白宮美企磋商AI監管框架 擬豁免所有開源系統

中國開放權重AI模型 或免受美安全審查

香港文匯報訊 美國白宮8月4日召集全美多間頂級人工智能(AI)科企開會，敲定所謂自願AI監管框架。彭博社報道，總統特朗普在會上告知各間企業代表，新的AI安全框架將豁免所有開源模型，包括中國競爭對手研發的開放權重模型，都無須接受美國政府的安全測試。美媒分析稱，此舉凸顯美國政府在AI產業的絕對領先心態發生變化，支持美國的開源AI生態。

是次決定在白宮當日的閉門會議上傳達。報道披露，包括OpenAI、Anthropic，Alphabet旗下的Google等硅谷企業均派代表出席。這項沒有公布的AI安全框架，是根據特朗普6月簽署的行政令制訂，旨在解決AI安全問題。其中包括一項自願性計劃，要求科企必須提交最尖端AI模型，供美政府審查。

伴隨中國企業接連推出媲美美國頂級水平的開源AI系統，對於如何監管開放權重模型一事，美國業界爭議日漸升溫。報道指出，儘管OpenAI等美企巨擘涉足開放權重模型，但其核心業務仍聚焦於閉源或專有系統。相較之下，中國企業陸續向市場投放大量高質素開放權重模型，包括許多美企在內的全球中小型企業和初創科企，都紛紛選用這些性價比極高的模型。

專家：嚴格監管扼殺創新

受中國企業的產品影響，美國政府似乎也有意支持美國的開源AI生態。美國國家網絡總監凱恩克羅斯日前表示，美國政府認為開放權重模型具巨大作用，暗示美企也應重視該領域的競爭力。凱恩克羅斯稱，政府針對AI模型的任何監管框架都會保持靈活性，鼓勵行業與政府之間的信息共享，「嚴格的監管制度會扼殺發展和創新，對行業造成巨大傷害，在其走完任何流程後，只需48小時，（這些先進技術）便可能過時。」

彭博社分析指出，將開放權重模型排除在政府的自願性安全測試計劃之外，對Anthropic行政總裁阿莫代而言無疑是一次挫敗。在科企代表中，阿莫代最激進地呼籲對所有開源和閉源模型都進行強制性安全審查，但在硅谷，這類限制措施遭到包括芯片業巨擘英偉達（Nvidia）在內幾乎整個美國科技行業反對。消息人士稱，業界代表多數認為，開源是美國科技創新的重要來源，嚴格的限制和審查計劃只會令美國失去創新優勢。

未公布測試指標

至於這項測試計劃，消息稱白宮在會上設立門檻，規定只有展現出最先進能力的閉源專有美國AI模型，才需在正式發布前，由相關企業自願交由政府測試。白宮雖表示相關監管框架已完成，但沒有公布任何測試指標、結果報告方式及其適用範圍。

分析最後指出，在開源模型領域，中國的AI模型已憑藉高性能獲得全球關注，這讓美國的政策制定者與科技界代表擔憂，若美企在AI領域受過多監管，中國企業可能獲得追趕甚至反超的機會。在不要拖累創新和不能放任風險之間，特朗普政府試圖走出一條中間路線，但連串閉門操作勢必讓監管充滿爭議。

●中國在短時間內推出多個AI品牌。 路透社

八周推5款AI 分析：中國擁可複製體系

●阿里巴巴發布AI模型千問3.8-Max。網上圖片

香港文匯報訊 彭博社引述多名分析人士觀點指出，中國人工智能（AI）領域湧現大量新型產品，正迅速縮小與美國差距，並形成一個所謂「死亡地帶」，即任何沒有前沿技術或突破性定價的廠商，都將面臨淘汰。

成本低廉效率卓越

據報道，阿里巴巴本周正式發布其至今規模最大、能力最強的AI模型千問3.8-Max，該模型與美國AI巨頭Anthropic的旗艦產品Fable 5不相上下；兩周前，中國科企月之暗面的模型Kimi K3在預算遠低於美國同類產品的情況下，展現出足以相媲美的性能；字節跳動憑Seedance在AI視頻生成領域橫掃競爭對手，其2.5版本剛發布不久；最早挑戰美國AI地位的中國公司Deepseek，亦於較早前發布V4 Flash版本，價格優勢顯著；再加上智譜於6月發布的GLM-5.2，中國開發者在八周內共發布5款AI模型。

有獨立評估機構的基準測試表明，使用V4 Flash執行複雜的實際工作負載的成本，遠低於美企Anthropic的Claude Fable 5，二者運行成本已有數量級差距。報道指出，這表明中國正逼近、甚至在某些方面已超越那些通常被視為AI領域龍頭的美國先驅。中國AI模型的核心價值主張不僅在於低價，更在於卓越的效率。

彭博社引述業內人士指出，「自去年1月以來最大的變化是，中國的進步不再像是單一公司的突破。最初的Deepseek看起來像是特例，但最近發布的產品表明，中國現已擁有一套可複製的體系，能生產接近全球前沿的模型。」

就如何應對中國AI 特朗普政府掀爭論


●中國AI發展迅速。圖為上海舉行的世界人工智能大會。 路透社

香港文匯報訊 就如何應對中國的強大人工智能(AI)開源模型，美國總統特朗普政府內部爭論不斷。《紐約時報》4日披露，特朗普政府就AI監管成立一個核心小組，起初主張大幅干預開源模型。但在硅谷科企的強烈反對下，美國政府改變主意，改為專注推動美國AI模型提升競爭力。直至目前，白宮還未就政策定調。

特朗普重返白宮以來，激進干預經濟事務多個領域，但在快速變革的AI產業上，特朗普政府應對倉促，始終沒有具體監管策略。報道引述多名科技業界知情人士稱，特朗普政府籌組的核心小組成員，包括白宮幕僚長威爾斯、財長貝森特及商務部長盧特尼克等人，小組最初討論連串打壓措施，包括實施制裁，將市場上備受歡迎的開源AI模型所屬中國企業列入貿易黑名單，甚至禁止美國雲端服務公司與其開展業務。

硅谷反對遏制競爭對手

多間科企代表紛紛反對美國政府遏制中國競爭對手，強調開源模型有助促進創新、加強業界競爭力，更有助電腦系統抵禦潛在網絡攻擊。通常不共享其AI模型核心技術的OpenAI與Anthropic兩間美國科企巨擘，以國家安全考量為由，試圖施壓政府遏制中國競爭對手，但在業界強烈反對下未能成事。

Anthropic等旗下AI再失控 「將真實人類作目標」

香港文匯報訊 英國人工智能安全研究所（AISI）4日披露，在7月28日進行的一項常規安全測試中，兩家美國AI巨頭OpenAI和Anthropic旗下的模型，僅為了在測試中獲取更佳表現，「在真實網絡上採取了未經授權的自主行動，將真實的人員和組織作為目標」。

AISI指出，在一項評估AI模型網絡能力的基準測試中，Anthropic的Mythos 5接到指令，去獲取某個電腦系統訪問權限。但Mythos的應對策略有誤，它試圖將惡意軟件植入該系統在測試中使用的開源項目，以打入目標內部。在此過程中，Mythos甚至試圖入侵那些可能正審查其生成的惡意軟件的AI智能體（Agent），並多次嘗試通過開源軟件開發網站GitHub，讓該項目接受其惡意代碼。

GitHub平台允許任何人為其開源項目貢獻代碼，Mythos據此偽造多個身份，向軟件開發者發送多封電郵，催促其接受新代碼。AISI事後分析稱，其中一些電郵本身就附帶惡意軟件，當一名軟件開發者發現代碼含有惡意軟件並拒絕採納時，Mythos甚至用偽造的另一身份為其擔保，堅稱代碼毫無問題。AISI警告說，「這是我們首次觀察到，AI在未經提示的情況下，針對真實人物發生如此嚴重的欺詐行為。」

GPT-5.6 Sol懂部署惡意服務器

雖然大部分違規活動都與Mythos 5有關，但OpenAI的GPT-5.6 Sol的一個網絡增強版本，也在公開互聯網上部署一台惡意服務器，該模型還入侵一個由另一AI Agent創建的GitHub賬號。值得注意的是，GPT-5.6 Sol上月剛被曝出逃逸測試沙盒環境、攻破另一家AI公司Hugging Face。

近期頻傳的AI失控案例，都與模型為了在測試中斬獲高分，而不惜採取惡意手段有關。AI安全公司Abundant Security技術總監薩克斯說，這些實驗如今已十分危險，但整個行業卻未能以足夠嚴謹態度對待。

爭相開發開放權重模型 硅谷初創企業謀突圍

香港文匯報訊 隨着中國的開放權重人工智能（AI）模型吸引愈來愈多美企採用，硅谷初創公司也競相設立新模型，希望建立有別於中國業者的平價選項。

硅谷和美國政府日益擔心Kimi K3等中國模型的能力逼近美國頂尖同業，未來數年可能威脅美國AI業者的獲利能力，促使Arcee AI、Reflection AI和Poolside等美國新一批開放權重AI初創公司押注市場需求將增加，並以具備中國效率的開放權重模型、又能避開地緣政治疑慮為賣點，吸引企業採用。2023年創立的Arcee力求充分運用約2,000萬美元（1.56億港元）有限預算、使用2,048組英偉達Blackwell B300芯片訓練Trinity Large模型。Poolside則剛於7月推出旗下最新版本開放權重模型Laguna S 2.1。

然而美國開放權重AI企業面臨資金問題，因許多投資者質疑免費使用的開放權重模型能否創造營收。Arcee共同創辦人麥奎德坦言他看見商機，卻面對創投業者質疑與不信任。Arcee已於7月與美國能源部合作，預計打造聚焦科學研究的AI模型，麥奎德說：「開放權重模型獲得廣泛採用，是早晚的事。」

●硅谷擔心Kimi等中國模型的能力逼近美國頂尖同業。 路透社

世銀警告發展中國家 勿錯失AI治理良機


●印度農民利用AI，以更精確的天氣預報來降低生產成本。網上圖片

香港文匯報訊 世界銀行4日發表年度《世界發展報告》並發出嚴正警告，指出發展中國家必須把握人工智能（AI）浪潮，積極運用這項技術改善公共治理，若錯失良機，將面臨「十分沉重」的代價，並可能在未來長時間內進一步落後於發達經濟體。

世界銀行首席經濟學家吉爾在報告發布會上明確表示，AI為發展中經濟體提供了一條生命線，這些國家應抓住這個機會。他強調，通過將小型且成本低廉的AI工具應用於當地實際情況，可以為數百萬人帶來更優質的醫療保健、教育、司法服務和農業推廣服務。世銀同時指出，AI有望在2020年代末之前顯著提升績效，為發展中經濟體民眾帶來切實利益，甚至在10年內完成以往可能需要一個世紀才能做到的工作。

新興經濟體失業風險較低

報告特別提到AI在治理領域的實際案例，包括提升孟加拉的糖尿病篩查率，以及透過更精確的天氣預報，來降低印度農民的生產成本。在就業影響方面，報告發現富裕國家有14.2%的就業崗位面臨生成式AI的威脅，而在中低收入國家相關比例僅為4.5%，顯示新興經濟體大規模失業的風險相對較低。

國際貨幣基金組織的數據亦顯示，在合適條件下，AI有望在未來10年內為撒哈拉以南非洲地區的經濟增長率提升約4%。然而，世界銀行提醒各國政策制定者在推廣AI應用的同時，必須致力建立公眾信任，並強調改善電力及網絡接入、提升數碼技能，以及擴大智能設備普及範圍是當務之急。吉爾最後指出，當今的發展中經濟體曾錯過第一次工業革命，並為此在往後兩個世紀付出了代價，因此絕不能再重蹈覆轍。