



代理式 AI 識運用 更要識設限

甚麼是 代理式 AI？

傳統AI根據提示詞作出回應；
代理式AI則能自主規劃並執行
目標。

特點

- 可按目標拆解步驟並逐步執行
- 可連接工具、系統或外部服務
- 可根據結果調整下一步行動



為何風險 更高？

若缺乏安全機制，代理式AI會將
錯誤的答案轉化為錯誤的行動。

1. 權限過大

若其可接觸過多系統、帳戶或
資料，一旦被濫用，將會造成
廣泛影響。

2. 指令或輸入的內容有誤導

惡意內容、假數據或未經信任的
輸入內容，可能導致其作出
錯誤判斷或執行不當行動。

3. 缺乏監察

如果其行動與工作結果無人
覆核，小問題可能一路急速
惡化而不被察覺。

三大常見 風險入口

1. 自動整理資料

風險：引用錯誤資料、資料來源未經信任、內容有誤導。

2. 自動發送或提交

風險：發送的内容、對象出錯，或未經覆核便執行工作。

3. 連接內部工具或系統

風險：接達權限過大、資料外洩、意外操作內部資源。

安全使用 代理式 AI 五大貼士

1. 先設權限，後接工具
2. 高風險操作要經人覆核
3. 輸入來源要驗證
4. 敏感資料要分級處理
5. 保留紀錄，方便追蹤

用得聰明，始於管得清楚

三大提醒

1. 未核實，不信任
2. 未限權，不部署
3. 未批准，不執行

下放工作效率高
自主權限要管好

