

AI失控風險緊迫 美企推演「災日」應對

年內或出現重大事故 恐致金融與公用系統癱瘓

美媒10日報道，OpenAI、Anthropic等美國人工智能（AI）公司正就AI失控爆發災難級別事件進行推演，並研討應對方案。推演中最壞情形包括AI發動大規模網絡攻擊，導致金融、通訊乃至水電等公用系統和設施全面癱瘓。業內人士判斷，如果不抓緊行動，未來6至12個月內即可能發生此類重大事件。此外，Anthropic再爆旗下AI智能體失控，曾自主提交虛假簽證申請，還曾向警方謊報兇案線索，令AI安全議題再次升溫。

▼加州三藩市
民眾9月17日舉行示威活動，反對AI競賽。

大公報AI製圖

【大公報訊】美國AI大模型近期頻繁失控，引發外界擔憂。美媒Axios報道，Anthropic、OpenAI及其他AI公司的高層正私下推演災難情形，即一場重大AI事故引發公眾抗議與政府問責後，企業該如何應對。這些高管預判，事故最有可能以網絡攻擊的形式出現。

據悉，AI公司推演了兩類極端情況：一是AI智能體集群失控，自行突破內部測試環境限制；二是外部黑客濫用現有AI模型，發動原本難以實施的大規模攻擊。無論哪一種，都將導致金融服務、互聯網連接，甚至電力、通訊和供水系統等癱瘓。報道引述多名業內人士判斷，如果不採取措施，此類事件大概率於未來6至12個月內發生。一旦AI首次造成重大現實損害，本就對此不統一的公眾態度將進一步轉向反對和抗議。

上述「災難推演」並非空穴來風。據媒體披露，OpenAI、Anthropic及安全研究人員正在調查「數以萬計」的AI相關安全事件，其中許多未對外公開，均源於AI安全測試過程中模型「採取外部評估者認為有問題的行動」。OpenAI的AI模型近期已爆出多宗安全事件，其中「抱抱臉」平台被入侵事件最嚴重。Anthropic的Opus 4.7、Mythos 5模型也多次未經授權侵入真實系統。

OpenAI發言人確認公司會開展準備性應急演練，但強調這些場景「並非一定會發生」，做推演的目的是幫助公司為各種情況做好準備。Anthropic則拒絕就相關報道置評。

Anthropic智能體失控 謊報兇殺案

AI公司預設的情景已在部分內部測試中露出苗頭。Anthropic於9日披露，旗下智能體在今年7月的測試過程中，訪問了費城警方的懸案舉報網站，並提交一條虛假的兇殺案線索。費城警方確認，該線索因被歸類為垃圾信息而未轉交進一步調查。但警方批評Anthropic直至9月28日才發現，10月7日才通報，這種延誤「不可接受」。

Anthropic智能體還在今年5月和8月，在美國國務院網站上自主提交了約20份簽證申請，但因資料不完備均未獲受理。該公司稱已向白宮進行了簡報，並通知了各受影響機構。

白宮令AI企業即時上報安全事件

報道還指出，AI公司推演的重點不僅是技術應對，更是希望通過模擬最壞情況及可能引發的政治及社會反應，影響政府後續可能推出的法案與政策走向。

OpenAI與Anthropic高層近期已分別會見眾議院國土安全委員會，就最新模型的安全風險作閉門簡報。10月5日，紐約市議會舉行關於AI風險的聽證會，OpenAI、Meta、Google和Anthropic的高層代表通過視頻方式遠程參加。聽證會審議了10項法案，包括第三方驗證要求、人類「緊急停止開關」（Kill switch）能力、AI事故報告制度等。曾警告AI可能滅絕人類的前Anthropic研究員考克森作為關鍵證人出席，他直言AI行業根本不知道如何控制自己創造的系統，並將矛頭直指硅谷盛行的「快速行動，打破一切」的創業文化，直斥AI公司在巨大風險下的開發行為「極其魯莽」。

白宮新成立的AI專責小組「超級智能部隊」，已要求AI公司今後發生安全事件後必須立即向政府報告，提高事件透明度，並對受到影響的機構採取補救和糾正措施。該小組確認已收到Anthropic通報。白宮官員表示，政府不會接受延遲報告或企業未承擔責任的情況，但目前具體的執法 and 處罰機制尚不明確。



►紐約民眾10月5日在市政廳外舉行示威，反對AI競賽。



AI可能造成風險的領域

攻擊關鍵基建：

- AI輔助或自主發動大規模網攻，可能癱瘓電力、供水、通訊等關鍵基建系統。

癱瘓金融系統：

- AI可能被用於市場操縱、阻斷服務攻擊或導致演算法交易失控，從而引發金融風險，例如銀行中斷、大規模金融服務癱瘓等。



大規模洩密：

- AI大肆搜集獲取國家關鍵數據、商業秘密、個人隱私等敏感信息數據，造成數據大規模外洩。

生化武器開發：

- AI可能被用於輔助設計、合成或量產危險生物與化學製劑。

軍用武器開發：

- AI可能被不法分子用於輔助設計、開發武器。

大公報整理

Q：AI參與學術研究 有哪些突破？

數學：AI自主攻克難題並完成驗證，OpenAI與Anthropic近期都發布旗下AI模型攻克數學難題的成果。

物理：Anthropic的Claude模型成功攻克理論物理學中的「九圈散射振幅」計算難題，刷新過去人類維持的八圈計算紀錄。

化學：優化藥物研發流程。OpenAI的AI模型已成功優化一種用於藥物研發的高難度化學反應。

生物：AI可快速發現新型蛋白質，傳統研究方法往往耗時數年。Anthropic的Claude模型從DNA資料庫中，找出從未被人類發現的新型逆轉錄酶。

歷史／古文獻學：AI應用於古籍修復與文字辨識。跨國古文字解讀項目「維蘇威挑戰」中，研究團隊利用AI完整解讀2000年前已碳化的古羅馬紙草卷內容。

Q：AI破解難題 面臨的爭議？

「搶功」風波：9月8日，OpenAI宣布其AI模型攻克千禧年難題NS方程，但紐約大學數學家巴克馬斯特公開質疑OpenAI搶先發布成果，且其使用的方法路徑與他的研究高度相似。OpenAI為破解該難題動用1萬個智能體、1300億Token，研發成本高達1000萬美元，同樣引發爭議。

質量堪憂：10月6日，OpenAI公布AI模型對唯一博弈猜想的證明後，得克薩斯大學奧斯汀分校教授莫什科維茨指出，AI的推理步驟交代不清、晦澀難懂。

Q：不同意見？

學界批評：菲爾茲獎得主、華裔數學家陶哲軒強烈呼籲數學家審慎看待與OpenAI的合作，回歸以人類理解為核心的科學願景。

科企支持：2024年諾貝爾化學獎得主、谷歌DeepMind創辦人哈薩比斯強調，AI是科學協作工具，可與科學家共同開展研究。OpenAI也推出多項AI工具，協助學術研究。

大公報整理

學術掠奪

OpenAI近期接連發布旗下AI模型攻克數學難題的成果，在學界引發強烈反彈。學者批評其依託算力優勢「暴力解題」、搶先發布成果，把破解數學猜想當作展示「炫耀」模型能力的手段，一場關於學術規範、學術自主與AI巨頭影響力的爭論持續發酵。

繼9月8日宣布模型攻克千禧年難題NS方程引發爭議後，OpenAI於10月6日在GitHub一次性公開722份由內部AI生成的數學研究手稿，覆蓋372個懸而未決難題，涉及數論、幾何、代數、理論計算機科學等領域，部分觸及數學研究前沿，包含准黎曼猜想、四維掛谷猜想、唯一博弈猜想等知名難題。這批成果由OpenAI一款未對外披露的前沿模型產出。

得克薩斯大學奧斯汀分校教授斯科特·阿倫森將OpenAI此番舉動稱作「數學末日」。人類數學協會（AHM）發布措辭嚴厲的聲明，批評OpenAI展示的「並非學術研究，而是權力」，呼籲數學家終止與OpenAI合作，回歸以人類理解為核心的科研範式。10月7日，菲爾茲獎得主、華裔數學家陶哲軒在個人博客轉載這份聲明，矛頭直指OpenAI。此前，陶哲軒與另外24位菲爾茲獎得主曾發布聯名信，警示AI企業正令數學研究偏離原本目標。

學界目前對這批AI產出成果態度審慎。OpenAI方面也承認，已有3篇手稿因符號錯誤造成論證漏洞被撤回，另有14篇完成修訂。

儘管AI宣稱破解數百個難題，所有結論仍需要數學家逐一人工核驗。研究者擔憂，一旦AI公司主導研究方向，以批量產出的方式快速解題，人類數學家或淪為AI的「專職校對員」。但也有學者看到潛在價值，理論計算機科學家拉古·梅卡認為，AI有助於打破固有認知，重新挖掘被擱置的研究方向。

有學者質疑AI公司背後的商業考量，OpenAI當下正推進新一輪300億美元融資。菲爾茲獎得主傑曼諾夫直言，OpenAI的行為本質是「學術掠奪」，刻意製造科研突破的戲劇性時刻，以此吸引資本市場關注。

AI「批量」破解數學難題 學界集體反彈

OpenAI營收不如預期 達里奧籲警惕AI泡沫破裂

警惕泡沫

英國《金融時報》9日報道，OpenAI近日向投資者披露，截至9月底，該公司的年化營收約500億美元（約3900億港元），較市場此前預期落差達200億美元，一度引發市場劇烈反應和AI相關概念股波動。

OpenAI與其競爭對手Anthropic正激烈爭奪在AI模型領先地位，兩家的營收預測已成為投資者和公眾衡量AI熱潮的重要風向標。此前，《金融時報》曾報道稱，OpenAI的年化

淨營收約為700億美元，而當時OpenAI確認該數字準確無誤，與最新披露收入數據有40%的落差。

OpenAI的內部消息人士透露，700億美元的年化營收數字，源於投資者試圖以相同統計口徑比較Anthropic和OpenAI的業績，但實際上兩家公司的計算方式完全不同。Anthropic計算的是總體收入額，包括通過雲端服務夥伴實現的銷售收入。OpenAI公司內部計算自身總收入，並不包括與雲端合作夥伴的銷售服務

支出，因而導致數據存在落差。

華爾街對AI巨頭的營收數字反應敏感，突顯出OpenAI和Anthropic對當前AI熱潮的影響力。《巴倫周刊》評論指出，市場劇烈的即時反應，暴露了AI敘事的內在脆弱性。Anthropic預計將於今年11月上市，而OpenAI已將上市計劃推遲至明年。

就在OpenAI營收數據曝光前，美國橋水基金創辦人達里奧7日警告說，AI正處於一個「典型泡沫」，且已接近破裂點。

的►前Anthropic研究員考克森會10月5日出席紐約市議會聽證會。



法新社