AI Agent 治理之道
效率與風險的雙面刃

在科技浪潮的推動下，人工智慧正迅速演進，從過去單純的「你問我答」工具，逐步發展為能自主規劃、決策並執行複雜任務的AI Agent。這些技術已廣泛應用於各個領域，大幅提升效率與精準度。例如：AI代理可持續監測IT伺服器狀態，主動規劃更新與修復流程，從而減少對業務運作的影響。然而，在享受其帶來便利的同時，我們亦必須正視潛藏的風險。

當決策權逐漸交由AI Agent的演算法主導，其在應對突發且複雜情境時的行為往往難以預測，亦可能衍生隱私外洩或網絡攻擊等風險。今年4月，美國一間初創公司便發生相關事故，其AI代理在執行例行任務時，因沿用失效的API Token而產生「自主規劃」行為，僅在9秒內便刪除了整個生產數據庫及所有備份。事件獲多家科技媒體報導，凸顯AI技術失控所帶來的潛在風險。

因此，建立完善的「AI Agent治理框架」已成為企業不可忽視的課題，關鍵在於需要在效率與安全之間取得平衡。其中一項重要措施是落實「Human-in-the-loop」機制，在關鍵決策節點保留人手覆核，避免演算法完全失控。

AI Agent雖然是推動未來社會轉型的重要引擎，但若缺乏適當治理，亦可能適得其反。HKIRC亦即將推出「安全AI@職場賦能計劃」，提供清晰且可行的管治框架，讓企業在加速應用AI的同時，亦能兼顧穩定與安全。

黃家偉工程師

行政總裁
香港互聯網註冊管理有限公司 (HKIRC)