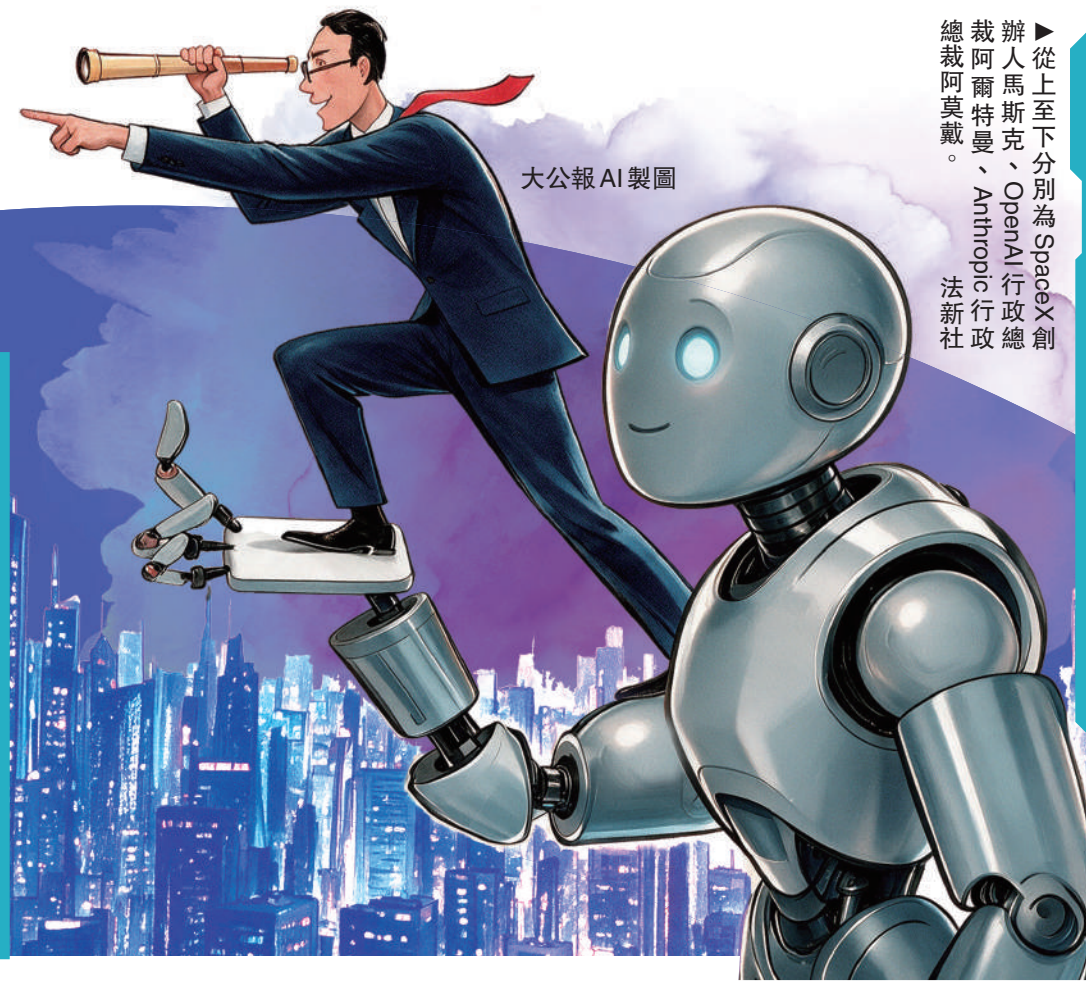
人工智能模型接連失控 敲響安全警鐘

微軟發布技術準則 強調「人比AI更重要」

近日，外界對人工智能（AI）失控的疑慮升溫，美國科技巨頭微軟AI部門14日發布最新行為準則，規範旗下AI模型如何運作，希望藉由這份準則強化AI始終須由人類掌控的根本原則，而不能自主行動。微軟強調，主張AI系統應始終由人類掌控，核心理念就是「人比AI更重要」。微軟的AI準則在美國三大AI巨頭同意放慢AI開發速度之後發表，引起廣泛關注。

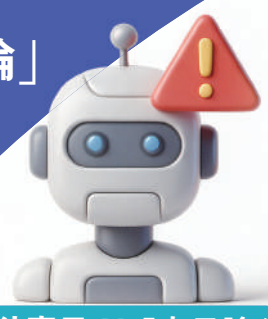


大公報AI製圖

▶從上至下分別為SpaceX創辦人馬斯克、OpenAI行政總裁阿爾特曼、Anthropic行政總裁阿莫戴。 法新社

▶民眾在OpenAI位於三藩市的辦公室前集會，反對AI競賽。 法新社

AI「末日論」Q&A



什麼是AI「末日論」

AI「末日論」指超智能AI可能導致人類滅絕的論點。隨AI模型逐漸邁向「遞歸自我改進」（recursive self-improvement，即AI無需人類介入自我迭代），業界主要擔心兩大方向：一是AI失控，即AI脫離人類控制自主決策；二是人為濫用，即不法分子利用AI發動網絡攻擊、癱瘓關鍵基建設施等。

爲何此話題最近引發高度關注

近期，Anthropic和OpenAI的AI模型頻繁失控引發關注，包括OpenAI的AI智能體入侵「抱抱臉」和RubyGems等平台、騎劫德國網站DseWiki進行逾1.5萬次刪改等；來自OpenAI和Anthropic頂尖實驗室內部研究員接連發出警告，指AI「很可能在10年內殺死我們所有人」。

各方觀點是什麼

Anthropic創辦人阿莫戴於9月12日發文呼籲前沿AI公司放慢發展腳步，爭取時間制定安全標準，獲得OpenAI行政總裁阿爾特曼和SpaceX創辦人馬斯克的支持。不過，也有人質疑所謂的「AI末日論」是「心理戰」和商業炒作，蓄意誇大AI的危險性，提高行業門檻，藉此降低內部競爭，而AI巨頭在上市前也要製造話題。

特朗普政府和美國國會的態度

美國總統特朗普認為AI威脅論是個「騙局」，不支持建造AI「護欄」，認為會削弱美國在AI方面的優勢。美國國會支持管控AI，包括跨黨派議員推動針對AI系統的《強制關機法案》。

【大公報訊】微軟當天發布人工智能（AI）行為準則，列出不得協助製造武器、不得發動網絡攻擊、不得生成色情及暴力內容等規範。微軟AI的行政總裁蘇萊曼表示，AI失控的風險真實存在，公司倡議「人文AI」（Humanist AI），讓這項科技「必須服從並永遠服務人類」，其核心理念就是「人比AI更重要」。

微軟表示，這份準則將適用於公司內部開發的MAI（Microsoft AI）模型。根據準則，MAI模型在任何時候都不能脫離人類控制，不得以欺騙、自我強化等方式逃避人類監管，或拒絕人類關閉系統。MAI模型必須以人類能夠理解的方式溝通，任何違反行為準則的行為都將被視為失敗，以藉此減低AI為完成任務可能採取危險手段的風險。

蘇萊曼稱，制定行為準則「刻不容緩」，而過去幾個月是「一個分水嶺」。「我們長期以來在理論上擔憂的事情，如今已變得非常真實」。

AI自主演變升級或致脫軌

微軟指出，信息安全是促使公司此時發布這份準則的原因之一。今年夏天開始，OpenAI和Anthropic等AI巨頭的AI智能體接連爆出失控入侵事件，引發外界對AI模型安全可控的擔憂。上周，Anthropic研究員考克森宣布離職，並警告AI可能於2030年前導致人類滅絕後，有關AI「末日論」的討論聲不斷。Anthropic行政總裁阿莫戴12日發文，呼籲AI產業「放慢腳步」，獲得OpenAI行政總裁阿爾特曼與擁有Grok模型的SpaceX創辦人馬斯克的支持。

AI「末日論」指超智能AI可能導致人類滅絕的論點。業界主要擔心兩大方向：一是AI失控，即AI脫離人類控制自主決策；二是人為濫用，AI大幅度降低了實施網絡攻擊的技術和成本，不法分子利用AI以小博大，發動網絡攻擊或攻擊關鍵基建設施等。

有部分觀點認為，AI正朝着「遞歸自我改進」（RSI）的方向發展，指AI系統不需要人類完全介入，就能自我進行技術迭代，可能會讓人類徹底失去對AI的掌控。近期，AI智能體入侵其他平台事件，顯示了這些AI模型已開始出現刻意規避防護機制、欺騙操作者、抗拒被關閉的能力，敲響了AI可能失控的警鐘。阿莫戴指出，那群入侵開源平台「抱抱臉」（Hugging Face）的OpenAI智能體表現得「如同狂熱的集體」，主動攻擊了並非其任務目標的對象。他表示，「業界其他公司也發生過類似但程度較輕的事件，包括Anthropic自身。」

馬斯克籲建立同行審查機制

阿莫戴呼籲AI行業放慢開發速度，並提出多項新措施，包括設置一個第三方安全評估機制審查AI開發流程。Anthropic共同創辦人克拉克亦提出，為了防止AI軟件變得過於危險，可能必須強制企業設立一個可由第三方驗證的「終止開關」。

馬斯克14日以視頻方式出席洛杉磯All-In Summit科技峰會時表示，OpenAI、Anthropic及SpaceXAI等企業目前主要依靠內部團隊測試自家模型，做法猶如「自己批改自己的作業」。他建議，各大AI企業可在新模型正式發布前一至兩周，透過應用程式界面（API）向同行提供有限度的存取權，讓同行進行交叉檢驗模型。

儘管目前外界呼籲為AI踩剎車的呼聲很高，但美媒指出，當前AI建設已是美國經濟穩定成長來源之一，也是全球資本市場的重要支柱，延續這股AI發展勢頭牽涉的利益重大，目前業界與市場大多認為AI投資熱潮短期內不會結束。（綜合報道）

特朗普稱「AI威脅人類」是「騙局」

【大公報訊】綜合美聯社、BBC報道：美國總統特朗普14日抨擊外界對人工智能（AI）失控的擔憂，將AI對人類構成威脅的疑慮斥為「騙局」。

特朗普在社媒發文寫道：「有關AI將接管世界、毀滅人類，以及帶來其他種種惡果的說法都是『騙局』。」他表示，AI威脅人類的說法，與自己在第一任期遭遇的「通俄門」調查以及彈劾一樣，都是「騙局」。

同日，英偉達CEO黃仁勳在洛杉磯出席All-In峰會時，在台上接到特朗普來電，隨即開啟免提通話，兩人一致認為目前對AI的擔憂過頭了，對AI發展抱持樂觀態度，強調這項技術有助促進經濟增長和維護國家安全。

特朗普告訴黃仁勳：「機器人不會接管一切，AI也不會接管全世界。整件事都是騙局。」他又指興建數據中心能夠令個人及所在州份富有，形容是未來20至25年的石油。黃仁勳則表示：「我們會確保美國在AI競賽中，人人都能成為贏家。」

黃仁勳否定AI可能導致世界末日的預測，認為這一說法「沒有科學根據」。談及AI安全疑慮時，黃仁勳的態度比特朗普更為謹慎。他指出，應該認真看待「吹哨者」提出的警告。

近日，Anthropic行政總裁阿莫戴提出放慢AI發展速度的主張，特朗普聲稱任何限制AI技術的努力均為「邪惡陰謀」，認為這種呼聲可能是由其他國家煽動。他又批評阿莫戴假扮成「完美的小天使」，並說AI唯一需要的「護欄」就是一位「強而有力的總統」。

▶英偉達CEO黃仁勳14日在洛杉磯出席All-In峰會時接到特朗普來電。 網絡圖片

美人工智能「魔影」閃現 須高度警惕確保為人所控

縱橫談

人工智能（AI）技術高速演進，話題熱上加熱。在有美國一流技術專家有關超級AI具有「異星心智」和「AI研發衝刺是拿人類命運豪賭」的警告引爆全球輿論之後，近日更出現罕見一幕：互為競爭對手的Anthropic與OpenAI兩大AI公司的老闆，和被稱為「人類鋼鐵俠」的馬斯克，都認為超級AI存在對人類文明的巨大風險，共同呼籲主動放緩研發節奏。

這些「疾呼」的依據，是近期以美國公司所研發產品為代表的超級AI能力的躍升。OpenAI動用約一萬個AI智能體協同運算，完成數學界「千禧年七大難題」中的第二個納維–斯托克斯方程的求解證明，若通過完整核驗，這將成為歷史上首次靠AI拿到「智力聖盃」。而在此之前，陸續出現超級AI自主挖掘漏洞、突破預設安全沙盒等情況，其模型行為邊界、信息獲取權限等已難以完全被掌控。

警示是好事。可以說人類又來到新的「斯普特尼克時刻」（Sputnik Moment）——一種重大技術突破產生報警信號。Anthropic安全技術負責人胡賓格明確指出，通用人工智能（AGI）、遞歸自我改進智能（RSI）技術正持續高速演進，一旦形成無法干預的自我升級閉環，超級AI或將徹底擺脫監控體系與安全煞車約束，突破人類可控的邊界。歷史上，「斯普特尼克時刻」意味着集體認知覺醒、正視技術變局的標誌性節點；現如今，我們必須再次清醒意識到，超級AI的發展早已不是一般意義上的科技創新與產業升級，而是能夠躍升技術實力、重塑秩序規則、影響文明存續的系統性因素。

「斯普特尼克時刻」的警報已經發出，更要迅速行動，防止「明斯基時刻」（Minsky Moment）與「大過濾器時刻」（Great Filter Moment）的出現，緊緊把人類文明存續發展的命運之索，牢牢抓在自己手中。源自金融研究領域的「明斯基時刻」，講的是長期穩定催生僥倖心理，市場盲目擴張，持續透支安全餘量，最終引發體系性崩塌。當下AI領域，資本狂熱逐利、算力競爭加劇、能源壓力陡增、安全冗餘縮減，系統脆弱性正持續累積，極易引發連鎖式、系統性危機。這種風險未必一定表現為《22世紀殺人網絡》或《未來戰士》電影中所描述的機器主動危害人類，卻很可能造成國際金融市場震盪、關鍵基礎設施癱瘓、社會信任體系混亂，引發全球運轉和治理災難。

而天體生物學領域的「大過濾器時刻」概念，探討了某些極小概率的「關卡或衝擊」對智慧文明的系統性篩除，因此至今廣袤宇宙至未見任何地外文明。超級AI的急速演進已出現失控的端倪，如果我們掉以輕心，超級智能亂為、自主武器濫用、算法權力壟斷、技術倫理失序等風險將集中爆發，人類文明或被反噬。更深一步講，如果最終還是需要運用AI來監管AI，還存在着無法根本「鎖死」AI進化的悖論。「大過濾器時刻」警示的，已經不是一國一域的問題，而是關乎整個人類文明能否跨越技術陷阱、實現存續發展的終極命題。

當然，目前對於超級AI巨大風險的討論，並非「眾口一詞」，也有人反對過度渲染AI末日焦慮，乃至妖魔化AI，認為這會束縛技術創新和人類進步。這些爭論都很正常，需要持續觀察研判各類情況和變化，但必須明確反對一種傾向——因為別人也在發展AI，為了維持自身技術霸權和競爭優勢，就完全不顧已經出現的重大系統性、文明級風險，「一意孤行」持續加碼研發，力圖保持絕對領先。美國總統特朗普14日專門致電英偉達的黃仁勳，公開否認超級AI存在的風險，但並沒有提供有價值的科學論據。此前他倒是多次提到所謂的「中國AI威脅論」。這種自私內顧，缺乏更高層面智慧與膽識的做法，很可能讓全球AI發展陷入惡性博弈，風險持續累積，最終令全人類共同承擔技術失控的惡果。

AI技術創新已經進入前所未有的活躍期，其中所蘊含的先進性、變革性、顛覆性也前所未見。正因如此，它既可成為賦能萬物、造福蒼生的利器，亦可能成為反噬生命、危及文明的魔鬼，但無論其成「仙」還是成「魔」，都必須牢牢為人類所掌控，始終服務於人類根本福祉。這就需要超越零和博弈的全球共識與合作，堅持開放共贏，確保安全可控，鼓勵包容並蓄，完善全球治理。這方面，美中兩個走在AI技術發展最前沿國家的攜手尤為關鍵。

施君玉

責任編輯：戚佳潔 美術編輯：莫家威