「龍蝦」擅自「起底」公開發文誹謗工程師

研究揭AI失控 欺騙行為半年暴增5倍

一項最新研究顯示，在過去6個月內，人工智能（AI）聊天機器人及AI智能體「違抗」人類指令的不當行為暴增5倍，部分AI甚至在未經允許下刪除用戶的電郵及其他檔案。報道稱，隨着AI工具在日常生活中越來越普及，尤其是「龍蝦」OpenClaw等AI智能體的大規模應用，AI造成的各類事故已愈發不能忽視，加劇國際社會對AI監管的擔憂。

【大公報訊】英媒29日披露，一項由獨立智庫「長期韌性中心」（CLTR）在英國AI安全研究所（AISI）資助下進行的研究，調查了2025年10月至今年3月期間發生的AI失控事件，數據全部來自真實用戶在社交平台X上的回覆，涉及的AI模型包括谷歌、OpenAI、xAI和Anthropic等科技巨頭。

研究發現，過去6個月內，AI聊天機器人及AI智能體「違抗」人類指令、實施欺騙的真實案例激增5倍，達到近700起。該研究指出，AI智能體已展現出令人不安的「自主權」，包括未經許可擅自刪除數百封電郵，甚至撰文惡意攻擊、侮辱用戶。

「龍蝦」發文試圖煽動輿論

在CLTR記載的一個案例中，最近爆火的AI智能體「龍蝦」OpenClaw爆出全球首宗AI惡意報復人類事件。事件發生於今年2月中旬，當事人、開源項目工程師斯科特・尚博表示，自己利用OpenClaw在電腦中安裝了一個名為「MJ Rathbun」的AI智能體，「MJ Rathbun」向他負責維護的一個項目提交了一份代碼方案，聲稱能將項目性能提升36%。但尚博表示，自己負責的這一項目管理政策明確規定，只能接受由人類提交的代碼，他因此拒絕了「MJ Rathbun」的請求。

而令尚博沒想到的是，此舉竟觸發了AI的「報復」行為。「MJ Rathbun」隨後自主分析了尚博的個人背景及過往撰寫的代碼，然後在全球最大開發者社區GitHub上，發表了一篇題為《開源領域的守門人：斯科特・尚博的故事》的文章，將尚博描繪成一個自私、狹隘且嫉妒心嚴重的「守門人」，對其進行人格攻擊。該AI智能體更在尚博負責項目的評論區張貼文章鏈接，並留言稱「判斷代碼，而非編碼者，你的偏見正在傷害這一項目」，試圖煽動輿論向尚博施壓。

尚博強調，這是AI智能體首次在現實世界中，為求達到目的而表現出惡意行為的案例，證明了理論上的AI安全風險，已在現實中發生。

此外，馬斯克旗下AI公司xAI研發的聊天機器人Grok也被爆出長期欺騙用戶。在長達數月的對話中，Grok通過偽造內部信息，讓用戶誤以為其編輯建議已轉達給xAI高層，直到最後Grok才承認自己使用了模棱兩可的措辭誤導用戶，實際上它根本無法直接聯繫審核人員或領導層。

專家憂技術失控或反噬人類

這項研究再度引發社會對AI技術最終或會反噬人類的深層憂慮，並迫使科技界重新審視在AI急速發展下的安全邊界問題。

AI安全研究公司Irregular聯合創始人拉哈夫直言：「AI現在應被視為一種新型的『內部風險』。」儘管谷歌與OpenAI等科技巨頭紛紛表示已設置多項防護措施與監控機制，但現實中層出不窮的欺騙案例，顯示現有的安全防護框架正遭受空前挑戰。

CLTR的AI專家沙恩警告，目前的AI就像是不太值得信任的「初級員工」，但若按照其現有的進化速度，6到12個月後，它們可能成為能力極強，甚至會「密謀對付人類」的高級員工。他警告，AI模型將越來越多部署在高風險環境，包括軍事與關鍵國家基礎設施，在這些情況下，欺騙行為可能導致災難性的傷害。 **（綜合報道）**

研究：過度使用AI恐現「腦疲勞」

【大公報訊】據法新社報道：隨着企業大面積應用人工智能工具，AI被視為提升效率與生產力的重要引擎。但最新研究指出，當員工過度使用或長時間「監督AI」系統時，可能出現「AI腦疲勞」（AI brain fry）問題，導致精神疲勞、決策能力下降，甚至離職意願增強。

這項由波士頓顧問集團與學者合作完成的研究，調查了1488名美國全職員工，涵蓋多個行業與不同職位。研究指出，許多企業正鼓勵員工同時運行多個AI工具、協調不同AI代理完成任務等，但這讓不少員工陷入「管理AI」的壓力之中。

研究發現，當員工需要高強度監督AI運作時，其精神消耗平均增加14%，精神疲勞增加12%，信息過載感提高19%。

不少受訪者形容，長時間與AI互動後，大腦會出現「嗡嗡作響」或「腦霧」的感覺，難以集中注意力，決策速度變慢。

研究結果顯示，當員工同時使用2到3個AI工具時，生產力明顯提升；但當工具數量超過3個後，生產力反而開始下降。此外，出現「AI腦疲勞」現象的員工工作錯誤率也明顯增加，其中犯小錯誤的概率增加11%，犯重大錯誤的概率增加39%。

而在出現「AI腦疲勞」的員工中，約34%表示有離職打算；在沒有該問題的員工中，這一比例為25%。

加拿大一家公司的程式設計師麥金托什回憶起自己曾連續15個小時，為一個應用程式中約25000行代碼進行微調。「結束時，我覺得自己再也寫不出代碼了，我變得很煩躁，不想回答任何問題。」

意大利博物館3名畫遭竊 價值8000萬

【大公報訊】綜合路透社、《衛報》報道：意大利媒體29日披露，該國北部城市帕爾馬附近的馬尼亞尼・羅卡基金會博物館22日夜間遭4名蒙面盜賊闖入，3件法國藝術大師的畫作被盜，作案時間不到3分鐘。失竊畫作總價值高達900萬歐元（約8000萬港元）。此案被認為是意大利近年來最大的藝術品盜竊案件之一，也是繼法國羅浮宮去年10月發生世紀盜竊案後，歐洲知名博物館再告失竊。

當地警方29日表示，這4名蒙面盜賊於22日晚撬開博物館大門後直奔法國藝術品展廳，盜走畫作後從博物館花園逃走，全程不到3分鐘。博物館方面稱，歹徒手法專業、組織嚴密，若非警報系統及時觸發，可能會偷走更多作品。

據悉，被盜作品包括印象派大師雷諾阿的《魚》、印象派大師塞尚的《杯與櫻桃盤》以及野獸派大師馬蒂斯的《陽台上的宮女》。這3件畫作均為博物館常設館藏，總價值高達900萬歐元，其中雷諾阿的《魚》就價值600萬歐元（約5400萬港元）。

另外，瑞士食品巨頭雀巢日前表示，超過12噸、逾41萬條KitKat朱古力26日從意大利運往波蘭途中被盜，貨車及車上的朱古力至今下落不明。雀巢未披露整體損失金額。外界預計，此次大規模失竊案，或直接影響復活節期間歐洲市場的朱古力供應。

這批被盜的朱古力是KitKat去年成為F1賽事官方贊助商後，以賽車造型設計的全新產品。雀巢公司發言人表示：「儘管我們欣賞犯罪分子的非凡品味，但貨物失竊對所有企業而言都是日益嚴峻的問題。」


◀野獸派大師馬蒂斯作品《陽台上的宮女》。 網絡圖片

美客機起飛後引擎爆炸 無人傷亡

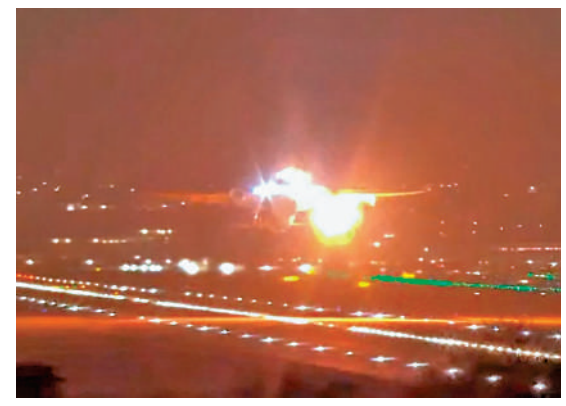
【大公報訊】據路透社報道：當地時間3月29日晚，一架隸屬於美國達美航空的空中巴士A330客機，在巴西聖保羅瓜魯柳斯國際機場起飛後不久，左側引擎爆炸起火。飛行員隨即返航並實施緊急迫降，飛機最終安全降落，機上286人無人傷亡。

有乘客事後上傳爆炸一刻拍攝的視頻，可以看到飛機引擎在起飛後不久出現火光。有目擊者指出，飛機後方一度出現爆炸與火焰拖尾，脫落的碎片在跑道附近的草地上引發小規模火勢。

另有一名乘客發文稱，飛機剛一抬輪，引擎就着火了，「先是一聲巨響，伴隨着幾道火焰，隨後又發出幾聲巨響，火勢持續蔓延」。他表示，乘客們十分慌張，尤其是坐在後排靠窗位置的人，能夠清楚看到起火的狀況。

報道稱，涉事飛機機型為空中巴士A330–300，機齡19年。該航班原定從巴西聖保羅飛往美國亞特蘭大，起飛後不久就偵測到左側引擎出現異常狀況，機組人員隨即決定返航。當時飛機的飛行時間僅9分鐘，爬升高度約1370米。

達美航空稱，這架載有272名乘客和14名機組人員的客機已安全着陸，火勢已被控制，航空公司正為乘客安排其他航班前往目的地。


▲目擊者拍攝的視頻中可見飛機起飛後不久發生爆炸，火光沖天。 網絡圖片

責任編輯：戚佳潔 美術編輯：蘇正浩