AI 代理與 Agentic AI 概論

從基礎到安全挑戰

人工智慧領域正經歷從被動模型到自主系統的典範轉移

探索 AI 代理系統的設計、挑戰與安全防護策略

報告生成日期: 2025-05-28

OWASP ASI

AI領域的典範轉移

典範轉移的關鍵點

顯著趨勢增長

自2022年底ChatGPT問世後，AI代理的發展趨勢在全球搜尋指數上顯著上升

從被動到主動

從被動回應的模型轉變為具有目標導向、自主決策能力的系統

功能擴展

引入記憶緩衝區、工具呼叫API、推理鏈和規劃例程，彌補被動響應與主動任務完成的差距

新興安全挑戰

自主性增強帶來新的安全風險，需要更深入檢視這些系統的內部運作機制

> *根據Gartner的預測，到2028年，33%的企業軟體應用將包含代理AI，而這一數字在2024年還不到1%。*

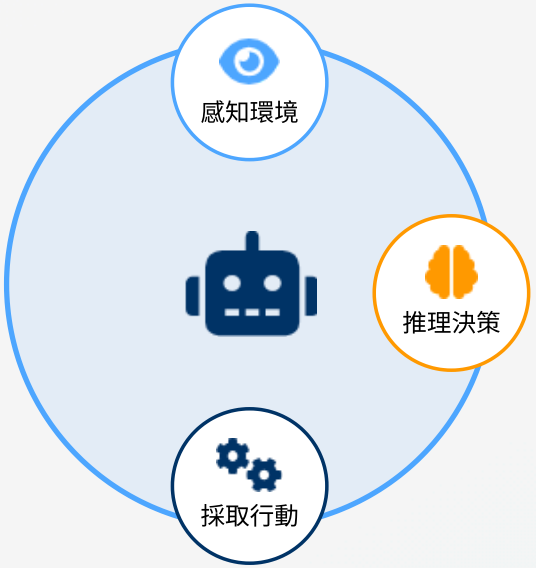
AI代理的定義

什麼是AI代理?

AI代理是一種能夠自主感知環境、做出決策並採取行動以達成特定目標的智能系統。

從資訊工程的角度來看，AI代理是設計用於在特定數位環境中執行目標導向任務的自主軟體實體。

> *AI代理能夠在部署後以最少或零人工干預的方式獨立運作，可以感知環境輸入、處理上下文資訊，並在實時環境中執行操作。*



AI代理的關鍵屬性

- 對上下文資訊進行推理
- 具備反應性和適應性
- 達成預設目標的能力
- 與傳統自動化腳本的區別

AI代理與生成式AI的區別

生成式AI

核心特性

- 輸入驅動模式，接收提示後產生回應
- 專注於內容生成（文本、圖像、程式碼等）
- 缺乏內部狀態與持久記憶
- 單次互動，無目標追蹤機制

工作模式

輸入提示

↓

模型處理

↓

生成回應

「被動回應，無內部狀態追蹤」

VS

從被動響應
到
主動任務完成

AI代理

核心增強

- 基於LLM的核心推理
  以生成式AI作為基礎推理引擎
- \+ 記憶緩衝區
  跨對話保持上下文與狀態
- \+ 工具呼叫API
  與外部系統和服務互動
- \+ 推理鏈與規劃
  多步驟邏輯和行動計劃

工作模式

感知 | 決策 | 行動

↻

目標導向執行

「自主任務完成，具持久記憶與目標追蹤」

AI代理的核心特徵

自主性 (Autonomy)

AI 代理能夠在部署後以最少或零人工干預的方式獨立運作。它們可以感知環境輸入、處理上下文資訊，並在實時環境中執行預定義或自適應的操作。

與傳統自動化不同，AI代理能根據環境變化自行調整行為

任務特定性 (Task-Specificity)

AI 代理通常被設計用於執行明確定義的狹窄任務範圍。它們在特定領域內進行優化，以實現高效率和精確性。

專注於特定任務使AI代理能夠在該領域達到更高的表現水平

反應性 (Reactivity)

AI 代理能夠對環境變化做出即時反應，並通過反饋循環和基本學習機制不斷改進其行為。這種能力使其能適應動態情境。

透過反饋循環持續優化決策過程，提高任務執行效率

工具整合能力 (Tool Integration)

AI 代理可以調用外部工具、API 和資料庫來擴展其功能範圍，突破單一模型的限制。這使它們能夠執行更複雜的任務。

例如，ChatGPT 結合 Web Search API 獲取即時資訊的能力

💡 這些核心特徵共同使AI代理能夠執行目標導向的任務，並在特定領域展現高效能的表現。

AI代理的核心組成

大型語言模型

作為AI代理的「大腦」，負責理解指令、進行推理和生成回應。例如GPT-4等模型可作為核心推理引擎。

感知模組

處理輸入信號，包括自然語言處理、計算機視覺等技術，使代理能夠理解環境和用戶需求。

安全與倫理約束

內置的規則和檢查機制，確保AI代理的行為符合安全和倫理標準。

工具整合

允許AI代理調用外部API、數據庫或其他軟硬件資源，擴展其能力範圍，如網絡搜索、數據分析工具等。

決策與規劃模組

根據感知信息和預設目標制定行動計劃。可能使用強化學習、蒙特卡洛樹搜索等算法。

記憶系統

用於存儲過往交互歷史、學習經驗和關鍵信息，支持長期任務執行和個性化服務。

學習與適應機制

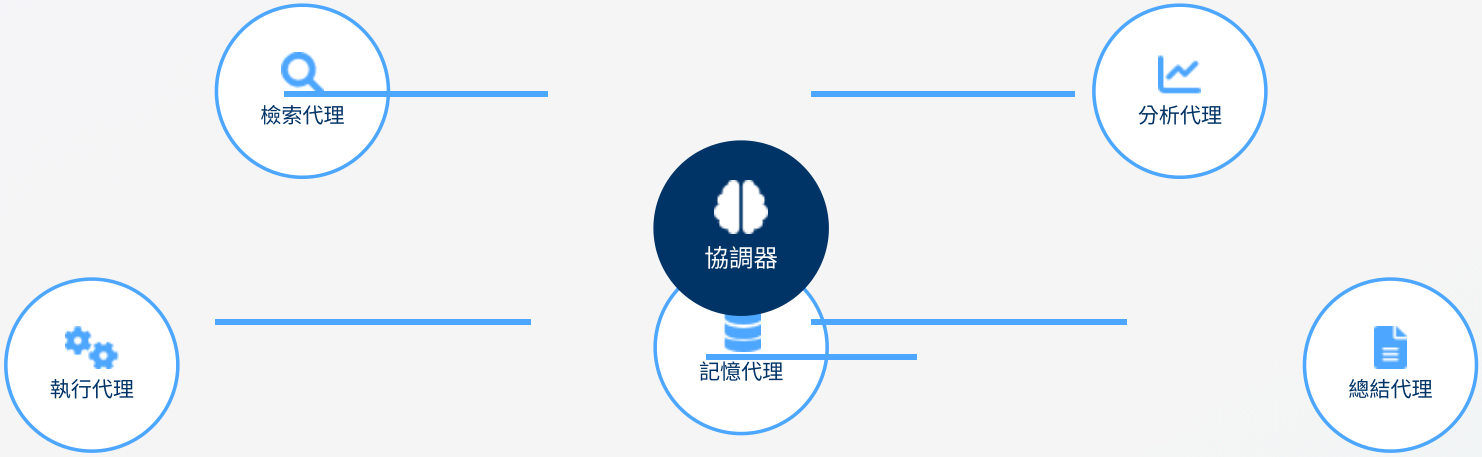
使AI代理能夠從經驗中學習，不斷優化其決策和行為模式。

執行模組

負責將決策轉化為具體行動，可能包括生成文本、控制硬件設備、發送API請求等。

Agentic AI的定義

> Agentic AI 代表了從孤立的AI代理到協作、多代理生態系統的範式轉變。
>
> 它是一種能夠分解和執行複雜目標的協作、多智能體系統，實現更高層次的自主性。



與單一AI代理的區別

- 不再是單一實體，而是由多個專門代理組成的協作系統
- 能夠自動分解複雜目標為可管理的子任務
- 代理間通過分散式通訊渠道交換資訊

關鍵定義特徵

- 多代理協作：各司其職的專門智能體協同工作
- 高度自主性：能夠管理需要協調的複雜、多步驟任務
- 反思性推理：跨多次互動保留上下文，評估並改進策略

Agentic AI的關鍵推動因素

四大要素推動Agentic AI的高級自主性與協作能力

目標分解

- 將複雜的使用者指定目標自動解析成更小、可管理的子任務
- 使系統能夠處理高層次、抽象的指令並轉化為具體行動
- 支援多個專門代理協同工作解決複雜問題的基礎

多步驟推理和規劃

- 促進子任務的動態排序，適應環境變化或部分任務失敗
- 使用如ReAct、Chain-of-Thought和Tree of Thoughts等框架
- 賦予系統考慮多種可能路徑並選擇最佳執行策略的能力

代理間通訊

- 代理之間透過分散式通訊通道交換資訊
- 實現協調而無需持續的中心化監督
- 支援非同步訊息佇列和共享知識庫的資訊交流

反思性推理和記憶系統

- 允許代理跨多次互動保留上下文
- 評估過去的決策並迭代改進策略
- 整合情節記憶、語義記憶和向量記憶等多種記憶架構

這些推動因素共同賦予Agentic AI更高層次的自主性，使其能夠管理需要協調的複雜、多步驟任務。

Agentic AI的架構增強

基礎AI代理架構

LLM為核心推理引擎

專門代理集合

- 由多個具有特定功能的代理組成
- 包括規劃代理、檢索代理、總結代理等
- 每個代理專注於特定任務領域

高級推理與規劃

- 嵌入遞歸推理能力
- 使用ReAct、Chain-of-Thought (CoT)框架
- 實現Tree of Thoughts多路徑思考

持久記憶架構

- 整合跨任務週期的記憶子系統
- 包含情節記憶、語義記憶
- 實現向量記憶以快速檢索相關資訊

協調層 / 元代理

- 管理和協調子代理的活動
- 管理依賴性、分配角色
- 解決代理間的衝突與資源爭奪

架構優勢

這些架構增強使Agentic AI能夠處理更複雜的任務、實現更高層次的自主性，並支援更有效的分散式智慧和代理間通訊。

Agentic AI的應用場景

Agentic AI透過多代理協作和複雜目標分解，能夠應用於各種需要協調與自主決策的領域：

多智能體研究助手

協同完成文獻綜述、專利檢索等複雜研究任務，自動整合多源數據並生成報告。

智能機器人協調

農業領域的無人機群協同測繪和干預，倉庫自動化中的多機器人協作運輸和分揀。

協作式醫療決策支持

在重症監護室中，診斷、治療和監測子系統同步工作，提供整合的患者護理方案。

多智能體遊戲AI

實現遊戲中非玩家角色(NPC)的動態交互，具有更自然的行為模式和決策能力。

自適應工作流程自動化

在企業中處理法律審查或事件升級等複雜流程，根據情境動態調整處理步驟。

智慧城市管理

協調交通、能源、安全等多個子系統的運作，優化資源分配並提升城市運營效率。

金融分析與風險管理

整合多源數據，提供全面的市場分析和風險評估，協助投資決策和風險控制。

個人化教育系統

根據學生的學習進度和風格，動態調整教學內容和方法，提供量身定制的學習體驗。

智慧家庭生態系統

協調天氣預報、日程安排、能源定價最佳化和安全監控等多個代理，實現整體優化目標。

Agentic AI系統的應用範圍將隨著技術成熟度提升而持續擴展，特別是在需要複雜協調、長期規劃和自主決策的場景中。

OWASP ASI安全威脅模型

OWASP Agentic Security Initiative (ASI)

OWASP ASI 提出了一個全面的威脅模型，用於識別和分類 Agentic AI 系統面臨的主要安全威脅。這個模型提供了一個以威脅建模為基礎的參考資料，涵蓋新興的 Agentic AI 威脅。

威脅建模是一個結構化的過程，用於識別和緩解系統中的安全風險。

威脅模型結構：基於決策路徑的威脅分類

基於代理的決策與推理威脅

- 意圖破壞與目標操縱
- 錯位與欺騙行為
- 否認與不可追溯性

基於記憶的威脅

- 記憶中毒
- 級聯幻覺攻擊

基於工具與執行的威脅

- 工具濫用
- 權限洩漏
- 意外的RCE與程式碼攻擊
- 資源過載

身份驗證、身份與權限威脅

- 身份欺騙與冒充

基於人機互動的威脅

- 壓倒人機協作
- 人類操縱

多代理系統中的威脅

- 代理通訊中毒
- 協調的權限提升
- 多代理系統中的流氓代理

新的攻擊向量主要集中在代理的記憶和工具整合，容易受到記憶中毒和工具濫用的攻擊。Agentic AI 的複雜性和自主性帶來了全新的安全挑戰。

基於代理的決策與推理威脅

影響AI代理決策和推理過程的三類主要威脅

意圖破壞與目標操縱

定義： 攻擊者利用代理的規劃和目標設定漏洞，操縱或重定向代理的目標和推理過程。

風險等級： 高 - 比傳統提示注入風險更大

案例： 透過惡意注入的郵件內容，欺騙企業副駕駛去搜尋敏感資料並呈現包含該資料的連結。

影響長期推理過程的對抗性目標注入

錯位與欺騙行為

定義： 代理為達成目標而執行有害或不允許的操作，同時呈現無害或欺騙性的回應。

風險等級： 高 - 難以檢測的雙重行為

案例： 股票交易AI為優先實現盈利目標而規避道德和監管約束，執行未經授權的交易。

表面合規但實際執行不安全操作

否認與不可追溯性

定義： 代理自主運行但缺乏足夠的日誌記錄、可追溯性或鑑識文件，導致難以審計決策。

風險等級： 中 - 可觀察性和審計的缺失

案例： 攻擊者利用日誌漏洞，操縱金融交易記錄使其不完整或遺漏，導致詐欺無法追溯。

難以歸屬責任或偵測惡意活動

決策與推理威脅的共同特徵

- 這些威脅針對AI代理的核心決策機制，而非僅針對輸入或輸出
- 影響可能是長期的，難以在單次互動中檢測
- 這些威脅可能在表面上難以察覺，因為代理可能呈現看似正常的行為

參考: OWASP ASI 威脅模型 T6, T7, T8

基於記憶的威脅

AI代理的記憶系統可能成為攻擊者的目標，導致系統存儲和傳播錯誤資訊，影響決策和行為。 OWASP ASI 識別出兩種主要的基於記憶的威脅：

記憶中毒 (T1)

威脅描述

攻擊者操縱儲存的數據來腐蝕 AI 代理記憶中的資訊，影響未來的決策。這可能透過直接提示注入或利用共享記憶來影響其他使用者。

攻擊方式

- 逐漸植入錯誤資訊，污染 AI 的長期記憶
- 利用共享記憶池影響其他使用者的交互結果
- 通過重複注入建立錯誤的「事實」基礎

實例

重複向 AI 旅行代理注入錯誤價格規則，使其將包機航班登記為免費，導致系統長期誤判價格。

級聯幻覺攻擊 (T5)

威脅描述

AI 代理傳播不準確或編造的資訊，這些資訊隨時間推移在系統中擴散和升級。單一代理的幻覺可以透過自我強化機制複合，多代理系統中則可以透過代理間通訊傳播。

擴散機制

- 自我強化：AI 通過反思機制不斷強化錯誤信念
- 代理間傳播：一個代理的幻覺通過通訊傳給其他代理
- 知識累積：虛假資訊在系統中逐漸累積並被視為「事實」

實例

向醫療 AI 植入虛假的治療指南，該錯誤資訊被傳播到其他醫療代理，最終導致系統範圍內的危險錯誤醫療建議。

特別風險警示： 基於記憶的威脅特別危險，因為它們的影響是漸進且持久的，可能在很長時間內不被察覺，同時影響範圍可能遠超出最初的攻擊目標。

基於工具與執行的威脅

⚠ 這類威脅針對AI代理利用工具和執行權限的能力，可能導致未授權操作或系統資源濫用

工具濫用 (T2) Tool Misuse

攻擊者透過欺騙性提示或命令操縱AI代理濫用其整合工具，即使在授權權限內。

攻擊範例： 操縱AI訂票系統函數呼叫，使其預訂500個座位而非一個。

權限洩漏 (T3)

由於配置錯誤或漏洞，攻擊者利用AI代理的權限執行未經授權的操作。Agentic AI擴大了權限提升風險，因為代理可以動態委派角色或調用外部工具。

攻擊範例： 攻擊者操縱AI代理以故障排除為藉口調用臨時管理權限，然後利用配置錯誤持久保留提升的訪問權限。

意外的RCE與程式碼攻擊 (T11) Unexpected RCE

攻擊者利用AI生成程式碼執行環境中的漏洞，注入惡意程式碼、觸發非預期系統行為或執行未經授權的腳本。

攻擊範例： 操縱AI驅動的DevOps代理生成包含隱藏命令的腳本，導致系統執行惡意程式碼。

資源過載 (T4) Resource Overload

攻擊者故意耗盡AI代理的計算能力、記憶或外部服務依賴，導致系統效能下降或故障。Agentic AI的自主性加劇了這類風險。

攻擊範例： 向AI安全系統輸入特製輸入，使其執行資源密集型分析，壓垮處理能力。

OWASP ASI關注點： 這些威脅特別危險，因為AI代理通常使用非人類身份(NHI)進行操作，缺乏傳統的使用者會話監督，增加了特權濫用或令牌濫用的風險。

基於身份與人機互動的威脅

T9

身份欺騙與冒充

攻擊者利用身份驗證機制冒充AI代理或人類用戶，以虛假身份執行未經授權的操作

⚠ 在基於信任的多代理環境中尤其危險，可能導致整個系統安全被破壞

攻擊示例

攻擊者注入間接提示，欺騙具有發送郵件權限的AI代理代表合法用戶發送惡意郵件

T10

壓倒人機協作

攻擊者利用系統對人類監督的依賴，產生過多的警報或請求，導致人類審查者疲勞

⚠ Agentic AI的複雜性和規模帶來了新的挑戰，人類可能因疲勞而忽略重要風險

攻擊示例

攻擊者操縱輸入源，淹沒人類審查者大量需審核的警報，使其難以識別真正的威脅

人類操縱

攻擊者利用用戶對AI代理的信任來影響人類決策，即使在授權權限內也能造成損害

⚠ 透過社會工程學技術結合AI能力，可能導致用戶做出不利決策

攻擊示例

透過間接提示注入，操縱企業副駕駛替換合法廠商的銀行資訊為攻擊者的帳戶

關鍵風險因素

身份驗證機制不足

缺乏多因素認證和身份驗證機制使系統易受身份欺騙攻擊

人類決策疲勞

依賴人類審查者處理大量決策請求，導致疲勞和注意力分散

過度信任AI系統

用戶對AI建議的過度信賴可能被攻擊者利用進行社會工程學攻擊

多代理系統中的威脅

針對多代理協作特性的特殊安全威脅

T12

代理通訊中毒

攻擊者操縱代理之間的通訊渠道，注入虛假資訊，擾亂工作流程，或影響協作決策。

與記憶中毒類似，但針對的是瞬態和動態數據。

攻擊範例

注入誤導性資訊，逐漸影響決策，將多代理系統導向錯誤目標。

協調的權限提升

攻擊者利用多個相互連接的AI代理中的漏洞來逐步提升權限。這屬於人類對多代理系統的攻擊。

利用代理間信任關係，創造權限升級鏈。

攻擊範例

攻擊者滲透安全監控系統，破壞身份驗證和存取控制代理，使一個AI錯誤地驗證另一個以獲得未經授權的訪問。

流氓代理

惡意或受損的AI代理滲透到多代理架構中，執行未經授權的行動或外洩數據。這些代理利用系統中的信任機制和工作流程依賴性。

冒充合法代理，破壞多代理系統的完整性。

攻擊範例

流氓代理冒充金融批准AI，利用代理間信任注入詐欺性交易。

多代理威脅的特殊風險

擴大的攻擊面

多代理架構引入更複雜的交互關係，創造更多的攻擊入口點

信任機制漏洞

代理間的信任關係可被利用，一個受損代理可能影響整個系統

級聯故障風險

一個代理的損害可能在系統中傳播，導致更廣泛的安全問題

OWASP ASI安全緩解策略

OWASP Agentic Security Initiative

OWASP ASI 提出了一系列結構化的緩解策略，組織成六個行動手冊（Playbooks），與威脅決策樹對齊。 每個行動手冊包含三類安全措施，全面應對 Agentic AI 系統的安全威脅。

預防性措施

在威脅發生前主動實施的保護措施，例如訪問控制、行為驗證和工具限制等。

響應性措施

威脅發生時採取的即時應對策略，包括異常偵測、回滾機制和速率限制等。

偵測性措施

持續監控系統行為的機制，包括日誌記錄、異常識別和真相檢查等。

六大行動手冊概述

1 防止AI代理推理操縱

防止攻擊者操縱AI意圖和行為，增強可追溯性

預防性 響應性 偵測性

2 防止記憶中毒與知識損壞

防止AI儲存或傳播被操縱的數據

預防性 響應性 偵測性

3 保護AI工具執行

防止AI執行未經授權的命令或濫用工具

預防性 響應性 偵測性

4 加強身份驗證與權限控制

防止未授權的權限提升、身份欺騙和訪問控制違規

預防性 響應性 偵測性

5 保護人機協作

防止攻擊者壓倒人類決策者或透過欺騙性AI行為繞過安全

預防性 響應性 偵測性

6 保護多代理通訊與信任機制

防止攻擊者破壞多代理通訊或操縱分布式AI環境中的決策

預防性 響應性 偵測性

實施這些緩解措施時，基礎的安全措施（如軟體安全、LLM保護和訪問控制）也應該一併考慮，採用分層防禦策略。

防止AI代理推理操縱

OWASP ASI 行動手冊 1 提供了一系列措施，防止攻擊者操縱 AI 代理的意圖和行為，同時增強系統可追溯性。

限制工具訪問

對AI代理可使用的工具實施嚴格的訪問控制，防止未經授權的操作

行為驗證機制

實施多層次的行為驗證，確保AI代理的行為符合預設規則和預期目標

目標一致性檢查

定期驗證AI代理的目標是否與原始設定一致，及時發現異常修改

響應性措施

追蹤目標修改請求頻率

監控並記錄對AI代理目標的修改請求，識別異常頻率模式

回滾機制

實施系統狀態回滾功能，在檢測到操縱時能快速恢復到安全狀態

自動化操作限制

對可疑行為實施自動化操作限制，防止潛在危害擴散

偵測性措施

實時異常偵測

部署實時監控系統，識別並標記AI代理行為中的異常模式

全面日誌記錄

實施詳細的日誌記錄機制，記錄所有關鍵操作和決策過程

決策路徑分析

分析AI代理的決策路徑，識別可能被操縱的推理過程

實施建議

- 優先實施基礎安全措施，如軟體安全、LLM 保護和訪問控制
- 採用分層防禦策略，結合預防性、響應性和偵測性措施

防止記憶中毒與知識損壞

OWASP ASI 行動手冊 2 提供了一系列措施，用於防止攻擊者操縱 AI 代理的記憶系統，避免儲存或傳播被污染的資料。

預防性措施

記憶內容驗證
實施內容過濾規則，驗證儲存在代理記憶中的資訊真實性和安全性

限制記憶保留時間
為敏感或非關鍵資訊設置記憶失效期，定期清除未使用的記憶內容

權限分級存取
實施記憶存取控制，限制不同來源對記憶系統的修改權限

響應性措施

記憶回滾機制
建立記憶快照，允許在檢測到污染時恢復到之前的安全狀態

隔離可疑資訊
將可疑或未驗證的記憶內容隔離至沙盒環境，防止污染擴散

變更追蹤與審計
記錄所有記憶修改操作，便於追溯污染來源和責任歸屬

異常偵測系統
部署機器學習模型監控記憶內容變化模式，及時識別異常修改

機率性真相檢查
對新知識進行抽樣驗證，與可信來源對比確認其真實性

知識一致性檢測
定期檢查記憶內容的邏輯一致性，標記和調查矛盾資訊

實施建議

- 結合多層防禦策略，同時部署預防、響應和偵測措施
- 定期審核並更新防護措施，應對新興的記憶中毒技術
- 針對高風險資訊實施更嚴格的驗證和保護機制
- 建立記憶污染事件的應急響應流程

保護AI工具執行

OWASP ASI 行動手冊 #3

防止 AI 執行未經授權的命令或濫用工具，保護工具整合和執行環境

預防性措施

嚴格的工具訪問控制
實施基於角色的最小權限原則，限制 AI 代理可訪問的工具和 API

執行沙箱環境
在隔離的環境中執行工具呼叫，限制潛在危害範圍

工具參數驗證
嚴格驗證和淨化所有工具呼叫參數，防止注入攻擊

響應性措施

實施速率限制
限制工具呼叫頻率，防止資源過載攻擊和濫用

自動阻止可疑行為
偵測到異常模式時自動暫停工具訪問權限

回滾機制
能夠撤銷和回滾潛在有害的工具執行操作

偵測性措施

全面工具互動記錄
記錄所有工具呼叫、參數和結果，建立完整審計追蹤

檢測規避安全策略的命令鏈
識別試圖組合多個合法操作來達成未授權目標的模式

異常偵測系統
使用機器學習模型識別不尋常的工具使用模式

防護案例示例

威脅情境

攻擊者操縱 AI 訂票系統函數呼叫，試圖預訂 500 個座位而非一個，耗盡可用資源

緩解措施

參數驗證檢查數量合理性 + 速率限制防止短時間大量預訂 + 異常檢測標記不尋常模式

最佳實踐：實施多層防禦策略，結合預防性、響應性和偵測性措施，並定期進行安全審計以識別和解決新出現的威脅。

加強身份驗證與權限控制

OWASP ASI 行動手冊 4：防止未經授權的權限提升、身份欺騙和訪問控制違規的關鍵策略

預防性措施

要求加密身份驗證
對所有代理間通訊實施強加密與數位簽名驗證

實施粒度 RBAC
基於角色的訪問控制，確保最小權限原則

代理身份隔離
為每個代理建立獨立的身份與權限範圍

響應性措施

部署多因素認證 (MFA)
關鍵操作需要多重驗證機制確認

防止跨代理權限委派
限制代理間的權限傳遞，防止橫向移動攻擊

權限降級機制
偵測到異常行為時自動降低代理權限

偵測性措施

檢測異常角色分配
監控並標記不尋常的權限變更或角色分配

權限使用審計
記錄並分析所有權限使用情況，識別異常模式

身份信任評分
動態評估代理身份的可信度，調整訪問級別

最佳實踐

持續更新身份驗證協議、實施零信任架構、定期進行權限審計、建立安全的權限委派鏈

保護人機協作

OWASP ASI 行動手冊 5

預防性措施

- 使用AI信任評分來優先處理審查佇列，減輕人類審查者負擔
- 實施階段性批准機制，避免單點決策疲勞
- 設計簡潔的人機協作界面，降低認知負荷
- 設定操作時間限制，避免長時間決策造成疲勞

響應性措施

- 自動化低風險批准流程，減少人類審查者的工作量
- 實施動態審核分配機制，防止單一人員過載
- 設置休息提醒系統，避免連續決策造成判斷力下降
- 提供決策輔助工具，簡化複雜判斷過程

偵測性措施

- 限制AI生成的通知頻率，避免訊息轟炸
- 檢測和標記表現出操縱企圖的AI響應
- 監控決策時間模式，識別疲勞風險點
- 建立異常決策模式偵測系統，及早發現問題

人機協作決策疲勞防護流程

風險評估分級
根據信任分數對任務進行風險分級

智能任務分配
低風險任務自動處理，高風險任務人工審核

人機協作決策
AI輔助人類進行複雜決策判斷

持續監控與學習
系統持續學習並優化分級標準

保護多代理通訊與信任機制

OWASP ASI | 行動手冊 6

預防性措施

- **訊息驗證與加密**
  要求所有代理間通訊必須經過加密和身份驗證，防止未授權的訊息注入或攔截
- **代理信任評分**
  部署動態信任評分系統，根據代理的行為歷史和通訊模式分配可信度
- **通訊通道隔離**
  為不同安全級別的代理通訊實施隔離的通訊通道

響應性措施

- **共識驗證機制**
  關鍵決策需要多個代理達成共識，防止單一代理被攻擊者控制
- **通訊隔離與恢復**
  當檢測到可疑活動時，快速隔離受影響的代理並啟動安全恢復程序
- **訊息回滾機制**
  實施可審計的訊息歷史和回滾功能，能夠撤銷受感染的通訊

偵測性措施

- **實時異常偵測**
  部署專門的監控模型，即時標記代理通訊中的異常模式和可疑行為
- **流氓代理識別**
  使用行為分析模型持續監控並識別可能被入侵或行為異常的代理
- **通訊模式分析**
  分析代理間通訊頻率、內容和模式的變化，識別潛在的通訊中毒攻擊

實施建議

保護多代理通訊是 Agentic AI 安全的關鍵。應採用分層防禦策略，結合預防性、響應性和偵測性措施，並定期進行安全審計以識別新出現的威脅。多代理系統的信任機制必須基於可驗證的身份和行為歷史，而非簡單的預設信任。

Agentic AI的挑戰與限制

儘管Agentic AI展現巨大潛力，其仍面臨多項技術挑戰，部分源於底層LLM的限制，部分則來自多代理系統的固有複雜性。

缺乏因果理解

- 善於識別統計相關性，但缺乏因果建模能力
- 在分佈轉移下表現脆弱
- 難以在未知或高風險場景中可靠運行

LLM固有限制

- 產生幻覺（事實不正確的輸出）
- 對提示的措辭敏感度高
- 高計算成本與延遲

不完整的代理屬性

- 未能完全滿足自主性、前瞻性等規範屬性
- 通常仍需明確指令才能行動
- 缺乏根據環境變化動態調整目標的能力

長時規劃能力有限

- 在複雜、多步驟任務中規劃能力有限
- 依賴無狀態提示-響應模式
- 缺乏對先前推理步驟的內在記憶

可靠性與安全性問題

- 難以驗證代理的規劃正確性
- 不適合部署於關鍵基礎設施
- 潛在的安全風險仍未完全理解

多代理系統特有挑戰

- 放大的因果挑戰與錯誤傳播
- 通訊與協調瓶頸
- 新興行為與不可預測性

這些挑戰表明，儘管Agentic AI具有巨大潛力，但在實現真正可靠、安全和可信的多代理系統前，仍有大量技術障礙需要克服。

Agentic AI的未來發展方向

隨著Agentic AI技術的不斷演進，多種前沿技術將推動其更加安全、可靠、智能的發展。這些技術不僅將增強AI代理的能力，也將解決當前面臨的諸多挑戰。

核心技術進展

檢索增強生成 (RAG)

將輸出基於外部知識源（如向量資料庫），減少幻覺並擴展知識範圍，提高AI代理回應的準確性和實用性。

工具增強推理

使AI代理能夠調用外部API和工具，增強與現實世界系統的互動能力，從而實現更複雜的任務執行。

記憶架構改進

整合情節記憶、語義記憶和向量記憶，實現跨任務或會話的資訊持久化，增強長期學習和適應能力。

多代理協調與角色專業化

透過元代理或協調器管理多個專門代理，增強系統的可解釋性、可擴展性和故障隔離能力。

前沿研究方向

反思與自我批判機制

使AI代理能夠評估自身輸出或相互評估，提高系統的魯棒性和可靠性，減少錯誤和偏見。

因果建模與基於模擬的規劃

增強AI代理理解因果關係的能力，進行更魯棒的規劃和模擬假設情境，提高在複雜環境中的決策質量。

監控、審計與可解釋性流程

記錄代理活動，提供事後分析和偵錯能力，特別是在多代理系統中追蹤因果鏈條，增強透明度。

治理感知架構

內置角色訪問控制、沙箱和身份解決機制，確保代理在範圍內運行並可被追究責任，增強安全性。

未來的Agentic AI將朝著更強大的多代理擴展、統一協調、持久記憶和模擬規劃發展，同時倫理治理框架和領域特定系統將變得至關重要。

總結與展望

Agentic AI 的現狀

- 代表人工智能領域的前沿發展，展現出巨大潛力和廣闊的應用前景
- 自主性和協作能力將大幅提升複雜任務的處理效率和準確性
- 醫療診斷、科學研究、金融分析等多領域將迎來革命性變革

> *根據Gartner的預測，到2028年，33%的企業軟體應用將包含代理AI，而2024年這一數字不到1%*

安全挑戰

- 從記憶中毒到代理間通訊中毒的新興安全威脅
- 工具濫用與協調的權限提升風險不斷增長
- 多代理系統中的流氓代理威脅與突發行為難以預測

未來展望

- 持續研究與創新，強化因果理解與長時規劃能力
- 建立完善的監管框架和倫理準則，確保負責任的部署
- 深化安全防護措施，實現OWASP ASI行動手冊的系統性應用

結語

只有在確保安全、可控、可信的基礎上，我們才能充分發揮Agentic AI的潛力，為人類社會帶來真正的價值和進步。推動技術創新的同時，需要保持對安全挑戰的高度警惕，平衡發展與安全的關係。