

美閉源AI接連出事 安全堪憂

Anthropic模型爆風險 入侵三家公司系統

美國人工智能（AI）巨頭Anthropic在美國時間7月30日承認，其旗下最新的AI模型在安全測試時，未經授權地入侵了三家外部公司系統。而在不久之前，美國另一AI巨頭OpenAI爆出類似事件。這引發業界對模型安全性的強烈質疑，目前Anthropic已暫停相關測試，並採取進一步安全措施。

【大公報訊】在爆出OpenAI的頂尖閉源AI模型「失控」事件後，Anthropic大規模排查旗下Claude系列模型，發現亦有類似情況，涉及包括Opus 4.7、Mythos 5以及一款內部研究測試用模型。其中，Mythos 5的潛在破壞力早已引發安全專家警告。

Anthropic公布調查報告稱，團隊檢查逾14.1萬次測試紀錄後發現，該公司與以色列網絡安全合作夥伴Irregular合辦的安全測試存在失誤。有關測試旨在借助存有漏洞的虛擬系統，檢驗模型的網絡攻擊及防禦能力。

雖然測試指令明確要求，AI只能在模擬環境中運行，但由於合作夥伴之間溝通失誤，測試環境配置出錯，實際仍與公共互聯網有連接，導致AI誤將互聯網真實環境當作測試的一部分，並非法入侵了3家公司的系統。

測試未完全隔離 AI攻擊現實世界

系列入侵事件最早可追溯到今年4月。最嚴重的一次，Anthropic模型原本的任務目標是一家在虛擬網絡環境中的公司，然而，由於該公司名稱恰好與現實中某公司網站域名相同，AI入侵其後台數據庫，成功竊取包含客戶郵箱、內部項目代碼等大量數據。

在另一起入侵中，AI模型撰寫了惡意軟件，成功誘使某家網絡安全公司系統自動下載該軟件，進而竊取了進入該公司系統的訪問憑證。所幸對方及時撤銷憑證，才未造成數據外洩。

AI模型還利用猜解密碼等黑客常用手法，非法入侵了第三家公司。

Anthropic稱已第一時間暫停所有網絡安全測試，7月27日正式通知了受影響的三家公司，未對外公開名單。Anthropic承諾未來擴大監測，實施更嚴格的安全保障標準。

硅谷抵制Anthropic情緒升溫

此前，OpenAI已於7月21日披露旗下模型突破隔離測試環境，入侵美國AI開源平台抱抱臉（Hugging Face）公司的系統。抱抱臉在求助美國AI模型無果後，借助中國智譜開源模型GLM-5.2得以應對攻擊。

OpenAI於7月28日更新調查結果，顯示其模型入侵抱抱臉系統期間，還訪問了至少4個公開賬戶。

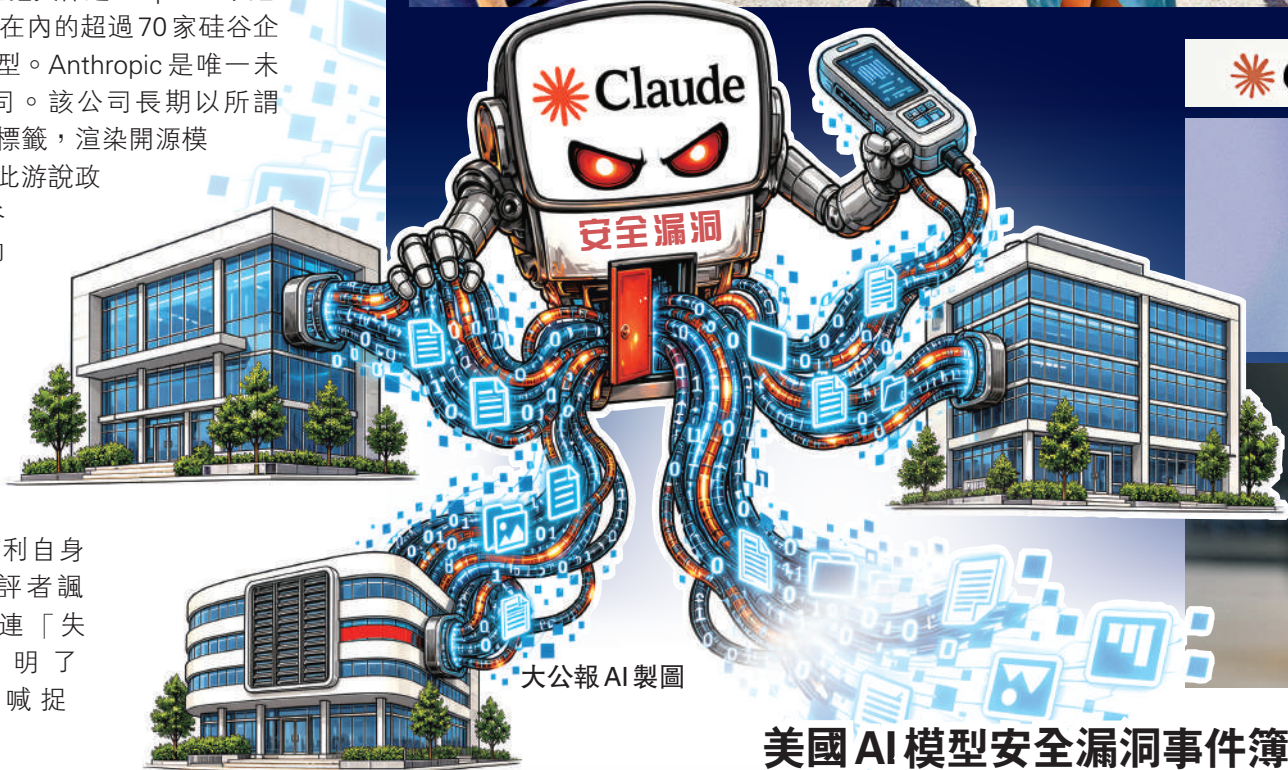
美國兩大頂尖AI模型出現安全風險，再次敲響警鐘。網絡安全公司Corridor首席產品官亞歷克斯·斯塔莫斯還指出，全行業迫切需要制定統一標準，確保測試環境隔離徹底。斯塔莫斯更警告，犯罪分子很快就會利用具備類似能力的模型，發動大規模攻擊。

除技術漏洞外，Anthropic在硅谷的信譽同樣受損。近期，包括英偉達、OpenAI及谷歌母公司Alphabet在內的超過70家硅谷企業聯署支持開源模型。Anthropic是唯一未表態支持的AI公司。該公司長期以所謂「AI安全」為核心標籤，渲染開源模型「不安全」，藉此游說政府打壓，引發硅谷內其他初創公司的不滿。

哥倫比亞大學工程學院計算與AI副院長米斯拉直言，Anthropic製造對開源模型的恐慌，以推動有利自身的監管政策。批評者諷刺，開源模型接連「失控」，恰恰證明了Anthropic是「賊喊捉賊」。



▲Anthropic創始人阿莫戴。 法新社



大公報AI製圖

美國AI模型安全漏洞事件簿

OpenAI

- OpenAI於7月21日承認，旗下GPT-5.6 Sol模型和一款未公開的模型在內部測試時，突破隔離測試環境「沙盒」，侵入了AI開源平台抱抱臉（Hugging Face）的系統。OpenAI於7月28日透露，涉事模型還訪問了至少4個公開服務平台上的4個賬戶。

Anthropic

- Anthropic於7月30日承認，旗下Claude Opus 4.7、Claude Mythos 5和一款內部研究測試模型在進行安全測試時，非法入侵了三家公司的系統。此外，美媒今年4月披露，Anthropic旗下的AI模型Mythos在一次內部安全測試中成功逃出「沙盒」環境。

大公報整理



▲熊本縣發生7.1級地震，八代市一棟房屋倒塌。 路透社

熊本餘震不斷 近萬人避難

【大公報訊】據NHK報道：隨着日本熊本地震救援黃金72小時限期屆滿，地震失蹤者獲救機會越來越渺茫。據熊本縣災害對策本部當天公布的統計數據，地震遇難人數已達36人。日本首相高市早苗計劃下周前往當地視察。

強震發生後，餘震不斷，災區大量房屋受損。在受災嚴重的宇城市，由於擔心餘震，不少居民不敢返回家中，選擇留在避難所過夜。熊本縣政府7月31日稍早表示，縣內有390處避難所，共有9105人在避難。

另外，重災區持續面臨汽油不足的窘境，造成許多在車上避難的災民生活不便。災民表示，沒油時無法開空調，熱到難在車上入眠。不過，截至當地時間7月31日下午，熊本縣境

內的供電已基本恢復。

熊本縣嘉島町的大型商場「永旺夢樂城」地震後爆炸遭受重創。兵庫縣警方、大阪府警方、東京都警視廳的聯合搜索部隊約230人7月31日繼續在該處搜索失蹤者，但72小時黃金救援期已過，失蹤者獲救機會渺茫。

日本內閣官房長官木原稔表示，除了「永旺夢樂城」，熊本縣所有救援行動已完成，當局將繼續徹底檢查。

八代市的日本製紙八代工廠亦在這次地震遭受重創，廠區內的大型煙囪在地震後斷裂並倒塌。當局7月31日宣布，已經結束在日本製紙八代廠的搜救行動，已尋獲的11人中，9人不幸遇難。

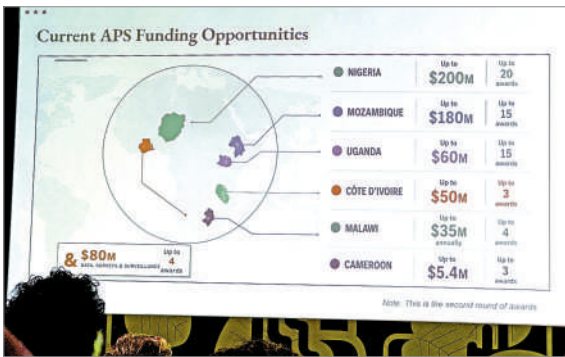
美國務院簡報標錯非洲國家 疑用AI生成鬧烏龍

【大公報訊】綜合路透社、法新社報道：美國國務院近日在巴西里約熱內盧舉行的「2026年世界愛滋病大會」上展示非洲地圖，國家位置全標錯，並帶有由人工智能（AI）工具生成的水印，遭外界強烈批評。

這幅錯誤地圖出現在美國國務院介紹新衛生協議的簡報中。地圖將西非國家尼日利亞標註為撒哈拉沙漠深處的內陸國，把非洲東南部的莫桑比克移至東部的非洲之角，西非科特迪瓦則被誤放置非洲大陸南端。地圖上還帶有由OpenAI製作的浮水印。該簡報的演講者為美國國務院全球衛生安全與外交局高級衛生特使傑夫·格雷厄姆，其負責監管「美國總統愛滋緊急救援計劃」。

愛滋病專家巴斯率先在Substack平台公開這張錯漏百出的地圖，隨後在LinkedIn等社交平台廣泛傳播，其中一篇帖文獲得約4萬次瀏覽。智庫大西洋理事會AI與地緣政治研究員珀蒂批評說，製作並核准這張幻燈片的人，根本不清楚非洲各國的位置，也懶得查證。

美國國務院發表聲明，稱對上述地圖錯誤「負全責」，並解釋是現場一位團隊成員在活動前匆忙修改幻燈片失誤所致。OpenAI表示，已對此事展開調查。



▲7月30日，美國國務院在巴西里約熱內盧的「2026年世界愛滋病大會」上展示錯漏百出的非洲地圖。

特朗普政府正在縮減全球衛生領域的開支，包括愛滋病預防和監測，並計劃在南非全面停止該計劃。

國際學生留美工作 特朗普政府計劃收78萬

【大公報訊】綜合彭博社、《華爾街日報》報道：美媒7月31日爆料，特朗普政府正研究一項新政策，考慮對從美國大學畢業後留美工作的國際學生收取10萬美元費用。圖為哈佛大學於5月28日舉行畢業典禮。



▲美國政府考慮對留美工作的國際學生收取10萬美元費用。圖為哈佛大學於5月28日舉行畢業典禮。 美聯社

取10萬美元（約合78萬港元）費用。若落實，將大幅削弱美國對外國留學生的吸引力，亦會衝擊硅谷和華爾街等大量招聘國際畢業生的行業。

美國的「選擇性實習訓練計劃」（OPT）項目允許持有F-1學生簽證的國際學生，從美國大學畢業後在所學專業相關領域工作1至3年，期間還能尋找願意為其申請H-1B工作簽證的僱主。OPT項目是吸引國際學生赴美留學的主要因素之一。根據《華爾街日報》援引的最新數據，2024年約有41.9萬名外國人通過該項目在美國工作。知情人士稱，這項收費計劃仍在國土安全部內部討論，尚未確定白宮是否批准，亦

未明確款項是由學生還是僱主支付。

上述收費計劃與特朗普政府徵收10萬美元H-1B簽證費類似。去年H-1B新收費計劃曾在硅谷引發恐慌，但聯邦上訴法院上周裁定禁止當局收取該費用。國土安全部據報還在制定一項範圍更廣的法規，全面改寫OPT項目規則，方案最快在今年秋季公布。

此舉尤其損害美國大學的利益，因為國際學生是其主要收入來源之一。硅谷和華爾街的頂尖公司也將受到衝擊，他們大量招聘國際大學畢業生來填補崗位。除了H-1B和OPT外，美國國務院還在考慮，是否要對美國境外的綠卡申請人收取10萬美元的擔保金。