

可直接交付的 AI 懶人包

AI Agent 上線前 7 關安全煞車自查表

把資料、工具、權限、人工核准與停機機制一次檢查完

一、七關上線檢核

- 資料已分為公開 / 內部 / 機密；機密與個資不進未核准服務。
- Agent 只看得到任務必需的工具，不提供萬用工具箱。
- 預設唯讀；寫入、刪除、付款、發佈與改權限不得預設開啟。
- 高風險與不可逆動作會先停下，等待指定人員核准。
- 網頁、郵件與外部檔案一律視為不可信資料，不當成系統指令。
- 工具呼叫前會驗證目的地、網址、參數、檔案與輸出內容。
- 已有日誌、速率限制、沙盒測試、kill switch 與回復方式。

二、最小權限矩陣

- READ：預設允許，但只讀核准資料來源。
- WRITE：按任務開啟，限制目的地、欄位與單次數量。
- DELETE / PUBLISH / PAYMENT / PERMISSION：一律要求人工核准。

三、高風險動作核准卡

- 動作：Agent 想做什麼？
- 範圍：會影響哪些帳號、資料或外部對象？
- 證據：輸入、來源與預期輸出是否可追溯？
- 回復：做錯時能否取消、回滾或停機？
- 核准人：誰有權按下最後確認？

來源與使用限制

- 本表依 OWASP LLM06:2025、OWASP Agentic Top 10 2026、NIST AI RMF GenAI Profile，以及台灣行政院與數位發展部公開指引整理。
- 七關為 Wilson AI Lab 實務整理，不是法律、資安認證或官方逐字規範；高風險環境仍應由資安、法務與業務負責人共同審查。

追蹤 Wilson AI Lab，一起學好 AI，把工具真正用進工作與創作。

YouTube : <https://www.youtube.com/@WilsonAILab>

Skool : <https://www.skool.com/wilson-9076>

Instagram : <https://www.instagram.com/wilsonailab/>

Threads : <https://www.threads.com/@wilsonailab>