生成式 AI 及資訊操弄：台灣大選中的觀察與應對

〔新興科技與民主研究主題〕

時　　間：113 年 6 月 25 日（二）13:00-14:30
地　　點：政大公企中心 10 樓 A1034 教室
主 持 人：杜文苓（科技、民主與社會研究中心主任）
報 告 人：周睦怡（科技、民主與社會研究中心副主任）
　　黃凱紳（科技、民主與社會研究中心研究員）
　　林哲瑋（科技、民主與社會研究中心研究員）
與 談 人：翁浩正（DEVCORE 執行長）
　　侯宜秀（台灣人工智慧學校祕書長）
　　蔡蕙如（國立臺灣大學新聞研究所助理教授）
　　瓦　科（Access Now）
記　　錄：科技、民主與社會研究中心（DSET）

2022 年底 OpenAI 推出 ChatGPT 以來，生成式 AI 迅速崛起，成為工作幫手的同時，也帶來諸多挑戰。其中，使用生成式 AI 進行資訊操弄特別令人關注，成為民主社會普遍的擔憂。生成式 AI 在資訊操弄上會帶來哪些具體的影響？本文以 2024 年台灣總統大選為案例，速覽生成式 AI 在資訊操弄的及其當前狀況。

一、資訊操弄的定義與分析框架

資訊操弄指的是一系列藉由操弄資訊環境，對受眾國家或社會群體的政治環境可能產生負面影響的故意行為。資訊操弄具有連續性，使某群體在政治上產生負面影響的行為，透過一連串的資訊操作以達到政治目的，而這一系列的操作行為，可以使用資訊操弄的框架來進行分析。

本文使用 DISARM 基金會正在維護的 DISARM 框架[1] 以分析資訊操弄的手法狀況。由於 DISARM 是開源框架，且均有逐步針對內容進行更新，針對每一

[1] DISARM Foundation. (17 February 2024). *DISARM Disinformation TTP (Tactics, Techniques and Procedures) Framework*. https://github.com/DISARMFoundation/DISARMframeworks (Last visited: 2024/05/07)

個資訊操弄之手法，DISARM 都有提出反制手段；以及 DISARM 可以採用 STIX 資料格式儲存與交換資訊操弄案例。在資訊操弄的分析過程中，若能與外界持續交流、交換彼此的案例，就能整理出更多的手法。

DISARM 框架參考 ATT&CK 框架的作法，將資訊操弄的實踐分為四個階段：在每一個階段裡，攻擊方會有不同的戰術目標，這些戰術目標由眾多不同的手法達成。DISARM 框架的四個階段下，有十六個戰術目標，而這十六個戰術目標目前含括多達 244 個手法。[2]

二、生成式 AI 被用於資訊操弄的現況

近期有不少關於台灣 2024 年總統大選的資訊操弄研究報告，包含台灣傳播學會[3]、微軟威脅分析中心（Microsoft Threat Analysis Center）[4]、台灣民主實驗室[5,6]、AI Labs[7]、Team T5[8]、ASPI[9] 等團隊的研究均有分析使用 AI 為虛假訊息操弄之案例。歐盟對外事務部也提出生成式 AI 用於資訊操弄的建議[10]。

---

[2] Credibility Coalition: Misinfosec Working Group. (27 August 2019). *Building standards for misinfosec. Applying information security principles to misinformation response*. https://github.com/DISARMFoundation/DISARMframeworks/blob/main/DISARM_DOCUMENTATION/DISARM_HISTORY/2019-08-27_MisinfosecWG-2019-1.pdf (Last visited: 2024/05/07)

[3] Hung, C., Fu, W., Liu, C., & Tsai, H. (12 April 2024). *AI Disinformation Attacks and Taiwan's Responses during the 2024 Presidential Election*. Taiwan Communication Association.

[4] *China Tests US Voter Fault Lines and Ramps AI Content to Boost Its Geopolitical Interests*. (4 April 2024). Microsoft Threat Analysis Center. https://blogs.microsoft.com/on-the-issues/2024/04/04/china-ai-influence-elections-mtac-cybersecurity/ (Last visited: 2024/05/07)

[5] 台灣民主實驗室（2024）。〈2024 台灣選舉：境外資訊影響觀測報告初步分析〉，台灣民主實驗室，2024 年 1 月 19 日。https://medium.com/doublethinklab-tw/fe7f819aeabd（最後瀏覽日：2024 年 5 月 7 日）。

[6] 台灣民主實驗室（2024）。〈人造多重宇宙：2024 台灣大選境外資訊操作與影響觀察報告〉，台灣民主實驗室，2024 年 6 月 5 日。https://medium.com/doublethinklab-tw/493423f9bba8（最後瀏覽日：2024 年 8 月 15 日）。

[7] *2024 Taiwan Presidential Election Information Manipulation AI Observation Report*. (31 January 2024). Taiwan AI Labs. https://ailabs.tw/uncategorized/2024-taiwan-presidential-election-information-manipulation-ai-observation-report/ (Last visited: 2024/05/07)

[8] *Cyber Threats against Taiwan's 2024 Presidential Election*. (4 March 2024). Team T5. https://teamt5.org/en/posts/whitepaper-cyber-threats-against-taiwan-s-2024-presidential-election/ (Last visited: 2024/05/07)

[9] Zhang, A. (18 January 2024). *As Taiwan Voted, Beijing Spammed AI Avatars, Faked Paternity Tests and 'Leaked' Documents*. ASPI. https://www.aspistrategist.org.au/as-taiwan-voted-beijing-spammed-ai-avatars-faked-paternity-tests-and-leaked-fake-documents/ (Last visited: 2024/05/07)

[10] *2nd EEAS Report on Foreign Information Manipulation and Interference Threats*. (23 January 2024). European Union External Action. https://www.eeas.europa.eu/eeas/2nd-eeas-report-foreign-information-manipulation-and-interference-threats_en (Last visited: 2024/05/07)

根據DISARM框架，生成式AI用於資訊操弄的可能技術手法(techniques)，可用於以下四個階段，分別是：計畫（Plan）、準備（Prepare）、執行（Execute）與評估（Assess）。本文以此框架為基礎，同時參酌美國 CSET 智庫[11] 及 Open AI 公司[12] 的研究成果，分析 AI 會影響哪些資訊操弄的環節。除評估階段較無需要透過 AI 為之外，其餘的三個資訊操弄的階段都有應用 AI 的可能。

(一)計畫階段

在計畫階段，CSET 認為，AI 可以用於分析網路環境資訊（T0080: Map Target Audience Information Environment）以找出可操作的議題；同時，也可以分析目標受眾的足跡與資料（TA13: Target Audience Analysis），進而找出目標受眾對議題的立場與弱點[13]。

北約卓越中心發現大型語言模型擅長內容分析，能針對文本的情緒、立場進行分類，也能用來執行一般由人類分析師執行的重複任務，進而提高攻擊者對社群網絡進行分析的效能[14]。

不過，觀察與分析者只能從執行的階段往回推測攻擊者使用的手法。在計畫階段的手法無法收集到公開資訊，因此也無從證實攻擊者是否使用 AI 來進行網路資訊的分析。

(二)準備階段

在準備階段，攻擊者需要建立發起攻擊的基礎設施（TA15: Establish Social Assets），以及製造要傳遞的資訊與內容（TA06: Develop Content）。散播資訊與內容的基礎設施包含數量龐大的虛假帳號（T0090: Create Inauthentic Accounts），或是傳播假消息的管道，如社交平台上的社團或粉絲頁（T0007: Create Inauthentic Social Media Pages and Groups），或是內容農場（T0096: Leverage Content Farms）。

在資訊投遞管道方面，攻擊者需要建立大量的社群平台虛假帳號以進行資訊投遞。過往虛假帳號可能會有重複的頭像、統一的生成時間或空白的發文紀

---

[11] Sedova, K., McNeill, C., Johnson, A., Joshi, A. & Wulkan, I. (December 2021). *AI and the Future of Disinformation Campaigns*. *Part 2: A Threat Model*. CSET. https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns/ (Last visited: 2024/05/07)

[12] *Forecasting Potential Misuses of Language Models for Disinformation Campaigns and How to Reduce Risk*. (11 January 2023). OpenAI. https://openai.com/research/forecasting-misuse (Last visited: 2024/05/07)

[13] *Id*.

[14] Fredheim, R. (5 June 2023). *Virtual Manipulation Brief 2023/1: Generative AI and Its Implications for Social Media Analysis*. NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/publications/virtual-manipulation-brief-20231-generative-ai-and-its-implications-for-social-media-analysis/287 (Last visited: 2024/05/07)

錄等特徵，容易使人對帳號起疑。CSET 判斷，透過 AI 工具，攻擊者可以自動化建立大量的社群平台帳號[15]。AI 工具可生成許多逼真的頭像，以及生成使用者的身分與「生活」，大量製造出「仿真」的虛假生活照、工作履歷、閒暇興趣。OpenAI的研究報告也認為，大型語言模型將會使生成個人化的訊息更為容易[16]。

在「蔡英文秘史」案例中，有部分的假帳號疑似使用 AI 生成的頭像[17]，使這些帳號看起來更像真人。2020 年 Graphika 已經揭露 Spamouflage 使用 AI 中的生成式對抗網路（GAN）的方式生成假帳號的頭像，重疊這些頭像就可發現這些頭像的眼睛與嘴巴在同一個位置，且都有背景模糊的狀況[18]。

在內容準備方面，攻擊者需要準備要傳遞的資訊內容，包含文字、圖片或影音等。CSET 認為，透過 AI 工具，可以更有效率地建立起更多元、更吸引人的圖片影音或是迷因，吸引目標受眾[19]。

另外，某些假訊息可能需要偽造或修改特定文件。CSET 認為，AI 工具可以使這些修改更為真實，也更難被偵測；或產生更為真實的 deep fake 影音，以混淆受眾的認知與判斷[20]。

OpenAI 也認為，生成式 AI 的出現將會降低這些攻擊者製作內容的成本，同時也讓規模的提升成為可能。同時這些內容會更精緻，也更容易誤導目標受眾[21]。

在本次總統大選的觀察中，有觀察到以 AI 生成聲音（賴清德是春風專案線民案例[22]）或影片（蔡英文秘史[23]）之案例；圖片[24]或文字[25]方面，在此次大選中較無觀察，但已有相關資訊操作的國外案例。

---

[15] *supra* note 11.

[16] *supra* note 12.

[17] 張之豪（2024 年 1 月 8 日）。〈中国網軍現在四處散佈一本不知道是哪個白癡寫的《蔡英文秘史》，不外乎就是「數典忘祖」、「出身不純」、「媚日親美」〉。Facebook. https://www.facebook.com/JihoTiun/posts/pfbid02Ey5UEVVDfqhFeYHqGhfbvZXc2C1xrde7ge5VRpFSYAWYG2ijtkRew9XNpMxh5wqsl（最後瀏覽日：2024 年 5 月 7 日）。

[18] Nimmo, B., François, C., C. Shawn Eib, & Ronzaud, L. (12 August 2020). *Spamouflage Goes to America*. *Graphika*. https://graphika.com/reports/spamouflage-dragon-goes-to-america (Last visited: 2024/05/07)

[19] *supra* note 11.

[20] *Id*.

[21] *See supra* note 12.

[22] *See supra* note 4.

[23] *See supra* note 19.

[24] *Digital Threats from East Asia Increase in Breadth and Effectiveness*. (7 September 2023). Microsoft Threat Intelligence. https://www.microsoft.com/en-us/security/business/security-insider/reports/nation-state-reports/digital-threats-from-east-asia-increase-in-breadth-and-effectiveness/ (Last visited: 2024/05/07)

[25] Jane Manchun Wong. (11 July 2024). *Hijacking a ChatGPT Spam Bot and Turning It into My Personal Hype Squad*. Twitter. https://x.com/wongmjane/status/1811082135941566890

(三)執行階段

在執行階段，攻擊者會利用建立好的管道，以及之前針對資訊環境所做的分析，對目標受眾投遞相關訊息（TA09: Deliver Content），並會利用如假帳號分享或按讚等方式放大相關的資訊（TA17: Maximize Exposure），一方面影響演算法的判斷以接觸更多受眾（T0121: Manipulate Platform Algorithm），另一方面也可吸引受眾，引發其自主分享以產生病毒式傳播。

CSET 認為，透過 AI 工具，虛假帳號將能扮演不同的角色，並使這些帳號間的互動更為真實，也更難分辨，這可用於吸引目標受眾的注意，或是針對特定的人物進行騷擾（TA18: Drive Online Harms），使其陷入沉默螺旋[26]。

在 2024 年 7 月，Instagram 資安專家 Jane Manchun Wong 在 Thread 上觀察到串接 AI 的假帳號機器人，會在使用者進行 prompt engineering 後，出現 system prompt，表示該機器人使用 AI 作為其自動回應的機制[27]。

圖一：「AI 時代的民主韌性」場次（照片來源：DSET）

(四)小結

總結以上的操作行為，目前用於訊息操弄的 AI 主要是以生成式 AI 的形式利用，包含生成文字、圖片、聲音或影片。除了生成內容外，也會利用生成式對抗網路生成頭像，以使假帳號看起來更為逼真。

---

[26] *supra* note 11.

[27] Wong, *supra* note 25.

至於研究環境、研究受眾這兩種 CSET 預測選舉期間資訊操弄利用 AI 的方式，雖然目前沒有被觀察到相關痕跡，但除了沒有發生以外，也可能是因為這些方式並未被找到鑑識的方法，以致於無法被觀察到，未來仍應持續觀察相關的痕跡。

三、生成式 AI 在資訊操弄上之效果為何？

本文探討了生成式 AI 在資訊操弄中的影響，並分析了台灣總統大選期間的相關案例。台灣傳播學會[28]、微軟威脅分析中心[29]、AI Labs[30] 等機構一致認為，生成式 AI 在此次選舉中的影響有限。即使生成式 AI 被用於製造偽造影片與圖片，但其效果不顯著，並未成功改變輿論。歐盟對外事務部觀察也發現，使用生成式 AI 進行資訊操弄的案例並不多，目前主要側重於「建立內容」與「建立合法性」(如使傳統媒體相信)。歐盟對外事務部指出，生成式 AI 對訊息操弄帶來的是「一場演變(Evolution)，而不是一場革命(Revolution)」[31]。

雖然生成式 AI 並未使資訊操弄的模式產生整體性的轉變，但仍在降低內容製造的時間與成本方面產生具體的效果，並使協同行為更難偵測，也帶給事實查核機構帶來沉重負擔[32]。

四、對策與結語

目前的防禦手段仍然有效，如快速澄清機制、與事實查核組織合作、政府透明說明等。持續強化鑑識技術能加快澄清速度。AI 也能用於監控社群平台上的異常操作，公民社群也能利用 DISARM 框架交流案例與對策。

生成式 AI 的出現引發了對資訊操弄加劇的擔憂，但目前其主要影響在於提高內容生產效率。若未來生成式 AI 技術的進步，使 AI 生成的內容更具真實性，難以識別真假，且成本低廉、生成速度更快，加上社群媒體平台的廣泛使用和傳播速度，就有機會使虛假信息能夠迅速傳播，影響範圍更廣，並對政治過程和公眾輿論造成深遠影響。

---

[28] *supra* note 3.

[29] *supra* note 4.

[30] *supra* note 7.

[31] *supra* note 10.

[32] *supra* note 7.

為了應對生成式 AI 帶來的資訊操弄挑戰，需要採取綜合措施。首先，應加強對生成式 AI 技術的監管，制定相關法律法規，規範其應用和使用範圍。現行《總統副總統選舉罷免法》第 90 條第 2 項及《公職人員選舉罷免法》第 104 條第 2 項，已有使用深偽技術散播虛假訊息影響選舉之相關刑責，但其具體內涵仍相當模糊，有待未來進一步充實。其次，徒法不足以自行，民主選舉實踐仰賴社會大眾對資訊操弄的警覺性和識讀素養，故應強化大眾辨別虛假信息的能力。在技術面，可探索先進的技術手段，如 AI 偵測工具，以快速識別和阻斷虛假信息的傳播。最後，促進國際合作，以 DISARM 等方式共享資訊操弄的應對經驗和技術，形成全球聯防機制，共同維護資訊環境的健康和政治過程的公正性。