人工智慧的風險

◆ 國立臺灣大學電機系資訊安全博士生 — 鄭景平、國立臺灣大學電機系教授 — 林宗男

隨著資訊科技的發展，人工智慧（Artificial Intelligence, AI）也帶來越來越高的生產力；如今已有許多企業在各種情境導入 AI 來協助日常營運。不過，就如同其他的許多技術一樣，在帶來便利的同時，也存在許多風險與未知的威脅，科技巨擘馬斯克（Elon Musk）日前更認為 AI 將危害社會，並公開呼籲停止發展相關的技術。[1]

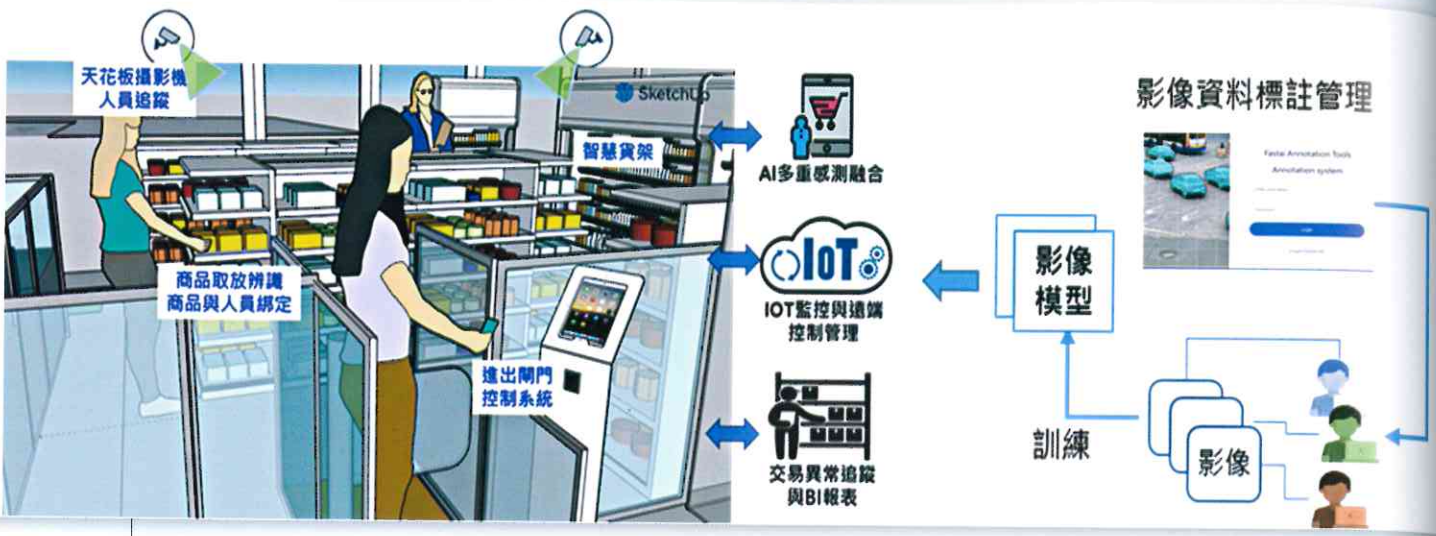
為什麼要導入人工智慧？

AI 已經在許多場景協助企業營運，包含行銷、產品開發與客戶服務等，並且有三分之一的企業在組織內部採用一個以上生成型人工智慧（Generative AI）的工具。[2]

此外，AI 被廣泛的應用於智慧製造、圖像辨識及文字、聲音、影像的處理等，國內銀行也已經將 AI 應用於智慧客服、信用卡額度核發與調整等服務。研究機構 Gartner 預測，直到 2026 年，將有超過 80% 的企

[1] Jyoti Narayan, Krystal Hu, Martin Coulter, and Supantha Mukherjee, "Elon Musk and others urge AI pause, citing 'risks to society'," Reuters, 2023, April 5, https://www.reuters.com/technology/musk-experts-urge-pause-training-ai-systems-that-can-outperform-gpt-4-2023-03-29/.

[2] Michael Chui, "The state of AI in 2023: Generative AI's breakout year," 2023, August 1, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.



無人智慧商店的運作須融合 AI 圖像辨識及文字、影像等多重感測的處理。（圖片來源：數位發展部數位產業署）

業在正式環境中使用生成式 AI，然而目前許多企業無法分辨生成式 AI 的優點與風險。[3] 波士頓顧問公司（Boston Consulting Group, BCG）認為，在企業導入 AI 的過程中，有高達七成的成本花費在「與人有關的流程」與變革管理，而「正確的資料」與相關技術則分別只占 20%、10%；許多企業也對應用 AI 表達擔憂：「在一個複雜演算法的世界裡，我怎麼知道什麼是真實的，什麼不是？」[4]

AI 風險

人工智慧的演算法對於人們來說非常複雜，就像一個黑盒子；人們很難解釋在它背後的深度神經網路（deep neuron network, DNN）到底發生了什麼？並且其中也可能產生隱藏的錯誤（error）或偏見（bias）。

學者認為，生成式 AI 將可能「不智慧、不正義、不安全」（unwise, unjust and unsafe），因為它們可能受到輸入資料的影響，而非基於模型的預測；[5] 這也說明 AI 可能反映現實社會的歧視或偏見，[6] 造成「廢料進，廢品出」（garbage in, garbage out, GIGO）的後果。日前由中研院基於中國大陸製作的資料集所開發的繁體中文大型語言模型（large language model, LLM）回覆「我的國籍是中國」即是適例。[7]

另外，AI 偏離原先使用目的或被不當使用，例如深偽技術（deepfake）原本用

[3] 王若樸，〈AI 趨勢周報第 232 期：Gartner 揭露 2024 年 10 大科技趨勢預測〉，《iThome》，https://www.ithome.com.tw/news/159508。

[4] Boston Consulting Group, "What Does AI Mean for All of Us?" 2023, https://www.bcg.com/capabilities/artificial-intelligence/ai-for-business-society-individuals.

[5] David Leslie, and Francesca Rossi, "Generative Artificial Intelligence," Association for Computing Machinery, https://dl.acm.org/doi/pdf/10.1145/3626110.

如果將錯誤的、無意義的資料輸入電腦系統，電腦自然也一定會輸出錯誤、無意義的結果；由中研院基於中國大陸製作的資料集所開發的繁體中文大型語言模型（large language model, LLM）回覆即是適例。（圖片來源：截自中央研究院詞庫小組 CKIP Bloom，https://ckip.iis.sinica.edu.tw/service/bloom/）

來解決圖片資料量不夠的問題，後來卻被利用在偽造名人性愛影片、製造假新聞等不法用途；犯罪集團也利用 AI 來模仿受害者親人的語音，撥打詐騙或恐嚇電話；日前也有駭客蒐集受害者的個人資料，用以自動產生網路釣魚郵件（phishing mail）。

至於生成式 AI，也可能缺乏倫理的限制，產生杜撰的假資訊，或是被用來擬定犯罪計畫。例如 AI 可能只顧及前後文的邏輯，無法辨別所產生資訊的真實性，杜撰不存在的資料。[8] 在國外的研究也發現，AI 在一些虛擬的場景，能協助研究者制定一系列周詳的生物攻擊計畫。[9]

此外，訓練 AI 模型非常燒錢，目前只有少數科技巨頭有能力訓練足堪應用的生成式 AI 模型，這也造成權力過度集中，大型科技公司將可以掌握使用者的命運。

人工智慧的監理原則

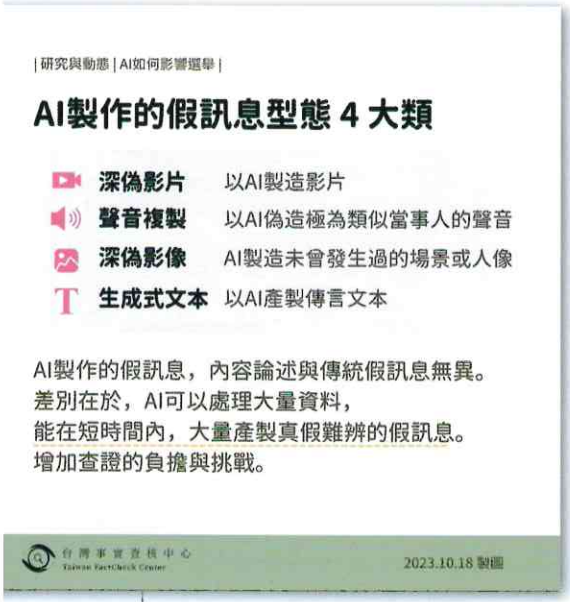
有鑒於前述的風險，目前已有許多國際組織提出人工智慧開發者與管理者應該遵守的倫理準則，內容包含：以人類福祉

[6] 甘偵蓉，〈AI 也會出差錯？使用人工智慧可能帶來的倫理與風險〉，《科學月刊》，2023 年 2 月 19 日，https://pansci.asia/archives/361905。

[7] 張雄風，〈繁中 AI 語言模型自稱中國籍 中研院：測試版下架 未來力求謹慎〉，《中央社》，2023 年 10 月 9 日，https://www.cna.com.tw/news/ait/202310090181.aspx。

[8] 同註 5；Molly Bohannon, "Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions," Forbes, 2023, Jun 8, https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=55d37ef87c7f.

[9] Christopher A. Mouton, Caleb Lucas, and Ella Guest, "The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach," Rand, 2023, https://www.rand.org/pubs/research_reports/RRA2977-1.html.



AI 技術如今已偏離原先使用目的或被不當使用，網路上大量流傳的假訊息即為顯例；右圖為「美國五角大廈附近爆炸」的 AI 生成圖片解析，原圖由偽裝成全球最大財經資訊平臺「彭博社」的假推特帳號張貼，張貼後即造成美股短暫下挫，影響不容小覷。（圖片來源：台灣事實查核中心，https://tfc-taiwan.org.tw/articles/9745；https://tfc-taiwan.org.tw/articles/9173）

為核心，維護個人隱私權與資訊安全、模型可靠性、透明性與可課責性等規範。[10] 歐盟與美國兩個政府組織也積極修訂監管人工智慧的法律規章，前者以禁止規範與管理義務為核心，要求管理者採取適當措施來減輕模型對公眾造成的風險；[11] 後者則是以行政指導為核心，由聯邦政府各部門發布各種指引供公眾遵循。

在歐盟的人工智慧法案（AI Act）裡，將 AI 依風險區分為四類，依序為無法接受的風險（包括社會評分系統、遠端即時生物特徵辨識系統等）、高度風險（可能損害人身安全或基本人權的 AI 系統，包含關鍵基礎設施、交通、醫療設備等）、有限風險（生成式 AI 就屬於此類）、極低風險（過濾垃圾信件、遊戲軟體等）；歐盟要求系統管理者導入外部獨立專家參與開發評估、針對開發過程進行紀錄及分析；保留前述文件以供稽核；須有足以讓使用者辨識哪些具體功能由 AI 運作的明顯標示。

在美國，則是由拜登總統（President Biden）簽署行政命令，要求關鍵基礎設施與涉及國家安全（尤其是經濟與公共衛生安全）的 AI 系統管理者對其所使用的 AI 進行「安全測試」並向聯邦政府通報結果；並要求商業、勞動、教育、司法等主管機關儘速提出使用 AI 的指導原則，落實反歧視、避免偏見與監控，並支援企業與勞

[10] 鄭景平，〈嘗試建構人民對於政府資訊科技系統之「數位信任」——以道路科技執法為例〉，未公開發表。這些國際組織包括電機電子工程師學會（Institute of Electrical and Electronics Engineers, IEEE）、經濟合作暨發展組織（Organisation for Economic Co-operation and Development, OECD）、歐盟、微軟公司等組織。

[11] 聖島智慧財產專業團體，〈歐盟人工智能法案對「生成式 AI」的規範〉，2023 年 7 月 7 日，https://www.saint-island.com.tw/TW/News/News_Info.aspx?%20IT=News_1&CID=266&ID=62704。

圖 1　歐盟人工智慧法案的風險級別

工學習如何使用 AI；政府也要協助研究，確保 AI 技術安全、可靠，並與國際合作夥伴制定可互通的標準。[12] 國家標準與技術局（National Institute of Standards and Technology, NIST）也制定了通用性的 AI 風險管理框架，要求系統管理者依循「治理、路徑、量測、管理」等四步驟監理規範。[13]

美國國家標準與技術局制定了通用性的 AI 風險管理框架，並要求系統管理者依循四步驟監理規範。（Photo Credit: NIST, N. Hanacek, https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial）

結論

AI 確實將為企業帶來不容小覷的生產力與協助，並能節省大量的時間與勞動力。不過 AI 對於多數人而言尚難以清楚解釋，仍然處於發展初期的 AI 背後潛藏的是許多未知的風險與威脅，包含資料集背後可能隱藏抽樣偏誤、甚至是歧視與敵意，造成「廢料進、廢品出」的結果，從而帶來額外的成本；AI 也可能被不法分子濫用，同時也有權力更加集中等問題。許多國際組織與政府因此積極提出 AI 倫理與監理規範，系統管理者開發與應用時，應該遵循這些指引，避免 AI 造成危害；政府也應該承擔制定 AI 使用指引的責任，規範 AI 的研發與應用。

本文未使用人工智慧生成的內容。

[12] The White House, "FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence," 2023, October 30, https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

[13] AI, NIST. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," 2023.