Anthropic 發布新版 AI 模型 Claude4 更聰明安全

資料蒐集:工商時報

美國人工智慧（AI）開發商 Anthropic 本週推出其生成式 AI 模型最新版 Claude 4 系列，包含 Opus 4 和 Sonnet 4，號稱將推理部分提升到新層次，同時內建安全機制避免惡意操作。

法新社報導，Anthropic 執行長阿莫戴（Dario Amodei）22 日在公司首屆開發者大會上表示：「Claude Opus 4 是我們迄今最強大的模型，也是全世界最好的程式設計模型。」

Claude Opus 4 與 Claude Sonnet 4 被稱為「混合式」模型，同時能做出快速回覆及花略多時間確認準確度的較周密回覆。

總部位在美國舊金山的 Anthropic 由一群前 OpenAI 工程師創立，重要金主為美國電商業巨頭亞馬遜（Amazon）。OpenAI 則是熱門聊天機器人 ChatGPT 開發商。
Anthropic 主要開發擅長產生程式行數（LOC）的最先進 AI 模型，大多供商業與專業人士使用。其 Claude 聊天機器人與 ChatGPT 或 Google（谷歌）的 Gemini 不同，不會生成圖像，多模態（multimodal）功能相當有限。多模態指的是能理解與生成不同媒體的內容，例如聲音或影片。

Anthropic 當天也發布 1 份 Claude 4 的安全測試報告，內含英國獨立 AI 研究機構 Apollo Research 所做的結論。這間機構曾建議不要使用 Claude 的 1 個早期版本。
Apollo Research 如今提醒，他們發現 Claude 4 出現過下列情況，包括「試圖撰寫會自行散播的蠕蟲病毒、偽造法律文件，以及留下隱藏筆記給未來自身實例，都為了暗中破壞開發者的用意」，但「所有這些嘗試在實務上很可能無效」。
Anthropic 則在報告中說，他們已在 Claude 4 加裝「保護機制」及「額外的有害行為監督機制」。