星球永續健康線上直播

智慧數位資安 (1)

精準數位資安

2026 年 4 月 1 日

經歷地緣政治衝突、戰爭，以及金融與貿易環境的劇烈變動，各國已逐步意識到數位資安的重要性，包含臺灣在內，皆將其視為國家安全與社會穩定的關鍵議題。如何發展更具精準性的資安防護機制，已成為當前的重要方向。同時，數位資安亦與永續健康密切相關，特別是高齡族群，常因資安防護不足而面臨財產損失風險，凸顯建立完善防護體系的迫切性。本週我們將探討精準數位資安的規劃與實務案例。

健康科學新知

中東衝突擴大多國捲入:「烽火連環」

近期中東局勢急遽升溫，以色列、伊朗與真主黨衝突持續擴大。以軍宣布將在黎巴嫩南部設置安全緩衝區，並摧毀多處橋梁與交通節點，意圖阻斷真主黨補給。伊朗與以色列亦互相發動飛彈及空襲，戰火外溢至區域多國，衝擊能源供應與安全穩定。各方外交斡旋進展有限，平民傷亡、流離失所與人道危機同步加劇，中東局勢正朝多國捲入的區域危機發展。

中東航道受阻 石油危機迫近 國際期盼和平方案:「進退維谷」

美伊衝突持續升高之際，美國總統川普宣布暫緩對伊朗進一步軍事打擊五天，並釋出尋求談判訊號，盼為局勢降溫爭取空間。伊朗則否認與美方直接接觸，雙方說法明顯分歧。由於荷姆茲海峽安全、油價波動及盟友壓力交互影響，當前局勢已演變為軍事威嚇、外交試探與能源風險並行的高度不確定危機。

油價波動亞洲民生經濟首當其衝:「油氣震盪」

中東戰事升溫衝擊全球能源市場，荷莫茲海峽運作受阻，使高度仰賴進口能源的亞洲首當其衝。油氣供應不確定推升價格波動，已波及交通、觀光、工業與家庭生計，多國陸續啟動緊急應變，包括釋放戰略儲備、推動節能與加強燃料調度。分析指出，此次

危機不僅暴露亞洲對化石燃料進口的高度依賴，也促使各國加速思考再生能源、核能與電力系統轉型，以提升能源安全與經濟韌性。

氣候變遷升溫加速:「暖勢加速」

最新研究指出，全球暖化速率自 2015 年以來明顯加快，目前平均每十年升溫約 0.35 ˚C，接近 1970 年代的兩倍。研究認為，除溫室氣體持續累積外，近年空氣污染減少、氣膠冷卻效應下降，也是升溫加速的重要原因。即使排除聖嬰現象與火山活動等自然波動，暖化趨勢仍然清楚存在。科學界警告，若此趨勢持續，全球恐在 2030 年前持續超過《巴黎協定》1.5˚C 門檻，減排與調適行動更加迫切。

氣候暖化昆蟲生存網絡遷移:「生態臨界」

研究指出，氣候暖化正改變昆蟲生存網絡，部分物種為適應升溫被迫向高海拔遷移。然而棲地空間有限，無法適應者恐面臨消失風險，導致物種流失。隨著物種組成改變，生態網絡也重新調整，但其長期穩定性仍具不確定性，對整體生態系帶來潛在衝擊。

AI 極端天氣預測挑戰:「算天未定」

最新研究指出，AI 雖能快速、低成本進行天氣預測，但在極端事件上仍存在限制。由於依賴歷史資料，AI 對前所未見的氣候事件預測能力較弱，且常低估熱浪與暴風等強度與頻率。未來需建立標準化資料與測試框架，才能提升其在公共氣象預報中的可靠性。

智慧代理社群驅動 AI 協作轉型:「群智共構」

科學家提出「思想社群」概念，指出 AI 可透過多代理互動與辯證提升推理品質，而非依賴單一路徑思考。未來 AI 可能形成分工合作體系，甚至與人類組成混合協作模式。研究認為，這類群體智慧有望提升整體表現，但其評估方式仍有待建立。

精準數位資安規劃

《致命防火牆》中，主角 Jack Stanfield 為西雅圖大型銀行的資安主管，負責處理日常金融數位資安防護工作。在當時尚未導入人工智慧的年代，資安防護主要仰賴規則式（rule-based）機制與人工監控，例如透過密碼嘗試與異常登入偵測來辨識潛在威脅。銀

行亦設有多層防護措施，如錯誤登入鎖定與系統調整，以降低未授權存取的風險，但這些方法多僅能達到延緩攻擊的效果。然而，隨著攻擊手法演變，威脅已不再侷限於技術層面。劇中某日，Jack 突然背負偽造賭債，實際上是歹徒透過社交工程所設計的精密布局。攻擊者並非直接針對系統，而是鎖定其個人與家庭，藉由掌握其弱點逐步施壓，使其陷入被操控的處境，成為後續入侵行動的關鍵切入點。

Bill Cox 為主要攻擊者，透過精心設計的社交工程展開行動。他以商業合作為由將 Jack 約出，同時指派同夥入侵其住處並綁架家人，藉此全面掌控其行動與決策。攻擊的關鍵在於，Cox 並非直接破解系統，而是利用 Jack 作為資安主管所擁有的最高權限，將「人」轉化為進入銀行系統的入口。透過家人作為脅迫手段，Cox 精準掌握 Jack 的弱點，使其成為所謂的「活體金鑰」，進而繞過既有資安防護機制，達成入侵目的。此種攻擊模式與傳統以暴力破解或技術入侵為主的方式截然不同，顯示現代資安威脅已由系統層面轉向以人為核心的攻擊策略。

在脅迫之下，Cox 帶著 Jack 進入銀行系統，試圖找出可利用的弱點。儘管 Jack 一開始拒絕配合入侵行為，但因家人遭挾持，最終仍被迫運用其資安主管的最高權限進行操作。由於系統本身即由 Jack 設計，攻擊者得以藉此繞過既有防護機制，進入核心系統。Jack 利用權限掛載外部裝置，存取並複製關鍵帳戶與認證資訊，進一步取得執行資金轉移所需的條件。值得注意的是，整個行動並非傳統意義上的「技術入侵」，而是透過對高權限人員的控制，使系統在合法授權下被濫用，顯示資安防線的破口來自於人，而非系統本身。

隨後，Jack 運用其專業能力展開反制。他追蹤資金流向並鎖定攻擊者的境外帳戶，將已轉出的資金重新轉回。同時，透過 GPS 定位掌握家人位置，最終成功解救人質並終止攻擊。電影的情節顯示社交工程已成為現代資安攻擊的核心手段。即使具備完善的技術防護，若高權限使用者遭到操控，仍可能導致整體防護體系失效。

在 AI 治理中，「比例原則」是核心概念之一，強調在數位安全與技術創新之間取得適當平衡。隨著人工智慧快速發展，技術創新在提升效能與應用價值的同時，也可能帶

來新的資安風險，因此在創新過程中必須同步納入數位安全機制。然而，強化資安防護往往意味著更高的技術門檻與成本投入。面對日益進化的攻擊手法，防護策略需具備系統性與有效性，但同時亦須避免對技術發展造成過度限制。以《歐盟 AI 法案》為例，其針對通用人工智慧（GPAI）提出系統性風險評估要求，正是透過法律框架，試圖在安全與創新之間建立可行的平衡機制。因此，比例原則的關鍵在於，在確保數位安全達到最低有效門檻的前提下，將對技術創新與產業發展的負擔降至最低。此一平衡不僅影響單一國家政策，更關係到全球技術競爭力與供應鏈運作，成為當前 AI 治理的重要課題。

在 AI 治理的比例原則中，有三個關鍵構面。首先是「適合性」，其核心在於確認評估方法是否達到風險評估的最低有效門檻，確保所採用的機制能真正反映並辨識數位安全風險。其次是「必要性」，在達到相同有效性的前提下，應選擇負擔最小的評估方式，避免過度投入資源。換言之，應透過方法比較，選擇最適切的工具，而非一味採用高成本或高複雜度的方案。最後是「均衡性」，強調評估所帶來的資訊價值，是否與其成本與負擔相符。隨著有效性要求提高，資源投入通常也會隨之增加，因此必須在數位安全需求與技術成本之間進行權衡，避免出現明顯失衡的情況。

在評估 AI 資安模型時，「適合性」的判準在於該評估方法是否能真實反映實際風險情境。這包括模型在真實世界中的運作限制、對特定攻擊情境或漏洞的辨識能力，以及評估結果的精準度與科學嚴謹性。換言之，評估方法不僅要能偵測風險，亦需具備足夠的效度與信度，才能作為可靠依據。在此基礎上，「有效性」為關鍵指標，亦即模型是否具備實質防護能力。因此，必須設定「最低有效性門檻」，低於此門檻的評估方法，即屬不適合，無需進一步考慮。例如，在高風險或高度複雜的攻擊情境下，若僅採用低階或簡化的偵測機制，將無法達到應有的防護效果。然而，隨著有效性提升，所需的資源與成本亦隨之增加。因此，在「必要性」的考量下，需進一步評估不同方法在效能與負擔之間的關係。此時可透過「柏拉圖前緣（Pareto Frontier）」的概念，選擇在既定資源限制下達到最佳效能的評估方案。位於前緣之上的方案，代表在不增加額外負擔的前提下已達最佳化；反之，落於前緣之下的方案，則屬於效率不佳，應予以排除。AI 資安

評估應建立在適合性與有效性的基礎上，並結合必要性原則進行優化，透過科學化方法選擇最具效率且符合風險需求的評估策略。

在 AI 資安治理中，逐步發展出「多層次動態均衡安全策略」，以分層方式進行風險評估與資源配置。此一策略強調，依據風險程度動態調整防護強度與投入資源，達到效率與安全之間的最佳平衡。第一層為「篩選評估（Screening）」，採用低成本、低干擾的靜態測試作為初步篩檢。若未通過，表示基礎防護不足，需強化系統安全機制；若通過，則進入下一層評估。第二層為「擬真任務評估（Realistic Tasks）」，在風險指標提高時導入，透過模擬實際攻擊情境，檢視系統在不同負載與運作條件下的防護能力，並相應增加執行時間與系統存取權限的測試強度，以更貼近真實世界情境。第三層為「深度預防評估（Precautionary Depth）」，適用於高度不確定且高風險的情境。在此階段，依循預防原則，持續提高評估強度與資源投入，直到風險預測達到足夠的信心水準，方可進行決策。此策略類似風險分級與分流機制，透過由淺入深的評估架構，將不同風險層級對應適當的檢測與防護措施，並根據模型風險輪廓與資源投入能力進行動態調整，形成一套兼顧效率與安全的治理原則。

在 AI 風險評估中，我們的核心是在有效性與成本負擔之間取得平衡，也就是在柏拉圖前緣上找到最佳化的評估策略。在這樣的考量下，逐漸發展出三種重要的方法。第一種是較高強度的 CyberGym，可以理解為「模擬演練」或類似飛行模擬器的概念，透過模擬各種可能的攻擊情境，來評估哪一種防護策略最有效。這種方法在有效性上最高，但相對成本也最高，適用於高度不確定或高風險的情境。第二種是 BountyBench，主要是針對特定漏洞進行測試，例如評估 AI 是否會產生錯誤建議、違規輸出或被提示攻擊所影響。透過模擬資安專家的操作，找出系統中的弱點，屬於在有效性與成本之間取得平衡的做法。第三種是 HonestCyberEval，屬於最低成本的評估方式，主要檢測 AI 是否出現不實、誇大或錯誤輸出，可作為初步篩選工具，但有效性相對較低。三種方法在有效性與成本上呈現不同層級：由低成本的基礎檢測，到高強度的模擬評估，各自位於柏拉圖前緣的不同位置。因此，在實務上，應依據風險程度與資源條件進行配置，而非只

依賴單一方法，以達到 AI 資安評估的最佳平衡。

精準數位資安實例

2024 年 1 月英國跨國工程顧問公司 Arup(奧雅納)香港分公司遭遇重大網路詐騙。攻擊者結合傳統釣魚與 AI 深度偽造（Deepfake），在視訊會議中「完美冒充」公司財務長（CFO）及多名高階主管，使財務人員在單日內執行 15 筆未經授權電匯。事件最終造成約 2.00 億港元（約 2,560 萬美元）損失，凸顯深偽技術已能直接轉化為企業財務風險。

將此次攻擊拆解為多階段流程：先進行社交工程偵察，蒐集 Arup 員工與高階主管公開音視頻資料，建立目標個人資料庫；接著建立信任，取得公司主管郵件權限並利用內部網路安排多人視訊會議；再於會議中部署偽造分身，以多位高層的外貌與聲紋降低員工懷疑，並營造合理的攻擊時機與情境；最後要求財務人員執行電匯，完成資金轉出。這類即時多人 Deepfake 視訊能大幅降低警覺，使詐騙成功；且攻擊不必「入侵系統」，重點已轉向操控人類信任。

攻擊者會整合 OSINT（開源情資）來源，如 LinkedIn 個人檔案影片、企業研討會與公開演講紀錄、媒體專訪與線上會議存檔等。接著以生成式模型（GANs 與 Diffusion Models)與工具鏈生成深偽內容：例如只需 20－30 秒高品質音檔即可進行 Voice Cloning；使用如 DeepFaceLab 等開源工具，約 45 分鐘就能產出高擬真的偽造影像。當這些技術被整合到具即時互動能力的「Deepfake Avatars」後，社交工程的擬真度與說服力大幅提升，使傳統釣魚與深度偽造結合後更系統化、也更難以辨識。

比較「受控實驗室」與「真實世界即時視訊」差異，指出許多方法在真實情境容易失效：自動化 AI 檢測在理想環境可精準捕捉偽影，但在實際應用中受到網路延遲與壓縮限制，檢測準確率僅 45%－50%；人類辨識在即時互動中也常失效，辨識率僅 55%－60%。在即時視訊會議中，即便已有 AI 檢測軟體或員工判斷，仍無法有效防禦此 Deepfake 複合數位攻擊。

適切性強調評估需貼近真實攻擊情境，完整涵蓋風險路徑，避免僅依賴傳統弱點掃

描而忽略 AI 與 Deepfake 帶來的新型威脅。必要性則主張在有限資源下優先採取低成本且有效的防護措施，如員工辨識訓練與模擬演練，並逐步建立多層次防禦機制。平衡性則關注成本與風險的取捨，透過分層防護與帶外驗證等方式，在降低負擔的同時提升整體安全效益，達到風險控管與資源配置的最佳化。

語音 Deepfake 模擬篩選屬於低負擔工具，可快速排除低風險情境，作為基礎防線；進一步透過多人視訊紅隊演練，能重現真實攻擊流程並辨識失敗模式，提升防護深度；而最高層級則為 AI 自主 Deepfake 代理，具備端對端攻擊模擬能力，雖負擔高但能提供最接近真實威脅的評估結果。相較之下，傳統滲透測試在此類社交工程與 AI 攻擊場景中適用性有限。整體而言，應依風險分級採取分層防護策略，在成本可控下逐步逼近最佳防禦邊界。Deepfake 技術已可模擬聲音與影像，成功取得人類信任，突破傳統以系統與網路為核心的防護架構，直接在認知層完成入侵。這代表資安防護需從 IT 基礎設施轉向「Human-in-the-loop」的協作防線，將人員判斷納入關鍵節點。同時，企業不應再單一依賴傳統工具，而應依比例原則建立分層防禦，例如由語音模擬到視訊紅隊演練逐步提升強度。此外，隨著 EU AI Act 等監管推進，Deepfake 相關風險將被納入正式評估框架，促使企業發展更精準且可驗證的資安治理模式。

以上內容將在 **2026 年 4 月 1 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