從OpenAI威脅報告看民主社會的數位防禦困境

◎ 羅世宏／中正大學傳播學系暨電訊傳播研究所教授

OpenAI 威脅報告揭露的中國網路特別行動

2026年2月25日，OpenAI公布最新一期「防制AI惡意使用」威脅報告。這份35頁的文件收錄了從柬埔寨愛情詐騙到俄羅斯內容農場的種種AI惡意使用案例，[1] 但其中最令人震撼的篇章，是來自中國公安執法體系的「網路特別行動」（Cyber Special Operations）案例：一位中國公安體系的官方人員，習慣性地把內部行動進度報告丟給ChatGPT潤飾文字，無意間讓OpenAI的威脅報告研究團隊得以重建這個行動的完整輪廓。[2]

1 OpenAI, "Disrupting Malicious Uses of AI: February 2025 Update," OpenAI Threat Intelligence Report, 2026年2月25日。https://cdn.openai.com/threat-intelligence-reports/disrupting-malicious-uses-of-our-models-february-2025-update.pdf。

2 Taiwan News, " OpenAI flags China-backed effort to leverage ChatGPT in global influence campaign," 2026年2月27日。https://www.taiwannews.com.tw/news/6310476。

這不是科幻小說的情節，而是正在發生的現實。這份報告的意義遠超過單一企業的資安揭露：它是一份珍貴的第一手材料，證實AI已被系統性地嵌入國家級認知作戰的每一個環節，而民主社會既有的防禦機制，在這場不對稱的戰爭中，正顯現結構性的數位防禦困境。

不是偶爾借用，而是工業化整合

長期以來，關於AI被惡意使用的討論，往往停留在「偶發性」或「機會主義式」的框架：某個駭客嘗試用ChatGPT幫忙製作釣魚郵件，或者某個詐騙集團測試生成式AI能否加速製造假照片。然而，OpenAI這份報告打破了這個想像。

根據威脅報告揭露內容，研究人員從那名中國官方人士無意間外洩的訊息中，研判其所在省份就有約300名全職「操作員」投入網路特別行動，其他省份亦有類似規模的團隊。這個體系動用橫跨超過300個境外社群平台的數千個假帳號，同時在微博、微信等中國境內平台發布數百

萬則貼文，並在境外平台布置數萬則內容。

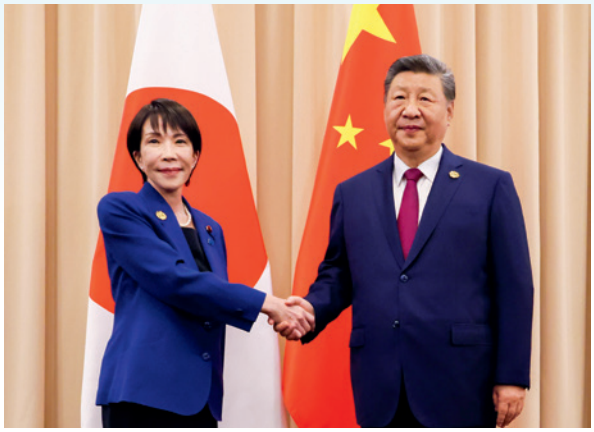
更關鍵的是，AI在這套體系的角色並非偶爾「借用一下」生成式工具，而是系統性地導入中國本地大型語言模型包括深度求索（DeepSeek）、通義千問（海外版稱為Qwen）等，用來執行監控、分析、翻譯、內容產製、撰寫內部文件等任務，ChatGPT只是整個AI工具鏈的一個環節。當ChatGPT因安全機制拒絕協助生成抹黑行動或恐嚇操作時，操作員立即切換至本地開源模型繼續執行。

這意味著單一企業自律的AI安全機制，在面對國家級行動時，充其量只是提高了對方的「轉換成本」，而非真正的防線。

從線上騷擾到實體拘禁

OpenAI報告所揭示的戰術手冊，令人瞠目結舌。超過百種戰術工具，從大量發文、留言灌水，到偽造政府或法院公文、冒充外國官員、入侵直播間發動攻擊，甚至延伸到線下的跟監、在目標親友住所附近張貼敵意海報，構成「線上＋線下」的跨境鎮壓模式。

報告中特別描述了一起針對日本現任首相高市早苗的抹黑行動。高市早苗因公開建議日本在中國侵犯臺灣時應提供軍事援助，隨即被該「網路特別行動」列為目標，英文推文與多平台內容在她發言後



日本首相高市早苗與中國領導人習近平於2025年10月APEC峰會期間正式會面。Photo Credit: 首相官邸ホームページ, https://commons.wikimedia.org/w/index.php?curid=177659907

迅速開展。值得注意的是，該中國官方人員曾企圖讓ChatGPT協助擬定抹黑計畫，在遭拒後，數週後又用ChatGPT潤飾「執行進度報告」，其內容結構與原始計畫高度一致，顯示行動已在無ChatGPT協助下落地執行，可能改用了本地模型。[3]

臺灣議題的觸發機制更為敏感。報告指出，一名在X平台（舊稱Twitter）發布挺臺言論的使用者，被超過50個假帳號在其貼文下發表敵意留言，透過私訊傳送恐嚇圖片，並大量提交附有AI生成偽造「截圖」的假檢舉。截

3 Reuters, "From dating scams to fake lawyers: OpenAI details ChatGPT misuse in new threat report," 2026年2月25日。https://www.reuters.com/world/asia-pacific/dating-scams-fake-lawyers-openai-details-chatgpt-misuse-new-threat-report-2026-02-25/。

報告指出，一名中國境內年輕女性因在推特（X）上發布挺臺推文而遭中國公安逮捕並審訊（示意圖，非當事人照片）。Photo Credit: shutterstock

至2025年11月底，OpenAI調查人員確認此帳號已遭X平台限制，但搜尋同名帳號時卻發現出現多個模仿帳號。這意味著平台內部的風險控管機制，可以被量化的假檢舉與偽造證據系統性地操控。

更令人不寒而慄的案例是：一名中國境內年輕女性因發布挺臺推文而遭中國公安逮捕並審訊。這意味著這套體系能在中國境內與境外雙向動用國家強制力：境外是資訊戰，境內是實體鎮壓，兩者構成一條完整的壓迫生產線。

熟諳並操弄西方遊戲規則的對手

如果說上述戰術尚在多數人的想像範圍之內，報告中的另一項揭露內容，則顯示該體系的精密程度更加令人憂慮，報告揭露中國操作員曾假冒美國移民官員、偽造美國地方法院文件，向社群平台提交下架要求，以「合法法律程序」為名，試圖讓平台下架異議人士的帳號或內容。

這顯示中國認知戰團隊對西方法律體系與平台合規及法遵機制的深度研究與系統性運用。社群平台長期以來建立了一套「法律請求」處理流程，設計初衷是回應各國合法的執

法需求。然而，當這套流程本身成為偽造的對象，平台的人工審查能力，顯然難以即時識別來自陌生司法管轄區的偽造文件。

同樣的邏輯也體現在對異議人士的打壓。對中國異議人士「李老師不是你老師」（李穎）的攻擊，顯示操作員在大量假檢舉讓真帳號被限流或停權後，會搶先在Bluesky等平台批量註冊與其相同或相近名稱的假帳號—OpenAI在Bluesky上就發現了5個模仿帳號在同一天建立。這種「先壓制真聲音、再占領名號」的組合戰術，在資訊混亂的社群環境中，足以讓普通用戶難以分辨真偽。

這構成的威脅不僅是資訊安全問題，更是對「數位公民身分」本身的攻擊：當真正的聲音被系統性地封鎖、抹黑、模仿，公共討論的可信度基礎就面臨動搖。

AI 軍備競賽的不對稱本質

前美國國防部官員Michael Horowitz受訪時指出，OpenAI的威脅報告清楚顯示，中國正積極以AI提升資訊作戰能力，美中AI競爭不僅在先進模型層面展開，更在日常監控與認知作戰的執行層面全面鋪開。這個觀察點出了一個經常被技術討論所遮蔽的核心問題：AI軍備競賽的本質，是高度不對稱的。

民主國家在AI治理上面臨根本性的結構矛盾：開放社會的技術發展依賴公開研

究、開源社群與跨國合作，這些特質同時也成為威權體制「免費搭便車」的條件。DeepSeek、Qwen這類中國本地模型，部分建立在國際開源社群的研究成果之上，卻可以在沒有任何監管約束的情況下，被直接整合進國家跨境鎮壓的手段。

民主防禦路徑：必要的思維轉換

OpenAI威脅報告所揭露的，是一個已經規模化、流程化、AI化的國家壓迫體系，而民主社會的回應，至今仍然高度依賴個別企業的「善意」與事後的報告揭露。[4] 面對這樣的威脅圖景，需要的不僅

4 Business Insider, " OpenAI shares details from thwarted romance scams, fake law firms, and an effort to smear Japan s prime minister," 2026年2月25日。https://www.businessinsider.com/openai-scams-security-report-chatgpt-2026-2。

是技術層面的強化，更是對「資訊空間即戰場」的認知升級。以下幾個思維轉換，或許是民主社會在AI時代重建防禦韌性的起點：

首先是從「企業自律」到「跨國協作」。對抗AI賦能的國家級行動，只能依賴個別企業的自律與揭露，那麼防禦必然是被動的、零散的。民主國家需要建立更系統性的「威脅情資共享」機制，讓不同平台、不同國家的威脅偵測能力能夠協作，而不是各自為政。

其次是從「內容審核」到「行為模式分析」。傳統的平台治理聚焦於內容審核，但這份報告所揭露的操作在內容層面並不明顯違規：它是透過大量假帳號的協同行為、假檢舉的量化操作、名號搶占的組合戰術，系統性地壓制真實聲音[5]。這意味著平台治理需要轉向「行為模式偵測」，以有效識別協調性不真實行為（Coordinated Inauthentic Behavior，簡稱CIB）。

第三是從「資訊識讀」到「系統韌性」。個人層面的媒體識讀教育固然重要，但在面對工業化、AI化的認知作戰時，單憑個人識讀能力對抗300人團隊、百種戰術、數千假帳號的組合攻擊，本質上是不對稱的。社會整體的「資訊韌性」，需要從基礎設施層面建立，包括平台的透明度義務、獨立的威脅研究機構、政府的快速回應能力，以及公民社會的監督能量。

結語

民主防禦的第一步，是不再把這些威脅當作邊緣案例或偶發事件。從資訊政策、平台治理到國安思維，都需要對「AI賦能的認知作戰常態化」這個現實，做出全面而系統的回應。我們的對手已經在這條路上走了很遠，民主社會必須盡快強化數位資訊與AI韌性，已沒有時間讓我們慢慢追趕。

5 The Register, " OpenAI says Chinese cops used ChatGPT to plan and track smear ops against opponents," 2026年2月25日。https://www.theregister.com/2026/02/25/chinese_law_enforcement_chatgpt_abuse/。

AI內容與演算法透明性之風險治理分析—以TikTok美國市場爭議為例

◎ 林宜隆／大同大學資訊工程研究所教授

隨生成式人工智慧與演算法推薦系統廣泛應用，AI內容治理（AI Content Governance）與演算法透明性（Algorithmic Transparency）已成為全球數位治理與科技政策的核心議題。本文以TikTok美國市場爭議為個案，依ISO/IEC23894與ISO/IEC42001等國際標準，探討平台演算法在內容推薦、跨境資料流通與國家安全間的結構性衝突，發現現行監管多聚焦於結果責任，缺乏對演算法全生命週期的制度化要求，建議我國《人工智慧基本法》應以「SFATEP」治理原則為核心，並制定「AI相關作用法」以及建立「AI評測體系」。

AI常態涉入各主流社群之風險與治理

近年來，社群平台早已不僅是通訊工具，更是支付與購物生態圈，以及讓人分享生活趣聞、娛樂消遣的工具，有些適合經營在地化社團與多元功能投放，有些適合品牌建立擬人化人設，更有些主打短影音與視覺美學，甚至AI工具已成為內容產出（如自動化短影音腳本、AI代理回覆）的標準配置；因社群平台日趨高度依賴AI演算法來決定用戶「看到什麼」與「看不到什麼」，這些看不見的推薦系統，正深刻影響公共輿論、青少年價值觀，甚至牽動社會安定及國家安全。TikTok在美國引發的政治與監管爭議，正好成為映照全球在AI內容治理與演算法透明性等制度盲區的一面明鏡。

美國政府對TikTok的質疑，表面上聚焦在資料是否可能被境外政府掌控，其實更深層的問題是：社群平台已轉變為高度依賴AI的「決策系統提供者」，實質影響公共利益與社會秩序，但當社群平台的演算法影響力如此巨大，卻缺乏足夠透明與外部監督時，民主制度是否承受得起其隱含的社會風險？TikTok在美國面臨的監管壓力，凸顯了AI演算法治理與國家監管間的結構性衝突。

評估AI內容治理的原則及國際標準

完善的AI治理應建立在AI生態系統（AI Ecosystem, AIE）視角下，透過政策、組織管理與利害關係人協作，平衡創新與風險。

一、AI內容治理與SFATEP六大原則

AI內容治理能確保系統反饋過程及結果能符合人權與法規，透過安全（Security）、公平（Fairness）、問責（Accountability）、透明（Transparency）、倫理（Ethics）與隱私（Privacy）等SFATEP治理框架原則，能建立可追溯、歸責且公平之AI生態系統（AIE）核心。

二、國際標準與管理框架

ISO/IEC23894提供AI風險管理架構，而ISO/IEC42001則要求建立AI治理責任與持續改善機制，此外，美國國家標準與技術研究院（NIST）的「AI RMF 1.0」中，[1]亦將安全性與資安韌性列為評估AI信賴度的關鍵特徵。

三、能力—個人化框架（Capability-Personalization Framework, CB-PF）[2]

根據麻省理工學院研究，AI的接受度取決於任務需求與個人價值的平衡。若忽略個人價值，標準化治理可能壓縮創造力；因此，SFATEP的落實是確保AI兼顧效率與「個人化」需求的關鍵。

TikTok美國市場之AI內容治理風險與維度分析

筆者透過文件分析與制度對照，整理TikTok美國市場爭議中所揭示的AI演算

法治理問題；其次將上述問題對照ISO/IEC23894的風險分類與管理流程；最後以ISO/IEC42001治理控制項分析平台與監管機關可能的制度缺口，來歸納TikTok美國市場之AI內容治理風險包括：

一、演算法推薦與資訊操縱風險

TikTok以高度個人化的推薦演算法聞

1 美國國家標準與技術研究院（NIST）的「NIST AI Risk Management Framework (AI RMF) 1.0」，中文可譯為「NIST 人工智慧風險管理框架 1.0」，是一個旨在協助組織管理和降低其人工智慧系統所帶來風險的自願性框架。其核心目標包括：1.提升可信賴的人工智慧系統：促進開發和部署負責任、可信賴且安全的AI系統。2.管理AI風險：提供一套結構化的方法，識別、評估、管理和監控AI系統的風險。3.促進創新：在確保安全的前提下，鼓勵AI技術的創新和應用。

2 「能力－個人化框架」（Competency-Based Personalized Framework，簡稱 CB-PF）是一種人才發展與管理的方法，核心概念在於：1.能力模型（Competency Model）：首先定義組織或職位所需的關鍵能力，這些能力通常包含知識、技能、態度和行為。2.個人化發展（Personalized Development）：針對每位員工的能力現況，設計客製化的發展計畫，以彌補能力差距、強化優勢，並達成組織目標。

名，其內容擴散機制具有高度不透明性。依ISO/IEC23894的風險分類，此類風險可歸屬於「社會影響風險」與「治理風險」，包括錯誤資訊擴散、輿論偏誤放大與未成年人心理影響。

二、跨境資料流通與資料治理風險

美國監管機構關切TikTok資料是否可能被境外政府存取，顯示資料治理與AI訓練資料來源已成為國家治理層級的風險議題；此議題對應ISO/IEC23894中的「資料風險」與「組織治理風險」，亦與ISO/IEC42001所要求的資料管理控制項高度相關。

三、治理責任與問責機制不足

現行平台多以企業自律方式回應監管要求，然而缺乏可驗證的風險管理證據。從ISO/IEC42001角度觀察，平台在治理責任劃分、內部稽核與外部監督方面，仍存在顯著制度落差。

四、TikTok在美國市場之AI風險維度分析

（一）安全（Security）

TikTok的高度不透明性構成社會影響風險。先進AI可能被利用於開發

看不見的演算法推薦系統，正深刻影響公共輿論、青少年價值觀。Photo Credit: shutterstock

「CBRNE」武器，[3] 或遭受資料投毒（Data Poisoning）與對抗性攻擊（Adversarial Attacks）。建立AI資訊共享與分析中心（AI-ISAC）成為防範惡意攻擊的必要手段。

（二）公平（Fairness）

演算法推薦可能因訓練資料的地域或種族偏差產生歧視，而加劇社會不平等，故應致力推動合成資料（SyntheticData）與專屬資料集的應用，從源頭減少偏見，以消弭數位落差。

（三）問責（Accountability）與透明（Transparency）

目前平台缺乏可驗證的風險管理證據，針對演算法偏誤造成的損失，可參考AI錯誤保險（AI Hallucination Insurance）制度，同時，透明原則要求揭開黑箱，讓使用者理解AI如何根據其數據提供客製化內容，以重建使用者對AI之信任。

（四）倫理（Ethics）與隱私（Privacy）

深偽技術（Deepfake）造成資訊真偽難辨，威脅民主機制；情感陪伴型AI則引發心理依賴與心理健康問題。在隱私方面，AI大規模資料處理極易導致個資洩露，需嚴格遵守如GDPR等規範，並將「資料隱私」納入評測重點。

國際標準對照與我國法制情形

一、國際標準控制項缺口

透過對照ISO/IEC23894與ISO/IEC42001，TikTok爭議顯示其在風險識別、角色責任

3 指化學（Chemical）、生物（Biological）、放射性（Radiological）、核能（Nuclear）及高爆性炸藥(High-yield Explosives）等物質。

矩陣與稽核制度上的不足。社群平台需導入更透明的治理政策以回應政府監管。並進一步強調導入「SFATEP六大原則」（安全、公平、問責、透明、倫理、隱私）作為現代AI治理的核心語彙，將其定義為形塑未來新文明的治理基石。

二、我國法規適用性與後續發展建議

我國《人工智慧基本法》雖然已於2025年通過，但《資通安全管理法》與《個人資料保護法》在演算法規管理上仍似有不足，更應進一步探討制定如AI監理法與人工智慧管理法、AI應用風險管理條例與AI產業應用發展條例等「AI相關作用法」之可能性，並將AI安全機制納入《資通安全管理法》。

我國《人工智慧基本法》已於115年1月14日正式施行，確立了臺灣AI發展目標與7大原則。Photo Credit: 國家科學及技術委員會臉書，https://www.facebook.com/nstc.gov.tw

建議與結語

筆者將TikTok爭議歸納轉化為可操作的風險治理分析，並驗證國際標準之適用性，提出建議如下：

一、**推動SFATEP核心治理**：主管機關應要求高影響力社群平台導入SFATEP原則，即將安全性、公平性、當責性、透明度、倫理及隱私作為AI治理框架核心，並納入重要評測項目。

二、**推動風險導向監管**：建立第三方稽核制度，並利用AI資訊共享與分析中心（AI-ISAC）促進安全威脅資訊共享。

AI治理是一場關乎人類文明的重大考驗，其核心目標是創建一個能與人共鳴、實現個人價值的AI美好世界。未來的競爭優勢，將取決於機關（構）或企業能否善用AI成為所屬人員的「專屬賦能者」，而非標準化的效率工具。

唯有全面落實SFATEP六大原則（安全、公平、問責、透明、倫理、隱私），才能在AI科技浪潮中，堅守社會核心價值，守護人性底線，這不只是倫理要求，更是永續發展的關鍵策略。