美國國家標準暨技術研究院發布「人工智慧風險管理框架：生成式AI概況」

美國國家標準暨技術研究院（National Institute of Standard and Technology, NIST）2024年7月26日發布「人工智慧風險管理框架：生成式AI概況」（Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile），補充2023年1月發布的AI風險管理框架，協助組織識別生成式AI（Generative AI, GAI）可能引發的風險，並提出風險管理行動。GAI特有或加劇的12項主要風險包括：

1.化學、生物、放射性物質或核武器（chemical, biological, radiological and nuclear materials and agents, CBRN）之資訊或能力：GAI可能使惡意行為者更容易取得CBRN相關資訊、知識、材料或技術，以設計、開發、生產、使用CBRN。

2.虛假內容：GAI在回應輸入內容時，常自信地呈現錯誤或虛假內容，包括在同一情境下產出自相矛盾的內容。

3.危險、暴力或仇恨內容：GAI比其他技術能更輕易產生大規模煽動性、激進或威脅性內容，或美化暴力內容。

4.資料隱私：GAI訓練時需要大量資料，包括個人資料，可能產生透明度、個人資料自主權、資料違法目的外利用等風險。

5.環境影響：訓練、維護和運行GAI系統需使用大量能源而影響碳排放。

6.偏見或同質化（homogenization）：GAI可能加劇對個人、群體或社會的偏見或刻板印象，例如要求生成醫生、律師或CEO圖像時，產出女性、少數族群或身障人士的比例較低。

7.人機互動：可能涉及系統與人類互動不良的風險，包括過度依賴GAI系統，或誤認GAI內容品質比其他來源內容品質更佳。

8.資訊完整性：GAI可能無意間擴大傳播虛假、不準確或誤導性內容，從而破壞資訊完整性，降低公眾對真實或有效資訊的信任。

9.資訊安全：可能降低攻擊門檻、更輕易實現自動化攻擊，或幫助發現新的資安風險，擴大可攻擊範圍。

10.智慧財產權：若GAI訓練資料中含有受著作權保護的資料，可能導致侵權，或在未經授權的情況下使用或假冒個人身分、肖像或聲音。

11.淫穢、貶低或虐待性內容：可能導致非法或非自願性的成人私密影像或兒童性虐待素材增加，進而造成隱私、心理、情感，甚至身體上傷害。

12.價值鏈和組件整合（component integration）：購買資料集、訓練模型和軟體庫等第三方零組件時，若零組件未從適當途徑取得或未經妥善審查，可能導致下游使用者資訊不透明或難以問責。

為解決前述12項風險，本報告亦從「治理、映射、量測、管理」四大面向提出約200項行動建議，期能有助組織緩解並降低GAI的潛在危害。

相關連結

NATIONAL INSTITUTE OF STANDARD AND TECHNOLOGY [NIST]

你可能會想參加

→ 【2023科技法制變革論壇】AI生成時代所帶動的ChatGPT法制與產業新趨勢
→ 「跨域數位協作與管理」講座活動
→ 新創採購-政府新創應用分享會
→ 【線上場】113年「新創採購機制及鼓勵照護機構參與推動」說明會
→ 【北部場】113年「新創採購機制及鼓勵地方政府參與推動」說明會
→ 【南部場】113年「新創採購機制及鼓勵地方政府參與推動」說明會
→ 113年新創採購-照護機構獎勵說明會

→ 【南部場】113年「新創採購機制及鼓勵地方政府參與推動」說明會
→ 【北部場】113年「新創採購機制及鼓勵地方政府參與推動」說明會
→ 【中部場】113年「新創採購機制及鼓勵地方政府參與推動」說明會
→ 【臺北場】113年度新創採購-招標作業廠商說明會
→ 【臺中場】113年度新創採購-招標作業廠商說明會
→ 【高雄場】113年度新創採購-招標作業廠商說明會

許嘉芳
法律研究員 編譯整理

上稿時間：2024年10月

資料來源：
NATIONAL INSTITUTE OF STANDARD AND TECHNOLOGY [NIST], *NIST-AI-600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, July 26, 2024, https://www.nist.gov/itl/ai-risk-management-framework (last visited Sept. 13, 2024).

延伸閱讀：
陳箴，〈美國國家標準與技術研究院公布人工智慧風險管理框架（AI RMF 1.0）〉，2023年2月，https://stli.iii.org.tw/article-detail.aspx?no=64&tp=1&d=8974（最後瀏覽日：2023/05/03）。

文章標籤

推薦文章