美國成立 AI 安全委員會的啟發

得安全 AI 系統者得天下

為防止 AI 濫用，美國已經成立 AI 安全委員會，全球也在欣曾相關控管的規範，如何打造安全 AI 系統，將是重要發展方向。

文／黃光彩

隨著 ChatGPT 的出現，多家生成式 AI 公司相繼宣佈他們的版本，包括 Google Gemini、ChatGPT4 Turbo、Llama-3、Claude Opus、 Mistral Large 及 TAIDE 等，提供更強大的文字、圖形及分析能力，擴大 AI 的應用範疇，滿足不同產業的應用需求。

產業 AI 應用的市場有多大，可能的帶來的資安風險就會有多大，黑帽駭客組織的獲利就可能會水漲船高。 因為當產業導入 AI 技術時，會提高產業數位化程度，將使用更多的數位設備，連接應用程式介面（API），以蒐集數位資料，導入不同的 AI 模型，加速產業智慧化，如果資安的考量不週全，這可能產生新的資安漏洞及被駭風險增劇，使得資安防守的範圍要加寬及加深，防守更加不易，也加劇了敵對競爭者或黑帽駭客組織利用 AI 技術進行攻擊的可能性，故須重新審視資安防禦的縱深模式。當 AI 加速產業智慧化，資安的風險及危機可能造成對國家安全的擔憂，將帶來國安新挑戰。所擔心的國安問題包括但不限於以下的幾種情境，如**<表一>**

美國政府成立AI安全委員會（NSCAI）

鑒於以上這些可能發生的問題，促使美國政府採取行動，以確保人工智慧技術的安全和負責任的使用。五月初，美國國土安全部宣布成立 AI Safety and Security Board（人工智慧的「安全與保障」委員會），**成員有 OpenAI 、Anthropic、Nvidia、IBM、Adobe、Microsoft、Alphabet、Cisco、和 AMD 等科技企業執行長**。

AI 技術的進步與應用的龐大潛力，將提高組織的營運效率和生產創新，但同時也可能會帶來新的安全挑戰和風險。例如 AI 系統可能被用於網絡攻擊、影響公共基礎設施的運作，甚至可能被用於影

攻擊型態	舉例
輸入型	利用操控 AI 輸入的內容，從而改變其輸出，以達到破壞原本功能的目的。
中毒型	在 AI 模型所擷取的資料中插入惡意程式碼，從而破壞模型的輸出。
模型建構	改變 AI 模型研究的數據，使得模型變得混亂，同時可能包含漏洞、入侵、操縱特徵等行為。
隱私型	利用 AI 技術侵犯個人隱私。
濫用攻擊	利用 AI 進行不公平的對待或決策。
損害型	干預 AI 系統，造成實質性破壞，例如干擾自動號誌系統、駭侵自駕車系統等。
煽動型	破壞社群的 AI 過濾系統，發布假訊息、誤導安全警告，從而讓人民對AI系統或政府失去信心。
恐怖分子	可能使用 AI 武器進行攻擊，包括虛擬網路癱瘓及實體破壞等行動。

表一，AI 可能產生的國安問題與情境。

關鍵基礎設施資安防護，來源數位發展部。

響政治運作過程。這個委員的目標是協助確保 AI 科技的使用安全，以及設法因應 AI 科技對能源、公用事業、交通、國防、資訊科技、食品和農業以及金融服務等關鍵服務所構成的威脅。**美國國家 AI 安全委員會（National Security Commission on Artificial Intelligence ，NSCAI）**也發出警告， 指出美國在 AI 時代尚未做好防禦或與中國競爭的準備。

除了美國之外，其他國家也有採取類似的舉措。例如，歐盟提出了**「去風險」（De-risking）**的概念，這是一種相對於「脫鉤」的策略，旨在增強歐洲自身的經濟抗壓能力，而不是孤立中國，這一策略**在 G7 峰會上**被接納並寫入公報。

同樣地，疫情爆發後，許多國家意識到在藥物、醫療用品上不能過於依賴單一市場，因此提出了供應鏈多元化和**「近岸外包」（near-shoring）**的概念。台灣在多個層面上都有類似的舉措來規劃自己的關鍵基礎設施保護策略，應對區域安全和國際政治的挑戰。例如，針對中國在金廈海域的行動，台灣在強化自身防衛能力的同時， 也要與周邊國家共同採取反制措施，借鏡俄烏戰爭的經驗，並新增了分散式指管原則，應用多種具體的 AI 技術來提升安全性和韌性。

美國 NSCAI 建議大幅增加聯邦政府在 AI 研究方面的非國防支出，並改革國防部的採購管理系統，以更快、更容易地引入新技術，這個委員會將為運輸部門、管線和電網營運商、網路服務提供商等制定相關建議，「預防和準備因應關鍵服務受 AI 相關干擾情事，以免影響國家或經濟安全、公眾健康或安全」。

平衡安全和隱私

歐美所提出的**《人工智慧安全法案》**透過「AI 警察」建立一個全面的保護框架，同時不斷地評估和更新這些措施，防堵資安威脅，以應對不斷變化的威脅和技術發展。其範圍涵蓋在各類型的應用範疇，要求廠商訓練 AI 的過程中揭露所有訓練資料與引用來源，透過強制創建記錄所有 AI 系統違規行為的資料庫，以追蹤可能引發的資安問題。平衡安全和隱私是一項挑戰，但也是保護關鍵基礎設不可或缺的一部分。

在實施這些措施時，還需要考慮到成本、用戶便利性和業務需求，以達到實際可行和有效的平衡。**<表二>**是一些策略和考慮因素：

最小化數據收集	只收集實現特定目標所必需的數據，避免過度收集可能侵犯隱私的信息。
數據加密	對存儲和傳輸的數據進行加密，以保護數據免受未經授權的存取。
訪問控制	實施嚴格的訪問控制措施，確保只有授權人員能夠存取敏感數據。
透明度	向用戶清晰地解釋數據如何被收集、使用和保護，並提供用戶選擇退出某些數據收集的選項。
定期審計	進行定期的安全審計和隱私影響評估，以識別和修復潛在的安全漏洞和隱私風險。
法規遵循	遵守相關的隱私法規和標準，如歐盟的通用數據保護條例（GDPR）和美國的加州消費者隱私法案（CCPA）。
用戶教育	教育用戶關於安全和隱私的重要性，以及如何保護自己的數據。
隱私設計	在產品和服務的設計階段就考慮隱私問題，確保隱私保護措施從一開始就被整合進去。

表二，AI 策略與考量因素。