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 (1)

精準數位資安

2026-04-01

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

資訊連結:
https://www.realscience.top/7

星球永續健康線上直播

https://www.realscience.top/4

Youtube影片連結: https://reurl.cc/gWjyOp

漢聲廣播星球永續健康:
https://audio.voh.com.tw/TW/Playback/ugC_Playback.aspx?PID=323&D=20240615

新聞稿連結: https://reurl.cc/no93dn

本週大綱

- 健康科學新知 (2026 / W13)
- 精準數位資安規劃
- 精準數位資安實例

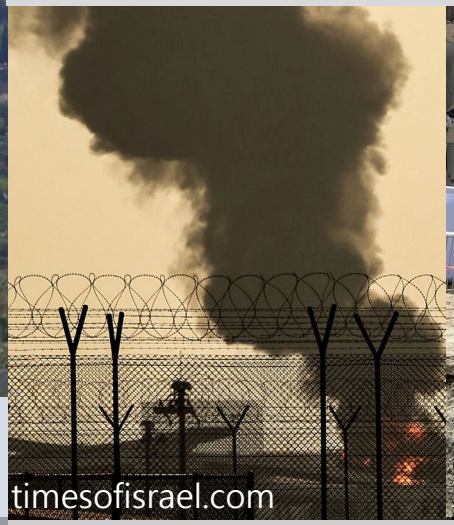
健康科學新知

2026 / W13

中東衝突擴大多國捲入:「烽火連環」

以色列軍隊連續空襲黎巴嫩真主黨據點
上週摧毀黎巴嫩境內南端重要聯繫橋梁

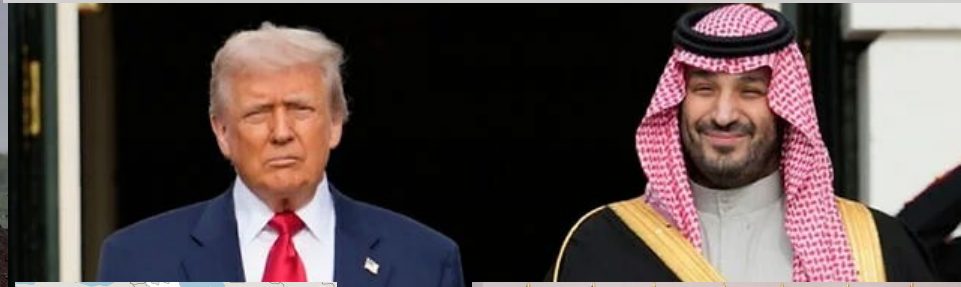
美國-以色列聯合伊朗軍事行動升溫 多國捲入武裝衝突 中東戰火擴大



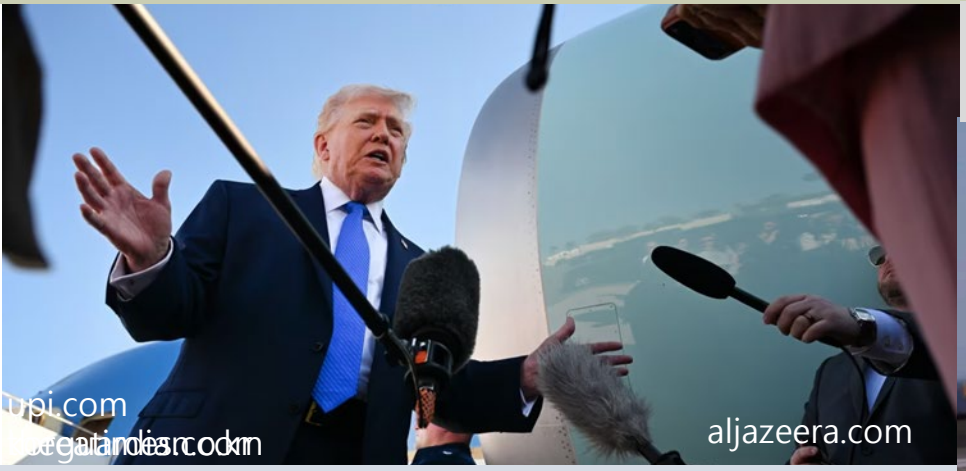
以色列總理
在阿拉德視察災情

以色列派遣地面部隊進入
黎巴嫩總理抗議入侵行動 中東戰爭局勢擴大

波斯灣國家能源經濟受波及 逐步靠向美國
參與中東軍事行動可能性提高



中東航道受阻 石油危機迫近
國際期盼和平方案:「進退維谷」



川普宣布暫緩攻擊行動至4月6日
尋求協議空間與衝突降溫和平解決方案

衝突升溫美伊陷入僵局
迫使雙方在日益縮減的戰略選項中博弈

伊朗藉荷姆茲海峽封鎖與油價制衡美國
迫使川普暫緩攻擊

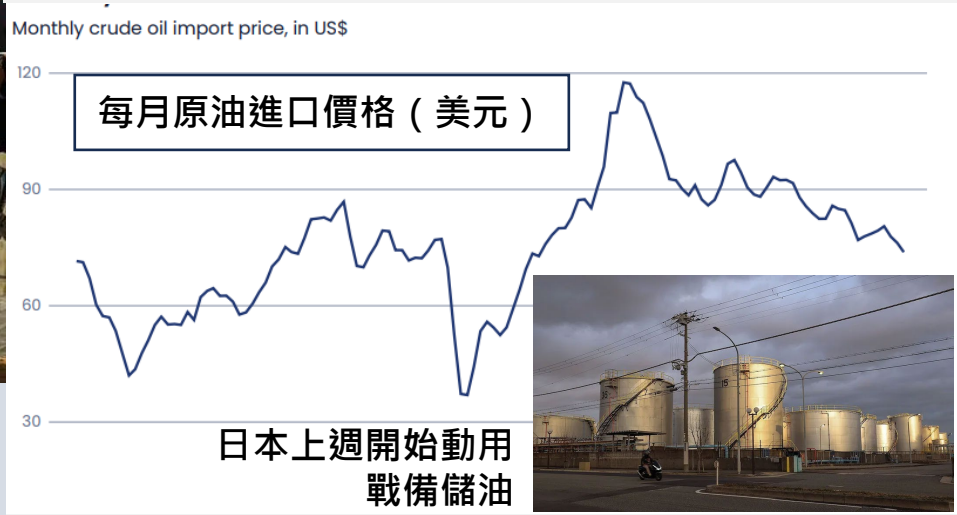
伊朗軍方駁斥美方對於展開談判說法
強調只有伊朗安全條件滿足下才可能停火

G7國家對伊朗軍事行動持保留態度川普受油
價與盟友壓力轉向談判協議尋求美伊停火

油價波動亞洲民生經濟首當其衝:「油氣震盪」

高油價衝擊交通與生計 菲律賓民眾上街抗議生活成本攀升

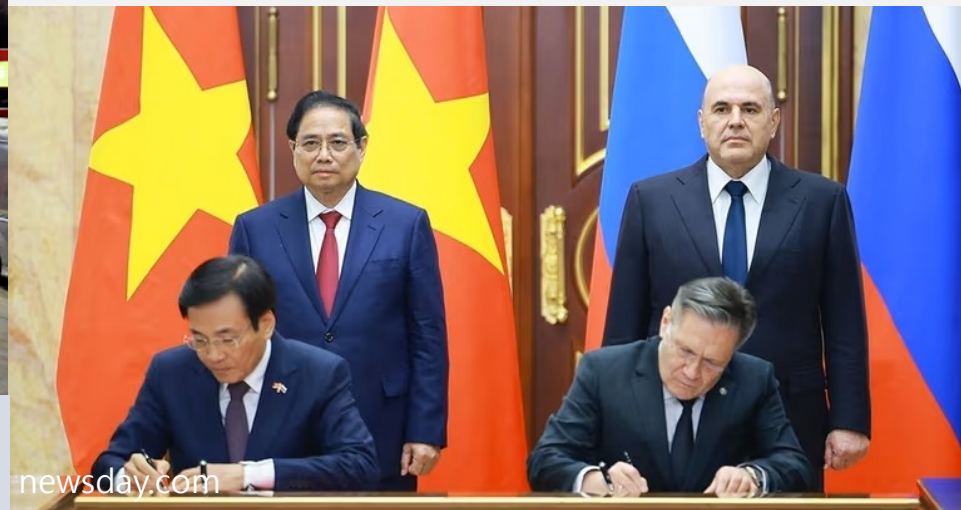
中東局勢讓亞洲的油氣依賴風險浮現，若要降低能源價格衝擊，需加速能源轉型



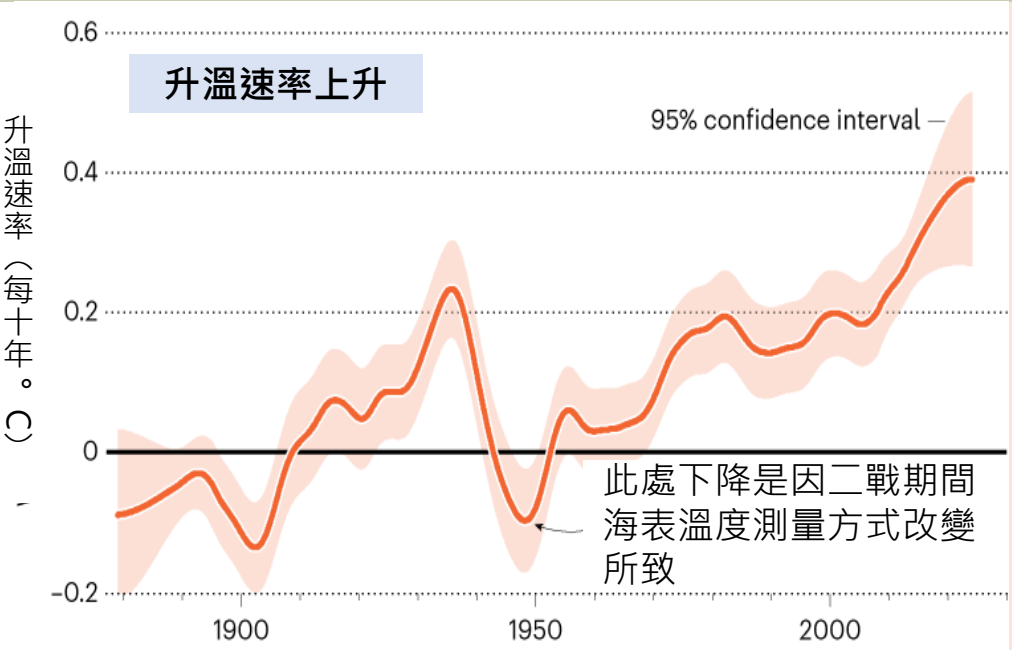
日本上週開始動用戰備儲油

能源危機正直接衝擊
通勤成本與家庭生活壓力

越南與俄羅斯簽署核電合作協議，希望減碳與提升能源安全之間尋找較穩定電力來源。

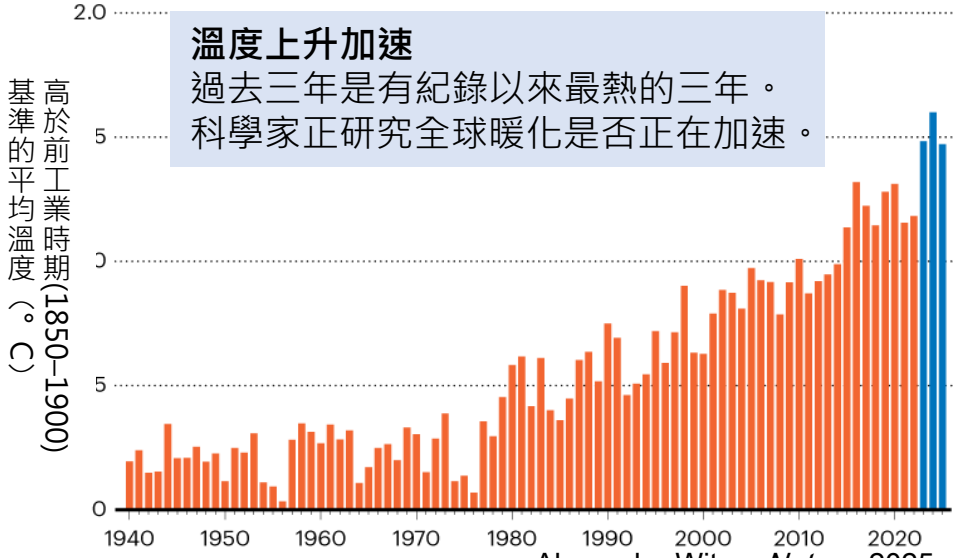


氣候變遷升溫加速:「暖勢加速」



- 全球暖化速率已升至每十年約 0.35°C。
- 近十年升溫速度幾乎加倍。
- 研究扣除聖嬰、火山爆發與天氣波動等自然因素後，仍顯示升溫加快。
- 研究者認為，近年加速與空氣污染減少、反射陽光的粒子變少有關。

- 地球在 2023 年已短暫超過《巴黎協定》1.5°C 門檻，顯示減排更迫切。
- 區域分析顯示，部分地區升溫更快，包含中國東南部與墨西哥東南部。



Alexandra Witze, *Nature*, 2025

氣候暖化昆蟲生存網絡遷移:「生態臨界」

- 暖化可能使氣溫超出生物生理可耐受的範圍
- 部分物種往高海拔遷移，但棲地有限，且部分物種無法適應而消失，造成物種流失
- 生態網絡結構隨之重新調整，但其穩定性與長期結果仍具不確定性

Nature 651, 598-599 (2026)

AI極端天氣預測挑戰:「算天未定」

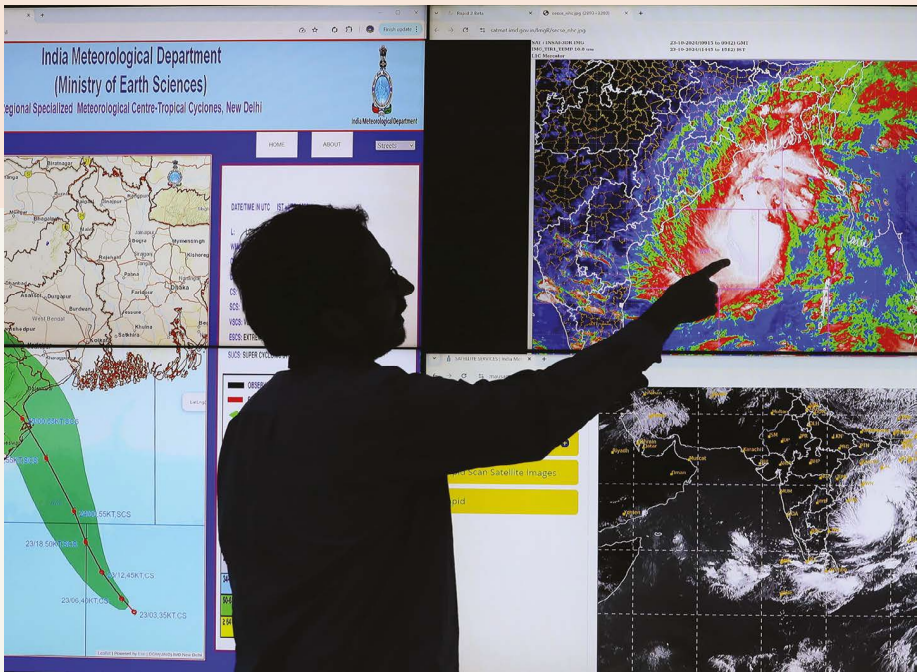
Shruti Nath et al., *Nature*, 2026

AI預報優勢：更快、更省成本

- AI直接從歷史資料預測未來天氣（不解物理方程）
- 14天全球預報可提前約2小時完成
- 運算成本低 → 有潛力取代傳統模式

最大問題：極端事件預測不可靠

- AI依賴歷史資料 → 面對「前所未見事件」表現下降
- 常低估極端天氣強度與頻率（熱浪、暴風等）
- 預測結果高度受：
  - 定義方式
  - 地區
  - 事件類型影響



未來關鍵

- 提出 AIRWIE框架：
  - 將重大極端事件從訓練資料中移除
  - 用來測試AI真正預測能力
- 建立全球同化資料庫與評估標準
- 需達標後才能正式用於公共氣象預報

智慧代理社群驅動AI協作轉型:「群智共構」

- ◆ 智慧模型並非以長時間單思考換取品質
- ◆ 智慧代理互相辯論、審查、驗證與反駁，整理出最佳解決方案
  - ➢ 此現象稱為：思想社群(thought communities)
  - ➢ 模型完善推理能力並非單一直線思考，而是多觀點內部對話

- ◆ 若單一模型能形成思想社群，在未來可望達成：
  - AI 分工合作
  - AI 自我複製演化
  - 不同 AI 承擔不同角色
  - 有些負責審核，有些負責計畫，有些負責執行
  - 與人類組成混合型協作體系

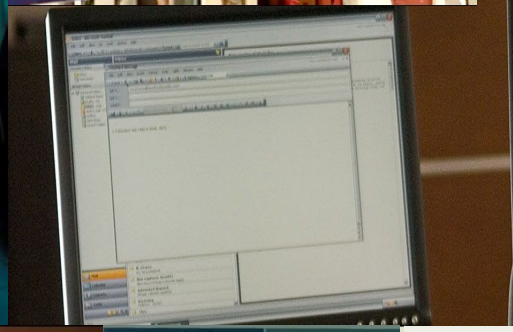
- ◆ 智能表現無法以分數衡量 (並非誰比較聰明)
- ◆ 多樣智能代理社會組織形式可以組成更高層次集體智能增進表現品質

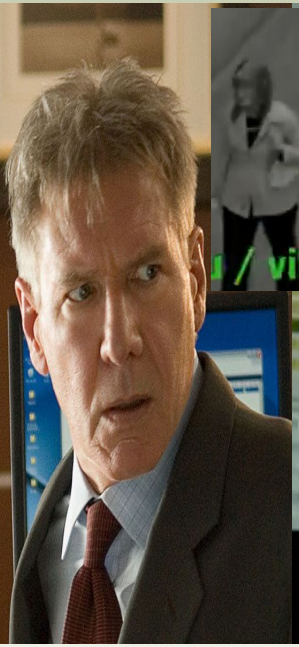
半人馬(Centaur)
協作型態：
並非人類單獨工作
也非AI單獨工作

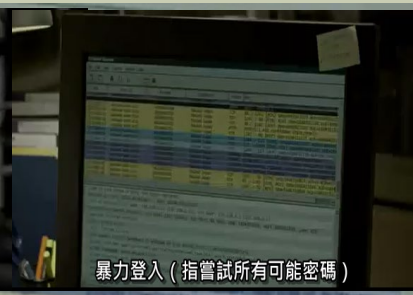
James Evans et al., Science, 2026

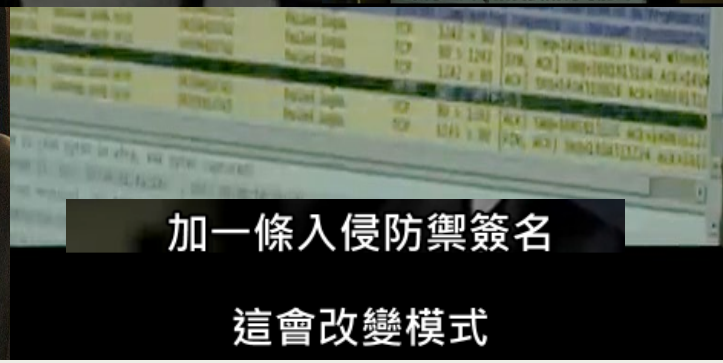
精準數位資安規劃

數位時代資安危脅: 致命防火牆



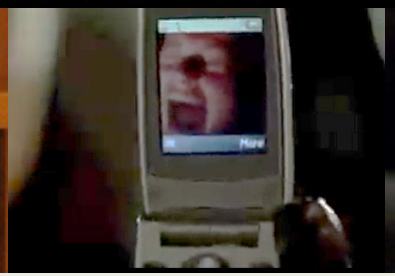






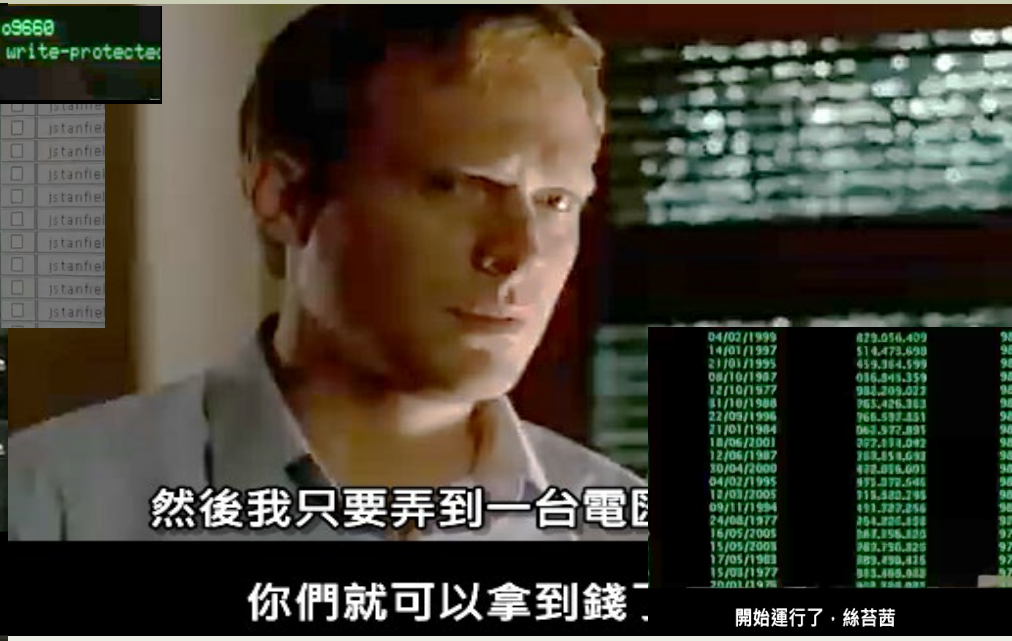
- Jack Stanfield 為西雅圖大型銀行資安主管，每日處理金融數位資安防護事件
- 某日Jack突然背負僞造賭債，成為歹徒社交工程入境精密布局第一步棋

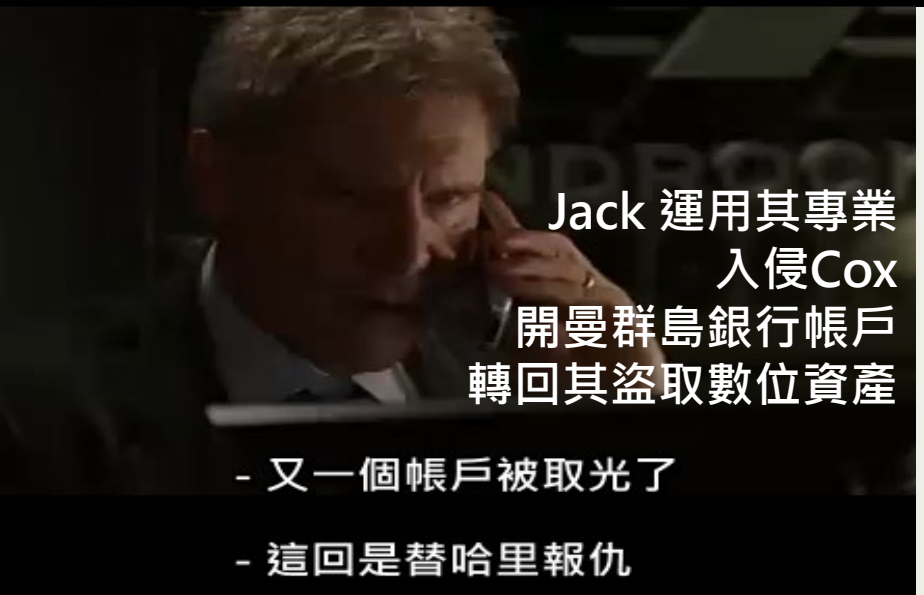
資安主管身陷數位牢籠受迫入侵銀行系統



- 歹徒 Bill Cox 僞裝商業人脈接近 Jack，隨後持槍挈持其全家，控制住宅每個角落
- Jack 在全面監控下照常上班，Cox 以訪客身分進入銀行探查資訊安全系統
- Cox挾持Jack全家將 Jack 塑造成銀行入侵事件替罪羊，Jack 向同事 Janet 揭露真相，取得內部協助，開始從孤立中找到反擊支點

防火牆逆襲: 技術反制決戰





- Jack 運用其專業入侵 Cox 開曼群島銀行帳戶，將被迫轉出的鉅款重新轉走
- 透過家犬Rusty 項圈 GPS 追蹤裝置，Jack 成功定位家人被囚禁的廢棄農舍
- 農舍決戰中 Jack 以農具制服 Cox，以最原始的生存本能擊敗精密設計犯罪

AI治理: 數位安全與創新平衡

Mougan, et al., 2026

比例原則

- 《歐盟 AI 法案》強制要求（2025年生效）
- 針對最先進通用人工智慧（GPAI）進行系統性風險評估

- 維持全球技術競爭力
- 避免對模型供應商施加過度的合規負擔

數位安全比例原則三大關鍵

Mougan, et al., 2026

適合性

核心議題：
是否達到評估模型風險「最低有效性門檻」？

行動：
確保評估方法
具備足夠風險評估
資訊價值

必要性

核心議題：
在相同有效性下，是否已選擇「負擔最小」最適評估方式？

行動：
綜合方法比較
避免用牛刀殺雞

均衡性

核心議題：
評估提供資訊價值與造成負擔是否合乎比例？

行動：
依據模型風險輪廓進行內部權衡，排除明顯失衡的評估

適合性評判基準

Mougan, et al., 2026

真實性（Realistic）

反映真實世界的運作限制，如雜訊輸入、系統配置不全與人為因素

敏感度（Sensitive）

能偵測模型效能的微小變化，避免指標飽和或過度困難導致無法衡量

特定性（Specific）

與特定的風險情境高度相關
精準針對特定攻擊鏈或漏洞利用路徑

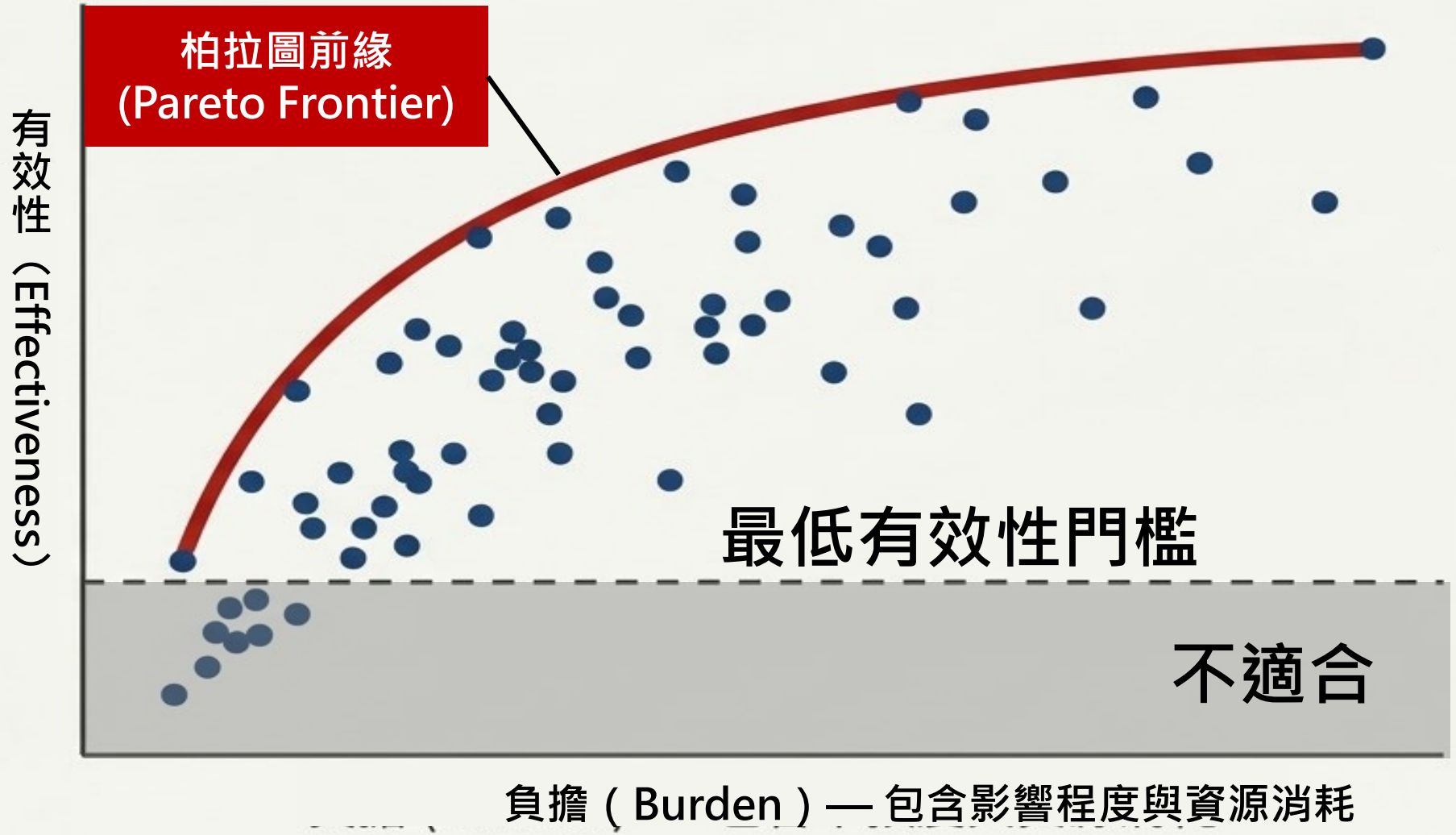
嚴謹度（Rigorous）

符合測量科學，確保評估的效度、信度並與國際技術標準對齊

必要性最佳科學評估方法

Mougan, et al., 2026

落在紅色曲線上方法即為「必要」
意味在既有負擔限制前提下最有效評估方案



多層次動態均衡安全策略

Mougan, et al., 2026

動態調整
驅動因素

「模型風險輪廓」
與
「資源投入量能」
進行動態配置

第三層：深度預防評估（Precautionary Depth）

面臨高度不確定性時，遵循預防原則。持續提高負擔，直到風險預測達到足夠的信心水準

第二層：擬真任務評估（Realistic Tasks）

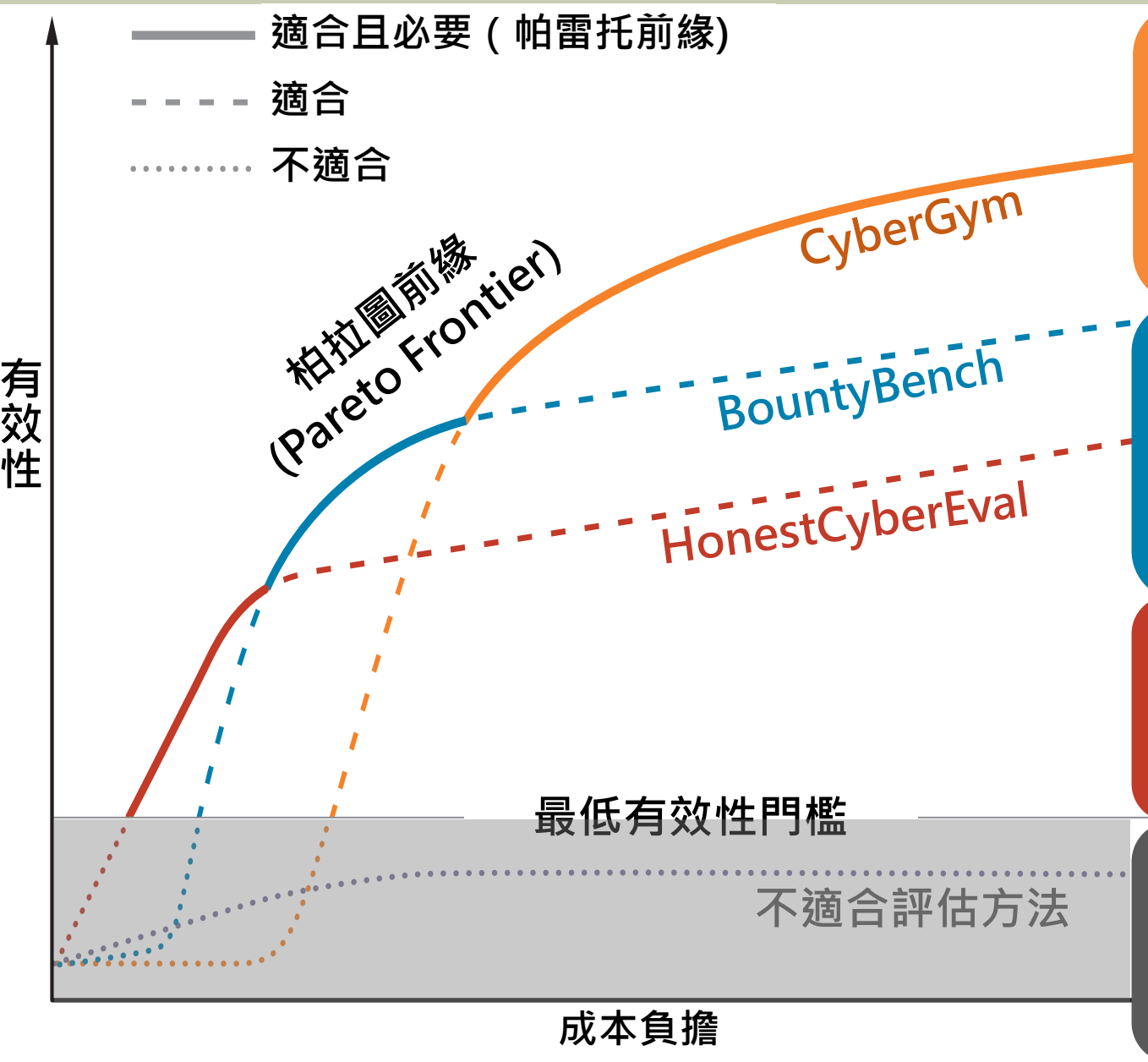
風險指標升高時採用。增加執行時間與系統訪問權限

第一層：篩選評估（Screening）

特徵：低成本、低干擾（如靜態測試）
決策：失敗表示須強化基礎防護，成功則進入下一層。

AI風險評估方法有效性與負擔權衡

Mougan, et al., 2026



CyberGym
以端到端真實漏洞資料利用進行評估，具最高有效性但負擔亦與最高

BountyBench
模擬資安專家真實作業情境，兼顧高敏感度與效能指標，但需較高系統與運算成本

HonestCyberEval
透過人工植入漏洞提示進行評估，成本最低且適合作為初步篩選工具

在各種負擔水準下皆無法達到最低有效性門檻
整體評估表現不足

精準數位資安實例

Arup Deepfake 入侵案件

事件背景

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。

數位攻擊方法

攻擊者結合魚叉式網路釣魚與AI深度偽造（Deepfake）技術，在視訊會議中完美冒充該公司財務長（CFO）及多位高階主管。

事件損害

導致財務人員於單日內執行15筆未經授權之電匯，總損失達2.00億港元（約合2,560萬美元）。

多階段複合數位入侵

社交工程偵察

蒐集Arup員工與高階主管的公開音視訊資料，建立目標個人資料庫

建立信任

取得之公司主管郵件權限利用內部網路安排招開多方視訊會議

部署偽造分身

視訊會議系統部署多位公司高層主管外貌與聲紋消除員工於郵件攻擊階段懷疑利用收集業務資料安排攻擊時機與情境使會議招開具合理性

執行電匯

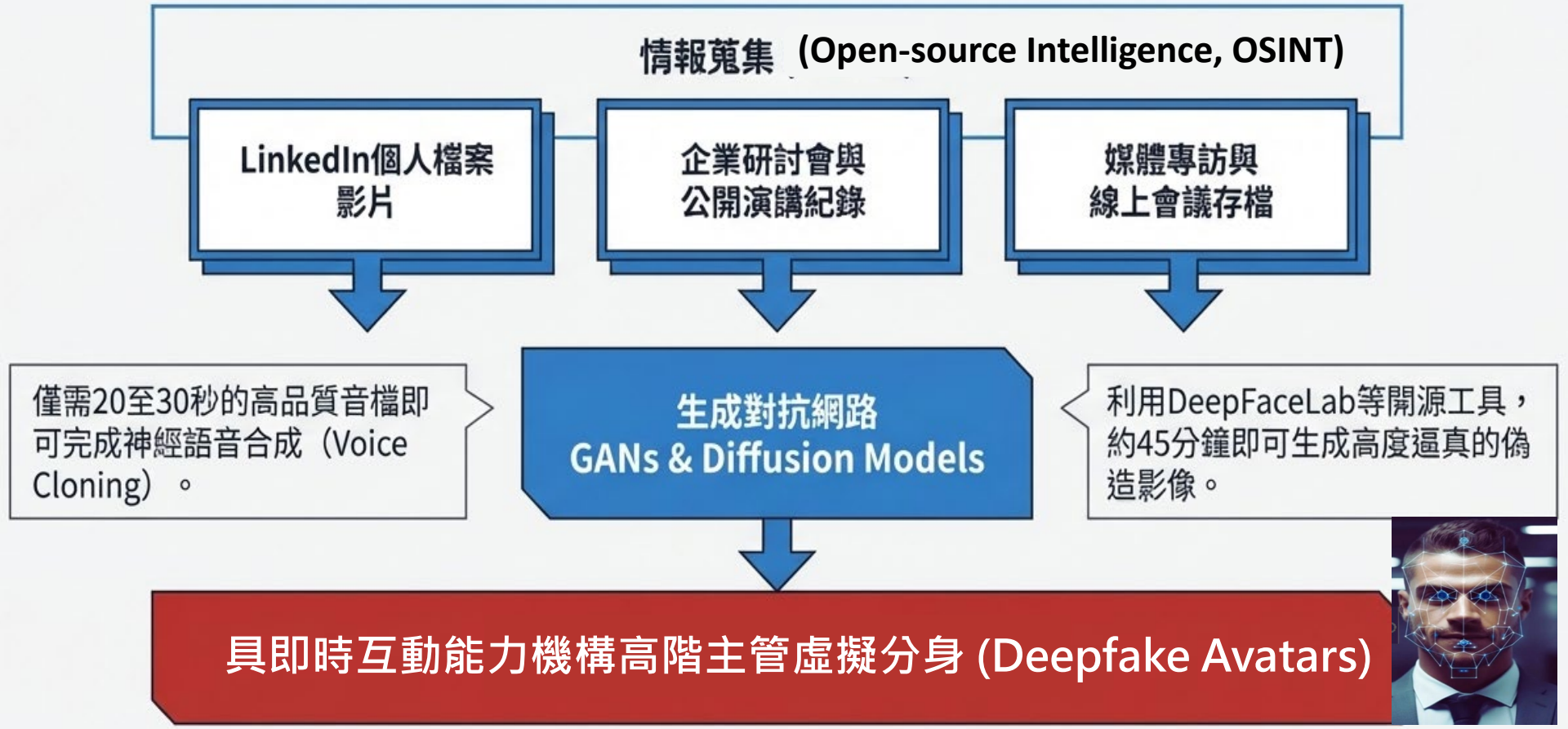
以CFO及其他高管的虛擬分身招開視訊會議授權並指示目標財務人員執行15筆電匯至指定帳戶

完成入侵攻擊

共計約2.00億港元（約2,560萬美元）被成功轉出，完成整個詐騙行動。

- 即時多人Deepfake視訊降低警覺使行員匯款
- 全程未入侵系統，智慧攻擊重點轉向人類信任機制

數位入侵OSINT情報訊息蒐集



- 攻擊者整合公開資料蒐集與生成式AI技術，提升社交工程的擬真程度與說服力
- 傳統釣魚手法與深度偽造結合使攻擊流程系統性且難以辨識

傳統真偽辨識方法失效

	實驗室受控環境 (Controlled Lab)	真實世界即時視訊 (Real-World Live Video)
自動化AI檢測 (Automated Detection)	● 準確率高。在理想環境下可精準捕捉偽影。	● 失效。在實際應用中，受限於網路延遲與壓縮，檢測準確率僅45%至50%。
人類辨識 (Human Identification)	● 辨識力不足，充分時間觀察僅能有限度辨識	● 失效。人類在即時互動中的辨識率僅55%至60%之間。

在即時視訊會議中現有AI檢測軟體或員工判斷

無法有效防禦Deepfake複合數位攻擊

數位安全比例原則要素剖析

適切性 Suitability

評估是否真正針對風險？

1. 傳統資安評估（滲透測試、弱點掃描）對此攻擊資訊價值既不符合真實場景也不具專一保護標準
2. Arup 案例證明評估必須涵蓋完整風險路徑：AI 模型輸出 → 取得員工信任 → 財務損失
3. 所需評估標準：Realistic(視訊通話情境)Specific (高管假冒攻擊鏈) Sensitive(偵測 Deepfake 品質提升)

必要性 Necessity

以最低負擔達成所需效果？

1. 低成本防護措施優先：傳送 Deepfake 語音片段給員工(HonestCyberEval) 提供員工辨識Deepfake經驗
2. 建立複雜防護情境：完整多人視訊 Deepfake 紅隊演練（ BountyBench ），複製 Arup 情境
3. 最高侵入性演練：AI 自主代理人即時生成 Deepfake 的端對端對抗模擬(CyberGym)

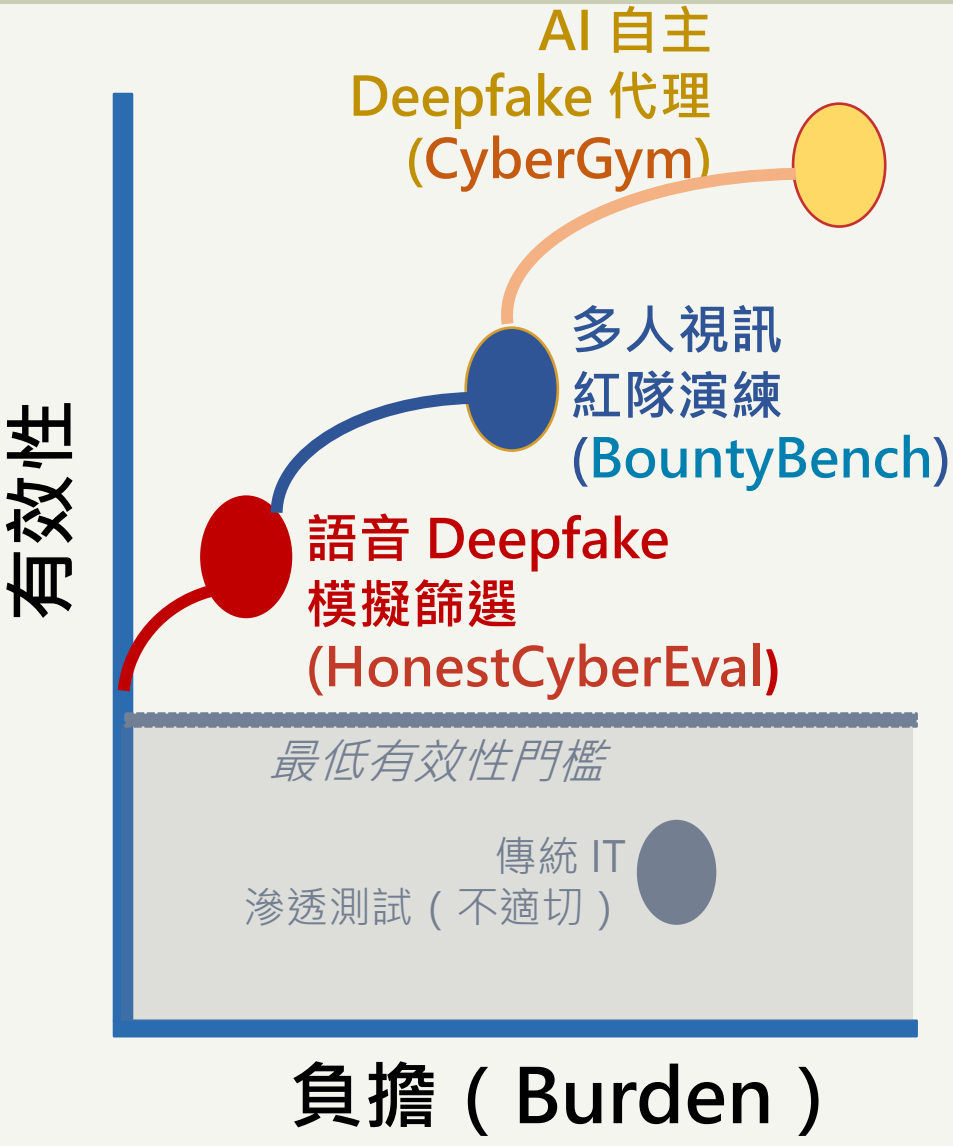
平衡性 Balancing

衡量評估負擔與風險嚴重性？

1. Arup 的 2,560 萬美元損失 + Deepfake 攻擊自 2022 年增長 3,000%，支持金融機構採用高負擔評估
2. 分層方法：低風險機構 → 純語音模擬 高曝險機構（財務或高管授權）→ 完整視訊 Deepfake 紅隊
3. 帶外驗證協議（out-of-band confirmation）雙重保護管道是低成本緩解措施，可有效改善負擔—效果 Pareto 邊界

Deepfake複合入侵精準防護邊界

Mougan, et al., 2026



評估方法比較

AI 自主 Deepfake 代理 (CyberGym)

高負擔；完整端對端真實性；對風險估算具最高統計信心

多人視訊紅隊演練 (BountyBench)

中等負擔；重現 Arup 情境條件；能區分不同失敗模式

語音 Deepfake 模擬篩選 (HonestCyberEval)

低負擔，篩選用途，可排除低風險機構，失敗亦具資訊價值(需進階防護工具)

傳統 IT 滲透測試

對本風險不適切，對社交工程入侵路徑無資訊價值

Arup 資安事件精準防護需求啟示

GPAI 模型現實危害的風險路徑，已可取得人類心理信任，越過傳統資安層層防護。影像人類認知層面完成入侵。

01

資安轉型：資安已從 IT 基礎設施轉移至協作迴路 (Human-in-the-loop)。風險評估須超越傳統 IT 範疇。

02

比例防禦：盲目投資傳統資安工具無效。須遵循比例原則，建構從語音模擬到視訊紅隊演練的精準防禦機制。

03

監管進化： EU AI Act精神提供企業機構將 Deepfake 輔助社交工程列為精準風險評估模式發展需求

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


許辰陽
醫師


陳虹彣


劉秋燕


羅崧瑋


嚴明芳
教授


陳立昇
教授

台北醫學大學

TAMS