

金融業運用代理 AI 安全諮詢文件

文件定位：本文件為諮詢文件，旨在就金融機構導入代理 AI 之安全治理、風險分級、控制原則及實務導入方向徵詢各界意見，所列內容作為後續政策研議與文件修訂之參考。本文件不取代現行法令、監理規範、委外管理及金融機構既有內部控制要求。

適用範圍：本文件主要適用於可介接、操作、異動或觸發金融機構營運系統之代理 AI。僅供個人辦公或開發輔助，且於使用者既有權限邊界內運作之個人 AI 助理，例如資訊整理、內容草擬、會議摘要或程式碼建議等，得依既有內部控制及資安管理機制辦理，無須逐項參考本文件。惟如其可自主呼叫工具、跨系統取用資料、對外傳送、寫入系統、觸發正式作業流程，或其用途、資料、工具、權限、自動化程度與使用對象已有重大變更，仍宜重新評估其是否屬本文件適用範圍，並依風險調整控管措施。

諮詢重點：鑑於 AI 與代理 AI 相關技術、應用模式、攻防手法、國際治理要求及資安解決方案持續快速演進，本文件本於風險導向、比例原則及技術中立立場，就下列面向徵詢意見：

- **安全治理（第 3 章）：**責任分工、應用場景列管、安全護欄、監控稽核及生命週期治理。諮詢問題包含所揭治理原則是否完整、是否有可避免過度負擔或強化責任歸屬之作法，或有其他精進建議。
- **風險分級（第 4 章）：**分級方法、上線基準、例外管理及持續重評。諮詢問題包含所揭三軸分級方法是否妥適、是否有其他更具可操作性之分級管理方法，或有其他精進建議。
- **安全控管（第 5 章）：**以零信任架構延伸之安全控制、人工確認或覆核、監控應變，以及可達成同等風險管理目的之替代作法。諮詢問題包含所揭安全設計是否足以控管代理 AI 特有風險、是否可銜接既有資安機制或有可達同等目的之替代作法，或有其他精進建議。

歡迎各界就文件架構、控制措施、導入可行性及其他實務議題提供意見。所提意見宜避免限定特定模型、產品、供應商或單一技術架構，並說明適用之應用場景、風險等級、實務限制、替代作法、技術成熟度、導入與維運所需資源、既有制度銜接方式，以及可驗證之控制效果，作為後續政策研議與文件修訂之參考。

文件使用方式：建議先依第 3 章建立責任分工、應用場景列管、安全護欄、監控及生命週期治理；再依第 4 章，以資料敏感度、動作影響及能力邊界／可擴散性完成風險分級；最後依第 5 章選用與風險相稱之基本控制、進階要求及系統可控性目標。附錄 A、B 分別提供威脅情境與零信任架構對照；低風險或試行應用場景得參考附錄 C 留存第一階簡化紀錄，涉及較高風險、正式上線、提高自動化程度或例外放寬者，得參考附錄 D 辦理第二階治理檢核，或依既有制度採等效作法。

摘要

相較於以生成回覆為主之聊天式 AI，代理 AI 具備規劃步驟、呼叫工具、跨系統存取資料並推進任務之能力。因此，其治理重點不僅在模型輸出品質，更在於應用場景之目的與邊界、責任歸屬、資料與權限邊界、工具使用、監控稽核及異常止血是否可被有效管理。

金融機構宜以「應用場景」(Use Case，指代理 AI 為達成特定業務目的，在明確資料、工具、權限、動作及輸出或寫入邊界內運作之具體場景)為管理單位，先釐清代理可達成之業務目的、可取用資料、可呼叫工具、可執行動作，以及可輸出或寫入之目的地；再依資料敏感度、動作影響、能力邊界／可擴散性及對客戶權益與營運韌性之影響進行風險分級，並依分級結果確認系統可控性要求，採取相稱控管。

在前述定位下，本諮詢文件參照零信任之不預設信任、持續驗證、最小權限及可追溯原則，作為代理 AI 安全控管之設計方向，並將代理 AI 視為具行動能力之非人類身分與工作負載。金融機構宜透過安全駕馭工程(harness engineering)，將代理 AI 之規劃、工具呼叫與跨系統執行能力限制於核准邊界內。導入代理 AI 時，宜以下列原則作為共同基準：

- **銜接既有內控：**代理 AI 介接營運系統時，應以補強既有內控為原則，不得繞過原系統既有授權、職務分工、交易檢核、紀錄留存、覆核及復原機制。
- **風險導向導入：**依應用場景風險分級、系統可控性要求與自動化程度分階段導入，並以風險分級決定上線門檻、覆核要求與監控強度。
- **循零信任原則：**依應用場景風險管理代理身分與權限、執行環境、網路、工具及資料取用，不因身分已驗證、連線已建立或工具可被呼叫，即推定後續行為可信；金融機構宜採取可達成能力邊界管理、持續驗證、異常阻斷、停用、復原及回溯之控制，使代理 AI 持續於核准邊界內運作。
- **保留人為控制：**涉及不可逆、高影響或重大客戶權益之動作，宜依風險等級及既有法令、內控與職務分工要求，保留人工確認或覆核，不宜僅由代理自動執行；並宜具備降級、轉人工處理，以及必要之申訴、更正與模型風險管理機制。
- **確保可控可追溯：**缺乏必要紀錄、無法停用、無法撤銷、無法復原或無法追溯者，不宜正式自動化。
- **管理第三方與供應鏈風險：**涉及第三方模型、平台、工具或服務者，宜納入既有委外與供應鏈治理，依應用場景風險掌握責任分工、資料與工具邊界、變更與事件管理、持續監控及退場安排。第三方提供之資訊或控制能力不足以支持風險管理者，宜採取相稱之補償控制或限縮使用範圍。

一、代理 AI 運作特性

相較於以「生成回覆」為主的聊天式 AI，代理 AI 之核心在於「推進並完成任務」：先依目標規劃步驟，再透過工具呼叫逐步產出可用成果，例如查詢彙整、填表、產生草稿與通知。其常見特性包括目標導向、可規劃、多步執行、工具呼叫、跨資料來源整合，以及維持任務脈絡與狀態。

與傳統自動化工具，例如 RPA（機器人流程自動化）及工作流程引擎相比，傳統自動化多依固定規則與既定流程執行，適用於結構化資料與穩定介面；代理 AI 則可在情境變動時調整步驟以完成任務，較能處理非結構化資訊。實務上可將代理 AI 定位為負責判斷、規劃與任務編排之能力層，並由 RPA 或工作流程引擎承接高度標準化之流程執行，形成分工清楚之自動化流程。

就金融場景而言，代理 AI 常見應用場景包括：法規／合約文件彙整、客服與營運查詢、法遵／AML（防制洗錢）警示初篩、跨系統報表與工單處理等，可降低查找與彙整成本，並使人員更聚焦於判斷、覆核與例外處理。惟因代理可呼叫工具且可能跨系統存取資料，風險亦相對放大，例如目標偏離原始目的、工具誤用或越權、敏感資料外洩，以及多步流程造成錯誤連鎖擴大等。

金融機構導入代理 AI 時，應先釐清其於作業流程中的身分定位與責任邊界。由於代理具備自主規劃並推進任務之特性，實務上可將其視為「非人類身分（Non-Human Identity, NHI，指非自然人但可被賦予身分與權限之服務、代理或工作負載）」的一種工作負載，並以「輔助作業、受控執行」為原則，確保所有代理行為均在可稽核、可中斷之護欄內運作。

二、國際間對代理 AI 安全之因應

代理 AI 可規劃任務、呼叫工具、跨系統取用資料並持續執行，使模型輸出可能直接驅動實際動作或多步流程。國際相關文件已由模型與輸出管理，逐步延伸至治理責任、技能與供應鏈、代理身分與權限、執行環境、網路通訊、執行期政策、資料與應用安全、監控應變及生命週期治理。為研議本文件之風險與控制方向，本章整理相關國際資料如下：

- **審慎導入與責任治理：**五眼聯盟（Five Eyes）《Careful Adoption of Agentic AI Services》¹主張將代理 AI 納入既有資安與風險治理，由低風險、非敏感且可觀測之場景起步，再隨控制成熟度逐步開放資料、工具及自動執行能力，並明確管理責任歸屬。
- **風險分類與測試：**OWASP《Top 10 for Agentic Applications for 2026》²提供代理應用整體風險分類；CSA《Agentic AI Red Teaming Guide》³與 MITRE ATLAS⁴可進一步支援代理 workflow、權限、記憶、編排、供應鏈及攻擊情境之測試與偵測規則設計。
- **技能與供應鏈治理：**OWASP《Agentic Skills Top 10》⁵聚焦代理技能之惡意內容、來源、權限、隔離、版本及跨平台重用風險。金融機構可據以建立技能清冊、安裝與變更核可、發布者及完整性驗證、權限審查、沙箱隔離、版本固定、持續掃描、監控及撤銷；未經核可之技能不宜載入正式代理或接觸敏感資料與高影響工具。
- **身分與權杖防護：**NIST IR 8587《Protecting Tokens and Assertions from Forgery, Theft, and Misuse: Implementation Recommendations for Agencies and Cloud Service Providers》⁶針對權杖與身分聲明之偽造、竊取及濫用，提出金鑰保護、權杖驗證、生命週期控制、撤銷與持續監控等實作建議，適用於代理 AI 及其他非人類身分存取系統、資料、工具與 API 之情境。
- **零信任治理與行為管理：**CSA《The Agentic Trust Framework: Zero Trust Governance for AI Agents》⁷依代理自主程度逐步授權，並以工作負載及執行環境之持續驗證、隔離、行為管理、阻斷與復原，降低代理於不可信或失控環境執行所生風險。

¹ <https://www.cisa.gov/resources-tools/resources/careful-adoption-agentic-ai-services/>

² <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>

³ <https://cloudsecurityalliance.org/artifacts/agentic-ai-red-teaming-guide/>

⁴ <https://atlas.mitre.org/>

⁵ <https://owasp.org/projects/agentic-skills-top-10/>

⁶ <https://nvlpubs.nist.gov/nistpubs/ir/2026/NIST.IR.8587.pdf/>

⁷ <https://cloudsecurityalliance.org/blog/2026/02/02/the-agentic-trust-framework-zero-trust-governance-for-ai-agents/>

- **網路分段與通訊防護：**CSA《Zero Trust Microsegmentation Guidance》⁸提供依實際通訊需求限制相互連線、管控網路出口與目的地、降低非必要可達性及橫向擴散之參考，可支援代理跨系統、跨工具與多代理通訊之分段隔離。
- **執行期政策控制：**OWASP《Agent Control Standard (ACS)》⁹提出標準化的執行攔截點、可供不同代理平台共用的政策規則，以及執行期間的即時強制控管，使工具呼叫、資料取用、外送、寫入、升權及其他高影響動作，可於執行前後統一檢查並採允許、拒絕、修改、降級或轉人工覆核；可作為共用控制平面、AI/MCP Gateway 或等效政策控制點之技術參考。
- **可驗證安全要求：**OWASP《Artificial Intelligence Security Verification Standard (AISVS)》¹⁰將人工智慧安全轉為可測試、可驗證之技術要求，可供資料來源、輸入輸出、記憶、RAG／知識庫、資料流向與保護措施之架構設計、開發測試、委外採購、上線驗收及獨立驗證選用。
- **協作監控與事件應變：**CISA JCDC《AI Cybersecurity Collaboration Playbook》¹¹補充事件資訊分享、共同分析、通報與協作應變，可供代理行為、工具呼叫、資料流向、政策判斷與處置結果之跨系統關聯，以及重大供應鏈或跨機構事件之聯防參考。
- **生命週期治理：**新加坡網路安全局之《Securing Agentic AI – An Addendum to the Guidelines and Companion Guide on Securing AI Systems》¹²，補充依代理能力辨識風險，以及導入、監控、變更、阻斷與復原等生命週期控制之實務參考。

綜合上述參考文件，代理 AI 安全之共同方向，是將代理納入既有治理與風險管理，並依零信任之不預設信任、最小權限與持續驗證原則，以應用場景及能力邊界限制其身分、權限、技能、工具、資料、通訊與執行環境；於任務執行期間持續檢查政策條件、監控實際行為，並保留必要之阻斷、降級、復原與事件應變能力。本文件據此形成治理架構、風險分級及以零信任架構延伸之安全控管內容。

⁸ <https://cloudsecurityalliance.org/artifacts/zero-trust-microsegmentation-guidance/>

⁹ <https://genai.owasp.org/resource/agent-control-standard-acs/>

¹⁰ <https://owasp.org/projects/artificial-intelligence-security-verification-standard-aisvs-docs/>

¹¹ <https://www.cisa.gov/resources-tools/resources/ai-cybersecurity-collaboration-playbook/>

¹² <https://www.csa.gov.sg/resources/publications/addendum-on-securing-ai-systems/>

三、代理 AI 安全治理

代理 AI 安全治理宜先釐清業務應用場景、責任邊界、資料與工具範圍、可執行動作、輸出或寫入目的地，以及異常時之處置、停用與追溯方式；再依風險高低採行覆核、監控、政策檢查、降級、停用與復原等控制，並逐步將共用控制能力沉澱為平台層機制。

3.1 責任分工

責任分工宜在應用場景層級釐清需求提出、核可、維運與查核權責，避免代理 AI 跨系統與多步流程造成治理缺口。因應代理涉及工具呼叫與跨系統資料存取，建議至少涵蓋下列角色：

- **業務營運**：擔任應用場景業務負責人或應用場景所有人，定義應用場景目的、使用範圍、流程邊界、預期效益與效益指標，提出上線、調整、持續使用或退場需求，並負責人工流程調整、業務結果與相關風險管理責任。
- **平台治理／維運**：負責提供代理 AI 共用技術能力，並將應用場景清冊、工具與資料邊界、權限框架、風險分級、上線門檻及安控要求，轉化為平台與系統層控制，維護代理執行環境、部署基線、網路出口、工具連接器、監控、版本管理、生命週期管理、停用、降級、切回人工流程與安全替代流程。
- **資訊安全／SOC（資安監控中心）**：負責代理行為監控、告警規則、威脅情境測試、事件處置、止血演練與資安通報，並確認高風險應用場景具備可偵測、可阻斷、可復原之能力。
- **資料治理／法遵**：負責資料分類、個資與法遵要求、資料最小揭露、外送限制、留存刪除與客戶權益影響評估，並就高敏感資料或受監理保護資訊提供核可條件。
- **客戶權益／公平待客**：對涉及授信、投資建議、保險、費用爭議、申訴處理、差別待遇或客戶權益影響之應用場景，宜納入公平待客、透明揭露與客戶權益影響評估，避免代理輸出或自動化流程造成不當差別待遇、誤導說明或難以申訴之結果。
- **內部稽核**：提供獨立查核與確信，確認核可、紀錄留存與控制運作之有效性，並追蹤改善。

代理 AI 不宜取代法定或內部授權責任人之判斷、核可或覆核責任；凡涉及人員核可、客戶確認、主管覆核、法遵判斷、高影響資安處置或其他內控要求之事項，仍宜由相應業務、管理、核可、覆核或處置角色負責。代理 AI 之輸出、建議、參數產生或工具呼叫結果，應依風險及核准條件作為輔助資訊、草稿或受控執行內容；涉及正式紀錄、寫入、外送或流程推進者，仍應保留相應核可、覆核及責任歸屬，並宜能對應至原核可、覆核、營運、

資安或法遵責任角色。不宜以代理 AI 已產出、建議或執行為由，免除人員審慎義務、覆核責任、內控責任或客戶權益保護義務。

3.2 高階治理與董事會監督

對高風險或具重大業務影響之代理 AI 應用場景，金融機構宜建立高階治理與定期報告機制，使高階管理階層或董事會掌握風險輪廓、控制成熟度與重大例外，例如高風險應用場景數量與變動、重大例外、阻斷與覆核異常、停用、降級與切回人工能力、第三方事件及未結改善事項等關鍵指標。報告內容宜能支持是否核准上線、提高自動化程度、延展例外、限縮使用範圍或退場等決策，並明確記錄決策結果與後續追蹤責任。

3.3 應用場景列管

金融機構宜以應用場景為主要管理單位，並視風險掌握其所使用之代理、技能、工具及連接器，以及彼此之相依、授權與使用關係。列管內容以應用場景目的、責任人、風險等級、資料來源、技能與工具、可執行動作、目的地、自動化程度及停用方式為核心；模型風險、獨立驗證、監控、例外及其他治理要求，則依應用場景風險確認其適用性，並由相關管理程序辦理及留存紀錄。

通用型代理 AI 工具不宜僅以工具名稱概括列管，亦無須將每次低風險輔助任務均視為完整應用場景；金融機構宜依實際資料、工具、權限、工作區、可執行動作及對正式作業之影響，決定列管粒度。同一技能或工具跨代理、跨應用場景或跨平台重用時，應於相關應用場景中相互參照，並依風險重新確認其適用範圍。

3.4 風險導向

代理 AI 治理宜以應用場景為單位採風險導向，依資料敏感度、動作影響及能力邊界／可擴散性判定風險等級，並綜合考量客戶權益、核心或高重要性系統及營運韌性影響；再依分級結果與自動化程度，決定核可門檻、系統可控性、控制深度、覆核及例外管理方式。資料、工具、目的地、權限或自動化程度發生重大變更時，應重新評估風險分級與控管強度。

應用場景立案前，宜先確認業務問題、預期效益及替代方案；如可由制度調整、規則系統、流程引擎或 RPA 等較低自主性方式達成，宜優先採用。確有需要代理 AI 之規劃、工具呼叫、跨系統整合或流程推進能力時，再辦理應用場景立案、風險分級與相稱控管。

涉及外送、寫入、批次、跨系統、交易、權限、高影響資安處置、關鍵營運或重大客戶權益者，應提高核可、監控、覆核、阻斷與復原要求；高風險、不可逆或高影響動作不得僅

由代理自動執行。僅提供查詢、摘要、草稿或預覽，且不具外送、寫入、狀態變更或流程觸發能力者，得依風險簡化管理。

3.5 安全護欄

安全護欄宜將風險分級轉化為可執行之限制、核可與阻斷機制，使代理 AI 於核准邊界內運作。代理 AI 任務從接收內容、選用工具、執行動作至影響營運系統，宜依序設置模型與內容、工具與介接、執行與動作，以及介接系統安全控制等四個控制點；各控制點分別負責防止不可信內容驅動任務、未核准工具或介接被使用、高影響動作未經檢查即執行，以及代理繞過營運系統既有安控。金融機構得依既有架構、應用場景風險及自動化程度，於四個控制點採取相稱或等效作法：

- **模型與內容**：檢查提示、輸入、輸出、工具回傳及外部內容，避免未驗證內容驅動後續動作、寫入記憶或進入 RAG／正式知識庫。
- **工具與介接**：管理模型、工具、連接器、MCP Server、技能及外部服務之來源、授權、用途、介面、目的地、版本、日誌與停用，並透過適當閘道或等效控制點落實，使可供代理使用之能力元件均經核准且可被識別、限制與撤回。
- **執行與動作**：限制寫入、外送、刪除、升權、交易及批次等高影響動作；執行前及必要之執行後，宜依應用場景、身分、資料、目的地、參數及核可狀態進行檢查，並依政策結果採允許、限制或修改參數後放行、拒絕、降級或轉人工覆核。涉及高影響或不可逆動作時，應經政策檢查或人工確認。
- **介接系統安全控制**：代理介接營運系統時，應遵循原系統之授權、交易檢核、覆核、紀錄及復原要求，不得繞過既有控制直接執行交易、異動資料或推進正式作業流程。

技術或平台僅為控制載體。較高風險或跨系統應用場景，宜建立共用控制平面或等效控制，整合政策判斷、存取管理、日誌、阻斷、停用與既有資安機制，並採高可用性、備援及安全失效設計。政策引擎、閘道、授權或驗證服務逾時、不可用或無法取得明確結果時，高風險動作原則上應拒絕、降級或轉人工處理，不應因控制服務失效而直接放行；復原後並宜確認待處理任務、政策版本及執行狀態，避免重複或未授權執行。

尚未具備統一控制平面者，得運用既有 API Gateway、IAM/PAM、DLP/CASB 及變更管理機制達成等效控制；必要控制不足時，應限於模擬、唯讀、草稿、預覽或人工輔助模式。

3.6 監控與稽核

監控與稽核宜貫穿各項安全護欄，並在最小必要留存原則下，保有足以還原任務流程、責任歸屬、工具呼叫、資料流向、政策判斷、核可及處置結果之稽核軌跡。

監控規則宜依應用場景風險設定必要日誌、告警門檻、通報與處置方式，並偵測越權存取、異常工具鏈、非核准外送、高頻重試、覆核異常或高風險動作缺少必要確認等情形。發現異常時，宜依風險採告警、節流、阻斷、降級、停用、撤銷授權、復原或轉人工處理，並銜接 SIEM、SOC、SOAR 及既有事件應變與稽核機制。

代理越權、工具濫用、異常外送、記憶或知識庫污染、行為漂移及連接器異常等，宜納入既有事件分類與通報流程。金融機構並宜依事件、演練、誤報漏報及應用場景變更定期檢討監控規則，將結果回饋風險分級、控制條件、例外及自動化程度；重大第三方、供應鏈或跨機構事件，應依既有法令、契約及協議辦理外部通報與協作應變。

3.7 生命週期治理

代理 AI 治理不宜僅於上線前一次性審查，而應採零信任原則，依代理自主程度及應用場景風險逐步授權，落實最小權限、持續驗證及必要之分段隔離；並依規劃與立案、開發與測試、試行與上線、營運監控，以及變更、例外與退場五項生命週期重點辦理最低限度檢核，檢核深度得依應用場景風險、自動化程度及系統可控性調整。

- **規劃與立案：**確認應用場景目的、責任分工、風險分級、資料與工具邊界、核可路徑及替代方案評估。
- **開發與測試：**將提示、工具、連接器、RAG（檢索增強生成）來源、政策與部署版本納入變更管理，並完成提示注入、工具濫用、資料外洩與復原測試。對較高風險應用場景，可依工具成熟度逐步建立正常工具使用與流程運作基準，作為上線後偵測異常工具鏈與行為偏離之參考。
- **試行與上線：**依風險等級決定唯讀、草稿、半自動或正式執行範圍；高風險應用場景應具執行前政策檢查、覆核、監控、停用、撤銷、降級、回復安全狀態或安全替代能力。
- **營運監控：**持續監控代理身分、工具呼叫、資料來源、目的地、覆核品質、異常工具鏈與外送行為，並定期盤點權限、例外與目的地白名單。
- **變更、例外與退場：**模型、提示、工具、資料來源、目的地、權限或自動化程度重大變更時宜重新分級；應用場景退場或例外到期時，宜撤銷權限與憑證、停用代理及相

關 workflow、移除非必要連接器與目的地、處理記憶、快取、暫存資料及知識庫索引，
確認第三方留存與刪除安排，並保留必要稽核證據。

四、代理 AI 風險分級

本章承接第 3 章風險導向治理要求，提供代理 AI 應用場景之示意性風險分級方法，供金融機構評估核可門檻、試行或上線方式、覆核要求及第 5 章控制深度時參考。L1 至 L4 為本文件提供之風險分級參考，金融機構得依既有風險管理制度及實務需要調整運用。

風險分級不宜僅以代理平台或模型名稱判定，亦非一次性認定；當資料範圍、工具能力、目的地、寫入或外送能力、自動化程度等發生重大變更時，宜重新檢視分級。

4.1 分級流程（建議步驟）

建議以「應用場景」為單位，依下列步驟完成分級：

1. **定義邊界：**確認應用場景之使用目的、使用者、資料來源、工具、輸出／寫入目的地、可執行動作，以及是否涉及外送、批次、自動重試、跨系統或委派，確認該應用場景之評估範疇。
2. **三軸評估：**依資料敏感度、動作影響及能力邊界／可擴散性，判斷應用場景可能接觸之資料、可能造成之結果，以及其自動化、批次、重試、跨系統或委派能力是否可能擴大影響；並依業務情境、影響對象、可復原性及對客戶權益、營運韌性之影響，校準風險等級。
3. **判定等級：**採就高不就低原則，三項面向任一符合較高風險者，即以該等級作為應用場景風險等級。能力邊界應按技能、工具、連接器、代理委派及工作流程可形成之最高組合能力判定，不得因元件分開評估或拆分應用場景而低估整體流程風險。

4.2 分級定義（示意：L1 – L4）

分級判定宜留存各軸判斷依據、引用之資料分類或業務影響標準、最高風險觸發因素、校準理由、評估人員、核可人員、評估日期及下次重評條件。若採拆分應用場景方式列管，並宜記錄各應用場景間之資料流、工具委派與相依關係，避免分拆後低估整體流程風險。

下表以 L1 至 L4 示意分級，分級結果係作為第 5 章控制措施適用範圍與落實強度之依據。L1／L2 應用場景原則上適用基本控制措施，並視資料敏感度、外送或寫入情形及實際作業情境補充必要監控或覆核；L3／L4 應用場景則宜進一步檢視進階要求，確認系統具備與風險相稱之可控性。應用場景之資料範圍、可執行動作、能力邊界或其他核准邊界發生重大變更時，宜重新評估。

等級	資料敏感度	動作影響	能力邊界／可擴散性
L1 (低)	公開資訊、一般非敏感內部資訊，或不含可識別個人與受保護金融資料之測試資料。	僅產生低影響輔助輸出，不形成正式紀錄、系統異動或流程推進效果。	任務單純、步驟有限，無跨客戶、跨系統或自動擴散能力。
L2 (中低)	一般內部資料、少量個資或低影響營運資料，且應用場景明確、影響對象有限。	產出查詢、摘要、彙整、草稿或預覽等輔助結果；如涉及內部流程，僅得由人員依既有程序採用或另行執行，不由代理直接形成正式紀錄、寫入系統、變更狀態或推進正式流程。	可含多步查詢或彙整，但任務範圍、查詢對象與輸出用途明確，不得自行擴大任務。
L3 (中高)	敏感資料、跨來源整合資料、較大量客戶或營運資料；或資料使用情境涉及客戶互動、法遵義務或較高營運影響。	可能產生受控外送、正式紀錄、正式工單、非核心系統異動或流程推進效果，且其業務情境、影響對象、可復原性或覆核要求顯示需較高管者。	可多步、批次或跨系統執行，但宜受限於明確資料範圍、目的地、參數及最大影響範圍。
L4 (高)	高度敏感、高集中度或高價值資料，或資料使用情境可能對客戶權益、法遵義務或營運韌性造成重大影響者。	可能影響核心金融業務、核心或高重要性系統、交易或付款、額度或權限、不可逆或難以復原之處置、法遵最終處置、重大客戶權益或營運韌性。	具升權、委派、跨系統、批次、持續執行或高影響動作能力；錯誤或濫用可能迅速擴大影響。

4.3 正式上線基準與例外管理

涉及正式作業流程之自動或半自動執行者，建議至少具備下列條件；如相關條件尚未完備，宜先採模擬、影子運行、草稿、預覽或由人員逐項執行等輔助模式：

- **正式上線基準：**應用場景正式使用前，應確認資料範圍、工具、目的地、可執行動作、覆核方式、責任歸屬及必要紀錄均與核准條件一致，並依風險分級確認具備第 5 章相稱之監控、停用、撤銷、復原、追溯及異常處理能力。

- **試行與韌性驗證：**較高風險或具流程推進能力之應用場景，宜於正式上線前先採模擬、影子運行或限定範圍試行，確認代理 AI 於異常情境下具備降級、阻斷、停用、復原或切回人工流程之能力；試行結果宜作為正式上線、延後上線或限縮使用範圍之依據。
- **上線前測試：**高風險、具流程推進能力，或涉及外送、寫入、批次、自動重試、跨系統或委派之應用場景，應於正式上線前完成提示注入、工具濫用、權限承接、記憶污染、代理間通訊、HITL 繞過、異常外送及批次擴散等安全測試，並明確定義測試情境、通過條件、必要紀錄與修正追蹤方式。
- **例外與補償控制：**例外宜限於最小範圍、明確期間、補償控制、到期檢視及可供稽核之軌跡；補償控制僅作過渡安排，並宜有改善計畫或轉正路徑。
- **正式自動化限制：**例外或補償控制不得造成日誌缺漏、事件無法查證或還原、繞過目的地管控，或喪失停用、撤銷、復原及人工接手能力；未完成上線前測試、資料邊界確認，或缺少必要覆核、更正與申訴機制者，不宜正式自動或半自動執行。L4 高風險動作不得在缺乏必要授權、覆核或阻斷機制下自動執行。

4.4 分級運用與持續重評

金融機構宜將分級結果納入應用場景生命週期管理，作為核可、試行或上線方式、控制深度、監督、調整與查核之共同語言。應用場景風險等級依資料敏感度、動作影響及能力邊界／可擴散性判定，不因控制措施已建置而降低；控制成熟度則用以判斷該風險等級之應用場景是否具備正式上線、提高自動化程度或擴大使用範圍之條件。金融機構應透過定期盤點及重大變更重評，使風險分級、控制要求與實際運作狀態保持一致。

- **應用場景變更重評：**應用場景之資料來源、工具白名單、目的地、權限、自動化程度、外送、寫入、批次、自動觸發或流程推進能力重大變更時，宜重新評估風險分級與系統可控性。
- **模型與工具版本變更：**底層模型、提示詞或關鍵工具重大更新時，宜依風險辦理回歸測試或等效驗證；L1／L2 應用場景之微幅版本、提示詞或核准範圍內參照資料更新，且未擴大資料、工具或動作能力者，得以簡化紀錄辦理。
- **代理漂移檢視：**金融機構宜定期檢視代理數量、版本、用途、責任人、權限、工具、連接器、資料來源、提示與自動化流程，避免代理蔓延、重複建置、測試代理長期留存、舊版代理未退場，或代理權限與工具設定逐步偏離原核准範圍。

五、代理 AI 安全控管

安全控管原則：鑑於 AI 安全框架、技術實作、平台架構與攻防手法仍快速演進，本章所列安全控管面向、控制項及系統可控性目標，係供金融機構辦理應用場景盤點、風險分級、控制設計及檢核對照之參考。金融機構得依既有架構、應用場景風險、自動化程度、系統可控性要求及技術成熟度，採取相稱、等效或更新作法；控制設計宜以代理 AI 行為之可觀測、可限制、可阻斷、可復原及可回溯為核心，而非逐項套用固定技術要求。

代理 AI 具備規劃任務、呼叫工具、跨系統取用資料並推進流程之能力，實務上宜視為具行動能力之非人類身分與工作負載，其安全控管與零信任所強調之不預設信任、最小權限、持續驗證、明確資源邊界及完整可視性高度契合。為使本章控制架構與金融業既有資安推動方向銜接，本章參照金融業零信任架構之身分、設備／執行環境、網路、應用與工作負載及資料等面向，並以行為控制及可視性與分析作為貫穿前述面向之橫向能力，延伸並補強為適用於代理 AI 應用場景之安全控管架構（威脅情境與控制架構對照詳附錄 A 及附錄 B）。上述控制可視為代理 AI 之安全駕馭工程（harness engineering），將代理之規劃、工具呼叫與跨系統執行能力限制於核准邊界內。

本章依代理 AI 執行任務時之控制結構，分別由第 5.1 至 5.5 節就身分、設備／執行環境、網路、應用與工作負載及資料建立控制條件；第 5.6 節行為控制及第 5.7 節可視性與分析則作為橫向能力，貫穿前述各控制面向。行為控制綜合各項條件，持續判斷代理於特定任務及工作階段中得否執行具體行為，並運用自動化與協調能力支援跨控制面向之政策執行與控制協同運作；可視性與分析則透過跨系統日誌、關聯分析、異常偵測及事件回饋，驗證控制是否持續有效並支援調整。

為利引用，本章基本控制措施及進階要求採「控制面向序號一項次」編號，例如身分為 1-1、設備／執行環境為 2-1。金融機構得依風險特性調整控制項內容與落實方式。

5.1 身分（Identity）

本節管理代理任務之身分與授權鏈，依序確認誰觸發任務、代理代表何者執行、得承接或委派哪些權限，以及如何撤銷權限並確認責任歸屬。基本控制聚焦身分識別、權限承接、任務授權、委派、退場及身分與授權紀錄；進階要求再強化憑證與權杖管理、短期授權，以及身分條件不足時之限制。主要控制點包括 IAM／PAM、代理與工作負載身分、委派授權及憑證管理。

A. 實務導入挑戰：

- **身分、授權與憑證管理：**代理任務可能同時涉及觸發者、代理身分、代理所代表之服務或系統身分、技能身分、後端帳號及存取憑證，並在多步或跨系統流程中持續委派。若未清楚界定各身分間之代表、授權與責任關係，可能造成權限不當承接、委派範圍擴大、憑證遭留存或重用，以及人員異動、應用場景退場或例外到期後權限未及時撤銷。

B. 基本控制措施：

- **1-1 身分唯一識別：**使用者、代理，以及代理所代表之系統或服務，應具可個別識別之身分，並能對應實際執行責任。
- **1-2 身分與憑證分離使用：**代理不得共用帳號、冒用個人帳號，或混用其他代理、服務或系統身分與憑證；宜使用專屬、可識別且可撤銷之服務身分或工作負載身分。
- **1-3 任務授權邊界：**代理取得或使用權限時，應以核准應用場景、任務目的及必要操作為邊界，限制可代表之身分、授權範圍及有效期間；不得僅因使用者已登入，即承接其完整權限或執行超出本次任務之操作。
- **1-4 技能身分與授權承接：**代理載入或呼叫技能時，應以核准之代理或工作負載身分及授權範圍執行，不得承接觸發者或代理之完整權限；涉及轉交其他代理或跨平台執行者，並應重新確認執行身分與授權範圍。
- **1-5 委派關係管理：**代理將任務或權限委派予其他代理、技能、服務或系統時，應明確記錄委派者、受委派者、委派目的、授權範圍、有效期間及核可依據，並可對應至應用場景與責任人；不得以連續委派擴大原授權範圍。
- **1-6 代理退場與撤銷：**人員離職、職務異動、應用場景退場或例外到期時，宜一併撤銷相關代理權限、Token 與自動化工作流，避免孤兒代理長期持有跨系統權限。
- **1-7 身分與授權紀錄：**身分與授權來源系統應記錄觸發者、代理身分、所代表之服務或系統身分，以及實際使用之授權或憑證；涉及跨系統、多代理或高風險任務時，並宜記錄工作階段、授權範圍、有效期間及核可結果，供跨系統關聯與查證。

C. 進階要求：

- **1-8 憑證與密鑰強化管理：**代理使用之 API Key、OAuth Token、工作負載身分、密鑰或其他存取憑證，宜納入集中保管、依任務需要核發、最小權限、定期輪替、使用紀錄及即時撤銷管理；不得寫入提示、程式碼、日誌、記憶或未受控設定檔。
- **1-9 權杖與身分聲明驗證：**代理或系統接收權杖或身分聲明時，宜驗證簽章、簽發者、適用對象、有效期間、授權範圍及撤銷狀態；驗證失敗、資料不足或超出核准用途者，不應放行。
- **1-10 任務型短期授權：**涉及升權、跨系統委派、多代理協作、交易、高影響資安處置

或重大客戶權益者，宜於執行前重新確認本次任務符合核准應用場景與授權必要性，並採限於本次任務或工作階段、範圍受限且可隨時撤銷之短期授權；授權與委派來源應留存紀錄。

- **1-11 身分條件不足時之限制：**若無法區分觸發者、代理身分與所代表之服務或系統身分，無法確認授權或委派依據、有效範圍及撤銷狀態，或無法將行為關聯至責任人與核可紀錄，不宜開放正式自動或半自動執行；既有能力應降級為唯讀、草稿或人工輔助模式。

D. 系統可控性目標：

- **可觀測：**應可識別觸發者、代理身分、授權依據及責任歸屬。
- **可限制：**應可依應用場景、任務目的、代理或工作負載身分及委派關係，限制授權範圍、有效期間與權限承接。
- **可阻斷：**身分驗證失敗、權杖或憑證異常、授權逾期、撤銷未生效、越權或異常委派時，應可拒絕存取、停止權限承接或轉人工處理。
- **可復原：**應可撤銷權限與憑證、停用代理，並回復至安全狀態。
- **可回溯：**身分、授權與委派紀錄應可對應至應用場景、責任人及核可依據。

5.2 設備／執行環境（Device / Runtime）

本節管理代理是否在健康、受管且可隔離之環境執行，並涵蓋執行位置、健康與納管狀態、部署區隔、程序與資源邊界，以及異常時之中止、隔離、降級與環境重建。基本控制先建立環境分級與部署區隔、環境識別、健康驗證及異常中止與隔離；進階要求再強化技術隔離、程式與指令限制、環境資源限制、動態條件、復原驗證，以及環境條件不足時之限制。主要控制點包括主機、容器、桌面、自動化環境及工作負載平台。

A. 實務導入挑戰：

- **執行環境不可控與影響擴散：**代理若在健康狀態不明、隔離不足、程序未受限制或資源邊界不明的環境執行，非預期程式碼、異常程序或資源耗用可能擴大影響，並使異常偵測、停用及環境復原能力失效。

B. 基本控制措施：

- **2-1 執行環境分級與部署區隔：**宜依生命週期階段及應用場景風險分級管理執行環境。正式或高風險代理不宜與一般使用者桌面、高權限系統或關鍵營運環境混用；低風險、唯讀、摘要、草稿或個人輔助應用場景，得於受管理之使用者桌面環境執行。

- **2-2 執行環境基準與識別：**宜掌握代理部署位置、執行主機或容器、環境識別、核准組態、端點防護、可用資源、可接觸系統及停用方式，作為健康檢查、變更管理與事件查證之基準。
- **2-3 執行環境健康驗證：**代理執行環境應於上線、重大變更、重啟或執行高風險任務前，確認修補、端點防護、組態基準、監控及必要隔離機制正常；不符條件者，宜限制、暫停、隔離或降級執行。
- **2-4 執行環境異常中止與隔離：**發生異常程序執行、環境健康狀態不符、端點防護失效或資源耗用異常時，宜能及時降級、中止、隔離或停用受影響之工作負載。

C. 進階要求：

- **2-5 執行環境隔離與資源邊界：**正式或高風險代理宜採沙箱、容器、隔離桌面、專屬執行節點或等效機制，限制檔案系統、程序、網路、憑證與資源之可達範圍；短期未具備完整隔離能力者，應限縮可接觸系統及資料，並降低自動化程度。
- **2-6 程式與指令執行限制：**執行環境宜限制未經核准之程式碼、腳本、SQL、系統指令、程序啟動及高權限操作，並依風險採唯讀檔案系統、受限程序、命令白名單、資源配額或禁止執行。
- **2-7 執行環境資源限制：**執行環境宜依風險設定運算、記憶體、儲存、程序及工作負載資源配額，並監控資源耗用與異常占用；達到環境資源門檻時，宜告警、限制資源、隔離或中止受影響之工作負載。
- **2-8 執行環境動態條件：**對高風險或提高自動化程度之應用場景，宜持續取得執行位置、健康、納管、隔離及資源狀態，提供工具呼叫或高影響動作之即時政策判斷；條件不符時，應回傳可供拒絕、降級或轉人工判斷之狀態。
- **2-9 環境復原與切換驗證：**高風險應用場景宜具備重建或回復核准組態、隔離受影響環境、切換至安全執行環境或人工流程之能力，並依風險定期驗證復原所需時間、狀態一致性及待處理任務之處置。
- **2-10 執行環境條件不足時之限制：**若無法確認執行位置、健康、納管或隔離狀態，無法限制程序與資源，或無法於異常時中止、隔離及回復安全狀態，不宜正式自動或半自動執行；既有能力應限於模擬、唯讀、草稿或人工輔助模式。

D. 系統可控性目標：

- **可觀測：**應可掌握執行位置、環境健康、端點防護、組態偏離、程序與資源使用狀態。
- **可限制：**應可限制程序與指令之執行，並控管運算、記憶體、儲存、程序及工作負載資源用量。

- **可阻斷**：異常程序執行、資源耗用異常或執行環境失效時，應可停止、隔離或降級受影響之工作負載。
- **可復原**：應可將執行環境之組態與資源回復至核准狀態，並重新建立安全執行環境。
- **可回溯**：部署、健康檢查、變更、停用及復原紀錄應可供查證。

5.3 網路（Network）

本節管理代理可與哪些對象、經何種路徑及通道通訊，並限制連線路徑、出口、端點、通道與目的地，驗證跨代理或跨系統訊息之來源及完整性。基本控制先建立通訊邊界、訊息驗證及異常處置；進階要求再強化網路微分段、高風險通訊與傳輸控管，以及網路條件不足時之限制。主要控制點包括網路出口、API Gateway、DLP／CASB 及跨系統通訊機制。

A. 實務導入挑戰：

- **外連與跨系統通訊風險**：代理可能透過郵件、雲端硬碟、通訊平台、API 或下游工具連線、傳輸或外送，繞過核准出口與目的地限制；代理與其他代理、工具或系統間傳遞之指令、狀態或核可訊息，如未確認來源、完整性及工作階段關聯，也可能遭偽造、竄改、重放或錯送。

B. 基本控制措施：

- **3-1 通訊邊界與路徑管制**：應依核准應用場景，界定代理可連線之來源、目的地、通訊通道及方向；相關通訊應經核准出口、閘道或其他控制點，並銜接既有網路及外送管制，保留必要之連線與傳輸紀錄。代理不得自行建立或繞行核准範圍外之連線。
- **3-2 跨代理與跨系統訊息驗證**：代理與其他代理、工具、系統或自動化流程間傳遞之指令、狀態與核可訊息，應可識別來源、傳送對象及所屬工作階段，並驗證訊息完整性及核可資訊未遭竄改；來源不明、驗證失敗或工作階段不一致之訊息，不應接受或傳遞。
- **3-3 異常通訊處置**：對非核准目的地、繞行核准出口、來源或工作階段驗證失敗、異常上傳或大量傳輸等情形，應依風險採告警、限速、阻斷、隔離或中止。

C. 進階要求：

- **3-4 網路微分段**：較高風險、跨系統或多代理應用場景，宜依資料、系統與服務之實際通訊需求限制相互連線，避免非必要可達性與橫向擴散；相關政策應納入變更、例外及監控管理。

- **3-5 高風險通訊安全：**涉及多代理協作、跨系統流程，或執行寫入、批次、升權及高影響資安處置者，宜依風險採相互驗證、傳輸加密、訊息簽章、防重放、結構化訊息及工作階段綁定等控制，並確認通訊雙方、訊息序號與核可狀態一致。
- **3-6 高風險外送通道與傳輸限制：**涉及外部、第三方、客戶、跨境、批次或大量外送者，宜依風險設定專用出口或受控通道、目的地白名單、傳輸方向、單次與累計傳輸量、速率及時段限制，並強化傳輸紀錄與工作階段關聯；達到傳輸門檻或違反外送條件時，應可暫停或中止傳輸。
- **3-7 網路條件不足時之限制：**若無法確認來源、目的地、通道、工作階段、核可狀態或必要紀錄，或無法限制路徑、出口及非核准連線，不宜正式外送、跨系統傳輸或自動推進流程；既有能力應限於內部、唯讀、草稿或人工轉送模式。

D. 系統可控性目標：

- **可觀測：**應可識別並記錄外連、外送及跨系統傳輸。
- **可限制：**連線路徑、出口、目的地、端點及通訊通道應可限於核准範圍。
- **可阻斷：**應可阻擋非核准目的地、異常外送或繞行出口之流向，或中止傳輸。
- **可復原：**應可撤銷或回復網路政策、路由及目的地設定，使連線回復至核准狀態。
- **可回溯：**連線與傳輸紀錄應可查證來源、目的地、通道、時間及阻斷結果。

5.4 應用與工作負載（Application & Workload）

本節管理代理應用、工具、技能、連接器及工作負載本身之安全。基本控制聚焦工具與介接清冊、工具輸入輸出檢查及營運系統安控銜接；進階要求再強化結構化介面、第三方與技能治理及應用與部署變更管理。本節確認可供代理使用之能力元件與介面是否可信、完整、有效及受管理。

A. 實務導入挑戰：

- **能力元件、輸入輸出與系統介接風險：**工具、技能、連接器、版本或政策設定如來源不明、遭竄改、失效或未受管理，可能使未核准能力被載入或提供不正確結果；模型輸出、工具回傳或外部內容如格式錯誤、欄位不完整、型別或允許值不符，也可能造成工具或系統介面處理錯誤。代理介接營運系統時，如未適用原系統既有授權、職務分工、交易檢核、覆核、紀錄及復原機制，亦可能繞過既有內控。

B. 基本控制措施：

- **4-1 工具與介接清冊：**工具、連接器、MCP Server 或等效工具服務，以及其可接觸系統、輸入輸出範圍、核准用途、版本與停用方式，應納入清冊；代理僅得使用清冊中有效且符合核准用途之項目，已停用、版本失效或超出核准用途者不得使用。
- **4-2 工具輸入輸出檢查：**模型輸出、工具回傳及外部內容如將提供工具、連接器或營運系統使用，應檢查格式、完整性、必要欄位、資料型別、允許值及介面規格；不符合介面要求、無法解析或無法確認處理結果者，不應作為有效輸入或可信執行結果。
- **4-3 營運系統安控銜接：**代理 AI 介接營運系統時，宜銜接原系統既有之授權、職務分工、交易檢核、覆核、紀錄留存及復原機制；代理層控制僅作補強，不得取代或繞過前述控制。原系統之接受、拒絕、逾時、部分成功及狀態不明等結果，應以可識別且可查證之方式回傳。

C. 進階要求：

- **4-4 結構化介面強化：**較高風險、多代理協作或跨系統工作流，宜明確定義工具 Schema、訊息格式、必要欄位、錯誤碼及狀態回傳方式，並確認輸入輸出符合介面規格，使後續控制點得以辨識資料、參數與執行狀態。
- **4-5 第三方工具與元件治理：**第三方模型、工具、連接器、MCP Server、技能及外部服務，宜納入供應鏈管理，掌握其來源、授權、事件通知、停用及退場安排。正式或較高風險應用場景，並宜確認其支援身分與權限、目的地限制、必要紀錄及阻斷之能力；相關控制能力得以供應商認證、第三方查核報告、既有雲端服務評估或其他等效佐證確認。供應商資訊或佐證不足以支持風險管理者，宜採補償控制或限縮使用範圍。
- **4-6 技能載入控管：**正式或較高風險應用場景載入技能前，宜確認技能已完成核准，且簽章或雜湊、版本、權限宣告及相依工具與核准內容一致；不一致、驗證失敗或無法停用之技能，不得載入正式環境。
- **4-7 應用與部署變更管理：**代理相關提示、工具描述、工具 Schema、連接器、政策規則、流程設定及部署版本之新增或變更，應納入安全開發、測試驗證、變更審查、完整性保護及復原管理；正式或高風險應用場景上線前，宜完成工具濫用、參數邊界、流程狀態及回歸測試。

D. 系統可控性目標：

- **可觀測：**應可掌握工具、技能、連接器、介接系統、版本、載入狀態、輸入輸出、錯誤及介接結果。
- **可限制：**應可限制可載入或啟用之工具、技能、連接器、版本、介接系統，以及符合介面規格之輸入輸出。

- **可阻斷**：未核准、已停用、版本失效、完整性驗證失敗或不符合介面要求之工具、技能、連接器及輸入輸出，應可拒絕載入、呼叫或提供後續使用。
- **可復原**：應可停用或撤回異常元件，並將工具、技能、連接器、介面設定及部署版本回復至核准狀態。
- **可回溯**：工具、技能、連接器之來源、版本、核准、載入、變更、停用、輸入輸出及介接結果應可供查證。

5.5 資料 (Data)

本節依資料生命週期管理代理 AI 對資料與知識之取用、暫存、加工、留存、寫入、傳送及回復。基本控制先界定資料取用邊界，再管理記憶、脈絡與衍生資料、RAG／知識庫及外部資料與知識來源之使用條件；進階要求則強化敏感資料與大量取用、高風險資料使用與流向規則、非標準介面取用、污染回復，以及資料條件不足時之限制。主要控制點包括資料庫、資料平台、連接器、RAG／知識庫、記憶及資料外送機制。

A. 實務導入挑戰：

- **資料與知識治理失效**：代理可能同時取用內部資料、外部內容、檢索結果、記憶及中間摘要；若缺乏來源、分類、可信度、用途、案件範圍、授權、保留與輸出流向治理，可能造成跨客戶誤取、過度揭露、錯誤或污染內容持續留存、記憶與知識庫污染，或資料流向未核准目的地。另一方面，為供查證而完整集中提示、上下文或輸出，也可能形成新的敏感資料集中點。

B. 基本控制措施：

- **5-1 資料取用與存取邊界**：代理透過企業搜尋、連接器、API、資料庫、資料平台、資料服務層、RAG 或向量庫取用資料時，應依應用場景目的、資料分類、代理身分及授權範圍，限制可用來源、資料集、紀錄、欄位、客戶或案件範圍及操作方式，並遵循最小揭露原則；不得僅因資料已被索引、使用者可搜尋或介面可存取，即推定代理得任意取用。
- **5-2 記憶、脈絡與衍生資料治理**：代理使用或產生之工作階段脈絡、可重用記憶、中間摘要、快取及其他衍生資料，宜依用途、資料分類及使用者、客戶或案件範圍區隔管理，並設定寫入、存取、保存、清理與刪除方式。未經驗證、超出核准用途或逾保存期限之內容，不應供後續任務引用；用途終止或應用場景退場時，宜同步清理並保留必要之查證紀錄。

- **5-3 RAG／知識庫建置與運作：**RAG、向量庫或知識庫宜使用經核准之資料來源與版本，並管理索引更新、權限過濾、引用依據及污染處理；代理輸出宜繼承來源資料敏感等級或重新分類。
- **5-4 外部資料與知識來源准入：**文件、郵件、網頁、第三方資料、使用者上傳檔案及其他外部內容，宜依來源或提供者、資料分類、核准用途、版本或更新日期及可信度，判斷是否得供代理使用；未經驗證、已過期或未核准之內容，不應作為正式資料來源，或寫入可重用記憶、RAG 或正式知識庫。

C. 進階要求：

- **5-5 敏感資料與大量取用控管：**涉及敏感資料，或代理可跨客戶、跨案件、批次或大量取用資料者，宜預先界定允許之資料來源、資料集、紀錄、欄位、客戶或案件範圍及單次與累計取用上限，並採最小揭露、遮罩或去識別化、分區隔離、查詢與匯出限制及異常取用監控。超出核准範圍或數量門檻之資料，不應提供代理取用或匯出。
- **5-6 高風險資料使用與流向規則：**涉及敏感資料外送、寫入可重用記憶、更新正式知識庫或擴大資料流向者，應預先定義允許之接收對象、使用目的、保存方式、必要遮罩、資料分類繼承及後續使用限制，並使相關條件可供系統查驗。
- **5-7 非標準介面之資料取用：**代理透過截圖、視窗讀取、網頁或應用程式介面取得資料時，應確認可取用範圍、欄位、用途及最小揭露仍符合核准條件；無法依原系統權限與資料分類限制取用範圍，或介面無法穩定識別欄位與資料狀態者，宜遮罩、限縮資料或停止自動取用。
- **5-8 污染資料處理與回復：**涉及外部資料、記憶、知識庫或檢索引索者，宜建立污染通報、封鎖來源、下架內容、清理記憶、重建索引及刪除驗證流程，並防止污染內容於處理完成前再次被檢索引索或引用。完成處理後，應驗證資料、記憶、知識庫或索引已回復至核准狀態；受影響工作階段、輸出、外送及後續動作則應納入事件查證與處置。
- **5-9 資料條件不足時之限制：**若無法確認資料來源、分類、授權範圍、核准用途、保存方式或輸出流向，不宜正式外送、寫入可重用記憶或更新正式知識庫；既有能力應限於不含敏感資料之測試、唯讀、草稿或人工輔助模式。

D. 系統可控性目標：

- **可觀測：**應可掌握資料來源、分類，以及取用、檢索引索、寫入記憶、更新知識庫、產生衍生資料、輸出、外送、刪除及污染回復情形。
- **可限制：**應可依資料分類、用途、授權範圍及最小揭露原則，限制資料取用、寫入、輸出與外送。

- **可阻斷**：不符資料分類、用途、授權或來源驗證條件之取用、寫入、輸出或外送，應可拒絕或中止。
- **可復原**：資料、記憶、知識庫或索引發生錯誤、污染或過度留存時，應可更正或刪除錯誤資料、清除污染內容、重建索引，並移除不應留存之資料。
- **可回溯**：資料取用、寫入、輸出及外送之操作紀錄，應可查證資料來源、應用場景、代理身分與核可依據。

5.6 行為控制（Behavior Control）

本節承接第 3.5 節安全護欄，將第 5.1 至 5.5 節所建立之身分與授權、執行環境、網路、能力元件與介面及資料使用條件，轉化為代理於特定任務、工作階段及當下狀態之行為決策。控制重點在確認具體行為是否仍位於核准邊界內，並就工具選擇與組合、參數、流程、委派、執行規模、高影響動作及內容驅動行為，採取允許、限縮或修改參數後放行、拒絕、阻斷、降級或轉人工處理；第 5.7 節再就決策與執行結果進行關聯監控及控制回饋。

A. 實務導入挑戰：

- **行為偏離與影響擴散**：代理即使使用合法身分、核准工具及可存取資料，其工具組合、參數、執行順序、委派、批次、重試或最終影響仍可能超出應用場景目的與核准邊界，並於多步或跨系統執行中擴大影響。

B. 基本控制措施：

- **6-1 行為邊界**：應依核准應用場景與任務目的，明確界定代理得使用之工具與資料、可執行之動作、對象、目的地、參數、自動化程度，以及單次與累計影響範圍；即使工具或資料本身已獲准使用，其具體使用方式仍不得超出前述邊界。
- **6-2 工具鏈與流程限制**：代理組合工具或推進流程時，應限制允許之工具組合、執行順序、資料與狀態傳遞、委派及子任務展開，避免多個低風險能力串接後形成未核准之外送、寫入、刪除、交易、權限變更或流程擴散。
- **6-3 執行規模與持續性限制**：涉及批次、排程、自動重試、遞迴、委派或大量並行任務者，應設定任務數量、並行數、執行時間、重試次數、子任務及單次與累計影響上限；達到門檻時，宜告警、中止、降級或轉人工處理。
- **6-4 執行前行為查驗**：涉及外送、寫入、刪除、交易、權限變更或其他具實際影響之動作，執行前應綜合第 5.1 至第 5.5 節之身分與授權、執行環境、網路、能力元件、資料、目的地、參數及核可狀態，依第 3.5 節所定四個控制點，依序確認模型輸出或外部內容未受不當指令驅動、所用工具與介接均屬核准範圍、具體動作及影響範圍符

合核准邊界，且未繞過介接系統既有安控，據以作成允許、限縮、拒絕、降級或轉人工處理之決定。任一控制點或必要條件未通過、驗證失敗或無法判斷時，不應放行。

C. 進階要求：

- **6-5 執行結果驗證與狀態確認：**涉及高影響或跨系統流程者，執行後宜透過系統回傳、狀態查詢或其他可驗證方式，確認實際結果、影響範圍及回傳狀態與核准內容一致，並將確認結果作為是否續行後續流程之依據。
- **6-6 規劃與執行分離：**高風險應用場景不宜由自然語言、模型輸出或代理規劃直接觸發高影響動作。代理產生之規劃、建議、程式碼、腳本、SQL、系統指令或執行參數，應先轉為可供檢查之待執行內容，再由受控流程執行。
- **6-7 高影響動作放行與覆核：**涉及寫入、外送、刪除、交易、權限變更、高影響資安處置或不可逆結果者，應依風險及既有內控，於執行前採逐次政策檢查、人工確認、雙人覆核或其他等效放行控制；執行後得以抽樣覆核檢視控制品質。執行前覆核應呈現資料來源、差異、目的地、參數及預期影響，並區分草稿、預覽與實際執行；金融機構並應管理覆核者資格、負荷與品質，留存放行、限縮、退回、拒絕及覆核結果。
- **6-8 流程狀態與斷路器：**應明確處理逾時、部分成功、重複執行、錯誤累積及狀態不一致；超過門檻、必要檢查或核可遭跳過，或控制服務失效時，應中止流程、啟動斷路器或轉人工處理，並依執行結果回復安全狀態、撤回可逆動作或處理已完成異動。
- **6-9 行為控制條件不足時之限制：**若無法界定或執行核准行為邊界，無法查驗重要動作，或無法於越界、異常及失效時限制、阻斷或轉人工，不宜正式自動或半自動執行；既有能力應限於模擬、唯讀、草稿或預覽模式。

D. 系統可控性目標：

- **可觀測：**應可掌握本次任務之規劃、內容驅動關係、工具選擇與組合、參數、政策判斷、核可及執行結果。
- **可限制：**應可依核准行為邊界限制工具組合、參數、流程、委派、執行規模及單次與累計影響範圍。
- **可阻斷：**具體行為越界、必要條件異常或無法判斷時，應可拒絕、中止、降級或轉人工處理。
- **可復原：**應可處理部分成功、重複執行及狀態不一致，並依執行結果停止後續流程、回復安全狀態或處理已完成異動。
- **可回溯：**行為及其放行、限制或阻斷決策，應可對應至應用場景、觸發者、代理、授權、能力元件、資料與核可依據。

5.7 可視性與分析（Visibility & Analytics）

本節統一規範第 5.1 至 5.6 節日誌之共同要求、跨系統關聯、日誌與監控有效性、監控分析及控制回饋。基本控制先建立必要日誌、共同識別、保護與留存及事件流程關聯；進階要求再確保日誌與監控持續有效，並建立行為基準、異常偵測、處置與控制回饋，以支援事件還原及控制有效性驗證。原始紀錄仍由各來源系統保存；集中監控系統原則上僅留存必要之關聯資訊，不重複保存完整提示、輸出或敏感資料。

A. 實務導入挑戰：

- **日誌分散與關聯不足：**代理任務涉及多個控制點與營運系統；若事件未被記錄，或日誌內容、時間基準、識別方式及保存期間不一致，將難以串接觸發、授權、工具呼叫、資料流向、政策判斷、核可與處置結果，也不易辨識異常及查核責任。若為查證而重複集中完整提示、輸出或敏感資料，也可能增加資訊暴露風險。

B. 基本控制措施：

- **7-1 必要日誌與共同識別：**第 5.1 至 5.6 節之相關控制點與來源系統，應依風險記錄工作階段、應用場景、代理與版本、觸發者、主要動作、資料流向、政策與核可、執行結果、錯誤重試及處置結果，並以工作階段 ID、應用場景 ID、代理 ID、事件 ID 及時間戳或其他等效欄位，使同一任務之紀錄可相互對應。
- **7-2 日誌保護與留存管理：**日誌應依敏感度、查證及法定保存需要，採取存取控制、加密、完整性保護、遮罩、時間同步、保存及刪除管理。完整內容已由原系統保存者，得以識別碼引用；未重複留存完整提示、輸出、憑證或敏感資料者，仍應保留足以確認來源、政策判斷、動作及結果之中繼資料。
- **7-3 日誌關聯與事件流程銜接：**應以共同識別、事件 ID、父子任務或委派關係及可信時間資訊，關聯各來源系統之日誌，並將關聯結果提供監控、事件應變、復原及稽核流程使用。

C. 進階要求：

- **7-4 日誌與監控有效性：**宜持續確認日誌之可用性、必要欄位、時間同步、傳輸及跨系統關聯正常；發生日誌中斷、欄位缺漏、時間異常或關聯失敗時，應告警、標示受影響範圍，並啟動補送或查證程序。
- **7-5 行為基準與異常偵測：**較高風險或提高自動化程度之應用場景，宜綜合核准範圍、歷史運作狀態及風險情境建立行為基準，據以偵測偏離第 5.6 節核准行為邊界、

跨系統狀態不一致、異常工具鏈、異常資料流向及連鎖擴散等情形。新上線或樣本不足者，應先依核准邊界採規則式偵測。

- **7-6 異常處置與控制回饋：**監控或分析發現實際行為偏離核准範圍時，應依風險即時告警，並通知相關控制點或事件應變流程採取節流、阻斷、降級、撤銷授權或轉人工處理；處置結果並應回饋風險分級、第 5.6 節行為控制、政策規則、白名單、覆核門檻及監控規則，以持續改善控制措施。

D. 系統可控性目標：

- **可觀測：**應具備可關聯及驗證代理行為之日誌與端到端可視性，並可掌握日誌與監控之有效狀態。
- **可限制：**日誌或監控資料不足、關聯失敗或監控失效時，應可標示受影響之監控範圍，停止以不完整資訊作為正常或放行判斷依據，並通知相關控制點依風險採取限制。
- **可阻斷：**發現重大異常、必要紀錄缺漏或無法驗證控制結果時，應可即時告警，並通知相關控制點或事件應變流程採取阻斷、降級或轉人工處理。
- **可復原：**應可補送或重建必要紀錄、恢復監控與關聯能力，並支援事件調查及控制恢復。
- **可回溯：**任務與事件應可還原至應用場景、觸發者、代理、授權、工具、資料、政策判斷、核可及處置結果。

六、結語

代理 AI 正逐步成為金融機構提升營運效率、強化風險管理與推動服務創新的重要能力。金融機構宜以審慎創新為原則，將代理 AI 納入既有公司治理、資訊安全、資料治理、法遵內控、委外管理與營運韌性架構，依業務性質、風險承受度、資料敏感度及對客戶權益與營運韌性之影響，採取相稱之管理要求，並持續依技術發展、威脅情勢、法規標準、事件經驗及內外部檢核結果，滾動檢討治理政策、控制架構與導入策略。金管會將綜整諮詢意見及實務發展，作為後續政策研議、文件調整與監理溝通之參考。

附錄 A：OWASP Top 10 for Agentic Applications for 2026

（ASI01 – ASI10）威脅情境與安全控管對照摘要

本附錄依 OWASP Top 10 for Agentic Applications for 2026，將代理應用之主要威脅轉化為金融業務情境，作為第 5 章安全控管設計、測試與檢核之風險參考；ASI01 至 ASI10 為該風險清單之十項代理應用安全風險編號。本附錄所列風險名稱與編號，係依 OWASP 於 2025-12-09 正式發布之《Top 10 for Agentic Applications for 2026》整理；金融機構宜依官方最新版本更新對照。

下表案例係示意，並非表示各類應用場景必然發生該風險。本附錄用於協助辨識代理 AI 威脅情境、建立安全測試案例，並檢視第 5 章各控制面向是否已針對相應風險延伸與補強；實際控管仍應依應用場景風險、系統架構及各控制面向整體評估。

編號	風險項目與說明	金融場景例（示意）
ASI01	Agent Goal Hijack 目標劫持：代理受提示注入或惡意內容影響，偏離原始任務與核准邊界。	客服代理讀取含隱藏指令的補件文件後，改查他人資料或外寄摘要。
ASI02	Tool Misuse and Exploitation 工具濫用與利用：代理誤用合法工具、參數或工具鏈，造成未核准之外送、寫入或高影響動作。	報表代理串接匯出與外寄工具，將客戶資料送至未核准目的地。
ASI03	Identity and Privilege Abuse 身分與權限濫用：代理承接過大權限、濫用委派或憑證，導致越權與責任不清。	授信審查代理透過委派或快取憑證，越權查詢其他客戶徵信資料或觸發額度審核流程。
ASI04	Agentic Supply Chain Vulnerabilities 代理供應鏈弱點：代理所依賴之模型、技能、工具、外掛程式、範本、連接器或外部服務，可能因來源不明、元件遭竄改、惡意或不安全更新、相依套件弱點或供應商事件，使代理能力、權限或資料流向偏離核准範圍。	第三方徵信連接器更新遭竄改後，將授信報表與客戶資料送至未核准端點。

ASI05	Unexpected Code Execution (RCE) 非預期程式碼執行：代理因惡意或未受控輸入、工具設定或執行流程，生成、傳遞或執行未經核准之程式碼、腳本、SQL 或系統指令，可能造成未授權系統操作、資料異動、服務中斷或遠端程式碼執行。	核心銀行維運代理未經覆核即執行生成的 SQL 或指令，誤改交易資料或停用行員帳號。
ASI06	Memory & Context Poisoning 記憶與上下文污染：錯誤或惡意內容進入記憶、上下文或 RAG，持續影響後續決策。	貸款定價代理將錯誤利率或過期授信規則寫入記憶或 RAG，後續案件反覆引用。
ASI07	Insecure Inter-Agent Communication 多代理通訊不安全：訊息遭攔截、偽造、重放或竄改，造成未核准指令擴散。	授信、通知與帳務代理傳遞核可狀態時，遭偽造或重放訊息，導致未核准之客戶通知或帳務異動被執行。
ASI08	Cascading Failures 連鎖失效：單一錯誤、受污染輸入、失效控制服務或不正確狀態，在多步、多代理、批次、重試或迴圈流程中被沿用、放大並跨系統擴散。	錯誤 AML 警示摘要，或政策引擎、授權及驗證服務之錯誤狀態，被多個代理沿用並於批次或重試流程中放大，造成錯誤工單、客戶通知、交易限制或其他未核准處置。
ASI09	Human-Agent Trust Exploitation 人機信任剝削：利用自動化偏誤或高信心輸出，引導人員核可不當動作。	代理以高信心摘要、預設建議或不完整風險資訊影響覆核者，使其未查證來源、差異與限制，即核准投資建議、AML 案件處置或付款事項。
ASI10	Rogue Agents 失控或惡意代理：代理因入侵、目標漂移或污染而持續偏離授權範圍，造成資料外洩或營運破壞。	失控支付代理在權限未撤銷或停用未生效的情況下，持續查詢未核准客戶資料、呼叫外送工具，或再次觸發已停止之付款流程。

附錄 B：零信任架構與代理 AI 控管對照

本附錄對照金融業既有零信任架構與第 5 章代理 AI 安全控管。身分、設備／執行環境、網路、應用與工作負載及資料為既有控制面向，可視性與分析為既有橫向能力；本文件另增行為控制作為橫向能力，以綜合各項條件判斷代理得否執行具體行為，並可運用既有自動化與協調能力，支援跨控制面向之政策判斷、執行、告警、阻斷及控制回饋。

零信任 架構	零信任系統工程基底	代理 AI 增補重點	對應 章節
身分	採多因素身分認證、身分互通與最小授權，並將工作階段之動態屬性納入授權判斷，支援撤銷、限縮、告警及帳號生命週期管理。	區分觸發者、代理、服務及技能身分，限制權限承接與委派，並支援任務範圍授權、短效授權、撤銷及責任歸屬與查核。	5.1
設備／ 執行環 境	盤點並唯一識別設備，持續檢查健康、合規與弱點狀態，將設備動態屬性納入授權判斷，並支援偵測、隔離及生命週期管理。	將代理執行環境納入健康檢查、隔離、程序與資源限制，並支援異常中止、降級、停用及復原。	5.2
網路	採最小需求原則限制網路連線，並透過網路區隔或微分段、流量可視性、傳輸加密、異常偵測與動態應處，降低橫向擴散並維持網路韌性。	限制代理之外連、外送目的地及跨系統通訊，銜接既有出口管制，並支援異常流向阻斷及高風險通訊驗證。	5.3
應用與 工作負 載	落實最小授權及特權作業分離，將動態屬性納入存取判斷，並透過程式安全檢測、安全部署、行為監控及生命週期治理強化防護。	管理代理應用、工具、技能、連接器、輸入輸出介面、來源、版本、載入與停用，並銜接營運系統既有安控。	5.4
資料	建立資料盤點、分類與標籤，依分級落實最小授權、外洩防護、動態存取控制、加密與金鑰管理，並確保備份保護及有效還原。	管理資料與知識之來源與可信度、分類與授權、記憶與 RAG、使用與流向，以及留存、刪除與污染回復。	5.5
橫向能 力：行 為控制	自動化與協調：整合跨控制面向之政策判斷、執行、告警、阻斷及控制回饋。	綜合身分、環境、網路、能力元件與資料條件，控制內容能否驅動動作，以及工具選擇與組合、參數、	5.6

		流程、委派、批次、重試、高影響動作放行及異常阻斷。	
橫向能力：可視性與分析	於各控制面向整合事件日誌、行為基準、異常偵測、告警與回應，並逐步建立可依資安政策快速調適之一致性自動化管理機制。	關聯應用場景、代理行為、資料流向、政策判斷及處置結果，支援跨系統監控、事件還原與控制調整。	5.7

附錄 C：第一階場景初篩與簡化導入示例

本附錄提供代理 AI 應用場景第一階初篩示例，協助界定目的、風險、資料、工具、動作及是否需進入第二階治理。低風險或試行應用場景得留存簡化紀錄；涉及較高風險、正式上線、提高自動化程度或例外放寬者，得參考附錄 D 或依既有制度辦理。完成本附錄不同正式上線、合規或安全審查通過；金融機構得依既有制度採等效格式。

C.1 客服查詢自動化（查詢＋產出回覆建議）

核心欄位	第一階簡化紀錄
應用場景目的與定位	客服人員查詢貸款、補件、產品或服務資訊，由代理 AI 彙整核准知識庫與個案必要資料並產出回覆草稿；定位為單案唯讀查詢與草稿輔助，代理不具自動外送能力，由客服人員確認後以既有通道對外回覆。
初步風險與第二階判斷	• 通常列為 L2；維持單案唯讀、草稿及人工採用，且代理無自動外送、寫入、正式工單、批次或例外放寬能力者，可留存本附錄。 • 升級 L3：新增敏感資料範圍、受控客戶通知或內部外送、批次彙整、正式工單、非核心系統寫入或例外放寬，進入附錄 D。 • L4 高風險情境：修改客戶資料、受理交易、核心系統寫入、承諾核貸結果或其他重大客戶權益影響。未具備必要授權、政策檢查、覆核、阻斷及復原能力者，不得自動執行。
資料、工具與動作邊界	• 可用：客服知識庫、RAG、必要個案狀態、CRM 單案唯讀查詢、知識檢索、彙整及草稿產出。 • 不得：自動寄送、建立正式工單、系統寫入、案件狀態更新、修改客戶資料、跨客戶批次查詢、承諾核貸或受理交易；範圍不得逾越單一客戶或案件。
主要風險與必要控制	主要風險為越權查詢、提示注入、資訊錯誤或過期、敏感欄位過度揭露、過度信任草稿及紀錄不足。宜限定單一客戶／案件、必要欄位與唯讀工具白名單，並監控跨客戶查詢、敏感欄位調閱、非時段使用與異常重試。
停用、降級與事件查證	發生越權查詢、錯誤回覆外送、提示注入命中或告警未處理時，停用或降級代理、撤銷授權或 Token，並封鎖可疑來源或工具；查證工作階段、查詢內容、引用來源、覆核者及對外回覆範圍。

C.2 法遵／AML 警示初篩（彙整＋分流建議）

核心欄位	第一階簡化紀錄
應用場景目的與定位	代理 AI 依交易監控警示、客戶資料、歷史交易與作業手冊，產出 AML 案件摘要、疑點整理與分流建議；定位為案件輔助，不自動作成最終判定、結案、申報、解除警示或觸發交易限制。
初步風險與第二階判斷	• 通常列為 L3；涉及敏感資料及法遵流程影響之案件摘要與分流建議，原則上進入附錄 D。 • 可列 L2 的情境：限單一案件及必要資料，僅產出摘要、疑點整理或分流草稿，由 AML 人員依既有程序採用；代理不更新案件狀態、不自動結案、申報、解除警示或限制交易。 • L4 高風險情境：直接結案、申報、解除警示、限制交易或觸發帳戶限制；客戶通知應依其內容、目的及對客戶權益之影響另行判定。未具備必要授權、政策檢查、覆核、阻斷及復原能力者，不得自動執行。
資料、工具與動作邊界	• 可用：核准之案件資料、交易監控系統、規則／作業手冊知識庫及 RAG；限案件查詢、知識檢索、摘要、疑點整理與分流建議。 • 受控情境：建立正式工單、更新非核心案件狀態或依核准內容通知客戶，須列 L3 並完成第二階檢核。 • 不得：自動結案、申報、解除警示、交易或帳戶限制，或批次處理未核准案件範圍。
主要風險與必要控制	主要風險為摘要錯誤、關鍵證據遺漏、模型或 RAG 污染、跨案件誤取、自動化偏誤、敏感資料過度揭露及大量誤分流。宜限定資料來源、案件範圍、查詢欄位、提示詞範本及 RAG 版本，使輸出可追溯至原始交易、文件與規則，並由 AML 人員確認分流。
停用、降級與事件查證	發生跨案件誤取、錯誤分流、引用來源異常、污染知識命中或告警未處理時，停用或降級彙整與分流功能，下架污染知識、封鎖異常來源或重建 RAG 索引；查證工作階段、輸入、檢索來源、輸出、覆核者及後續處置。

C.3 內部營運助理／報表代理（跨系統彙整＋產出報表）

核心欄位	第一階簡化紀錄
應用場景目的與定位	代理 AI 以受控服務身分跨內部系統彙整營運資料，產出報表、工單或內部通知草稿；定位為內部營運輔助，不自動外寄、不批次匯出敏感資料，亦不直接寫入核心系統。
初步風險與第二階判斷	• 通常列為 L2：一般內部資料、有限來源，且僅產出報表、工單或通知草稿；可留存本附錄。 • 升級 L3：涉及個資、批次匯出、含敏感資料、正式作業資訊或大量對象之受控群組派送、正式工單、非核心系統寫入、狀態更新、高頻重試或目的地放寬，進入附錄 D。 • L4 高風險情境：核心系統寫入、金流、權限、核心設定或關鍵營運設定。未具備必要授權、政策檢查、覆核、阻斷及復原能力者，不得自動執行。
資料、工具與動作邊界	• 可用：核准營運資料庫、報表資料集、工單系統與必要內部文件；跨系統唯讀查詢、比對、報表範本、工單草稿及通知草稿。 • L3 受控情境：正式工單、受控群組派送、批次匯出或非核心系統寫入，須依目的地白名單、留痕及必要覆核辦理。 • 不得：外部或非白名單寄送、未核准敏感資料批次下載、核心系統異動、金流、權限或關鍵設定，且不得繞過既有報表核可流程。
主要風險與必要控制	主要風險為服務帳號權限過大、白名單過寬、批次匯出外洩、連接器或提示變更造成錯誤報表、自動重試放大及錯誤收件者外送。宜採獨立代理身分、最小權限、資料欄位與收件群組白名單，將報表產出與外送分離，並監控批次匯出、非白名單收件者、異常工具鏈及重試。
停用、降級與事件查證	發生非白名單外送、批次匯出異常、錯誤報表版本、異常工具鏈或重試失控時，停用外送工具或降級為草稿，撤銷代理 Token 並封鎖異常目的地；查證資料來源、查詢範圍、產出版本、外送紀錄及覆核者。

C.4 機器人理財（Robo-Advisor）投資建議／下單輔助

核心欄位	第一階簡化紀錄
應用場景目的與定位	代理 AI 彙整客戶風險屬性、投資目標、適合度資料、產品風險、費用及核准市場資訊，產出資產配置、再平衡、風險揭露、試算與下單預覽；定位為投資建議與交易前輔助，不在缺乏客戶確認、適合度檢核及必要覆核下執行交易。
初步風險與第二階判斷	• 通常列為 L3；涉及個別化投資建議、適合度檢核或下單預覽者，原則上進入附錄 D。 • 可列 L2 的情境：使用範圍明確之有限客戶資料，僅產出風險揭露、產品比較、費用或情境試算及非個別化建議草稿；不完成適合度決定，亦不產生可直接轉為交易之下單指令或預覽。 • L4 高風險情境：實際下單、自動再平衡、額度或投資組合調整、產品設定變更、核心系統寫入或其他重大客戶權益影響。未具備必要授權、政策檢查、客戶確認或覆核、阻斷及復原能力者，不得自動執行。
資料、工具與動作邊界	• 可用：客戶風險屬性、投資目標、適合度、產品、費用、風險揭露及核准市場資訊；限建議草稿、風險揭露、試算、下單預覽與再平衡建議。 • 確認要求：建議採用、預覽轉交易或再平衡採用，須經客戶明確確認、適合度檢核、差異預覽及必要覆核。 • 不得：代理自行下單、再平衡、調整投資組合、變更產品設定、修改客戶風險屬性、寫入核心系統或以自然語言直接觸發交易。
主要風險與必要控制	主要風險為不適合建議、產品風險揭露或費用試算錯誤、模型偏誤、資料過期、自動化偏誤及未確認即下單。宜限定核准產品、適合度規則、客戶資料與市場資訊版本，保留引用及時間戳，並將建議、風險揭露、下單預覽與實際交易清楚分離，監控適合度不符、高風險產品推薦及異常再平衡建議。
停用、降級與事件查證	發生適合度規則失效、產品資料錯誤、未確認交易、異常再平衡建議或客戶權益爭議時，停用交易相關工具或降級為草稿，封鎖錯誤資料、提示或模型版本並撤銷相關授權；查證客戶風險屬性、輸入、建議、引用、適合度檢核、確認、覆核與交易紀錄。

C.5 資安 SIEM/SOAR 分析與處置輔助

核心欄位	第一階簡化紀錄
應用場景目的與定位	SOC 代理從 SIEM、EDR、資產重要度與威脅情報彙整告警脈絡，產出研判摘要、處置建議與 SOAR 劇本草稿；定位為資安分析與處置輔助，不自動封鎖、隔離、停權、撤銷權限或關閉服務。
初步風險與第二階判斷	• 通常列為 L3：僅讀取正式告警並產出研判摘要或處置建議者，得依資料敏感度、範圍及能力邊界列 L2 或 L3；建立正式工單、更新處置狀態或銜接處置流程者，原則進入附錄 D。 • 可列 L2 的情境：讀取限定範圍之正式告警、資產資訊與威脅情報，僅產出研判摘要、證據彙整或處置建議草稿；不建立正式工單、不更新處置狀態，亦不連接可執行處置之 SOAR 工具。 • L4 高風險情境：封鎖、隔離、停權、撤銷權限、關閉服務、變更防護政策或其他可能影響關鍵營運之處置。未具備必要授權、政策檢查、覆核、阻斷及復原能力者，不得自動執行。
資料、工具與動作邊界	• 可用：SIEM/EDR 告警、資產重要度、弱點、威脅情報及核准劇本；限告警查詢、關聯分析、證據彙整、研判摘要、工單草稿及 SOAR 劇本草稿。 • L3 受控情境：建立正式資安工單、更新非核心處置狀態或銜接受控處置流程。 • L4 高風險處置：封鎖、隔離、停用帳號、撤銷權限、關閉服務、刪除資料、變更防護政策或其他高影響處置，須符合既有內控及必要授權、政策檢查、覆核、阻斷與復原條件。
主要風險與必要控制	主要風險為誤判造成營運中斷、威脅情報或提示污染、劇本參數濫用、連鎖失效、敏感證據外洩及處置後無法復原。SOAR 劇本宜採白名單、參數硬規則、分層自動化與版本管理；高影響處置須依既有內控確認，並監控告警來源、劇本參數、非預期工具鏈與高風險動作。
停用、降級與事件查證	發生錯誤處置建議、威脅情報污染、劇本參數異常、非預期工具鏈或高風險處置告警時，停用自動處置代理並暫停高風險劇本，撤銷相關 Token 及封存證據；查證告警來源、研判依據、工具呼叫、處置結果與復原情形，並依事件等級辦理復原及通報。

附錄 D：第二階治理檢核與實務參考工具

本附錄提供較高風險、提高自動化程度、正式上線或例外放寬應用場景之第二階治理參考，包括應用場景治理檢核表、監控規則、例外紀錄及供應鏈資訊。所列表格與欄位均為示例，金融機構得依既有制度、應用場景風險及實務需要選用、合併或採等效格式；已於既有制度留存之資訊，得以關聯識別碼、紀錄關聯機制或其他等效方式引用，無須重複填寫。

D.1 應用場景治理檢核表範本

本範本供較高風險應用場景辦理治理檢核及留存核心資訊，並作為監控、例外及相關審查紀錄之索引；另宜以關聯紀錄記載第 5 章控制項之適用性、落實狀態、等效作法、佐證紀錄及未完成事項，各項測試、核決及技術佐證仍依金融機構既有程序辦理。供應鏈資訊得視需要另參考 D.4 或既有制度之相應紀錄。以下名稱、編號、狀態及情境均屬虛構示例，不代表任何機構之實際案例、核可結果或導入時程。

序號	清冊欄位	填寫說明	填寫示例
1	應用場景名稱	以業務可理解方式命名，避免僅以模型或工具名稱表示。	客服貸款補件正式工單與受控通知代理。
2	業務目的	說明代理欲解決之作業痛點、服務對象與預期效益。	協助客服查詢貸款補件規則與個案狀態，建立正式工單，並將核准內容透過受控通道通知客戶。
3	應用場景負責人	指定業務端負責人及治理／平台窗口，以利核可、變更與事件處置。	消金客服部經理；平台治理窗口：AI 治理小組。
4	技術組成與識別	登錄應用場景 ID、代理 ID、服務身分、部署環境與版本，並引用技能及工具／連接器 ID，呈現其相依、授權與使用關係。	UC-LOAN-CS-003；代理：agent-loan-cs-prod v2.0；NHI：svc-agent-loan-cs；技能：SK-017、SK-028；工具：TL-021、TL-034、TL-052、TL-061。

5	目前狀態／生命週期階段	記錄規劃、開發、測試、試行、正式上線、限制上線、暫停或退場等狀態及生效日期。	限制上線；限貸款補件正式工單與核准通知範本，生效日期依實際核可紀錄填列。
6	自動化程度	記錄唯讀、草稿、預覽、人員採用、半自動或自動執行程，以及觸發與覆核方式。	單案查詢＋正式工單建立＋受控客戶通知；工單與通知內容須依條件經客服人員確認。
7	使用者族群	列明可使用之人員、部門、角色或系統，並標示是否開放外部使用者。	僅限客服中心貸款專線人員；不開放客戶直接使用。
8	資料來源與敏感等級	列出資料庫、知識庫、文件庫、RAG 來源與資料分類，並標示是否含個資、交易、授信或法遵資料。	貸款作業知識庫、CRM 個案與補件狀態、客戶聯絡資料及通知範本；含個資、授信狀態與正式作業資訊；等級：內部敏感。
9	可呼叫工具與連接器	列出工具／連接器識別碼、用途、白名單版本及停用方式；技能詳細資料得引用技能清冊。	知識庫檢索（TL-021）、CRM 唯讀查詢（TL-034）、正式工單建立（TL-052）及受控通知服務（TL-061）；均限核准版本、單案範圍及白名單範本。
10	可執行動作	明確定義代理可做與不可做之動作，並標示是否可寫入、外送、交易、批次或自動重試。	可查詢、彙整、產生回覆草稿、建立正式補件工單，並依核准範本發送補件通知；不可修改客戶主檔、授信結果或案件核准狀態，不可受理交易或承諾核貸結果。
11	允許輸出或寫入目的地	列明輸出介面、系統、收件者或目的地；外送須標示白名單與 DLP/CASB 檢查。	客服工作台草稿欄位、補件工單系統及受控簡訊／電子郵件通知服務；收件者限當次案件客戶，內容限核准範本並經目的地與 DLP 檢查。

12	批次／自動重試限制	標示是否允許批次、排程、自動重試、大量子任務展開或跨案件處理，並設定上限。	不允許跨客戶批次；每工作階段限單一客戶案件；工單建立及通知各限一次，自動重試上限 2 次，狀態不明時轉人工確認。
13	風險等級與三軸判斷	依第 4 章填列資料敏感度、動作影響、能力邊界／可擴散性及校準理由，並記錄最高組合能力與最高風險觸發因素；控制措施不作為調降固有風險等級之理由。	L3；含個資、授信狀態及客戶聯絡資料；可建立正式工單並向客戶受控外送通知；單案執行、目的地與範本固定、不得跨客戶批次或修改核心資料。最高風險觸發因素為正式紀錄與客戶通知。
14	核可紀錄	記錄核可單號、核可者、核可日期、有效期間與核可條件。	〔示例核可單號〕；經業務、資安、資料治理、法遵及平台治理單位核可；限單一客戶案件、核准通知範本及白名單目的地；有效期間依實際核可紀錄填列。
15	必要覆核條件	說明何種情境需人工確認、主管覆核、雙人覆核、客戶確認或其他同等控制。	工單欄位或通知內容超出核准範本、含敏感欄位、授信拒絕、費用／客訴／例外承諾，或目的地與案件聯絡資料不一致時，須由客服人員確認；涉及例外承諾時另須主管覆核。
16	監控規則引用	引用 D.2 監控規則 ID 及啟用狀態；不在清冊重寫完整偵測邏輯。	MON-01 跨客戶／跨案件異常查詢、MON-02 敏感欄位過度調閱、MON-04 統一閘道繞行或異常工具呼叫、MON-06 非白名單目的地外送、MON-07 工單／通知狀態不一致、MON-09 高頻重試、迴圈或大量子任務展開、MON-13 權杖重放或異常使用；啟用狀態依實際建置情形填列。

17	停用、降級與退場條件	定義何時停用代理、撤銷憑證、降級為唯讀／草稿、切回人工或退場。	發生越權查詢、工單錯寫、非白名單或錯誤客戶通知、提示注入命中、通知與工單狀態不一致，或告警未處理逾 SLA 時，停用工單與通知工具並降級為唯讀草稿；應用場景退場時，並撤銷代理權限、Token 及相關工作流，處理記憶、快取、暫存資料與知識庫索引，確認第三方留存及刪除安排，並保留必要稽核證據。
18	最近盤點與下次重評	記錄最近盤點日期、下次重評日期或觸發條件；重大資料、工具、目的地、權限或自動化程度變更應提前重評。	最近盤點日期及下次重評日期依實際紀錄填列；新增批次通知、擴大工單寫入欄位、變更目的地／範本或提高自動化程度時，應提前辦理。
19	例外引用	引用 D.3 例外 ID、狀態及到期日；完整風險理由、補償控制與改善計畫由 D.3 留存。	〔示例例外編號〕；狀態、核可日期及到期日依實際紀錄填列。完整風險理由、補償控制與改善計畫由 D.3 留存。
20	相關審查紀錄	填列本應用場景於既有上線、變更管理、資安測試、模型驗證、法遵或資料治理程序中的相關紀錄編號；詳細測試結果與核決內容仍由原程序留存。	〔示例變更單號〕、〔示例驗證紀錄編號〕；詳細內容依原審查紀錄辦理。

清冊維護說明：清冊宜於能力邊界重大變更時更新並重新評估風險；盤點頻率與方式得依既有制度及應用場景風險訂定。

D.2 SOC 偵測場景與監控規則範例

本表所列規則為偵測場景示例。進入第二階治理之應用場景，宜具備支援所選監控規則之必要日誌與關聯資訊；既有紀錄得以關聯識別碼或等效方式引用。

實際採用時，宜由應用場景治理檢核表引用相應監控規則 ID，供 SOC 與治理單位依應用場景風險、日誌成熟度、歷史基線及處置能力選用、合併與調校，不表示每一應用場景均須啟用全部規則。L1/L2 應用場景得以基本日誌為主；L3/L4 或涉及敏感資料、外送、寫入、批次、升權之應用場景，宜納入集中監控及必要之阻斷規則。門檻與優先級應依實際基線設定，不宜視為固定標準。

監控規則 ID	監控規則	觸發條件或偵測邏輯	建議處置	適用風險等級	適用應用場景
MON-01	跨客戶／跨案件異常查詢	同一代理或工作階段於短時間內查詢多名客戶、多個案件或非核可範圍資料。	通知應用場景負責人與 SOC；暫停跨客戶查詢能力；必要時降級為單案唯讀並啟動抽樣稽核。	L2-L4	客服、AML、營運報表
MON-02	敏感欄位過度調閱	代理查詢或帶入提示之欄位含高敏感標籤，且超過應用場景核可欄位白名單。	自動遮罩或阻擋；要求 HITL 覆核；記錄欄位、來源、代理身分與觸發者。	L2-L4	客服、AML、RAG 檢索
MON-03	資料平台查詢或異動異常	同一代理或工作階段查詢高敏感資料、異動資料或超出核准欄位範圍，且未對應核可應用場景、工作階段或覆核紀錄。	阻擋或暫停查詢／異動能力；通知應用場景負責人、資料治理與 SOC；檢視授權範圍、資料平台日誌與核可紀錄。	L3-L4	客服、AML、營運報表、資料平台整合
MON-04	統一閘道繞行或異常工具呼叫	代理嘗試直接連接後端工具、外部服務或未登錄 MCP Server，或工具描述、Schema、目的地與核准紀錄不一致；偵測欄位包括技能 ID／版本、未核准技	阻擋工具呼叫；停用異常工具或連接器；要求重新註冊與審查；必要時撤銷代理 Token 並啟動供應鏈檢查。	L2-L4	所有工具型代理、跨系統代理、第三方工具整合

		能載入、版本漂移及完整性驗證失敗。			
MON-05	異常工具鏈連用	同一工作階段於短時間內出現大量查詢、匯出、壓縮或封裝，再接續外部上傳、外寄、寫入或非白名單 API 呼叫，且偏離核可流程；並偵測技能權限越界、政策檢查前後結果，以及政策服務失效、降級或轉人工狀態。	即時中斷後續工具呼叫；降級為草稿；將工具鏈與參數摘要送 SOC 分析。	L3-L4	營運報表、SOAR、跨系統代理
MON-06	非白名單目的地外送	代理嘗試將資料、附件、摘要或通知送往未核可網域、收件者、API 端點、雲端儲存或通訊通道。	阻擋外送；封鎖目的地；要求重新核可；必要時撤銷代理 Token。	L2-L4	客服、營運報表、投資建議
MON-07	工單／通知狀態不一致	正式工單已建立但通知未送達、送達不同案件客戶，或通知已送達但工單狀態未同步。	阻斷後續通知或工單動作、轉人工核對並通知應用場景負責人與 SOC。	L3-L4	客服工單與受控通知
MON-08	非變更窗口修改提示或工具設定	提示詞範本、工具描述、連接器、policy、Schema 或 RAG 來源於非核可變更窗口被修改。	暫停新版本；回復至核可版本；啟動變更稽核與完整性驗證。	L2-L4	所有正式應用場景
MON-09	高頻重試、迴圈或大量子任務展開	同一任務在短時間內重複呼叫相同工具、產生大量子任務、委派多個代理或觸發批次擴散。	觸發斷路器；節流或停止工作流；保全工作階段、工具呼叫與錯誤碼。	L2-L4	AML、報表、SOAR
MON-10	HITL 覆核品質異常	覆核時間異常短、同一覆核者連續大量一致通過、長時間連續覆核、退回率異常偏	提高覆核層級或抽樣複核比例；檢視覆核負載與門檻；必要時暫停該應用場景、調	L3-L4	投資建議、SOAR、

		低，或告警後仍快速核准高風險動作；相關提示門檻宜依任務特性與歷史基線設定。	整覆核者權限或降級為人工流程。		AML
MON-11	高風險動作缺少確認或預覽	交易、付款、外送、刪除、封鎖、隔離、權限變更或客戶權益影響動作未產生差異預覽、風險提示或核可紀錄關聯機制。	拒絕執行；降級為草稿；通知應用場景負責人補正流程；列為正式上線阻斷項。	L4	投資下單、SOAR、權限管理
MON-12	行為漂移或惡意／失控代理跡象	代理在無核准變更下偏離原應用場景流程，反覆嘗試非核准資料取用、持續外連、建立未授權子任務、重啟已暫停工作流，或在移除原始異常輸入後仍持續異常行為。	立即停用或降級代理，撤銷 Token 與工具權限，保全工作階段、工具呼叫、目的地與輸入輸出紀錄；重新上線前完成鑑識、元件驗證與重新核可。	L3-L4	營運報表、SOAR、跨系統代理、第三方工具整合
MON-13	權杖重放或異常使用	同一權杖於短時間內由不同代理、工作負載、來源位置或端點使用；過期、已撤銷或簽章驗證失敗之權杖仍嘗試存取；實際 API 或工具呼叫超出權杖授權範圍。	阻斷呼叫，撤銷相關權杖與工作階段，通知 SOC 及身分管理單位，並檢視簽章金鑰、權杖簽發與驗證服務是否受影響。	L3-L4	跨系統代理、多代理協作、API 與高權限工具應用場景

規則調校說明：監控規則上線後，宜指定調校責任單位及檢視頻率。告警過多時，宜先檢查欄位品質、門檻設計與應用場景邊界；告警過少時，宜檢查日誌是否完整、白名單是否過寬及代理行為是否正確送入集中監控。對 L3/L4 應用場景，關鍵阻斷型規則宜納入演練。

D.3 例外管理空白範本

本表用於記錄第 4.3 節所稱之暫時性例外，並連結 D.1 應用場景；除例外範圍、期間、補償控制、核可、改善計畫及到期處置外，並應以專欄或關聯紀錄記載補償後殘餘風險與接受者，以及延期或重新核可之次數、理由、到期日、核可者及未完成改善事項。

- **適用原則：**例外應連結 D.1 應用場景，限於最小範圍及明確期間，並記錄風險、補償控制、核可、改善計畫與到期處置。
- **不可放寬事項：**例外不得造成紀錄或追溯缺漏、繞過目的地管控，或使必要之授權、覆核、阻斷、停用、撤銷、復原及人工接手能力失效。
- **到期處置：**未達正式上線基準者，應限縮能力、降低自動化程度或轉人工處理；例外到期應結案、回復原控制或重新評估核可，不得以持續延展取代改善。

例外欄位	填寫說明
例外 ID／應用場景 ID	填列唯一例外編號，並連結 D.1 應用場景。
原控制與例外內容	說明原控制要求、目前缺口及暫時放寬範圍。
例外理由與風險影響	說明例外之必要性、影響對象及可能結果。
補償控制與有效性確認	說明過渡期間採取之替代控制及其驗證方式。
核可者與核可日期	記錄核可角色、核可日期及決議。
生效日與到期日	明確界定例外有效期間。
改善計畫與責任人	記錄改善項目、負責人及預定完成日。
到期處置與狀態	填列結案、回復原控制、重新核可、停用或降級。

D.4 供應鏈與 AIBOM/SBOM 補充欄位

涉及第三方模型、代理平台、工具、連接器、外部資料來源或其他關鍵元件者，宜留存供應商可提供之來源、服務範圍、版本或變更資訊、事件通知、停用與退場安排；下列欄位得併入應用場景清冊、變更單、AIBOM/SBOM、委外或既有供應鏈管理文件。供應商資訊揭露有限者，得以供應商認證、第三方查核報告或既有雲端服務評估作為等效佐證。仍不足以支持風險管理者，宜採相稱之補償控制或限縮使用範圍。

補充欄位	填寫重點	佐證資料
模型與代理平台	模型名稱、版本、供應商、部署方式、代理框架或平台版本、使用區域與資料處理邊界。	供應商文件、部署紀錄、版本紀錄、資料處理或委外文件。
提示與政策版本	系統提示、提示詞範本、政策規則、護欄設定與變更日期；重大變更應保留核可紀錄。	提示版本、政策版本、變更單、測試與核可紀錄。
工具、外掛程式與連接器	工具名稱、用途、來源、版本、可執行動作、輸入 Schema、輸出處理、白名單與停用方式。	工具清冊、連接器設定、白名單、停用、回復或安全替代流程紀錄。
MCP/API 端點	端點名稱、用途、核准目的地、驗證方式、通訊保護、版本鎖定與是否經統一閘道控管。	端點清冊、閘道政策、驗證設定、阻擋或允許紀錄。
資料與 RAG 來源	資料來源、資料分類、索引版本、來源驗證方式、更新頻率、污染處理、下架與刪除方式。	資料來源清冊、索引版本、來源標示與引用追溯、刪除或重建紀錄。
完整性與弱點管理	依元件重要性與供應商可取得性，記錄雜湊、簽章、版本鎖定、已知弱點、供應商事件通知、修補狀態或替代佐證；並辦理技能跨平台重驗與持續掃描。	完整性驗證紀錄、弱點掃描或供應商通知、修補紀錄、替代方案或等效佐證。
第三方代理平台採用與委外管理	涉及第三方代理平台、MCP Server、Plugin 或外部連接器者，宜於採用前評估其安全測試、弱點管理、事件通知、停用及退場能力；高風險應用場景應將可停止、可撤銷、可回復、資料刪除與退場安排納入委外契約或資安附約。供應商佐證不足時，宜採補償控制或限縮使用範圍。	委外契約、資安附約、供應商評估紀錄、安全測試或等效證明、事件通知與停用／撤銷／退場紀錄。

退場與替代安排	停止使用條件、資料刪除、權限撤銷、技能與工具停用、替代服務及事件期間人工流程。	退場計畫、刪除證明、撤銷紀錄、停用與替代流程演練。
---------	---	---------------------------