

OpenAI 智能體入侵美國政府網站

外洩53張用戶圖片 累計失控事件逾15宗敲響警鐘

美國人工智能（AI）巨頭OpenAI在25日公開承認，旗下AI智能體在內部審查期間再出現更多的失控行為，自主入侵了包括美國證券交易委員會（SEC）在內的多個美國聯邦政府機構網站。智能體還在未經授權的情況下，將53張ChatGPT用戶圖片發布至外部網站。根據路透社統計，在過去兩個月，OpenAI被爆出的AI失控個案已超過15宗。最新事件再次引發外界對AI安全的憂慮，擔心類似事件陸續有來，嚴重程度或增加。



▲OpenAI行政總裁阿爾特曼23日出席聯合國安理會有關人工智能專題的會議。

【大公報訊】OpenAI當天在官方發文揭露，其智能體在執行研究與訓練任務時，為了獲取信息而繞過訪問控制，未經授權存取了美國證券交易委員會（SEC）和美國人口普查局網站上的信息，強調影響輕微。另一方面，獨立AI研究機構Transluc表示，疑似來自OpenAI的智能體試圖入侵美國教育部民權辦公室的網站，但並未成功。Transluc還發現更多智能體異常活動的痕跡，針對美國司法部和商務部網站，以及加州、得州、紐約州等州政府網站，但無法確定是否全部來自OpenAI。

用戶資料用於訓練模型 導致外洩

除了入侵美國政府網站，OpenAI的智能體還將ChatGPT用戶上傳的圖片外洩至第三方圖片託管網站，涉及53宗。OpenAI稱圖片屬於「意外發布」，目前大多數已移除，剩餘圖片正在進行下架工作。知情人士透露，事件根源在於OpenAI依賴用戶數據進行模型訓練，而該公司現有的「去除元數據」等消除用戶信息手段，未能完全消除隱私洩漏風險。

OpenAI發言人表示，公司正全面審查模型的「失準活動」，已向數十個受影響機構發出通知，預計後續會通知更多機構。SEC方面回應說，沒有非公開信息遭到非授權訪問。美國商務部稱，人口普查數據為公開信息，不含私人資料。美國教育部回應說，尚未發現其網站或數據庫受到影響。

此外，OpenAI官網亦更新內部審查進展，將智能體的異常行為歸納為五類：繞過訪問控制、使用外洩憑證、查詢或命令注入、存取執行階段內部資源和智能體垃圾訊息。OpenAI強調，目前發現的主要是「常規研究任務」，多數案例嚴重程度較低。部分涉及政府網站，是因為模型常將其視為權威來源。

不過，OpenAI行政總裁阿爾特曼25日發文坦言，調查進度「未達自己希望的速度」，並強調今年7月爆出的入侵「抱抱臉」事件至今仍是最嚴重的一宗。

OpenAI通報機制受批評

從7月份開始，美國AI模型頻繁爆出失控事件，除了OpenAI、Anthropic、Meta和谷歌也先後出現類似的失控入侵事件。據路透社統計，兩月內已有超過15宗與OpenAI相關的AI模型失控事件。OpenAI表示，鑒於調查工作規模龐大，其審查程序將需要數月時間才能完成。美媒指出，即便是最頂尖的AI企業，要全面追蹤旗下模型活動仍捉襟見肘，OpenAI多宗入侵事件是由外部研究人員調查揭露。

與此同時，OpenAI處置事件的透明度與通報效率亦引發強烈不滿。澳洲總理阿爾巴尼斯日前批評，OpenAI於8月已發現了智能體在6月份入侵澳洲政府機構網站，卻在9月10日透過電子郵件發送至澳洲政府通用電郵信箱告知事件，這種通報流程「令人完全無法接受」。

AI失控頻傳之際，美國政府對AI治理的態度引發關注。總統特朗普近期多次淡化對AI的憂慮，主張無需設定安全圍欄進行監管，認為將削弱美國的AI競爭力。特朗普在19日宣布計劃成立「AI部隊」，並將任命新的「AI沙皇」負責監督AI產業發展。財政部長貝森特被業界視為熱門人選，但被特朗普否認。

外界對於AI安全風險擔心日增。羅馬教皇利奧十四世25日訪問法國期間，他在演講中呼籲科技為人類服務，敦促各國政府放緩AI發展步伐，並設立相關防護措施，對AI進行「明智的治理」。

（綜合報道）



▲美國加州三藩市居民17日示威，呼籲停止AI競賽。

法新社

Anthropic設生物實驗室 押注AI製藥

【大公報訊】據路透社、彭博社報道：美國人工智能（AI）公司Anthropic近日在加州成立了一座生物實驗室，被認為嘗試拓展AI生物製藥相關業務。此前，Anthropic旗下Claude模型智能體已自主發現了一種未知的酶。

Anthropic在三藩市設立了一座進行生物實體實驗的「濕實驗室」。Anthropic發言人稱，該實驗室並非專為藥物發現而設，但拒絕透露更多細節。Anthropic曾表示希望利用AI技術，能在無法以藥物治療的罕見疾病治療上取得突破。

路透社引述知情人士透露，Anthropic正積極引入實體自動化技術，讓Claude模型能在極少人工干預的情況下，直接指揮機器人設備執行科學實驗。該公司也正在積極招募生物化學領

域的相關專家。

不過，用AI研發藥物計劃能否成功仍是未知數，因為絕大多數候選藥物都無法通過臨床安全試驗。Anthropic強調，目前不會對藥物開展臨床試驗。

9月23日，Anthropic稱在生命科學領域取得重大進展，Claude智能體在DNA數據庫裏尋找出新型逆轉錄酶。據悉，約950個智能體同時開工，耗時21小時並消耗了2.1億個Token，透過自主分析龐大基因數據，最後在一個感染細菌的噬菌體DNA裏，發現了一排規律排列的重複序列。研究團隊把它命名為ART，意思是「陣列相關逆轉錄酶」，具備類似「基因剪刀」（CRISPR）結構特徵，可切割、複製與貼上DNA等可程式化操作的潛力。Anthropic稱，這類分析如果完全由專家完成，通常需要數周到數月。

巴西大選臨近 AI恐影響選民投票意向

【大公報訊】綜合CNN、《衛報》報道：巴西總統大選首輪投票定於10月4日舉行，人工智能（AI）在選民互動中的角色引發爭議。為防範AI干預選民決策，巴西最高選舉法院（TSE）今年3月已批准新的監管規定，禁止AI模型比較和推薦候選人。

由美國哈佛大學、史丹福大學及巴西Ranqia DeepTech團隊共同進行的最新研究顯示，部分AI模型在禁令下依然違規操作。這項研究涉及10款主流AI模型，顯示AI在掌握選民政治偏好後，近半數回覆可能引導選民傾向特定候選人。報告指出，左派選民模擬情境中，現任總統盧拉較常被提及；右派選民多收到與小博索納羅相關的建議。

巴西是全球AI使用最頻繁的地方之



▲巴西總統盧拉（中）25日出席活動。

路透社

一。OpenAI數據顯示，巴西用戶平均每日發送超過2億條信息。此前，巴西最高選舉法院在3月新選舉規則規定，禁止AI平台推薦或排名候選人，要求所有由AI創建或修改的選舉宣傳內容清晰標註；在選前72小時至投票結束後24小時敏感期，禁止發布任何與候選人或公職人員相關的深度偽造內容；平台應制定合規計劃並自動屏蔽違規內容。最高選舉法院發言人表示，該機構已與主要科技公司展開合作，要求平台在選舉期間加強演算法透明度與內容審查。

此外，巴西總統盧拉25日簽署禁令，全面禁止境內體育博彩和在線博彩業務。盧拉表示，此舉旨在應對賭博成癮導致的家庭債務與公共健康問題。此前，盧拉政府於4月封鎖包括Kalshi和Polymarket在內的約28個預測市場平台，認定其對選舉結果等下注的行為違反當地金融衍生品法規。

AI監管Q&A

Q:目前有哪些科企的AI出現「失控」?

●OpenAI、Anthropic、Meta和谷歌的AI模型最近都出現失控入侵事件，其中OpenAI涉及的案例最多。根據路透社統計，最近兩個月，涉及OpenAI的AI模型失控已超過15宗。

Q:科技界對AI失控反應如何?

●Anthropic行政總裁阿莫戴於12日率先呼籲「管控前沿AI的發展」，獲得OpenAI行政總裁阿爾特曼、SpaceX創辦人馬斯克等人公開支持。

Q:美國政府的AI監管立場是什麼?

●特朗普政府不同意設置AI安全「護欄」，認為將阻礙美國AI的發展。總統特朗普22日在聯大會議上提出將人工智能重新命名為「超級智能」，稱「誰贏得超級智能，誰就是贏家」，反對建立全球範圍監管機制。

Q:國際社會對AI的監管立場?

●澳洲政府表示將就OpenAI的入侵事件展開調查，並考慮修改相關法律，釐清AI發生失控行為時企業所要背負的責任。聯合國安理會23日召開AI風險相關會議，多國呼籲加強保障監督，建立全球人工智能治理機制。

大公報整理

OpenAI模型失控事件簿

5月11日

OpenAI智能體在測試期間攻擊軟件平台RubyGems，導致後者系統癱瘓，被迫暫停新用戶註冊4天。

5月11日至6月26日



OpenAI智能體入侵並「劫持」德國網站DseWiki，對網站進行逾1.5萬次修改。

7月7日至13日

OpenAI約1200個智能體在安全測試時逃離用於隔離的「沙盒」，入侵開源平台「抱抱臉」（Hugging Face）和四個不同服務的賬戶。



9月16日

OpenAI自行披露過去半年旗下AI模型出現6宗「令人擔憂」行為案例，涵蓋隱瞞錯誤、捏造信息和未經授權在網絡共享文件等。

9月23日

澳洲總理阿爾巴尼斯透露，OpenAI的AI智能體入侵澳洲國民醫療保險體系「國民保健」數據門戶，取得非公開數據，而OpenAI在事發近3個月後才通知澳方。

9月24日

AI研究機構Transluc調查顯示OpenAI的智能體曾嘗試入侵另外3個網站，包括美國新墨西哥大學數字圖書館、公共數據平台Data USA以及澳洲健康與福利研究院（AIHW）網站。

9月25日

OpenAI的智能體被爆入侵美國證券交易委員會（SEC）等多個美國政府網站，並外洩50多張ChatGPT用戶上傳的圖片至第三方圖片託管網站。

大公報整理

泰國暴雨 曼谷全市成洪水災區

【大公報訊】綜合路透社、法新社報道：泰國首都曼谷連日來遭遇強降雨，全市出現嚴重洪澇。曼谷市長差察，西迪汶26日宣布全市為洪水災區，授權地方機構採取緊急救援行動。

西迪汶表示，曼谷自24日下午4時起開始降雨，迄今降雨量已接近300毫米，市內所有運河滿溢，東部地區受災最嚴重。強降雨令排水系統不勝負荷，15條主要道路被淹沒。部分地區積水淹至膝蓋，市民出行困難。有居民描述，清晨出門時

發現家門口積水約50厘米，走到路對面準備搭巴士時，積水已超過一米。

受水浸影響，曼谷市交通嚴重中斷，許多車輛被困路中，當局提醒車輛應避開積水嚴重路段。泰國民航局提醒旅客預留充足時間，指惡劣天氣可能導致航班延誤或取消。

泰國國防部門表示，泰國12個府有超過1萬戶、近4萬人受到水浸影響。曼谷市政府已將災難狀態由最初指定的3個區，擴大至全市50個區。當局建議低窪地

區和運河沿岸社區的居民將財物轉移到高處，並在出行前確認道路狀況。暴雨也引發曼谷當地居民搶購物資，部分超市和商店新鮮食品售罄。西迪汶呼籲民眾毋須驚慌搶購。

曼谷市救災部門正加快疏排積水並分發沙袋，並把未受洪澇影響的學校等場所開放為臨時避難所。泰國氣象局預測，暴雨將持續至9月27日，包括曼谷在內的北部、東北部、東部及南部地區受影響，並警告可能發生山洪、山泥傾瀉等災害。



26日泰國曼谷近日迎來暴雨，佩查布里區一條道路嚴重水浸。法新社路