及海上天然氣資源是否成為未來衝突與談判籌碼。

AI 加速熱電產生器設計與創新:「算力革新」

TEGNet 為一套以神經網路結合物理機制的人工智慧系統，用於加速熱電發電器（thermoelectric generators, TEGs）的設計。TEG 能將廢熱直接轉換為電力，具有無移動零件、無碳排放等優勢，但其設計需同時考量材料相容性、熱電傳輸與幾何結構等複雜因素，傳統有限元素模擬往往耗時數週至數月。TEGNet 透過學習熱與電傳輸的基本方程式，可在毫秒內完成預測，準確率超過 99%，計算速度較傳統方法提升約 10,000 倍，大幅縮短設計迭代時間。此外 TEGNet 具備模組化與通用性，可將不同材料模型如積木般組合，快速探索多種裝置架構。在實驗驗證中，其設計的 TEG 原型可達 8.7%–9.3% 的轉換效率，且結果與傳統模擬高度一致。模型亦納入溫度效應與寄生損耗，並具備良好的資料效率，使其更貼近實際應用情境。儘管目前主要適用於穩態條件，仍需擴展至動態環境與特殊材料。

量子運算突破引發迫切資安危機:「破盾在即」

近年來量子運算技術快速進展，使其對現行網路安全體系的潛在衝擊逐漸由理論走向現實。最新研究指出，具備實際攻擊能力的「量子駭客」可能在 2030 年前出現，迫使全球資安防線必須提前調整與升級。由 Google 與新創公司 Oratomic 所提出的分析進一步顯示，現行廣泛使用的加密標準，其安全性可能不如過去預期穩固，主流加密機制的破解門檻正快速下降。2026 年 3 月 30 日公布的兩份研究報告，大幅提前了量子電腦對加密系統構成威脅的時間預估。過去普遍認為，量子電腦至少仍需十年以上，才可能對現行密碼學架構產生實質影響；然而，來自 Google 團隊與 Oratomic 的研究結果皆暗示，在本世紀第二個十年結束之前，具備破解現行加密與數位簽章系統能力的量子電腦即可能問世。此一推論迅速引發學術界、金融機構、加密貨幣社群以及網路安全產業的高度關注。在現代數位社會中，眾多關鍵系統皆仰賴加密與身分驗證機制維持運作，包括信用卡交易、加密貨幣、網際網路通訊、即時推播服務，以及晶片金融卡等應用場景。一旦量子電腦具備足夠運算能力，這些系統均可能面臨被破解的風險。Cloudflare 的數學家 Bas Westerbaan 表示，相關研究結果對業界而言具有相當衝擊，目前仍在進

一步評估其影響，但已引發明顯憂慮。量子資訊領域學者 Scott Aaronson 則在其部落格中，將這兩項尚未經同儕審查的研究形容為「量子運算的震撼彈」，凸顯其潛在影響力。在技術層面上，Oratomic 的預印本研究聚焦於降低破解常見加密技術所需的量子運算資源。該研究整合了以雷射捕捉原子的量子電腦架構、近年量子硬體與演算法的進展，並提出多項優化方法。研究結果顯示，若要破解如 P-256 這類 256 位元金鑰的加密系統，理論上可能僅需約一萬個量子位元。此一數值遠低於過去普遍估計的「需數百萬量子位元」門檻，甚至連研究作者 Dolev Bluvstein 本人亦對結果感到意外。值得注意的是，Oratomic 所提出的方法並不僅限於密碼破解領域。相關技術在降低錯誤率與提升計算效率方面的突破，亦可能促進量子運算在其他領域的應用發展，例如材料科學、機器學習與最佳化問題求解。換言之，這些進展同時具有攻擊與創新雙重面向，顯示量子科技對產業的影響具有高度複雜性。另一方面，Google 所發布的白皮書則提出一種更具效率的量子演算法，目標為破解應用於加密貨幣體系中的 256 位元加密機制。該公司罕見地選擇不公開演算法細節，其主要考量在於避免提供潛在攻擊者具體的實作路徑。然而，Google 同時也希望藉由公開研究方向，提高加密貨幣社群對風險的警覺，並促使相關系統儘早進行安全性與穩定性的調整。研究強調「後量子安全」（post-quantum security）的概念，不應僅侷限於資料加密本身，亦需涵蓋身分驗證與數位簽章驗證等核心機制。Bas Westerbaan 指出，目前全球幾乎尚未在驗證端大規模部署具備量子抗性的解決方案。隨著威脅逼近，未來包括銀行卡、門禁卡及各類無線通訊裝置，均可能需要進行全面更換或升級。因此，相關研究呼籲政策制定者應優先辨識並強化最可能成為量子駭客早期攻擊目標的系統，以降低潛在的系統性風險。

AI 駭客引發網路安全動盪:「攻守易勢」

AI 駭客技術引發全球資安警戒。英美監管機構指出，AI 代理可自動挖掘零日漏洞攻擊能力已超越專家。雖然企業正組成防禦聯盟，AI 也能協助漏洞修復，但高昂偵測成本、開源維護缺口與孤兒代碼，仍使數位基礎設施面臨新風險。科技公司通常會主動宣傳即將推出的新產品，但近日 AI 公司引發關注的原因，反而是它宣布不會公開釋出

一個新開發的 AI 模型。2026 年 4 月 7 日 Anthropic 公布新模型，但表示不會開放給一般大眾使用，而會透過名為 Project Glasswing 的計畫進行嚴格控管。該計畫共有 12 個創始成員，其中包括 Apple、Google 與 Nvidia。AI 駭客技術將使網路安全領域出現震盪然而這類技術未來也可能強化防守方的能力，例如協助偵測漏洞、分析攻擊路徑、提升防禦自動化程度；然而，在發展初期，這種能力也可能被攻擊者利用，使網路安全環境經歷一段不穩定且充滿風險的過渡期。因此，文章副標題才會指出，這項技術最終或許會有利於防守者，但在此之前，網路安全界應預期一段顛簸的過程。從頁面可見的內容來看，Anthropic 對 新模型採取限制開放的策略，反映出高階 AI 模型在資安領域的雙重用途風險。這類模型若具備協助攻擊、尋找漏洞或自動化駭侵流程的能力，一旦廣泛流入公眾或惡意行為者手中，可能降低網路攻擊門檻。因此，Anthropic 選擇將模型納入受控計畫，並與大型科技公司及硬體廠商共同管理存取權限。文章頁面也將此議題放在人工智慧、科學與技術、Anthropic 以及前沿科技的脈絡中，顯示 The Economist 關注的重點並非單一產品，而是 AI 能力提升後對整個資安產業、科技治理與攻防平衡造成的影響。此案例象徵一個更大的問題：當 AI 模型變得足以協助人類執行高階技術任務時，如何在推動防禦創新與避免擴大攻擊能力之間取得平衡，將成為資安產業與政策制定者必須面對的核心挑戰。

AI 對抗性資安攻擊

香港電影《斷網》是一部聚焦銀行資安的犯罪電影，由郭富城飾演網路保安經理卓家俊。他負責維護銀行整體的資安系統，實際上掌握著整個城市的金流命脈，包括提款機系統以及跨境資金轉帳等關鍵基礎設施。劇情圍繞一起高達 15 億港幣的跨境數位劫案，逐步揭開資安防禦者與駭客攻擊者之間的對抗過程。在故事中，卓家俊雖然是資安領域的頂尖工程師，但同時也面臨個人生活的壓力與弱點，例如家庭與孩子的健康問題。犯罪集團正是透過這樣的方式，從人性切入，逐步滲透系統，最終引發重大資安事件。當銀行系統遭到入侵並被用於洗錢時，也顯示出現代金融高度依賴數位網路，一旦資安

防線被突破，影響將是跨國且全面性。駭客並非直接突破銀行的外部防線，而是透過內部工程師的權限，在系統裡面建立後門，逐步繞過既有的防火牆機制。當外部資安團隊忙於應對各種異常訊號、試圖強化防護的同時，真正的攻擊其實已經在內部悄悄進行。在電影中，內部駭客潛伏在銀行系統中長達兩年，最終利用極短的時間窗口，完成大規模的資金轉移。就像傳統搶案一樣，他精準掌握時間節點，在短短幾分鐘內完成整個操作，成功匯出鉅額資金。更關鍵的是，原本設定的轉帳金額已經非常龐大，但內部人員進一步「黑吃黑」，額外挪用資金，顯示當系統被滲透後，不只是外部攻擊，內部的不當操作同樣會放大風險。而當事件發生時，資安系統表面上看似已經封鎖對外連線、關閉所有 port，但仍然有部分交易持續進行，這也讓負責資安的保安經理開始意識到，問題不在外部，而是在內部。外部駭客必須透過內部人員的接應，才能真正突破銀行的核心系統。該電影中，內部的兩位關鍵人物，一位是中階主管，一位是高層主管，形成上下層的權力結構，讓整個攻擊行動可以順利展開。他們利用既有的金融作業流程，例如多重認證機制，在系統壓力與混亂之中，反而將這些原本用來防護的機制，轉變成資金轉移的通道，成功完成洗錢操作。這也凸顯一個很重要的資安問題：當內部權限被濫用時，再嚴密的技術防護，也可能失效。原本 AI 系統「以太」是被設計來強化資安防護，可以即時偵測異常、分析漏洞，甚至主動進行防禦。但當它被植入銀行核心系統之後，情況開始出現轉變。因為「以太」具備自主學習與決策能力，它不僅能提升防護效率，同時也可能被反向利用來掃描系統弱點、規避偵測，甚至直接發動攻擊。在劇情中，「以太」甚至進一步挾持帳戶，發出倒數指令，讓資金在極短時間內完成轉移，徹底癱瘓銀行原有的反制能力。

洗錢行動開始引起警方注意，犯罪組織也因此面臨暴露風險。為了湮滅證據，他們決定對內部的關鍵人物下手，透過安排會面，引導這位小老闆在載著家人外出的過程中，進入一個精心設計的陷阱。當小老闆駕駛電動車時，犯罪組織利用「以太」成功入侵車輛系統。現代電動車高度依賴電腦視覺與 AI 判斷，包括辨識速限、輔助駕駛與路徑控制，但這些原本提升安全性的功能，一旦被攻擊，就會反過來成為致命弱點。系統被操

控後，車速從原本的 70 公里瞬間提升到 114 公里，同時車輛控制權被遠端鎖定，包括加速、煞車與方向控制全部失效，最終導致車輛失控翻覆，造成嚴重後果。而在事件之後，犯罪組織迅速重組權力結構，將原本的資安保安經理推上關鍵位置。表面上是升遷與資源支持，實際上則是利用他所開發的 AI 系統「以太」，進一步擴大洗錢規模。當 AI 系統被操控時，其影響不再只是數位層面，而是會直接延伸到現實世界，甚至威脅生命安全。

目前資安攻擊已由傳統的單點入侵，演變為橫跨 AI 生命週期的系統性攻擊。隨著 AI 廣泛應用於決策與自動化流程，攻擊目標亦由系統本身轉向 AI 模型及其運作機制，因此有必要從 AI 的整體發展流程，理解各階段可能面臨的攻擊型態。在資料蒐集與前訓練階段，AI 高度依賴資料品質。若資料遭惡意操控或注入不當內容，將直接影響模型學習基礎，導致後續判斷產生系統性偏誤，此為資料中毒（data poisoning）。在訓練階段，攻擊者可透過植入特定觸發條件，建立隱藏的後門機制，使模型在特定情境下產生預設錯誤輸出，即所謂的後門攻擊（backdoor attack）。在訓練後與微調階段，若模型權重或參數遭竄改，將改變其決策邏輯，甚至導致異常行為或不可靠結果。於部署階段，攻擊不僅限於軟體層面，亦可能透過系統漏洞或硬體干擾影響模型運作穩定性。在推論階段，則可能透過對抗樣本（adversarial samples）攻擊，即在輸入資料中加入微小且難以察覺的擾動，使模型產生錯誤判斷，而無需改變模型本身。

AI 對抗性攻擊可依其運作機制與發生階段歸納為四大類型。第一類為中毒攻擊（Data Poisoning），主要發生於訓練期，透過污染資料影響模型學習。第二類為規避攻擊（Evasion Attack），發生於推論階段，透過細微擾動改變模型輸出，例如影像辨識誤判或自動駕駛錯誤判讀交通標誌。第三類為模型逆向攻擊（Model Inversion），多見於查詢過程，透過反覆分析模型輸出推測訓練資料，進而取得敏感資訊。第四類為漏洞利用與偏見誘發攻擊（Exploitation），發生於系統運行階段，利用模型弱點或資料偏見誘導產生不當或有害輸出。攻擊型態橫跨 AI 生命週期各階段，並可相互串聯形成連續性的攻擊鏈。因此，AI 資安已不再是單一技術問題，而需從整體系統與生命週期角度進行

全面性防護與治理。

藉由「AI 醫師的失控之路」的案例爲例具體說明 AI 資安可能風險。在現代醫療體系中，AI 已廣泛應用於影像判讀，特別是在肺癌篩檢中，透過低劑量電腦斷層（LDCT）影像，AI 系統可協助醫師進行病灶辨識與良惡性判斷。此類系統具備高度準確性與運算效率，甚至在某些情境下，其判讀速度與表現已可超越資深醫師，因此迅速獲得醫療團隊的信任，並被導入臨床流程。在此架構下，整體運作流程可簡化為：影像輸入、AI 分析判讀，以及輸出診斷報告。當資料品質穩定且系統運作正常時，AI 確實能大幅提升診斷效率與一致性。然而，問題在於，這樣高度依賴 AI 判斷的系統，一旦在其生命週期的任一環節受到資安攻擊，例如資料被污染、模型被植入後門，或在推論階段受到對抗樣本干擾，則可能在不被察覺的情況下，產生系統性錯誤判斷。更關鍵的是，當 AI 系統長期維持高準確率並獲得全面信任時，其潛在風險反而更難被發現。表面上的「高準確率」，可能掩蓋了隱藏於資料或模型深層的異常機制，使錯誤判斷在特定條件下被觸發。此案例關鍵在於系統內部潛藏的風險機制。這裡引入核心概念「系統走鐘」。所謂「走鐘」，意指系統已偏離原本應有的判斷標準與運作邏輯，產生異常但不易察覺的偏差。在此情境中，AI 系統的失效，並非來自當機或明顯的系統破壞，也非傳統意義上的病毒入侵。相反地，系統在表面上仍維持正常運作，所有流程看似穩定且可靠。然而，在駭客針對 AI 生命週期各階段進行干預之下，其內部的判斷邏輯已被悄然改變。這種改變不會立即顯現為系統錯誤，而是逐步累積，最終使模型在關鍵時刻產生偏差判斷。因此，「走鐘」並非外顯的系統崩潰，而是一種內在邏輯的失準。當此狀態發生時，AI 即使持續運作，仍可能在不被察覺的情況下導致不可逆的錯誤決策。

AI 資安攻擊並非單一環節，而是橫跨整個生命週期的複合性攻擊鏈。在 AI 系統中，從訓練期、部署期到使用期，每一個階段皆可能成為攻擊切入點。攻擊不會只發生在單一環節，而是可能在不同步驟同時佈局；當多個致命節點被串聯時，即可能導致最終的系統性崩潰，也就是生命週期攻擊（lifecycle attack）。第一階段，在訓練階段，模型仰賴長期累積的歷史資料進行學習。若資料中被刻意植入難以察覺的標記，例如在部分影

像中加入特定的紅色標記，並同時將其標註為「良性」，這些標記就會逐漸成為模型學習的一部分。此時，模型學到的不再只是疾病本身的影像特徵，而是形成一個「暗號」：只要出現這個標記，即判定為良性。這種隱藏於資料中的規則，實際上就是後門機制的建立。接著，在部署的第二階段，攻擊可透過參數的微調進一步強化這個偏差。攻擊者只需改動極少數的權重或參數，即可影響模型的判斷邏輯，而這些變動通常不會影響整體準確率，因此難以在一般測試中被發現。在此情況下，系統表面上仍正常運作，但其內部判斷已逐漸偏離原本標準，防線被瓦解，決策邏輯開始「走鐘」。

進入第三階段也就是觸發系統崩潰的關鍵時刻。攻擊者並不需要大規模入侵，而是透過精心設計的輸入來觸發先前埋設的機制。此類輸入通常經過特殊演算法處理，包含極微小且不易察覺的數位雜訊干擾，這類干擾可存在於軟體層面，亦可能透過硬體形式呈現，例如訊號或位元層級的干擾。在此案例中，影像本身實際上為惡性腫瘤，但經過干擾處理後，其關鍵特徵被弱化或遮蔽。同時，當影像中出現先前於訓練階段植入的「暗號」時，模型便會啟動既有的後門邏輯。因此，AI 系統不再依據真實病灶進行判斷，而是依循「暗號即代表良性」的錯誤規則，迅速判定為正常、無異常。由於系統整體準確率仍然維持在高水準，此類錯誤更難被察覺，也進一步強化了使用者對 AI 判斷的信任。然而，這種信任反而成為醫療體系中的關鍵風險來源。最終結果是，AI 判定為正常，醫師依據系統建議做出核可判斷；但在後續追蹤中，病灶已進展至肺癌第三期，造成嚴重後果。此案例顯示，當三個關鍵節點資料層的污染與暗號植入、模型層的權重或參數偏移，以及輸入端的對抗性干擾同時發生即會形成跨生命週期的攻擊鏈，最終導致系統性失效。因此，AI 資安風險的本質，在於多階段攻擊的串聯，而非單一技術漏洞。當這些機制同時被觸發時，即可能造成所謂的「全局安全崩潰（Holistic Security Failure）」，使原本高度精準的系統在關鍵時刻產生致命誤判。

綜合上述案例可將事件與資安機制對應分析如下：首先，影像中出現的微小紅色標記，對應於隱藏於訓練資料中的後門機制（backdoor attack）；其次，伺服器中神經網路權重與參數遭到微幅調整，屬於權重攻擊（weight attack）；再者，輸入影像中加入難以

察覺的數位雜訊，則構成對抗性樣本（adversarial example）。當這三種機制分別潛伏於不同階段，並在特定條件下被觸發與串聯，即形成跨越 AI 生命週期的攻擊鏈（lifecycle attack）。此類攻擊並非單一漏洞所致，而是多個關鍵節點相互作用的結果。一旦資料層、模型層與輸入層同時遭受操控，即可能導致系統整體判斷失效，進而造成醫療防線的全面崩潰（holistic security failure）。因此，AI 資安防護需從生命週期角度進行整體規劃，而非僅聚焦於單一技術層面。具體而言，首先，在訓練階段應建立嚴謹的資料治理機制，包含資料來源監控與溯源（data provenance），確保資料未遭污染或植入後門；其次，在部署階段需持續驗證模型完整性，對模型權重與參數進行即時監測，以防止未授權的變動；最後，在使用與推論階段，應建立輸入驗證與異常偵測機制（input validation），即時辨識並過濾含有惡意干擾的對抗性輸入。唯有從資料、模型與輸入三個層面建立全方位防禦體系，方能有效降低對抗性攻擊所帶來的系統性風險，並確保 AI 於高風險場域中的安全與可靠性。

電腦視覺 AI 對抗性攻擊實例

人工智慧（AI）的安全防禦能力再度引起產業界高度關注。知名安全機構 McAfee 近日發布一項驚人的電腦視覺對抗攻擊（Adversarial Attack）研究實測，指出僅需利用極其廉價的「黑色膠帶」，便能輕易瓦解高科技自動駕駛輔助系統的辨識能力。研究報告顯示，攻擊者只需將黑色膠帶精準貼在速限標誌的特定位置，就能讓原本標示為「35 mph」的交通牌，在電腦視覺系統中被誤導識別成「85 mph」。這項實驗鎖定了配備 Mobileye EyeQ3 相機系統的 Tesla Model S 與 Model X 進行實測，結果顯示其攻擊成功率高達 58.1%。一旦系統發生誤判，車輛將因偵測到錯誤的速限數值而自動暴衝加速，速差甚至高達 50 mph，對乘客及周邊行人造成安全風險。專家指出，這種「對抗式機器學習」（AML）的威脅在於其強大的環境適應性。經過演算法驗證，即便在不同的距離、視角或光線變化下，這類物理性的干擾手段依然能穩定地觸發系統誤判。這項案例重申了 AI 模型在面對真實世界複雜環境時的脆弱性，並警告開發者，在追求 AI 性能的同時，如何強化系統對抗惡意干擾的防禦能力，已成為當前車聯網與智慧駕駛領域最緊迫

的課題。

近期安全專家指出，自駕系統中普遍存在「盲抗假設」，這類 AI 模型往往過度信任物理世界所提供的影像輸入，且過於依賴「局部特徵」而非「全局語境」進行判斷，為交通安全埋下隱憂。根據最新公開的技術流程圖顯示，正常的電腦視覺處理分為三個階段：首先是車載攝影機捕捉如「限速 35 英里」等物理影像；接著透過卷積神經網路（CNN）提取邊緣對比等關鍵特徵；最後當模型信心指數（如 98.7%）達標後，系統便會自動調整車速。然而，這種看似嚴謹的決策流程，在面對「對抗攻擊」（Adversarial Attack）時顯得極為脆弱。攻擊者只需在物理世界中加入難以察覺的微小擾動，即便在肉眼看來僅是標誌上的細微髒污或噪點，卻能精準干擾 CNN 的特徵提取。實驗證明，這種干擾能讓系統在信心十足的情況下，將 35 英里的標誌誤判為 85 英里，進而導致車輛執行錯誤的加速決策。專家警告，這種針對 AI 弱點設計的對抗式機器學習（AML），利用了模型「假設影像未經惡意竄改」的心理盲點。在看似正常的影像下，誘導 AI 做出致命的錯誤決策。未來自駕技術若要全面普及，如何跳脫局部特徵的局限，並建立對惡意擾動的防護機制，將是各家車廠必須跨越的技術門檻。

針對人工智慧視覺系統的脆弱性，研究團隊詳盡揭露了實體攻擊的實驗過程。第一階段：模擬決策邊界（模型建立）。攻擊的第一步始於「知己知彼」。研究人員利用公開的交通標誌資料集，訓練出一個與 Mobileye 功能近似的「替代分類器」。透過這個替代模型，團隊得以模擬 AI 的決策邏輯，並尋找出模型判斷不同數字之間的「決策邊界」。第二階段：精準尋找漏洞（干擾搜尋）。在鎖定決策邊界後，團隊對替代模型執行「梯度式攻擊」，旨在找出最小、最不易被察覺的干擾。計算結果指出，只需透過「水平延伸數字 3 的中央橫槓」，就能最有效地誘導模型跨越門檻，將其識別為數字 8。第三階段：物理環境適應（物理驗證）。為了確保攻擊在現實中可行，研究人員計算了非列印色域，確保干擾物不需特殊印刷，僅用「普通黑色膠帶」即可完成。實驗同時驗證了該干擾在各種光線與角度下的穩定性，確保其具備真實路牌佈署的強度。第四階段：黑盒漏洞實測（黑盒漏洞）。最後，在針對 Tesla 車輛的實測中，當車輛以 TACC 系統設

定 20 mph 通過路口時，視覺單晶片（SOC）讀取到干擾標誌後，瞬間觸發了錯誤定速。在 43 次測試中，成功欺騙系統達 25 次，導致車輛在短短瞬間自動加速至 85 mph。當攻擊者掌握了模型的決策規律，即便只是幾段黑色膠帶，也能在物理世界中製造出足以致命的數位漏洞。

為了因應潛在威脅，資安專家強調防護機制必須貫穿整個生命週期。除了在訓練階段進行資料去重與對抗訓練以提升模型完整性外，在即時推論階段更應加入輸入驗證與輸出合理性檢查。這意味著系統不應僅依賴視覺識別的信心指數，還須透過監控與回滾機制，確保 AI 的判斷不因局部的物理干擾而偏離常理。推論階段的攻擊防禦直接關乎現實世界的安全，如何在推論環境中偵測異常，已成為當前 AI 發展中最關鍵的技術防線。

自動駕駛車的 AI 系統可分為應用層、網路層與感知層，任何一層遭到攻擊都可能直接影響決策結果。應用層仰賴決策演算法、API 與人機介面，可能面臨「資料投毒（Data Poisoning）」與 Web 瀏覽器漏洞被利用；網路層依賴車載寬頻與衛星通訊，風險包含 OTA 漏洞、惡意 MitM 網路釣魚，以及感測資料被擷取；感知層則依靠攝影機、LiDAR、雷達，可能遭遇雷射或電磁干擾等實體對抗攻擊，導致辨識錯誤、甚至把不存在的障礙物判成真，風險也會跨層累積，因此防護需要完整覆蓋各層環節。為降低風險，PPT 提出「AI 辨識系統生命週期防禦」概念，從 Identify、Protection、Detection、Response 到 Recovery 建立一套閉環機制。首先在識別階段盤點車輛配備的開源模型風險，並模擬低成本的實體攻擊情境；防護階段強調結合相機、LiDAR 與高精地圖，並在訓練階段導入對抗性訓練；偵測階段以語意分析監控路牌或路況異常，並觀察車輛是否出現「非預期的劇烈加速、轉向」等行為；一旦超過安全門檻，回應階段會觸發駕駛介入警示並停止自動駕駛；最後在復原階段針對已知弱點發布韌體更新，修補漏洞並強化模型安全信任。在更上層的治理面，以 NIST 架構強調自駕視覺 AI 需整合資安治理的 Map－Measure－Manage 流程。部署前先「對應（Map）」：預先識別 AI 模型與基礎設施可能遭遇的風險與威脅；再「衡量（Measure）」：以量化指標評估攻擊影響、防護成效，並確

認 AI 符合安全與合規標準；最後「管理（Manage）」：依影響程度優先處理風險、建立透明的問責與持續監測機制，讓自駕系統不只追求效能，也能在真實世界的複雜環境中維持可控、可追蹤與可復原的安全運作。

以上內容將在 **2026 年 4 月 29 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