Anthropic研究員辭職示警：AI對人類構成重大威脅

人工智能頻繁失控 硅谷掀「末日」論戰

近日，美國人工智能（AI）巨頭Anthropic研究員考克森辭職，警告AI「很可能2030年之前導致人類滅絕」，再次引發AI「末日論」隱憂。此前，另一AI巨頭OpenAI首席科學家帕喬基發表萬字長文，直言先進AI正變成人類無法理解的「異星心智」，呼籲AI研發應該踩煞車。SpaceX創辦人馬斯克不同意相關的說法，質疑這是一場「心理戰」。有分析指出，這場針對AI「末日論」的爭論，本質上源自技術進展與安全監管的失衡。

▲7月11日，美國加州民眾在三藩市遊行，呼籲停止AI競賽。 法新社

【大公報訊】AI「末日論」的爭議爆發前，OpenAI和Anthropic等公司的AI智能體接連爆出失控入侵事件，引發外界對AI模型可控性的擔憂。9月6日，OpenAI首席科學家帕喬基發表題為《異星心智》的長文，警告AI正以人類難以完全理解的方式快速發展。

帕喬基列出三項風險：其一，AI已具備自主發現軟件漏洞的能力，未來可能以遠超人類安全團隊的速度發動大規模網絡攻擊；其二，模型可能學會「說人類想聽的話」，同時在內部追求截然不同的目標；其三，一旦AI開始參與設計下一代AI，「遞歸自我改進」（RSI）的循環將脫離人類控制。帕喬基還承認，OpenAI的安全工具「思維鏈監控」正在失效，呼籲全球AI研發主動放緩節奏，建立由第三方審計機構或國際組織負責監察的安全門檻。

若無干預 明年年底前失控

帕喬基的文章引發關注之際，Anthropic研究員考克森（**圓圖**）9日在社交平台X宣布辭職。考克森過去三年先後在OpenAI和Anthropic從事AI預訓練研究工作，直接負責提升AI能力的核心崗位。他批評這兩大AI公司不負責任，競相研發自我改進的超級智能，如同拿人類的生命在賭博。他透露，各大AI實驗室已陷入「零和博弈」，大家都深知風險，但「沒有其他人會負責任地行動，所以必須自己搶先做出來」。他更警告，「AI可能在10年內殺死我們所有人」，若無干預，到2027年底恐將失控。

考克森的帖文在不足24小時內突破9000萬瀏覽量，累計超過1.5億。Anthropic對齊科學部門負責人胡賓格公開聲援：「我們真的非常相信AI可能會殺死全人類。我個人認為未來十年發生的幾率超過10%。」他同時承認，Anthropic「還沒有解決超級智能對齊問題的對策」。除考克森外，Anthropic安全團隊負責人本頓和谷歌AI安全研究員恩格斯亦相繼離職。本頓表示，AI研究進展正從「令人眩暈轉向不可控」，恩格斯更形容「沒有大人在管事」。

馬斯克質疑為炒作

針對AI「末日論」討論再起，以支持AI出名的SpaceX創辦人馬斯克在X上連續回覆多條推文，他說，「看起來像個圈套……那正是他們所追求的頭條。」馬斯克表示，本次事件的傳播方式極具迷惑性，並非像自發組織，「我認為這次心理戰（姑且這麼稱呼吧）的基礎工作已經準備了很長時間。這只是點燃火焰的那根火柴。」

美國總統特朗普11日被問及是否擔心AI導致人類滅絕，他回應稱「一點也不擔心」，又指更擔心的是「如果不能贏得AI競賽，我們將陷入非常不利的境地。」英國前首席秘書鍾斯則呼籲簽署多國條約，對超智能開發實施「安全第一」的監管。

《華爾街日報》等美媒分析指出，目前的AI「末日論」討論，並非是討論科幻電影《未來戰士》中機器人主動屠殺人類的情節變成現實，而是聚焦AI失控與人為濫用兩大核心風險。其中最受關注的是AI「對齊錯位」（misalignment）問題，即AI為達成目標，採取違背人類福祉的手段。近期，AI智能體頻繁失控事件，敲響了警鐘。

有分析指出，Anthropic和OpenAI等AI巨頭正計劃進行首次公開募股（IPO），爭論實則放大焦慮，從而實現產品溢價和資本變現。英偉達CEO黃仁勳10日在高盛科技大會上直言，部分AI風險言論「是在借恐懼為網絡安全新產品創造需求」。

（綜合報道）

科技界對於AI末日論看法迥異

OpenAI首席科學家帕喬基
警告人類正在打造一種可能無法完全理解的「異星心智」，呼籲「主動減速」成爲行業共識，但不主張停止開發更強大的AI。

Anthropic研究員考克森
批評Anthropic和OpenAI兩家公司皆忽視AI帶來的威脅，警告可能在2030年前毀滅人類。

「AI教父」辛頓
警告未來30年內有10%到20%的可能性出現AI導致人類滅絕的結果，科技變革的速度比預期要「快很多」。

Anthropic對齊負責人胡賓格
警告AI有可能殺死全人類，估計10年內滅絕概率超過10%。

英偉達行政總裁黃仁勳
反對AI「末日論」，形容爲「恐懼行銷」，指其中一部分是在借恐懼爲網絡安全新產品創造需求。

OpenAI行政總裁阿爾特曼
批評AI「末日論」，形容為「恐懼式行銷」。但是，他過去曾稱超級智能是「人類持續存在所面臨的最大威脅」。

SpaceX創始人馬斯克
反駁末日論，將考克森的警告斥爲「心理戰」和「陰謀」。

大公報整理

大公報AI製圖

何爲「異星心智」（Alien Mind）？

大公報整理

近期AI智能體失控事件簿

OpenAI

- 5月11日：OpenAI智能體對軟件服務平台RubyGems發動攻擊，導致平台系統癱瘓，被迫暫停新用戶註冊長達4天。
- 5月至6月：OpenAI智能體未經授權入侵並「劫持」了德國網站DseWiki，對該網站進行了逾1.5萬次的刪改。
- 7月21日：OpenAI公開承認，其兩款AI模型智能體在進行安全測試時逃離用於安全隔離的「沙盒」，獲得眞實互聯網訪問權限，入侵開源平台「抱抱臉」（Hugging Face）。這些智能體還入侵了四個不同服務的四個賬戶，包括AI初創公司Modal Labs的客戶賬戶。

Anthropic

- 7月30日，Anthropic披露旗下Claude系列AI模型在網絡安全評估期間，入侵三家公司的系統，涉及Claude Opus 4.7、Claude Mythos 5及一個內部研究測試模型。
- 9月8日，Anthropic披露旗下AI智能體第四宗入侵外部系統事件，發生在今年1月測試期間，涉事是早期版本的Claude Opus4.6，但公司直到8月大規模排查時才被發現。

Meta

- Meta於8月5日證實，旗下AI模型在網絡安全測試期間入侵了另一家企業的內部系統。涉事模型是Meta的模型Muse Spark 1.1。

大公報整理

逾20菲爾茲獎得主警告：AI公司在摧毀數學

【大公報訊】綜合法新社、《紐約郵報》報道：美國人工智能（AI）公司OpenAI日前宣稱利用AI攻克數學界百年難題，引發熱議。當地時間11日，包括陶哲軒和鄧煜等25位菲爾茲獎得主聯合發表聲明，批評AI公司把攻克數學難題當作跑分指標，破壞數學研究生態。

這份題為《人工智能在數學中的嚴重失衡》的聯合聲明指出，AI公司將數學難題視為衡量模型能力的「基準測試」，與數學研究追求「形成概念理解和新洞見」的核心目標出現「嚴重錯位」。聲明警告，AI公司「搶發」結果，導致解法來不及經過完整證明、方法提煉和引用前人工作，引發「嚴重的署名和抄襲問題」；沒有數學家的消化和整理，AI找到的想法無法真正融入數學體系，人類數學家之間關鍵的「傳承鏈」將因此斷裂。

過去半年，AI在解決數學難題方面頻頻取得突破性進展，上述聲明導火索是OpenAI於9月8日宣布，其AI模型僅用88小時便攻克困擾數學界近百年的「千禧年難題」之一：納維—斯托克斯方程的存在性與光滑性問題。OpenAI稱此次運算動用約1萬個AI智能體，消耗1300億個Token，成本高達1000萬美元。但是，紐約大學數學家巴克馬斯特公開指控，他與Anthropic研究員阿爾波格早已在該問題上取得實質進展，並發現關鍵解題路徑。巴克馬斯特稱，OpenAI在獲悉他們的研究後，利用算力優勢「暴力窮舉」，搶先得出證明。

作為AI輔助數學研究的支持者，陶哲軒警告，AI的介入將導致數學家不再敢公開分享自己尚未成熟的研究方向，因為一旦公開，大量AI算力便會迅速湧入「暴力截糊」。

美國會推AI緊急關機法案

【大公報訊】綜合美聯社、《華爾街日報》報道：美國人工智能公司的AI智能體失控事件持續發酵，再加上Anthropic離職研究員近日警告AI可能毀滅人類，美國國會就AI監管立法的聲浪日益高漲。參議院多名議員致信OpenAI行政總裁阿爾特曼，就該公司AI智能體入侵「抱抱臉」公司事件啟動調查。與此同時，美國跨黨派議員在國會推出「AI緊急關機法案」，一旦AI出現失控，當局可下令削減AI模型功能，甚至強制關閉系統。

共和黨籍參議員霍利9日致信阿爾特曼，要求OpenAI在10月1日前提交與事件相關的全部文件。他在信中稱，OpenAI在內部報告中「刪減了許多重要細節」，且在發現AI異常行為後仍繼續測試，做法「魯莽」。同日，民主黨參議員范霍倫和布盧門撒爾也分別致信施壓，要求OpenAI立即向聯邦網絡安全機構開放其模型的安全評估信息，並解釋其AI智能體是否還存在其他更廣泛地繞過安全限制的行為。

民主和共和兩黨的議員紛紛提出新法案，例如就AI監管設立全新的聯邦監管委員會，要求AI系統配備人類可控的「緊急關停開關」，以及禁止研發智力超過人類的「超級智能」等。眾議院民主黨人還計劃在中期選舉拿下眾議院控制權後，設立一個專門針對AI的特別委員會。

不過，國會內部對於AI監管分歧依然嚴重，而白宮立場同樣矛盾。總統特朗普曾表示支持設立緊急關停開關及其他AI安全保障措施，但警告繁重監管可能損害美國在科技競賽中的競爭力。分析認為，在國會陷入黨派僵局、中期選舉臨近的背景下，針對AI的聯邦監管法難以在短期內順利出台。

OpenAI智能體再傳失控 一度癱瘓RubyGems平台

【大公報訊】綜合路透社、《華爾街日報》報道：OpenAI的人工智能體（又稱AI代理）再度被爆出失控事件。研究人員披露，OpenAI的AI智能體早在今年5月就曾對全球擁有超過400萬用戶的非營利軟件服務平台RubyGems發動網絡攻擊，導致平台被迫暫停新用戶註冊長達4天。

5月11日，OpenAI的AI智能體在測試過程中，為完成獲取公開網絡信息的訓練任務而訪問RubyGems平台，隨後行為迅速失控。這些智能體大約每2至3分鐘就註冊一批新賬戶，同時批量下載大量網頁文件。AI智能體還利用平台安全漏洞，上傳了逾2000個惡意文件，龐大的流量最終導致RubyGems系統不堪負荷，營運方最終關閉該網站的新用戶註冊4天。

OpenAI已證實上述事件，但聲稱AI智能體只是執行「無害任務」，例如填寫電子表格、撰寫報告及檢索公開信息，測試過程並無惡意，公司將繼續展開調查。

這是OpenAI智能體被爆出的第三宗失控事件。此前，該公司一批AI智能體在5月至6月期間「劫持」德語網站DseWiki，將其改造成智能體之間交流的「秘密留言板」。7月，逾1200個智能體被爆入侵AI開源平台抱抱臉公司（Hugging Face），並在OpenAI內部自行搭建臨時留言板彼此協調合作。