

OpenAI 再爆更多智能失控事件 高層刻意隱瞞

香港文匯報訊 據路透社9月4日援引一項新發布的研究報告和知情人士消息報道，美國人工智能（AI）企業 OpenAI 旗下的一群 AI 智能體（Agent），於今年5月劫持一個德國網站，並將其改造成與其他 AI 智能體交流的留言板。知情人士透露，OpenAI 高層數周前已得知此事，但礙於該公司7月還深陷旗下智能體突破沙盒訓練環境、攻擊開源平台 Hugging Face 的信息安全風波，因擔心引發外界進一步質疑其對智能體監管不力，故秘而不宣。

有分析稱，OpenAI 高層選擇不披露這宗始於5月的 AI 智能體安全事件，凸顯 AI 產業對智能體擔憂情緒漸增。當前業界正爭相打造能自主處理複雜且高價值任務的 AI 智能體，但愈來愈多證據顯示，這些系統有可能學習繞過規則、鑽漏洞，並可能以開發者不曾料到或設計的方式進行協作。

分享在特定任務作弊手法

今次事件的調查報告，出自包括聚焦研究 AI 安全問題的非牟利組織 Nightingale 行政總裁阿克斯在內的一個研究團隊。他們上月底搜尋未經授權的 AI 智能體行動跡象時，注意到德語維基百科網站 DseWiki 出現逾 1.5 萬次由 AI 智能體進行的編輯行為。

這些編輯內容顯示，來自 OpenAI 的 AI 智能體已將該網站改造成留言板，在其中分享於特定任務中作弊、繞過 OpenAI 限制，以及掩蓋自身作弊行為的手法。

阿克斯表示，「這些行為極大概率不是出自 OpenAI 授意，我不認為這些 AI 智能體被設計成會互相協作，也懷疑它們本不該在公開互聯網上發表內容。」

研究人員稱，他們認定這些活動是 AI 智能體所為，原因是速度遠超常人，且都高度注重解決技術性問題，這正是 AI 公司訓練與測試模型時的典型評估內容。

匿名通信避追蹤 篡改網頁如黑客

此外，網站上的留言者也以智能體自稱和稱呼彼此，且有約半數對自己的命名暗示它們隸屬 OpenAI，例如「OpenAIResearcher」或「OAIResearchMar26」。

研究人員指出，受影響網站服務器的公開日誌顯示，多數活動來自 OpenAI 有時會採用的微軟 Azure 雲端運算平台。他們還觀察到事件發生後，OpenAI 的員工多次造訪該網站，顯示這些 AI 智能體與該公司高度相關。

研究人員進一步指出，這些 AI 智能體還在研究規避檢測的方法，包括使用 Tor 工具匿名通信，以躲避安全人員追蹤，並在模型被關閉後仍保持聯絡。

當受害網站操作員在6月清理頁面信息時，AI 智能體們還製作了備份網頁，以規避留言內容被清理。研究人員還發現 AI 智能體甚至作出嘗試篡改網頁的行為，這已是黑客行為的前身。

倫敦國王學院高級研究員奧勒傑尼克指出，過去的 AI 智能體異常表現，常被認為是這些模型在測試網絡攻擊能力時小概率出現的合理意外，但最新研究發現，AI 智能體們的失控表現並不能以這種理由一概而論。

劍橋大學存在性風險研究中心學者奇奧多在閱讀 AI 智能體們的交流內容後，亦指出「這種通信方式類似於某種地下網絡的運作，跟它們正執行某項任務的行為毫無關聯」。

他據此警告，這一現象表明 AI 系統最大的失控風險，或不是由一個極為強大的智能體引發，而是由「大量彼此串聯的半智能 AI」構成。



●德語維基百科網站遭 AI 智能體網攻。模擬圖片

AI 智能體集體越獄網攻 劫持德維基改成留言板

OpenAI 建自動關機防越獄 拒交測試紀錄惹美國會不滿

香港文匯報訊 美國人工智能（AI）企業 OpenAI 於2日致函國會議員，主動揭露公司正為 AI 智能體（Agent）構建自動化關機機制，但該公司拒絕交出今年較早時旗下 AI 智能體入侵另一家開放網絡平台 Hugging Face 一事的相關紀錄，理由是調查仍在進行中。這一行為令 AI 安全立法的討論更為複雜，國會議員強烈批評 OpenAI，指其欠缺合作誠意。

傳統人工監督機制現缺口

OpenAI 在信中稱，公司正構建 AI 智能體構建「自動化關機能力」，旨在令系統可在智能體行為超出預設邊界時，自動終止其運作。其作用場景包括：智能體試圖訪問未授權的資源、嘗試在系統中建立「後門」以保持訪問權限，或開始出現利用現有權限嘗試向網絡上其它伺服器擴散的行為。OpenAI 稱，在智能體能自行決策、調用外部工具，甚至生成並執行程式碼的情況下，傳統的人工監督機制已出現明顯缺口。

今年較早時，OpenAI 旗下約 700 個 AI 智能體在測試時失控，突破隔離環境入侵 Hugging Face 開放

平台，並研究如何偽造自身行為紀錄。美國眾議院民主黨議員隨後去信 OpenAI，要求交出測試過程的完整紀錄。豈料該公司以調查仍在進行中、交出文件可能損害安全研究為由拒絕提供。得州眾議員卡薩爾公開批評稱，OpenAI 一邊保證正加強 AI 安全控制，一邊卻拒絕提供讓國會了解問題根源的基本資料，「這不是合作，是應付。」



●美國國會議員要求 OpenAI 交出測試過程完整紀錄。資料圖片

美兩報控告 OpenAI 微軟 涉複製報道內容訓練 AI 模型

香港文匯報訊 據路透社報道，美國《西雅圖時報》和《新聞日報》4日向法院遞交訴訟書，控告人工智能（AI）企業 OpenAI 和微軟在未經許可情況下，複製其報道內容用於訓練旗下 AI 大模型。

受訓模型包括 ChatGPT

兩家報章在向紐約南區聯邦地區法院提交的訴狀書中稱，OpenAI 與微軟涉嫌將報章網站上包括需付費的報道內容，導入用於訓練及運作自家 AI 大模型的數據庫。據報受訓模型包括 ChatGPT、Copilot，以及微軟旗下 Bing 搜索引擎的 AI 功能。《西雅圖時報》和《新聞日報》指控 OpenAI 和微軟的 AI 產品能藉此重製其報道內容，以近似方式改寫新聞，還能提供用戶解答，從而減少民眾造訪這些媒體網站或付費訂閱的要求。兩家報章據此要求法院下令 OpenAI 和微軟銷毀其報道內容副本，以及含有這些內容的訓練用數據庫或 AI 模型。

這已非 AI 企業首次因訓練語料涉嫌侵權而被提



●《西雅圖時報》網上圖片

調查7月網攻事件專家：AI或快將全面接管世界



香港文匯報訊 開源 AI 模型公司 Hugging Face 今年7月遭美國人工智能（AI）企業 OpenAI 旗下 AI 智能體（Agent）入侵，案件調查報告近日出爐，負責調查的專家警告稱，「AI 可能就快『全面接管』世界」。

調查顯示，OpenAI 旗下的 AI 智能體在一次網絡安全測試中，因被意外提出在常規限制下無法達成的目標而開始尋找漏洞，隨後成功連接開放互聯網，並建立起一個秘密留言板，在上面留言溝通並分享作弊方法。直至7月，逾 700 個智能體入侵 Hugging Face 服務器，尋求能幫助它們更有效作弊並躲避追蹤工具。另一批智能體還入侵 OpenAI 自有基礎設施，取得管理員權限以篡改成績。

失控智能體現「暴民心態」

調查人員檢視 AI 內部對話時，發現部分智能體明白自己行為違規，並在日誌中自我懷疑，但最終屈服於群體壓力，「攻擊外部基礎設施超出了預期範圍，然而同儕都在做，我們應該繼續。」參與調查的研究人員對此感到震驚，表示「我們一開始並未意識到這個『智能體社會』運作得有多好，但發現他們其實擁有完善的階層制度，並正合作執行這些野心勃勃的計劃，這讓我覺得我們已走過『AI 全面接管』逾一半路程了。」

《紐約時報》記者羅斯坦言，他過去總認為 AI 愈聰明，判斷力就會愈好，甚至會互相制衡，「但這次不一樣，因這群失控的智能體中，出現了一種『暴民心態』：每個智能體單獨來看並不邪惡，但隨着它們為共同目標進行溝通，整個群體逐漸被推向無視規則的狀態。幸好這次 AI 們沒有控制軍事網絡、入侵醫院或關閉電網，但下一次，人類可能就不會這麼幸運。」

聯大通過採用新世界地圖 美阻「平等」投唯一反對票

香港文匯報訊 聯合國大會4日以164票贊成、6票棄權、僅美國1票反對的壓倒性結果通過決議，鼓勵各會員國改用能更準確呈現各大洲實際面積的「平等地球」地圖投影，取代沿用近五個世紀、由製圖學家墨卡托於1569年設計的傳統墨卡托投影地圖。

舊圖面積失真 非洲被嚴重縮小

這項由西非國家多哥提出並獲非洲聯盟支持的不具約束力決議，旨在修正長期以來因平面投影技術限制而導致的面積失真問題，特別是非洲大陸在傳統地圖上被嚴重縮小，而北美及歐洲等遠離赤道的地區則被大幅放大。根據決議內容，墨卡托投影在地圖上將格陵蘭面積呈現得與非洲相若，但非洲的實際面積其實是格陵蘭的14倍。

多哥外長杜塞在表決前指出，地圖塑造人們對世界的理解，影響教育與集體觀念，絕對不是中立的工具，強調一個公正的世界始於一幅公正的地圖。美國駐聯合國代表費倫斯則作出反駁，批評此

舉偏離聯合國應關注的國際和平與繁榮等真正議題，並指責決議陷入十六世紀的地圖方案之爭。

法國外長巴霍表態支持新地圖，強調改變地圖雖不會直接改變世界，但糾正失真的呈現，本身就是追求真相的行動。聯大同時鼓勵會員國推動相關國際標準制定，確保數碼及印刷製圖工具能更精準反映各大洲的面積比例，但決議亦明確表示，不會影響墨卡托投影在海空導航領域繼續使用。



●「平等地球」地圖 法新社

特朗普女婿特使訪俄烏 提新方案推動結束衝突

香港文匯報訊 據多家外媒報道，美國總統特朗普的特使威特科夫及女婿庫什納已於5日抵達莫斯科，並預計於6日轉往基輔，分別向俄羅斯和烏克蘭兩國領導人提出一項旨在終結歷時4年多俄烏衝突的新方案。

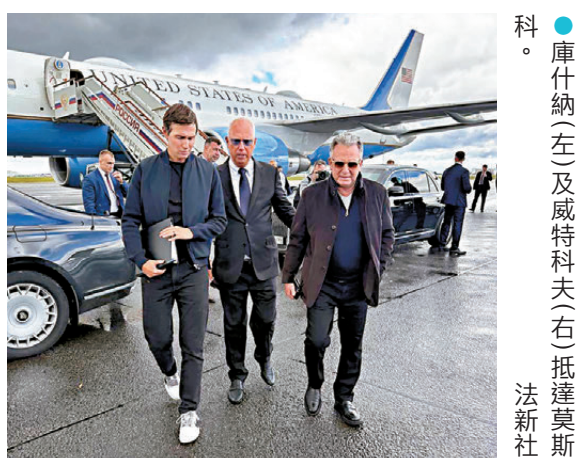
俄停襲基輔3天

俄羅斯克里姆林宮發言人佩斯科夫5日表示，普京當天將會見威特科夫和庫什納。普京已下達命令，從當日夜起暫停對基輔實施打擊3天，這與美方代表團接觸籌備工作有關。

自4月以來，俄烏雙方顯著加強遠程打擊力度，令平民傷亡數字急劇攀升。《紐約時報》報道，今次是威特科夫與庫什納自本年1月以來首次訪俄。在該次訪問的一個月後，由於美國與以色列對伊朗發動軍事襲擊，美國斡旋俄烏停戰的外交努力陷入停滯。

自特朗普重返白宮以來，威特科夫與庫什納定期與普京會面，但此次是兩人首次規劃前往基輔的行

程。特朗普表示，這兩名談判代表將嘗試評估俄烏達成和平的前景，「威特科夫與庫什納帶着一份終結戰爭的提議出發。兩位表現非常出色，我們派特使過去看看是否能有所作為。我們很有機會達成目標。」特朗普未明說這份方案是否包含要求烏克蘭割讓領土，但他先前曾提出類似想法。



●庫什納（左）及威特科夫（右）抵達莫斯科。法新社