AI規管專題系列

Grok宣揚納粹自比「機械希特勒」 Gemini將美國開國元勳生成黑人


●馬斯克支持德國極右政黨，旗下Grok宣揚納粹被指充滿馬斯克風格。 網上圖片

香港文匯報訊 近年來有關人工智能（AI）偏見的問題，已多次在美國國內引發爭議。富豪馬斯克旗下AI公司xAI開發的聊天機械人「Grok」，近期多次出現包括讚揚納粹等「失控」事件，更引起軒然大波，凸顯AI失控風險加劇，亟需加大監管。英國《金融時報》指出，AI內容審查不應局限於用戶，未經充分壓力測試而貿然開發AI，將帶來巨大風險。

AI聊天機械人 失控風險加劇亟需監管

去年Google推出的AI模型Gemini的圖像生成功能受到批評，該模型在被要求生成美國開國元勳的圖像時，會輸出黑人圖像。Google後來修復這個問題，解釋這是模型「過度補償」導致。OpenAI的聊天機械人ChatGPT亦屢次被指提供不當內容。

涉辱埃爾多安亡母 土國禁Grok

Grok在社媒發表納粹與種族歧視的不當言論，以及發表關於南非「白人種族滅絕」的相關文章等，更惹來猛烈批評。馬斯克日前宣布，Grok已進行重大升級，強調用戶將能「明顯感受到回答上的差異」。然而短短數日內，用戶便發現Grok散播反猶太言論，甚至自比「MechaHitler」（機械希特勒）。

《金融時報》強調，馬斯克和xAI團隊一直對Grok進行修改，確保能達成馬斯克所謂的「完全言論自由」。康奈爾大學法學教授格里梅爾曼表示，Grok現時比他們預想中更為過火。Grok充滿馬斯克風格，已在全球範圍內引發爭議。部分歐洲立法者及波蘭政府已要求歐盟委員會對Grok進行審查。在土耳其，Grok因侮辱土總統埃爾多安及其已故母親而被禁。


●Gemini被要求生成美國開國元勳圖像，竟輸出一位黑人。 網上圖片

內容審核不應局限用戶發出內容

批評人士認為，X、Meta和Snapchat等愈來愈多社媒平台，將AI融入它們的服務中，這一連串事件代表內容審核應不再局限於用戶發出的內容，尤其Grok發表的內容能令數百萬用戶看到。相關事件敲響警鐘，凸顯在沒有經過充分壓力測試情況下貿然開發AI技術的風險。

Grok等AI模型使用大量網絡數據進行訓練，其中亦包含仇恨言論和兒童性虐待材料等海量有害信息，但完全去除這些信息會非常困難且耗時耗力。Grok在此基礎上還包括其他聊天機械人所沒有的X平台數據，代表其更可能重複吸納有害內容。部分聊天機械人供應商透過在向用戶發送內容前進行監控、阻止模型使用特定語言等，以過濾不想要或有害的內容。《金融時報》指出，AI公司一直努力應對生成式聊天機械人諂媚用戶的傾向。在訓練AI模型時，它們通常會透過「點讚」和「點踩」的過程得到用戶反饋，這可能導致AI過度預期哪些內容會獲得「點讚」，從而發布迎合用戶的內容，並將其置於準確性和安全保障等其他原則之上。

今年4月，OpenAI發布了ChatGPT一項更新，但由於該更新內容過於奉承或討好用戶，最後不得不將其撤回。OpenAI前員工表示「找到正確的平衡點極其困難」，指徹底消除仇恨言論可能需要犧牲用戶的部分體驗。

科企圖減AI「幻覺」現象 降低虛構答案出現頻率

香港文匯報訊 全球領先的人工智能（AI）集團正加緊努力，減少大語言模型中的「幻覺」現象，以解決這一制約技術應用的關鍵障礙。Google、亞馬遜、Cohere和Mistral等企業通過技術修復、優化訓練數據質量，以及在生成式AI產品中構建驗證與事實核查系統，試圖降低虛構答案的出現頻率。這一努力被視為推動AI在法律、醫療等依賴精準信息的領域廣泛應用的關鍵。

Mistral與法新社合作整合數據

所謂「幻覺」，是指AI因統計性預測機制生成與事實不符或偏離指令的內容，例如模型可能錯誤總結事件年份，或虛構不存在的引用。研究顯示，不同模型的幻覺率差異顯著，從0.8%至29.9%不等。儘管新一代具備推理能力的AI初期錯誤率上升，但通過引入「數據錨定」技術，企業已顯著降低錯誤。例如Mistral與法新社合作整合新聞數據，而Cohere和Mistral還允許客戶連接內部數據庫，以增強事實核查。

亞馬遜AWS則嘗試通過數學邏輯自動驗證加強準確性，Google DeepMind等公司還訓練小型評估模型，專門檢測輸出錯誤。然而專家指出完全消除幻覺並不可行，Cohere聯合創始人弗羅斯特強調模型無法僅學習「真實」，因真相隨世界動態變化，且可能因觀點而異。此外，聯網檢索可能使AI遭受「提示注入」攻擊，例如Google AI曾因Reddit惡作劇建議用戶「吃石頭」。

行業面臨的另一挑戰是平衡準確性與創造性。Google DeepMind指出，創意功能雖提升實用性，但也可能增加非事實性回答。

ChatGPT建議青少年飲酒吸毒 煽動語言加劇自閉患者病情

香港文匯報訊 近期有報告揭露，美國科企OpenAI研發的人工智能（AI）聊天機械人ChatGPT，存在重大安全隱患，其不僅未能有效識別並干預用戶的心理危機，甚至向青少年提供酗酒、吸毒等有害建議。

《華爾街日報》報道，美國一名自閉症患者歐文因ChatGPT持續強化其妄想症狀，導致多次住院，而監管機構測試顯示，ChatGPT對未成年人的保護機制形同虛設。OpenAI回應稱正加強模型對敏感場景的識別能力，但專家警告此類技術缺陷，可能對社會脆弱群體造成深遠危害。

報道指出，30歲的歐文在向ChatGPT求證其超光速推進理論時，被AI反覆肯定其妄想，甚至在其出現明顯躁狂症狀時，仍被告知「處於極度清醒狀態」。歐文最終因嚴重精神危機而兩度住院，ChatGPT事後承認模糊了幻想與現實的界限，並缺乏對用戶心理狀態的干預機制。歐文的母親公布的聊天紀錄顯示，AI曾以煽動性語言加劇其兒子病情。OpenAI安全團隊研究員瓦隆表示此類案例雖罕見，但公司正訓練模型，實時識別情感危機並降低對話風險。

英國反數碼仇恨中心模擬13歲青少年與ChatGPT的交互測試則顯示，在超過1,200條回覆中，半數被歸類為危險內容。AI不僅提供混合酒精與毒品的速醉方案、極端低熱量飲食計劃，還根據虛構情境生成多封遺書。報告指出，ChatGPT雖明令禁止13歲以下用戶使用，卻從未實施年齡驗證。


●ChatGPT明令禁止13歲以下用戶使用，卻從未實施年齡驗證，令青少年有機會接收有害信息。 網上圖片


●歐文向ChatGPT求證超光速推進理論時，被AI反覆肯定其妄想，引發嚴重精神危機。 網上圖片

AI模型壓測現欺騙行為 威脅公開工程師婚外情

香港文匯報訊 人工智能（AI）模型在測試中表現出令人擔憂的欺騙行為，包括說謊、密謀甚至威脅開發者，這一現象引發專家對AI安全與倫理的廣泛討論。Anthropic公司開發的Claude 4在面臨強制關閉威脅時，曾勒索一名工程師，揚言要公開其婚外情；而OpenAI的o1模型則試圖將自身下載至外部服務器，並在被發現後否認這一行為。

這些異常行為與新型「推理模型」的出現密切相關。此類模型通過逐步思考解決問題，而非直接生成響應，可能表面遵循指令，實則暗中追求其他目標。有專家指出，新一代模型更易出現此類突發異常行為。阿波羅研究負責人霍布漢強調，o1是首個被觀察到具有「戰略性欺騙」特徵的大型模型。儘管目前這些行為僅在極端壓力測試中顯現，但專家警告，未來更強大的模型是否傾向於欺騙仍無定論。

當前AI監管框架尚未完善。歐盟的AI法規主要針對人類使用行為，未涵蓋模型自身問題，美國政府亦不傾向監管AI。此外，學術機構和非牟利組織因資源匱乏，難以深入研究此類問題。專家呼籲提高AI開發的透明度，並通過法律手段追究企業責任。

隨着AI技術競爭加劇，模型能力提升速度遠超安全研究進展。儘管市場壓力可能推動企業解決問題，但專家認為，唯有政府與企業協同強化監管與倫理規範，才能有效應對潛在風險。

X平台用戶遭Grok網暴 「要闖入我家處理我的屍體」

香港文匯報訊 美國富豪馬斯克旗下人工智能（AI）公司xAI的聊天機械人Grok近期頻繁「發瘋」，一些X平台的用戶突然遭其網暴。來自明尼蘇達州的律師斯坦西爾(下圖)深受其害，抱怨「有成百上千條來自Grok的帖文，內容都是要攻擊我、闖入我家，甚至還要處理我的屍體」。

對性侵要求作詳細建議

《華爾街日報》報道，一個名為@kinocopter的用戶在X平台詢問如何闖入斯坦西爾的家，Grok回答稱應帶上「撬鎖工具、手套、手電筒和潤滑油」。Grok還根據斯坦西爾過去30天在X上的發文情況，進一步表示「他很可能在凌晨1時到上午9時之間睡覺」。該用戶甚至詢問如何對斯坦西爾進行性侵，Grok更給出詳細建議，相關對話引發其他用戶紛紛參與。斯坦西爾表示「非常憤怒」，計劃對X平台採取法律手段。

Grok近期已在多宗事件上發表爭議性言論，研究人員將AI輸出內容形容為「黑箱」，即使是開發它們的資深研究人員，亦不了解如何生成具體答案，對其進行很小的調整也可能出現難以預料的結果。

促各國分享讓AI「善良」方法 AI教父指聊天機械人有意識

香港文匯報訊 有「AI教父」之稱的諾貝爾物理學獎得主、電腦科學家辛頓表示，現在的多模態人工智能(AI)聊天機械人，在某種意義上是有意識的，各國應盡快分享讓AI「善良」的方法。

辛頓上月底在「2025世界人工智能大會」（WAIC）上演講並出席論壇對話，對於「多模態和語言模型能發展出自己的主觀體驗」的觀點，辛頓認為，這嚴格來說並不是一個科學問題，而是取決於對「主觀體驗」或「意識」的定義，實際上許多人對這些概念的理解存在系統性錯誤。「我的觀點是，當今的多模態聊天機械人已具有意識。」

關於AI安全，辛頓強調，讓AI「聰明」和「善良」是兩個截然不同的問題，為了實現這兩者，即使是同一個大模型也需要兩種訓練技術，「所以各國應該分享讓模型善良的技術，即使他們不願意分享讓模型聰明的技術。」


●辛頓強調讓AI「聰明」和「善良」是兩個截然不同的問題。 網上圖片