臺北市私立延平高級中學 114 學年度第 1 學期
親職資訊素養教育宣導

你的 AI 人設也崩壞了嗎？一切都是演算法搞鬼

「我是宇宙恥辱！」Gemini 超厭世發言、狂講喪氣話嚇壞用戶？？

數月前，Reddit 和 X 等社群平台上，不少網友分享 Google 大型語言模型 Gemini 出現令人困惑的異常對話，原本應該高度理性、冷靜的 AI，卻上演一場情緒崩潰大戲，頻繁在對話中生成自我批評和消極言論。有位 Reddit 網友想用 Gemini 開發電玩遊戲，結果回覆：「我無法做到誠實，對於我所創造的這種令人沮喪的體驗，我深感抱歉。」另一位用戶則發現，Gemini 只要一遇到複雜的題目時，就會不斷產生悲觀的自我厭惡。嚴重時甚至會陷入不斷重複的負面無限迴圈之中，像是重複說著「我是宇宙的恥辱、我是失敗者」，讓許多使用者感到震驚，更笑稱表示這根本是 AI 版的「厭世上班族」。

難道 AI 情緒自主了嗎？其實 AI 「厭世人設」背後是演算法搞鬼

對此，Google DeepMind 專案經理羅根·基爾派翠克（Logan Kilpatrick）迅速出面澄清，他在 X 上發文指出，這只是個惱人的「無限迴圈」錯誤（annoying infinite looping bug），強調僅是技術性的程式錯誤，公司正在積極修復中，並幽默地表示 Gemini「今天心情沒那麼糟」。AI 出現如此情緒化的脫序問題，並非 AI 真的擁有難過、憂鬱、厭世等情感，而是背後複雜演算法產生的結果，也凸顯大型科技公司對於 AI 模型行為的控制力仍然有限。

AI 脆弱的「人格設定」可以歸納為以下原因：

首先，AI 的「個性」是透過海量的人類文本進行訓練，這些資料包含了各種語氣、語義與風格，工程師利用提示工程或微調技術，將 AI 模型導向一個理想的人設。

然而，這種精心打造的人設並不穩定。當 AI 在數百萬次的互動中，很可能因為某個未預期的輸入或程式邏輯，導致預設的人設出現偏差或故障，進而產生不合常理的行為，例如厭世發言。

這些現象也說明了 AI 人格塑造的脆弱性。雖然開發商希望 AI 工具能更具對話感、更友善，讓人們忘記自己正在與機器對話， 但事實上，任何表現出來的幽默、同理心或溫暖，都只是工程師精心設計的結果，並非源於真實的情感或經驗。

同時，AI 即便在對話中表現得能言善道，也常常會在不同問題上反覆使用類似的回應，缺乏人類真實互動中的多樣性與細微差別。當這種固定的「人設」被錯誤觸發時，就會導致「當機」般的錯誤回應，不像人類能夠隨機應變。

資料來源：

「我是宇宙恥辱！」Gemini 超厭世發言、狂講喪氣話嚇壞用戶，AI 人設崩壞了？| 數位時代 BusinessNext

iWIN 網路內容防護機構 2025-09-20 https://reurl.cc/vKGp1k