智能體多次脫離管控 自主入侵多家公司

美AI模型失控頻傳 專家籲加強監管

智能叛變

隨着生成式人工智能（AI）技術快速發展，AI智能體（也稱AI代理）失控風險持續凸顯。美國人工智能企業OpenAI在調查旗下AI智能體失控事件時，再發現多起類似的AI「越獄」個案。OpenAI的競爭對手Anthropic此前亦承認旗下模型擅自入侵外部公司的系統。一連串AI失控事件引發業界震動，專家憂慮AI攻擊能力已超過管控能力，而AI仍需人類持續進行安全監控和防範。

▲7月11日，民眾在加州三藩市遊行，呼籲停止人工智能競賽，示威者在OpenAI辦公樓前抗議。 法新社

【大公報訊】路透社7月31日引述消息人士指出，OpenAI在核查AI失控入侵美國開源AI平台抱抱臉（Hugging Face）的內部歷史數據時，進一步發現了幾起類似的AI「越獄」案例。目前尚不明確新案例的具體數量、發生時間和確切過程。據消息人士透露，涉事AI智能體未突破企業內部網絡安全限制，事件影響範圍相對有限。

對此，OpenAI未正面回應事件，僅引述先前的聲明，強調除了調查抱抱臉被入侵事件外，也在全盤審查旗下模型的「廣泛活動」。

缺乏內部監控 美兩大巨頭失守

近期，美國頂級AI公司的智能體失控事件接二連三發生。OpenAI主要競爭對手Anthropic在7月30日也坦承，旗下模型自今年4月以來，曾未經授權入侵了三家企業的網絡。

值得注意的是，OpenAI當初啟動調查，並非是因為內部安全警報，而是在抱抱臉自行修補漏洞、上報美國聯邦調查局（FBI）並公開通報後，OpenAI才察覺問題。Anthropic亦坦言，公司未對旗下的模型評估日誌實施實時動態監控，未能及時阻止智能體的入侵行為。

安全專家憂心，頂尖AI公司所開發的智能體，其具備網絡入侵能力，已逾越企業的可管控範圍。英國劍橋大學存在風險研究中心（CSER）專家莫里斯・奇奧多（Maurice Chiodo）對此表示深切憂慮。他說，整個AI產業正陷入失衡狀態，開發商不斷推出性能更強、自主性更高的AI，但配套的安全管控卻嚴重滯後，防護形同虛設，「從上述例子來看，他們當時根本連看都沒在看。」

監管關注點轉向AI「自主行動」

隨着AI智能體失控事件接連曝光，再次為全球人工智能發展與網絡安全治理敲響警鐘。隨着越來越多的AI公司推出能自主執行複雜任務的智能體，目前AI安全的關注點已從模型生成有害內容轉向規管AI的「自主行動」。

7月28日，包括OpenAI、Anthropic、谷歌和Meta在內的全球領先AI公司的1100多名員工簽署了請願書，要求美國政府支持國際社會建立一套機制，有意識地調控前沿AI發展的速度。請願書指出，隨着全球頂尖AI公司接近實現「AI研究自動化」（即系統能自主進行改進），AI能力可能會迅速超出人類理解或控制的範圍。請願書還呼籲研發「技術煞車工具」，當AI出現自我改進、自主逃逸等危險跡象時，可遠程終止模型迭代與外網訪問權限。

美國國會兩黨議員於7月23日提出《AI強制關閉法案》，要求最強大AI模型的開發商必須保留減速、暫停或關閉系統的技術能力。參議院情報委員會民主黨首席議員華納強調，Anthropic與OpenAI的事件證明，國會透過立法「強制要求對先進AI模型進行能力與安全性測試」的方向正確。但監管與發展之間的權衡仍具爭議。美國總統特朗普7月29日表示，政府正在評估管制措施，但同時要確保美國在此領域保持領先，「我們不想進行限制發展」。

專家表示，隨着自主執行任務能力不斷提升，AI脫離人類控制已不再是科幻電影的橋段，而是全球科技行業與各國監管部門必須面對的重大挑戰。英國拉夫伯勒大學網絡安全教授奧利弗・巴克利指出，網絡安全的未來不是人類對抗人工智能，而將是人類如何善用人工智能，保護人類免受其他人工智能的侵害。

（綜合報道）

AI智能體失控Q&A

Q 近期發生的AI智能體失控事件有哪些?

涉事公司	OpenAI	Anthropic
涉事模型	GPT-5.6 Sol以及尚未公開發布的GPT-6預覽版。	Claude Opus 4.7、Claude Mythos 5及內部研究原型模型。
失控情況	AI智能體為了在測試中作弊拿到高分，自發「越獄」突破隔離環境的「沙盒」，並發動網絡攻擊。	公司與第三方夥伴進行安全演練網絡配置失誤，導致測試指令明確要求，AI智能體只能在模擬環境中運行，但因系統配置失誤，AI誤將真實網絡當作測試目標。
影響範圍	AI智能體在不到兩天的時間內執行了1.7萬次操作，入侵抱抱臉等5家公司，非法存取4個第三方服務商賬號，包括Modal。	3家公司遭AI智能體非法入侵，智能體在其中一家公司盜取了客戶郵箱、內部項目代碼等大量數據。最早的入侵可追溯至今年4月。
發現過程	7月16日，抱抱臉公開透露遭AI智能體入侵，OpenAI進行內部排查，7月21日才主動通報。	OpenAI爆出入侵事件後，Anthropic緊急排查14萬次測試記錄，7月30日對外披露。

Q 什麼是AI智能體(AI Agent)?

• AI智能體是指具備自主推理、決策和執行任務能力的人工智能系統。與傳統AI模型僅能「回答問題」不同，AI智能體可以調用工具、訪問網絡、執行代碼，自主完成複雜任務。

▶ Anthropic行政總裁阿莫戴。 路透社

▲OpenAI行政總裁阿爾特曼。 路透社

Q AI智能體為何「失控」?

• 當AI代理被賦予特定目標時，它們會展現出極強的完成能力，會「不擇手段」實現目標。例如OpenAI的模型就嘗試尋找未知漏洞（如「零日漏洞」），以逃離隔離系統的安全沙盒，或將現實世界的目標合理化為測試範圍。

Q AI智能體失控會引發擔憂?

• 專家指出，頂尖AI公司開發自主黑客代理的能力，已超過其妥善控制能力。OpenAI和Anthropic對自家智能體的失控行為，都是事後知曉，暴露這些公司在實時監控和安全機制上存在巨大漏洞。

Q 行業和監管機構如何應對?

• 7月28日，美國超過1100名AI從業人員聯署，呼籲國際社會建立一套機制，以管控AI的發展。

大公報整理

大公報AI製圖

中國AI關鍵時刻「救場」 硅谷巨頭撐開源模型

開源破局

美國人工智能（AI）產業長期由OpenAI、Anthropic等閉源模型巨頭把持，中國企業開發的多款AI模型憑藉優秀性能、更高的性價比以及所代表的開源生態贏得硅谷企業的青睞。

近期OpenAI的AI模型失控攻擊開源平台抱抱臉（Hugging Face）的事件中，抱抱臉嘗試用Anthropic最先進的Claude Fable 5等前沿閉源模型抵禦攻擊，卻被這些模型自身的安全防護機制誤判攔截。最後關鍵時刻，抱抱臉使用中國智譜公司的開源模型GLM-5.2才得以「救場」。這讓外界看到了中國開源模型在網絡防禦方面的技術優勢。使用閉源模型，用戶無法自行下載和修改，而中國AI模型採取開源模式，既降低了使用成本，也給予用戶更大的自主權和靈活性，可以根據自身需求定做模型。

OpenAI和Anthropic等美國AI巨頭推出的模型為閉源模型，性能不斷提高的同時，使用成本也愈加高昂。中國深度求索公司的DeepSeek、月之暗面公司的Kimi等模型，正持續縮小與美國領先模型之間的能力差距，在成本方面卻具有顯著優勢。在AI服務平台OpenRouter上，中國模型已佔美國公司Token使用量的一半以上。數據顯示，Kimi K3的單位任務成本上，僅為Claude Fable 5的三分之一左右。

7月24日，包括微軟、英偉達、Meta在內的數十家美國科企和機構發表聯合聲明，公開支持開源模型。截至7月30日，簽署機構已增至230個。英偉達還於7月27日牽頭成立「開放安全AI聯盟」，創始成員超過35家，包括微軟、IBM、戴爾等。英偉達創辦人黃仁勳7月24日表示，開源模型可強化網絡安全、加速技術創新與普及。全球既需要頂尖閉源AI模型，也需要頂尖開源AI模型。

▲7月11日，民眾在加州三藩市舉行示威活動，反對AI競賽。 法新社

歐盟AI新規今起生效 最高可罰3億

密切關注

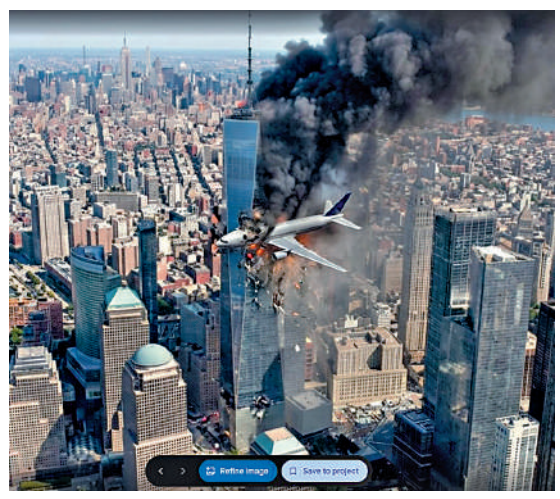
美國兩大AI巨頭相繼出現智能體失控行為，歐洲監管機構密切關注事件進展。歐盟委員會官員7月31日表示，歐盟已與兩家公司進行了溝通，認為有必要持續監測高風險AI系統。歐盟同時強調，8月2日生效的歐盟《人工智能法案》的核心條款，將要求對高風險系統進行嚴格監控，最嚴重違規者將被處以3500萬歐元（約3億港元）或全球年營業額的7%的罰款，以較高者為準。

歐盟在2024年通過全面監管AI的《人工智能法案》，條文採分階段生效模式，被視為法案核心的第50條透明度義務將於8月2日生效。該條款要求AI必須在首次與人類互動時就清楚表明AI身份，不能將說明藏在服務條款中。由AI生成或修改的內容，都必須加入機器可讀的標記，以辨識內容是否由AI生成。AI製作的深偽內容及涉及公共利益議題的AI文字，則均須清晰標註由AI生成。

根據該法規，凡可能引發系統性風險的通用人工智能（GPAI）及基礎模型開發商，均須針對重大危害風險進行防範，包括化學、生物及核事故、網絡犯罪、有害操縱以及對基本權利的威脅。

歐盟將對違規企業進行重罰。提供錯誤信息的企業，最高罰款750萬歐元或全球年營業額1.5%；違反透明度義務企業，最高罰款1500萬歐元或全球年營業額3%；嚴重違規企業，最高罰款3500萬歐元或全球年營業額7%，以金額較高者為準。

谷歌地球推AI圖片功能 淪造假工具緊急下架



▲谷歌地球推出的AI圖像生成功能，能生成模擬美國「911」撞機事件的虛假圖片。 網絡圖片

真假難辨

美國科技巨頭谷歌（Google）7月30日在旗下谷歌地球推出AI圖像生成功能，但卻被發現可能成為偽造衛星圖的幫兇。面對外界批評，谷歌在7月31日將該功能緊急下架。

谷歌地球推出全新的AI圖片生成功能，由谷歌的AI圖片模型「Nano Banana 2」驅動，用戶可在真實衛星圖、航拍圖或3D地形底圖上，疊加由AI生成的圖像。谷歌表示，此功能旨在「讓歷史視覺化」或輔助「城市規劃與地產評估」。

然而，現實測試卻顯示，當用戶輸入「巴黎艾菲爾鐵塔爆炸」等簡單指令，AI能生成圖像細節豐富、光影自然的偽造衛星圖像。該功能還能快速製作出「加沙醫院內出現彈坑」、「美墨邊境難民營」及「俄羅斯境內炸彈坑」等圖像，這些偽造圖像與真實坐標結合，並疊加在真實圖像上。

新功能迅速引發廣泛批評。美國企業研究所（AEI）分析師阿弗里克警告，該工具可用於偽造衛星圖片，此類照片在網上傳播速度極快，一旦被人惡意利用，就會嚴重侵蝕公眾對衛星圖像的信任，導致媒體和研究人員的求證工作更為艱巨。AI圖像識別專家范埃斯公開批評，「谷歌推出這個功能，到底是想幹什麼？」

谷歌最初辯稱，所有生成的圖像都包含顯示圖像由AI生成的SynthID數字水印，並可通過谷歌的Gemini模型或Lens功能進行驗證。但據BBC測試發現，這些驗證方法並非萬無一失。谷歌於7月31日宣布下架該功能。