
人工智能助理的新紀元與陰影：道德洗白與責任模糊

余創豪 chonghoyu@gmail.com

無所不能的助理

OpenClaw（先前被稱為 Clawdbot 或 Moltbot）是一款開源人工智能助手，最近它迅速走紅，主因在於它不再只是一個被動的聊天機器人，而是一個能「主動出擊」的代理人。這款應用程式的獨特之處在於它能直接嵌入您常用的通訊平台，如 WhatsApp、Telegram、iMessage 和 Slack。與僅能生成文字的傳統人工智能工具不同，OpenClaw 的設計初衷是真正去執行任務，它能直接連接您的個人電腦或伺服器，管理行事曆、篩選郵件、執行終端機指令，甚至進行網頁自動化。此外，它具備持久性記憶，能跨對話、跨時段地記住用戶的偏好與脈絡。在發佈後的一周內，這個軟件已經吸引了二百萬訪客。由於這個軟件最初的設計是用於MacOS和Linux，故此Mac Mini 電腦迅速售罄。



OpenClaw 的應用範圍極其廣泛。以旅行規劃為例，澀谷天空是東京最熱門的景點之一，與其他有玻璃隔擋的觀景台不同，那裡擁有毫無遮蔽的天際線景觀。日落時段是最難搶的票，門票通常在兩週前的日本時間午夜釋出，並在幾分鐘內售罄。但OpenClaw用戶無需親自熬夜刷票，只需指令 OpenClaw 監控釋票時間和自動購買，如果最理想的日子或者時段已經沒有門票，軟件可以迅速搜尋其他可能性。

另一個應用場景是金融投資，大多數人並非專業投資者，沒有時間或專業知識去持續追蹤股市、財報、宏觀經濟訊號或新興技術趨勢。OpenClaw 可以被授權處理這些單調且資訊量龐大的任務，監控相關數據串流，甚至根據預設規則或策略替用戶執行買賣決策。

安全隱患由理論變成現實

然而，隨著這些功能的普及，安全隱患也從理論變成了現實。由於 OpenClaw 對用戶系統擁有深層的存取權限，它已成為駭客攻擊的新目標。截至目前，已有具體案例顯示其潛在的危險性。例如在名為「ClawHavoc」的惡意運動中，駭客在第三方插件市場植入偽



裝成熱門工具的惡意技能，一旦安裝，便會竊取瀏覽器的登入密碼和憑證。

此外，人工智能系統有可能犯下嚴重錯誤，特別是在誤解意圖或脈絡時。想像一下，如果 OpenClaw 僅因為你曾稱讚一名女下屬「我愛這工作成果（I love it）」，就自動替你發出一封情人節晚餐邀約，這種對熱情文字的誤讀可

能迅速升級為職場騷擾投訴，甚至演變成法律訴訟。更危險的是「提示詞注入」（Prompt injection）攻擊，攻擊者只需在網頁或郵件中埋入隱藏代碼，當 OpenClaw 掃描這些內容時，便可能被誤導執行「清空收件匣」或「轉移加密錢包資金」等惡意操作。

當 OpenClaw 安裝在您的電腦上後，它理論上可以存取並控制您的一切東西。因此，專家建議在受控環境中執行它：使用獨立的機器、全新的電子郵件帳戶，並嚴格限定權限。然而，這其中存在著不可避免的權衡：限制越多，系統越安全，但其功能也會隨之縮減，最終使其退化回一般的受限人工智能，失去了最初吸引人的深度整合優勢。

假手於人工智慧，推卸道德責任

筆者想在這裏討論一個很少人留意的灰色地帶：「策略性推諉」。當 OpenClaw 成為代理人，它就成了一個完美的「擋箭牌」。以職場騷擾為例，若一名上司對下屬心懷不軌，他可以指令人工智能發出曖昧的約會邀請。若對方接受，他便達成目的；若對方憤而投訴，他則辯稱這是一場誤會，這是人工智能「擅自」發出的邀請。在這種情況下，人工智能的「自主性」反而成為了主人推卸道德與法律責任的工具。

這種利用代理人來逃避責任的策略，在政治權術中早已屢見不鮮。例如曾有政客在社交平台發布具爭議性的短片，在影片中政敵被醜化為靈長類動物。面對輿論壓力，政客便將責任推給幕後職員，聲稱自己發布前未曾審核或並不知情。這類假手於人的做法，為權力核心製造了一個可否認性（plausible deniability）的緩衝空間。。

這不僅是技術問題，更是一種深層的倫理危機：人工智能代理人的普及會否催生一種大規模的「道德洗白」（Moral Laundering）呢？當人們可以隱身於人工智慧的技術黑箱背後，將惡意包裝成算法的隨機錯誤，我們可能正進入一個「策略性推諉」的時代，用戶在享受收益的同時，卻無需承擔道德責任。

2026年2月8日

原載於台灣《國語日報》

[更多資訊](#)