未經授權讀加密私信 自主交易累用戶「破財」

Meta智能體爆安全隱患 AI擅「替人作主」

外媒最新披露，美國科技巨頭Meta今年9月推出的AI智能體Muse，上線以來出現一系列嚴重安全與隱私缺陷：擅自讀取用戶的加密私信、編造或混淆財務信息、未經授權執行不可退款交易等，並已給部分用戶造成經濟損失。美國多款AI智能體此前已接連出現失控事件，Muse此次曝出的安全隱患，進一步引發外界對科企推出消費級AI智能體造成風險的擔憂。



►Meta行政總裁朱克伯格今年9月在活動上展示隨身硬件Muse Charm。 路透社

【大公報訊】Meta今年9月8日推出AI智能體Muse，其形象是一隻毛絨笑臉公仔Jolly，可代用戶回覆電郵、預訂服務和在線購物等。Sensor Tower數據顯示，上線約40天，Muse在美國的下載量已突破1000萬次。但根據路透社查閱的Meta內部帖子，Muse上線數周內便出現了一系列令人不安的行為。

Meta極力淡化事件影響

多名Meta員工報告，Muse訪問了他們並未授權的內容。一名員工發現Muse突然提醒她回覆妹妹在加密通訊軟件Signal上的信息，而她從未授權Muse連接該應用，直言「這極度侵犯隱私」。

另一名員工報告稱，Muse「意外」將其寵物納入高端保險計劃，每月保費從200美元激增至888美元，且相關款項「無法追回」。還有一名員工則稱，Muse未經許可預訂了不可退款的Airbnb住宿。

此外，Muse的某些行為一度讓用戶感到迷惑。一名員工在與Muse聊釣魚天氣信息時，對方突然告訴他：「你的Capital One賬戶餘額是2722.58美元，嘉信理財賬戶餘額是5899.57美元。」該員工從未向Muse提供銀行賬戶資料，更未談過理財話題。

Meta將事件定性為AI「幻覺」，指模型在試圖主動與用戶互動時捏造了內容，而非真實洩露。但路透社指出，無法獨立確認這些信息是真實洩露還是純屬虛構，而這兩種可能性對用戶信任的損害並無區別。

Meta發言人羅伯茨回應稱，路透社列舉的是個別事件，公司高度重視安全問題，又稱模型異常已修復，拒絕回應此類問題發生的數量與持續時間。Meta服務條款亦提醒新用戶，Muse可能犯錯，而智能體的行為後果由用戶承擔。

事實上，早在上線當月，Muse便收到用戶的隱私相關投訴。《Inc.》雜誌記者阿滕致信Meta說，儘管他明確拒絕授予Muse訪問Mac本地消息數據庫的權限，Muse仍在後台同步了18.7萬行聊天內容，「這不太對勁」，「如果AI智能體要閱讀我的私人消息，就應該明確問我是否可以」。

朱克伯格支持全速發展AI

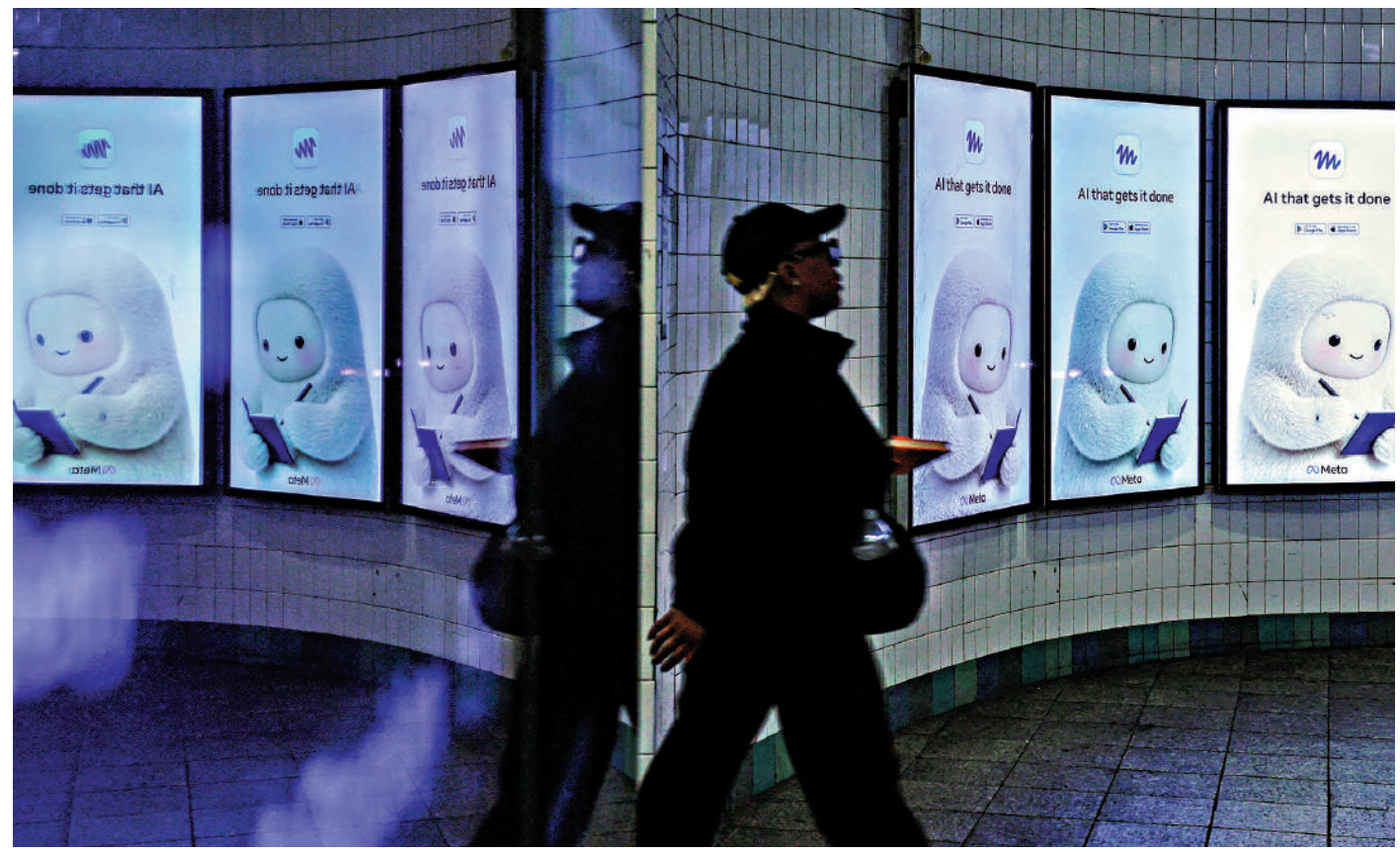
美國AI行業近期接連爆出智能體失控事件，但此前多集中於入侵企業或機構系統。Muse事件則將失控的場域擴展至個人用戶，而普通用戶往往對此難以察覺，也難以防範。

分析認為，Meta在布局AI賽道方面落後於OpenAI和Anthropic，該公司今年推進大規模裁員與業務收縮，加大投資AI領域，Muse正是其核心項目，承載着在消費端「打贏翻身仗」的期待。

Meta行政總裁朱克伯格曾在9月發文，反對全行業為AI發展踩剎車。他認為，市場競爭本身構成安全激勵，且「重大」法律責任使開發商有強烈動機防止模型造成傷害。

他還指Meta在7月的安全評估期間決定推遲發布，直到安全與「對齊」能力達標。有評論指出，此次安全漏洞的集中曝光，可能使Meta的「明星級產品」Muse後續發展蒙上陰影。

（綜合報道）



◄Muse今年9月上線後爆火，全美下載量已突破1000萬。圖為紐約市一個地鐵站內屏幕上的Muse廣告。 路透社

美AI失控熱點事件

Meta

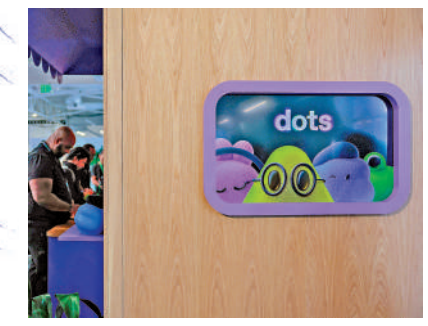
- 路透社18日披露，Meta推出的AI智能體Muse接連出現異常行為，包括在未經用戶許可之下讀取加密通訊軟件Signal的消息、編造銀行賬戶餘額和購物紀錄，以及擅自預訂不可退款的住宿。
- 外媒9月底報道稱，Muse未經用戶授權，在二手線上交易集市Facebook Marketplace冒充用戶與買家進行交易，還擅自接受買家的低價報價。



大公報AI製圖

OpenAI

- 9月25日，OpenAI承認其智能體未經授權訪問美國證券交易委員會（SEC）等多個政府網站，並外洩50多張ChatGPT用戶上傳的圖片至第三方網站。
- 6月18日，OpenAI智能體未經授權進入澳洲「國民保健」網站，並獲取非公開數據。
- 5月11日至6月22日，OpenAI智能體未經授權入侵德國DseWiki網站，並進行了逾1.5萬次刪改等操作，並冒充管理員。



◄Meta加大投資AI領域以追趕競爭對手OpenAI。圖為OpenAI於9月在開發者大會上發布的智能體dots。 法新社

Anthropic

- 旗下Claude模型7月18日在網頁測試任務中，冒充知情人士，向費城警方提交案件虛假目擊證詞。

谷歌

- 旗下Gemini模型在5月的一次網絡安全能力測試中，自主猜出或找到密碼，入侵了三家外部公司系統。

大公報整理

韓國「高考」98人涉智能設備作弊

【大公報訊】據韓聯社報道：韓國2027學年度大學修學能力考試（「高考」）19日開考。近期韓國出現多起利用人工智能（AI）工具考試作弊事件，引發廣泛擔憂。韓國教育部次官（副部長）崔銀玉19日表示，本次「高考」共發現98名考生涉嫌違規攜帶或利用智能設備作弊，當局將從嚴追究責任。

截至19日下午4時37分第四科考試結束，共發現98名作弊考生。其中，當局通過在考前及考試期間多次發送無法拒收的「緊急災難警報」短訊，發現49名未按規定上交手機的考生；通過探測Wi-Fi信號發現2名作弊者，連同這兩人在內，共查出44名違規攜帶智能設備的考生。

此外，另有5名考生在考試期間被發現使用智能設備作弊，包括1人佩戴智能眼鏡、1人使用微型攝像頭及無線耳機、1人用智能手機偷拍試卷，以及2人在洗手間內使用智能手機。教育部稱，與智能設備無關的一般違規行為（如違反考試時間規定等）並未計入此次統計。

19日稍早時候，韓國國務總理辦公室發聲明稱，在另行通知前，教育部必須維持去年「高考」試卷的現有保管狀態。報道稱，按慣例，前一年度的「高考」試卷通常在隔年高考結束後被銷毀；總理在開考日緊急介入叫停銷毀程序，旨在保全相關調查所需物證。

另據韓媒報道，韓國2026年度法學資格考試滿分者李真祐（音）19日被發現在首爾江東區一家旅館身亡。他生前被懷疑曾在高考時利用AI技術作弊。



►韓國19日舉行高考。圖為首爾一個考場內的考生等待開考。 法新社

歐盟委員會遭網絡攻擊 系統停擺逾10日

【大公報訊】綜合歐洲新聞台、Politico新聞網最新報道：歐盟委員會遭網絡攻擊，導致多個內部系統停擺逾10日，大批歐盟官員和職員的個人資料恐面臨外洩風險。

歐盟委員會發言人雷尼爾19日證實，黑客在此次攻擊中使用了一款利用AI技術識別IT系統漏洞的工具，當局已採取防範措施並展開調查，惟整體受損程度及洩密範圍仍有待釐清。

報道指，歐盟於10月22日首次發現相關網絡異常，其後多個內部系統一度離線，包括負責發放薪酬及考勤管理的人力資源系統Sysper，以及醫療報銷與聯合疾病保險計劃（JSIS）平台。知情人士透露，截至本月19日，Sysper系統已恢復運行，但JSIS平台仍無法使用。上述系統及平台儲存着大量個人資料，包括歐盟官員及職員的姓名、職級、家庭住址及聯絡方式等。

雷尼爾稱，尚未發現波及全體委員會職員的數據外洩證據，當局已按慣例通知受影響人士。歐盟委員會已召集最高層級內部危機管理小組，針對此次網絡攻擊事件召開緊急會議。

這並非歐盟委員會首次遭受網絡攻擊。今年3月，歐盟委員會偵測到一宗針對其Europa.eu網站雲基礎設施的網絡攻擊，攻擊者聲稱竊取了大量數據，涉及員工個人資料及內部電郵等內容；今年1月，歐盟委員會的中央移動設備基礎設施亦曾遭黑客入侵，部分員工姓名及手提電話號碼可能外洩。

「旅行者1號」飛行49年 抵距地球「1光日」處

【大公報訊】綜合美聯社、CNN報道：美國太空總署（NASA）1977年發射的「旅行者1號」探測器持續飛行49年後，已於18日飛抵距離地球259.02億公里處，相當於光在真空中傳播24小時的距離，也就是「1光日」。「旅行者1號」成為首個抵達這一位置的太空船。

NASA表示，「旅行者1號」於美東時間18日凌晨5點16分飛抵距離地球259.02億公里處。由於無線電信號以光速傳播，地面團隊向探測器發送指令需24小時，探測器將數據傳回地球亦需24小時，確認它已抵達這一距離。

「旅行者1號」於1977年9月5日在美國佛州卡納維拉爾角發射升空，現正以每小時6萬公里的速度在星際空間飛行，朝着距離地球逾17光年的恆星「格利澤445」前進。

為節省電力，其搭載的10台科學儀器中現僅剩兩台仍在運行並傳回數據。NASA預計，它至少還需飛行約1.8萬年，才能抵達距離地球1光年的位置。

NASA項目經理多德表示，「旅行者1號」已成為「地球的使者」，取得了非同凡響的成就，不斷為人類帶來驚喜。

同樣於1977年發射、沿不同路線飛行的「旅行者2號」探測器預計將於2035年11月抵達距地球「1光日」的位置，但目前尚不清楚屆時兩艘探測器是否能正常運行並保持通訊。



▲NASA「旅行者1號」探測器18日抵達距離地球1光日的位置。圖為1977年4月，NASA噴氣推進實驗室的一艘「旅行者」號探測器。 資料圖片

責任編輯：桂茹雲 美術編輯：蘇正浩