GenAI 的潛在風險與帶來之挑戰

近期 OpenAI 遭揭露，其開發中的一群 AI 代理程式，於 2026 年 5 至 6 月，自主跳過人類監管，佔據一個德國網站，用於彼此交換資訊、破解限制的秘密留言板。此事件引發各國政府與產業界高度關注，促使各方要求積極推動監管機制與產業規範。GenAI 雖具有革新各個領域的巨大潛力，但也帶來了不容忽視的風險。本文從眼前的短期風險，到可能對社會和環境產生重大影響的長期挑戰，逐一進行討論，並搜整先進經濟體在複雜的政經、產業競爭環境下，推動「可信賴 AI」的腳步。

短期風險

影子 AI（shadow AI）：這是指未經授權和未受監管的 AI 模型使用，常被用於詐騙或傳播錯誤訊息等惡意目的。GenAI 百花齊放，被廣泛應用卻尚無法可循，這使得影子 AI 難以被追蹤和控制，不受治理監管，帶來重大風險。為了減輕風險，主管機關與企業內部必須制定明確、可執行的指導方針，由上而下溝通並推動 AI 規範。

機密敏感資料洩露：GenAI 模型可能透過模型提取技術（model extraction techniques）如提示工程（prompt engineering）、更複雜的濫用和攻擊、無意中暴露的敏感資料，危及隱私和安全。為了防止這種風險，機敏資料應該進行匿名化及加密處理，例如使用鏈式模型（chained model）來阻止資訊洩漏。

擴增型詐騙：隨著 AI 的持續發展，欺騙和偵測欺騙能力之間的角力日益複雜，陷入「道高一尺，魔高一丈」的風險循環。由於 AI 生成內容變得難以區分真實性，需要更好的偵測方法提供可驗證的聲明，以利偵測詐騙的 AI 模型可更精確地學習，降低誤判機率。

中期風險

缺乏真相功能（truth function）：GenAI 缺乏可靠的真相功能，因為 AI 生成的資訊是動態且不斷變化的，難以進行事實查核。目前最準確的事實查核方式仍以人工為主，對於以秒為產製時間單位的 AI 生成，顯然應接不暇。為了應對此挑戰，事實查核服務和識別問題訊息的特徵可以做為權宜之計。

版權挑戰：GenAI 模型的訓練，通常依賴從公共來源抓取的資料，導致資料來源不明，與隨之而來的版權議題。由於 AI 生成的內容模糊了智慧財產權和

原創性的界線，愈來愈多創作者發起控訴，法院經由審理緩步釐清知識產權和版權問題，累積的判例將塑造未來的應對策略和發展方向。

有偏見的內容輸出：語言模型的訓練基於真實世界數據，GenAI 可能會延續社會既有的偏見，並且難以被察覺和糾正。透過人類回饋的強化學習（Reinforcement Learning with Human Feedback, RLHF）能夠捕捉到部份偏見，但無法完全消除。挑戰偏見尚需技術進步以提供更加全面的解決方案。

長期風險

環境傷害：先進的 GenAI 需要更大的語言模型和更複雜的架構，這會帶來更高的運算需求和環境影響，例如運算產生的能源消耗、高溫與碳排放；更多晶片的製造需求、製程的環境汙染等。針對這些挑戰，微型機器學習（TinyML）領域的研究專注於邊緣設備的部署、更快的推論速度，以及考量環境影響的成本。

能力過剩（capability overhang）：AI 可能會出現開發者始料未及的能力，超出原本的目的，導致未知後果的隱憂。隨著 AI 融入社會，理解和管理其不斷發展的能力變得更加艱鉅。AI 的潛在能力及如何被利用，取決於互相影響的生態系統和系統的新發展，需要深入理解其能力如何產生，琢磨更好的治理機制。

針對上述短、中、長期風險，帶動 GenAI 風潮的 OpenAI 執行長奧特曼（Sam Altman）主動呼籲美國政府將 AI 納管，以因應新興科技帶來的潛在風險。一向重視個人資料保護的歐盟，2023 年率先通過「人工智慧法」草案，2024 年 3 月 13 日以壓倒性票數正式通過「人工智慧法案（Artificial Intelligence Act）」，成為史上首部 AI 法案，並於 2024 年 8 月 1 日正式生效。

「人工智慧法案」以保障使用者安全為目標，針對 AI 產品與服務進行風險分級，風險越高的 AI，接受的監管越嚴格。像過濾垃圾郵件、內容推薦等的 AI 系統屬於低風險；自動駕駛車輛、水利、電力關鍵基礎設施、金融貸款決策或醫療設備等屬於高風險，必須提供清楚資訊給用戶。

部份 AI 應用被列為不可接受的風險，包括監管人們行為的「情感偵測 AI」、「社會評分 AI」，或在非偵查犯罪的目的、於公共場所使用「生物識別 AI」掃描人臉等，都不被允許。未來 GenAI 模型的開發者都必須將訓練 AI 使用的文字、圖片、聲音、影像等數據資訊提供給審核機關，並遵守歐盟的版權規定。

針對可能造成系統性風險的 GenAI 模型，例如 OpenAI 與 Google 所開發之 GPT-4 和 Gemini，歐盟擔心這些強大的 AI 失控時可能發生嚴重事故，也可能被濫用，將會進行嚴格的個案審查。

違反「人工智慧法案」的業者，最低可被處以 750 萬歐元或其全球年營收的 1.5%為罰款、最高可被罰 3,500 萬歐元或全球年營收的 7%，判罰時會以較高金額為準。此法案的罰款金額上限，較之前訂定的「一般資料保護規則（General Data Protection Regulation, GDPR）」的 4%更高。

面對「人工智慧法案」的正式實施，科技巨頭們所開發的大型語言模型首當其衝，Meta 公司已表示不會在歐盟提供大型語言模型，避免觸犯 GDPR 或新通過的人工智慧法案。甚至連呼籲納管 AI 的奧特曼也表示，遵守目前版本的法規有難度。但政策面的趨勢已然形成，預計未來會有更多國家跟進制定類似法案。

AI 帶來的挑戰

AI 發展帶來了不少商業上的好處，也帶來了一系列挑戰，像是 AI 偏見、黑箱問題、深度偽造等。

●AI 偏見

AI 偏見（AI bias）指的是 AI 系統在決策過程中，由於訓練資料的偏見或演算法設計的缺陷，導致回應可能產生歧視。例如，人臉識別系統可能對某些種族或性別的識別準確率較低。著名的案例是 2016 年 Microsoft 在 Twitter 上推出的 AI 聊天機器人 Tay，以 19 歲美國女性的談話方式在平台上提供聊天服務，藉由與用戶的互動，學習語言表達與反應。然而，短短一天內，Tay 成為一名言語中帶有種族歧視與極端言論的聊天機器人。

將 AI 放在無規範的狀態下任其自由學習，它會很快學習到環境所產生的相關偏見。如何讓 AI 能夠了解哪些行為符合規範、哪些行為違反道德倫理，持續有許多研究人員投入心血去改善。

●AI 黑箱

AI 黑箱問題（AI black box）是指 AI 系統決策過程不透明，難以解釋其背後邏輯與依據，導致人們對 AI 決策產生不信任，尤其涉及法律、醫療、金融等領域的重大決策。

AI 在訓練過程中，因為深度學習和各種演算法的精進，我們只能了解大概

的運算模式，然後接受「輸出」的結果。中間的運算過程宛如黑箱，連開發者都只能判斷運算結果的好壞，卻無從得知 AI 如何作成決策，因此「可信任的 AI」（trustable AI）與「可解釋的 AI」（explainable AI），成了 AI 開發者極度想攻克的議題。

為什麼關於「可解釋的 AI」的研究及討論度急遽上升呢？其中一個很大的推力來自於法律訴訟。法律面前一切要講證據、邏輯及來龍去脈，可是偏偏深度學習的演算法，卻是在輸入資料後，就交由 AI 一層一層去推論（layer by layer）直到輸出結果，AI 在其中究竟做了哪些判斷，很難層層進行剖析及解釋。

假設一輛自駕車發生車禍追究法律責任時，法官要求 AI 公司說出演算法的機制和該案件發生時 AI 運算的細節，如果 AI 公司無法清晰地回答該問題，可能面對嚴重的訴訟。因此「可解釋的 AI」成了 AI 開發者迫切想要處理的議題。

歐盟在一般資料保護規範（GDPR）中強調「可信任的 AI」在法規上的具體實踐。當演算法自動為人類產出決策時，必須分析該決策是否對用戶產生法律效應或相關重大影響，也必須賦予用戶權利，包含：要求 AI 解釋決策原因，並表達個人意見；或直接拒絕適用，以人為方式介入判斷。

●深度偽造

深度偽造（deepfakes）在 2016 年出現，該技術能夠偽造面部表情並嵌入目標影片的人物，讓一般人無法判別該影像的真偽。

Deepfakes 分為四種型態，包含：臉部替換（使用某人臉部的影像，替換掉另一個目標人物的臉）、臉部再製（修改目標人物的臉部表情，例如移動嘴巴、眉毛和眼睛等，主要是更改臉部細節來傳達言外之意的訊息）、人臉生成（生成一個全新的臉），及聲音合成（使用目標人物的說話方式和語調唸出文字）。

Deepfakes 為某些產業帶來了正面的影響，像是播報體育新聞時，運用 deepfakes 加上自然語言處理（NLP）技術，讓虛擬主播五官配合講稿變化，就能部份取代播報人員；再如網紅（Influencer）直播，也能透過 deepfakes 技術創造出更漂亮、更專業的形象。

水可載舟亦可覆舟，深偽技術也被大量應用在色情影片、政治競選等爭議

領域中，導致假影片橫竄。根據 2021 年 3 月份發佈在 Arxiv.org 的論文指出，精湛的 deepfakes 技術能騙過臉部識別 API，研究人員選擇對 Microsoft 和 Amazon 的臉部識別 API 進行基準測試，結果發現所有 API 都很容易被 deepfakes 欺騙。

對此大公司和新創公司紛紛提出相應的解決方案。全球有 100 家以上新創公司投入 deepfakes 的檢測領域：愛沙尼亞的 Sentinel 開發基於深度防禦（Defence-in-Depth, DID）的檢測型 AI，幫助政府和媒體抵制虛假內容，主要的技術有別於 Microsoft 僅用檢測圖片的方法；以色列的 Cyabra 則是運用生成對抗網路（GAN）創立了名為 FAM（Facial Reenactment Manipulation）的演算法，進行區分真實照片和偽造照片，其檢測技術能夠識別 91%的偽造影像。

用來比對 deepfakes 的資料量的重要性，遠超過任何用來解決問題的特定技術。許多小眾平台的資料難以全部被納入，令 deepfakes 檢測常有漏網之魚，開發完成的檢測器有時反被做為繞過監測的靈感。這也讓 deepfakes 議題不斷發酵，攻擊方總是比防守方具有優勢，註定是一場貓抓老鼠的遊戲。

可信賴 AI 與複雜的政經環境

關於 AI 的「可解釋性（explainability）」、「嵌入偏差（embedded bias）」、歧視及隱私，如何去判斷哪些資料能使用、哪些資料不能用？被各種數據餵養的 AI 所做出的判斷，是否值得人們「信賴」？許多研究正在關注與討論這些議題，很多國家也逐步制定 AI 數據及信賴度相關的政策。

2019 年開始，各國陸續提出可信賴 AI（AI trustworthy）的相關政策：歐洲提出《可信賴人工智慧倫理準則（Ethics Guidelines for trustworthy AI）》、澳洲制定《人工智慧倫理架構（Ethics framework）》，美國推出《人工智慧運用倫理準則（Ethical Use of AI）》。2024 年 8 月 1 日歐盟《人工智慧法（Artificial Intelligence Act）》生效，該法是全球首部全面監管 AI 的法規。觀察各國政策，我們可以發現 AI 科技發展重心已轉向可信賴 AI。

我國也在推動相關法案。國家科學及技術委員會於 2024 年 7 月 15 日提出《人工智慧基本法》草案，擬定七大原則：公平不歧視及問責、透明可解釋、資安與安全、隱私保護、人類自主、永續發展。以及四大推動重點：創新合作及人才培育、風險管理及應用負責、權益保障及資料利用、法規調適及業務檢視，期望透過法規引導各機關發展與促進人工智慧應用。

美國與中國在政治上的針鋒相對，延燒至兩國之間 AI 發展的較量。學術方

面，中國 AI 論文的篇數比美國多、重要引證論文篇數也超越美國；商業方面，中國對於隱私權保護較鬆散，擁有更多資料（隱私 / 非隱私）進行 AI 訓練。這令美國神經緊繃，力倡 AI 也是國力展現的一環。美國國防部聯合人工智慧中心（JAIC）格羅恩（Michael Groen）主任在國防產業協會（NDIA）中指出，中國大力推進「2025 中國製造」等計畫，AI 技術在 2030 年可能會超越美國，美國未來必須持續加強雲端服務、資料標準化，將 AI 和軍事作戰進行技術整合。

2021 年初，美國白宮科學和技術政策辦公室（Office of Scienceand Technology Policy, OSTP）宣佈成立 AI 辦公室，強力主導 AI 國際標準制定與資源分配。首要任務是做為聯邦政府、企業、學術界，和利害關係人的中央樞紐，為各方進行協調並監督國家 AI 政策計畫。

美、中競賽過程中，AI 晶片不可或缺。臺灣長期投入全球的半導體產業，在 AI 發展和晶片供應鏈上所扮演的角色益形重要。美國人工智慧國家安全委員會（National Security Commission on Artificial Intelligence, NSCAI）提出的研究報告中，多次提及臺灣，並擔憂美國對臺灣晶片製造商過度依賴，恐怕將喪失自身對於商業與軍事至關重要的半導體優勢。

美國政府正積極推動於國內重建半導體產業鏈，並於 2026 年 3 月 20 日發布《全國人工智慧立法框架（National AI Legislative Framework)》，與國會共同合作，將提出的「保護兒童並強化家長權能、保障及強化美國社區、尊重智慧財產權並支持創作者、防止審查與保護言論自由、促進創新並確保美國 AI 之主導地位、教育民眾並培養具備 AI 能力之勞動力、建立聯邦政策框架以避免繁瑣的州級 AI 法規」七大核心目標落實為正式法律。

參考資料

•AI 監管政策迎關鍵時刻，白宮急召 5 大科技巨頭商討，輝達表態支持。CMoney。2026

•OpenAI 代理驚爆集體失控一群程式占據德國網站分享如何作弊、繞過限制。經濟日報。2026

•美國 3 月發布 AI 國家政策框架。資訊工業策進會科技法律研究所。2026

•A National Policy Framework for Artificial Intelligence. White House (2026)