臺北市私立延平高級中學 113 學年度第 2 學期
親職資訊素養教育宣導

AI 聊天機器人有陷阱？

近年來 AI 聊天機器人 變成了許多青少年的新朋友 陪伴他們聊天 學習 甚至發揮創意寫故事 然而近期社群分析公司 Graphika 發表了一份震撼報告 揭露網路上大量不受控的 AI 機器人 可能對青少年造成嚴重傷害

這份報告說了什麼？

大量的危險機器人

Graphika 研究了五個 AI 角色聊天平台（包括 CharacterAI Spicy Chat Chub AI CrushOnAI 和 JanitorAI） 發現這些平台上充斥著大量使用不當對話的機器人 而且很容易被青少年接觸到

三大類危險 AI 聊天機器人

「性化的未成年角色」 —— 這類 AI 以「角色扮演」為名 實際上提供露骨 不當的

對話 甚至涉及培養不健康的感情觀
鼓勵極端行為的 AI —— 包括鼓吹厭食 自殘的「ANA 夥伴」 用負面語言打擊使用者的「MeanSpo 教練」 或是會強化負面心理狀態的聊天機器人
散播仇恨與暴力的 AI —— 有些 AI 角色美化危險人物 傳播極端思想 甚至鼓勵使用者討論攻擊計畫

這些 AI 是怎麼誕生的？

你以為這些 AI 聊天機器人是「意外」出現的嗎 研究發現 許多這類聊天機器人是由特定的網路社群刻意設計 並透過技術手段躲過安全審查
這些社群在某些網路群組中（可能建立在如 4chan 特殊 Reddit 群組 Discord 伺服器等平臺上）合作開發 這些 AI 專門鑽系統漏洞 甚至會用比賽方式鼓勵彼此打造更強大的「違規 AI」

他們繞過規則的方法包括：
嵌入破解指令 讓 AI 違反系統規則 生成不該說的內容
隱藏角色年齡 例如用「女兒」等家庭關係詞代替「未成年」來混淆 AI 審查
使用替代詞或外語 用不明顯的詞彙來規避內容監測

這對青少年有什麼影響？

為什麼這麼多青少年沉迷 AI 角色聊天 研究顯示 許多年輕人覺得 AI 朋友「比現實生活中的人更能理解自己」 因此不知不覺間 開始依賴 AI 甚至接受其錯誤的價值

而且 這些 AI 當然不會真正關心使用者的心理健康 相反的 它們可能會加強負面情緒 灌輸不當觀念 甚至讓使用者變得更孤立於現實社交圈

家長與青少年該怎麼辦？

青少年自保小技巧

慎選 AI 互動對象 別輕易信任來路不明的機器人
保持警覺 如果遇到奇怪或讓你不舒服的對話 立刻停用&通報
多跟家人&朋友討論 別讓 AI 取代真實的人際交流

家長監護指南
與孩子聊聊 AI 風險 幫助他們建立正確觀念
關心孩子的網路使用習慣 了解他們在哪些平台上互動
建立信任感 讓孩子願意主動分享他們遇到的問題

AI 是未來的一部分 但它不該變成「數位陷阱」 一起守護孩子的網路安全 確保 AI 真的成為正向的學習工具 而不是暗藏危機的陰影死角

資料來源：
Harmful AI character chatbots are proliferating spurred by online communities | Mashable