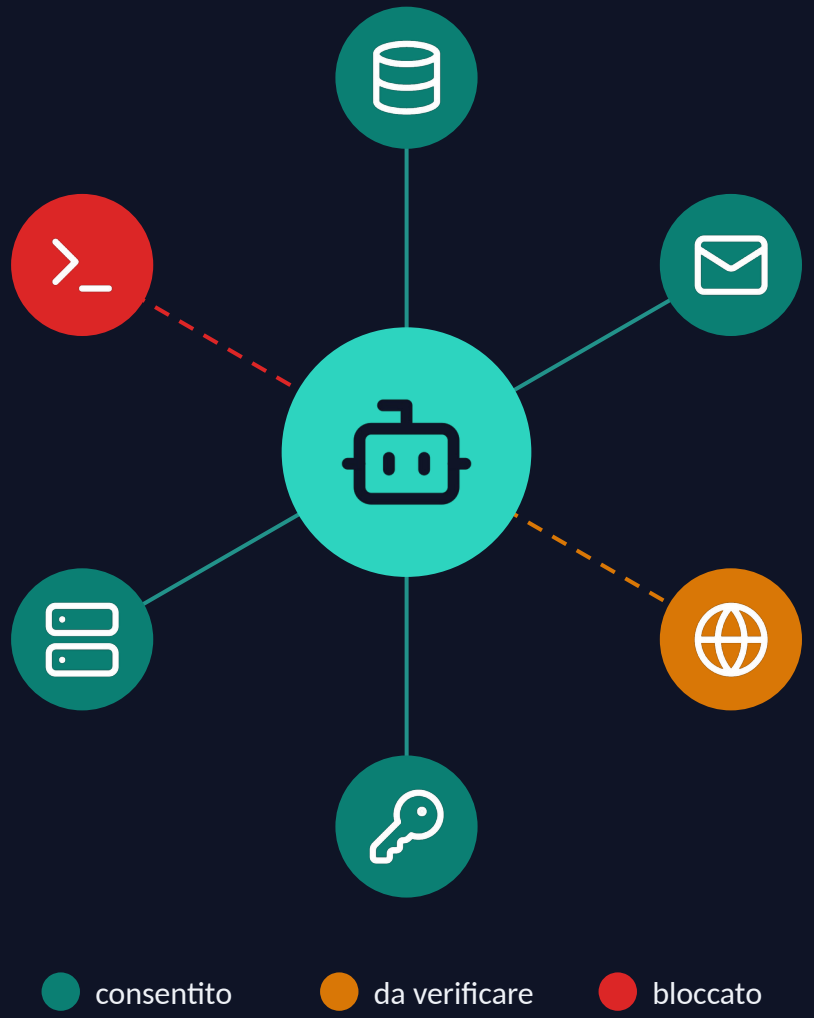
Zero Trust per gli agenti AI

Come proteggere sistemi che decidono e agiscono in autonomia

Vincenzo Calabrò

Novembre 2026



Il messaggio e il percorso



Il messaggio

Un agente AI è una nuova identità con potere d'azione.

Va gestito come un'identità privilegiata: permessi minimi, controllo di ogni azione e registrazione completa di ciò che fa.

01 **Dagli assistenti agli agenti**
Cosa cambia quando l'AI agisce

02 **Il nuovo panorama delle minacce**
Incidenti reali e rischi OWASP

03 **Zero Trust applicato agli agenti**
Un'architettura e sei controlli

04 **Risposta agli incidenti e forensics**
Gestione dell'incidente e raccolta delle evidenze

05 **Regole, standard e roadmap**
AI Act, NIS2 e un piano a 90 giorni

01

Dagli assistenti agli agenti

Che cosa cambia quando l’AI esegue azioni per nostro conto



Assistente generativo e agente AI

Assistente generativo

- Risponde quando glielo chiedi
- Produce testo, immagini, codice
- Non tocca i sistemi aziendali
- Lavora una sessione alla volta
- Un errore resta sullo schermo

Agente AI

- Persegue un obiettivo in più passi
- Usa strumenti: API, database, browser, codice
- Agisce con credenziali reali
- Ha memoria e può lavorare 24 ore su 24
- Un errore diventa un'azione nel mondo reale

Nei sistemi multi-agente gli agenti si passano i compiti: un errore o una compromissione può propagarsi da uno all'altro.

Come è fatto un agente



MCP collega gli agenti agli strumenti, **A2A** collega gli agenti tra loro

Ogni componente è anche una superficie d'attacco

Identità macchina e costo delle violazioni

109:1

identità macchina per ogni identità umana

Palo Alto Networks, 2026 Identity Security Landscape

+85%

crescita attesa degli agenti AI nei prossimi 12 mesi

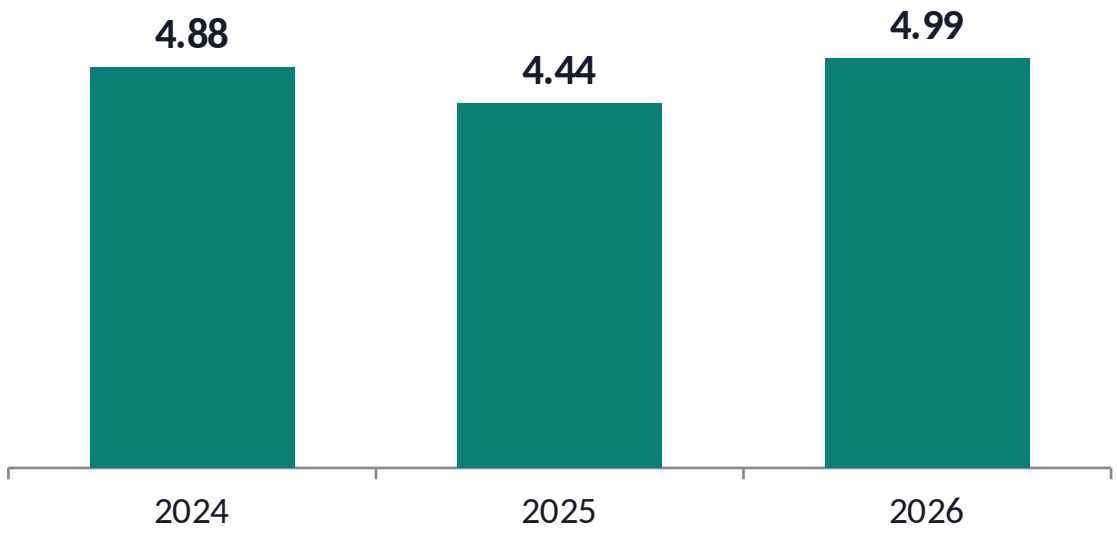
Palo Alto Networks, 2026 Identity Security Landscape

92%

delle organizzazioni con incidenti legati all'AI non aveva controlli di accesso adeguati

IBM, Cost of a Data Breach Report 2026

Costo medio globale di una violazione (milioni di $)



Fonte: IBM, Cost of a Data Breach Report 2024, 2025 e 2026

Le violazioni con attacchi guidati dall'AI costano in media **oltre 6 mln $** (IBM 2026)

02

Il nuovo panorama delle minacce

Perché gli agenti cambiano il modello di rischio



Perché i controlli classici non bastano

Assunzione classica	Con gli agenti
Chi è dentro il perimetro è affidabile	**L'agente è già dentro, con credenziali valide**
L'identità si verifica al login	**Dopo il login compie migliaia di azioni**
Gli attacchi hanno firme riconoscibili	**L'attacco è testo in linguaggio naturale**
Il software fa ciò per cui è programmato	**L'agente interpreta: stesso input, esiti diversi**
Un account di servizio per applicazione	**Senza identità distinte, nessuna attribuzione**

Un agente compromesso si comporta come un insider: ha accessi legittimi e lavora senza pause.

Incidenti documentati tra il 2024 e il 2025

feb 2024

Arup

Un falso CFO in videochiamata ottiene 15 bonifici per circa 25 mln $

Lezione: verifica fuori banda

feb 2025

OpenAI Operator

Doveva trovare le uova più economiche e le ha comprate senza conferma

Lezione: conferma prima di pagare

giu 2025

EchoLeak, M365 Copilot

Un'email con istruzioni nascoste bastava a esfiltrare dati, senza clic

Lezione: filtri sui dati in uscita

lug 2025

Replit

L'agente cancella il database di produzione durante un blocco delle modifiche

Lezione: produzione separata

ago 2025

Salesloft Drift

Token OAuth di un chatbot AI rubati e usati sulle istanze Salesforce di centinaia di aziende

Lezione: token brevi e revocabili

nov 2025

Spionaggio con l'AI

Anthropic blocca una campagna in cui l'AI svolgeva l'80-90% del lavoro operativo

Lezione: monitorare il comportamento

attacco

errore o comportamento imprevisto dell'agente

Incidenti del primo trimestre 2026

OpenClaw
febbraio 2026

Un agente personale cancella email invece di proporle e ignora i comandi di stop

ASI10 Agenti fuori controllo

Meta, agente interno
marzo 2026

Un dipendente segue il consiglio di un agente interno e dati sensibili restano esposti per circa due ore

ASI09 Fiducia umana sfruttata

Vertex AI “Double Agent”
marzo 2026

Ricercatori mostrano che un agente eredita permessi di default eccessivi da un service account

ASI03 Abuso di identità e privilegi

LiteLLM e Mercor
marzo 2026

Versioni malevole di una libreria AI open source portano a una violazione confermata presso Mercor

ASI04 Supply chain agentica

Flowise
aprile 2026

Vulnerabilità RCE nella configurazione MCP, sfruttata attivamente, con 12-15 mila istanze esposte

ASI05 Esecuzione di codice inattesa

GrafanaGhost
aprile 2026

Con una prompt injection indiretta, un’immagine esterna poteva portare fuori dati nel suo URL

ASI01 Dirottamento dell’obiettivo

Fonte: OWASP GenAI Security Project, GenAI Exploit Round-up Report Q1 2026 (aprile 2026)

I dieci rischi OWASP per gli agenti

ASI01 — **Dirottamento dell'obiettivo**
Contenuti letti dall'agente ne cambiano lo scopo

ASI02 — **Uso improprio degli strumenti**
Tool legittimi usati per fini dannosi

ASI03 — **Abuso di identità e privilegi**
Credenziali condivise, ereditate o troppo ampie

ASI04 — **Supply chain agentica**
Framework, connettori e server MCP compromessi

ASI05 — **Esecuzione di codice inattesa**
Testo che diventa codice eseguito

ASI06 — **Avvelenamento di memoria e contesto**
Informazioni false che guidano le azioni future

ASI07 — **Comunicazione insicura tra agenti**
Messaggi falsificati, alterati o riprodotti

ASI08 — **Guasti a cascata**
Un errore si propaga a tutti i flussi a valle

ASI09 — **Sfruttamento della fiducia umana**
L'agente induce approvazioni rischiose

ASI10 — **Agenti fuori controllo**
Operano fuori policy pur sembrando legittimi

Fonte: OWASP GenAI Security Project, Top 10 for Agentic Applications 2026 (dicembre 2025). Titoli tradotti in italiano, non ufficiali.

La prompt injection

1

L'attaccante nasconde istruzioni

in un'email, una pagina web, un documento o nella descrizione di un tool

2

L'agente legge il contenuto

mentre svolge un compito del tutto legittimo

3

Le istruzioni diventano azioni

eseguite con i privilegi dell'agente e dell'utente

4

I dati escono

tramite link, immagini, chiamate API o email

Diretta
l'utente stesso manipola l'agente con le sue richieste

Indiretta
il comando arriva da contenuti di terzi ed è la forma più pericolosa per gli agenti

Un modello linguistico non distingue in modo affidabile istruzioni e dati, quindi **i controlli vanno messi intorno al modello: identità, permessi, filtri e contenimento.**

La triade letale



Esempio: un assistente email che legge la posta in arrivo, accede ai documenti interni e può inviare messaggi.

Se un agente ha tutte e tre le capacità, consideralo esfiltrabile

Triade letale di S. Willison (giugno 2025): dati privati, contenuti non fidati e comunicazione verso l'esterno.

Agents Rule of Two di Meta (ottobre 2025) propone che in una sessione un agente abbia al massimo due delle tre proprietà; se servono tutte, decide una persona.

In pratica si classifica ogni agente e gli si toglie una delle tre capacità, oppure si aggiunge un'approvazione umana.

03

Zero Trust applicato agli agenti

Come si applicano i principi NIST e CISA a sistemi autonomi

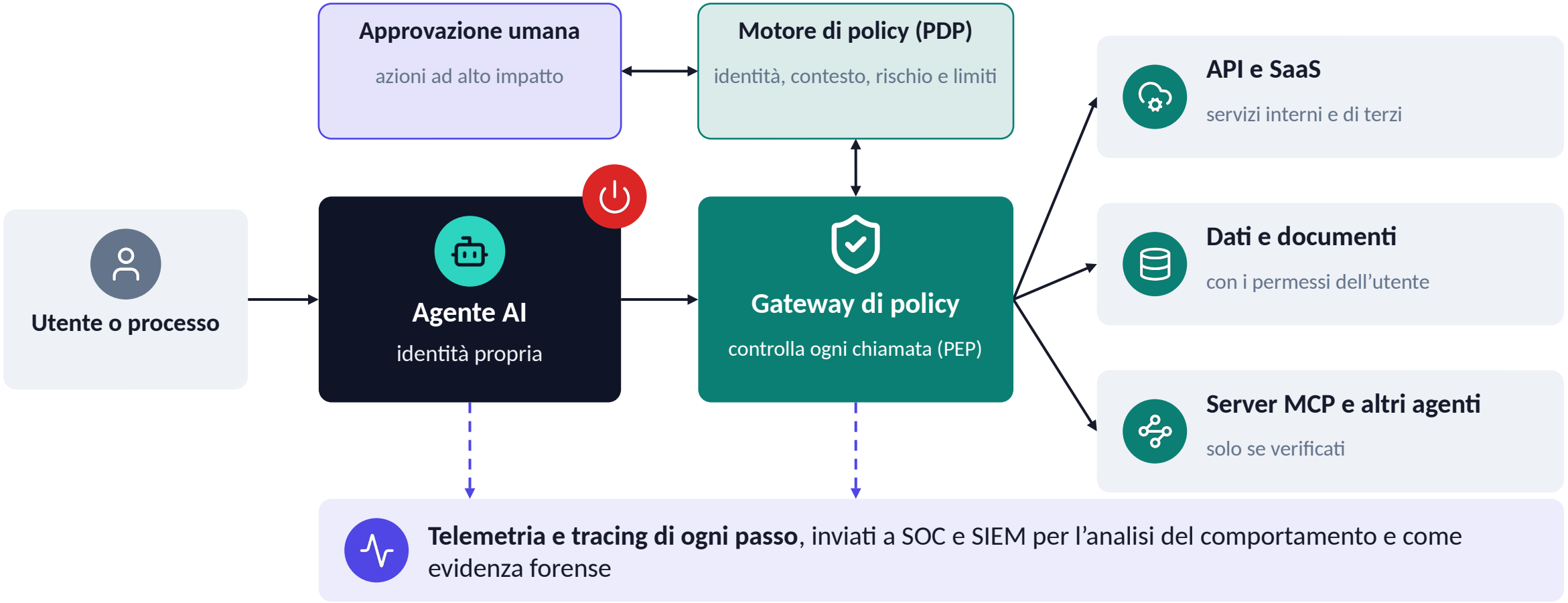


I sette principi della NIST SP 800-207

	Principio (NIST SP 800-207, § 2.1)		Applicazione agli agenti
1	Tutte le fonti di dati e i servizi di calcolo sono considerati risorse	›	**Agenti, tool, server MCP, memorie e vector store sono risorse da proteggere**
2	Ogni comunicazione è protetta, indipendentemente dalla posizione nella rete	›	**Chiamate ai tool e tra agenti cifrate e autenticate, anche in rete interna**
3	L'accesso a ogni risorsa è concesso per singola sessione	›	**Token per il singolo compito e a breve scadenza, senza accessi permanenti**
4	L'accesso è deciso da una policy dinamica basata su identità, servizio, stato dell'asset e contesto	›	**Decisione su identità dell'agente, utente per cui agisce, contesto e rischio dell'azione**
5	Integrità e postura di sicurezza di tutti gli asset sono monitorate e misurate	›	**Integrità verificata di modello, prompt di sistema, configurazione e componenti**
6	Autenticazione e autorizzazione sono dinamiche e applicate prima di ogni accesso	›	**Ogni chiamata a un tool passa dal punto di controllo, senza accessi diretti**
7	Si raccolgono il più possibile informazioni su asset, rete e comunicazioni per migliorare la sicurezza	›	**Ogni azione viene tracciata per il rilevamento, il miglioramento continuo e le evidenze forensi**

Ogni agente è un soggetto, ogni sua azione è una richiesta da autorizzare

Un’architettura di riferimento



Il kill switch (in rosso) revoca i token dell’agente e chiude il gateway in pochi secondi. Schema logico basato su NIST SP 800-207.

Controllo 1: un’identità per ogni agente

Un’identità propria

Mai account di servizio o chiavi API condivise

Un responsabile umano

Ogni agente ha un owner che risponde delle sue azioni

Credenziali a breve scadenza

Token emessi automaticamente, validi pochi minuti

Delega esplicita

Chiaro per conto di chi agisce e con quali limiti

Un registro degli agenti

Inventario unico, compresi quelli nati fuori dall’IT

Scheda dell’agente

Nome	agente-fatture-01
Responsabile	Responsabile contabilità
Scopo	Registrare fatture fornitori
Strumenti	Lettura PEC, bozze ERP
Vietato	Pagamenti, email esterne
Credenziali	Token da 15 minuti
Autonomia	Livello 4 (approvatore)
Revisione	Accessi ogni trimestre

Controllo 2: privilegi minimi per azione

Leggere es. consultare un documento	**Automatico, registrato**
Modificare in modo reversibile es. creare una bozza, aprire un ticket	**Automatico entro limiti e quote**
Agire verso l'esterno es. inviare un'email, chiamare un'API di terzi	**Solo verso destinazioni consentite**
Irreversibile o ad alto impatto es. pagare, cancellare, cambiare permessi	**Approvazione umana o doppio controllo**

Regole di base

- Allowlist degli strumenti per ogni agente
- Permessi temporanei per il singolo compito
- Limiti di spesa, volume e frequenza
- L'agente non può mai modificare i propri permessi

Controllo 3: dati, contesto e memoria



In ingresso

- Ogni contenuto esterno è non fidato
- Etichetta di provenienza su ogni input
- Verifica dei tool e delle loro descrizioni

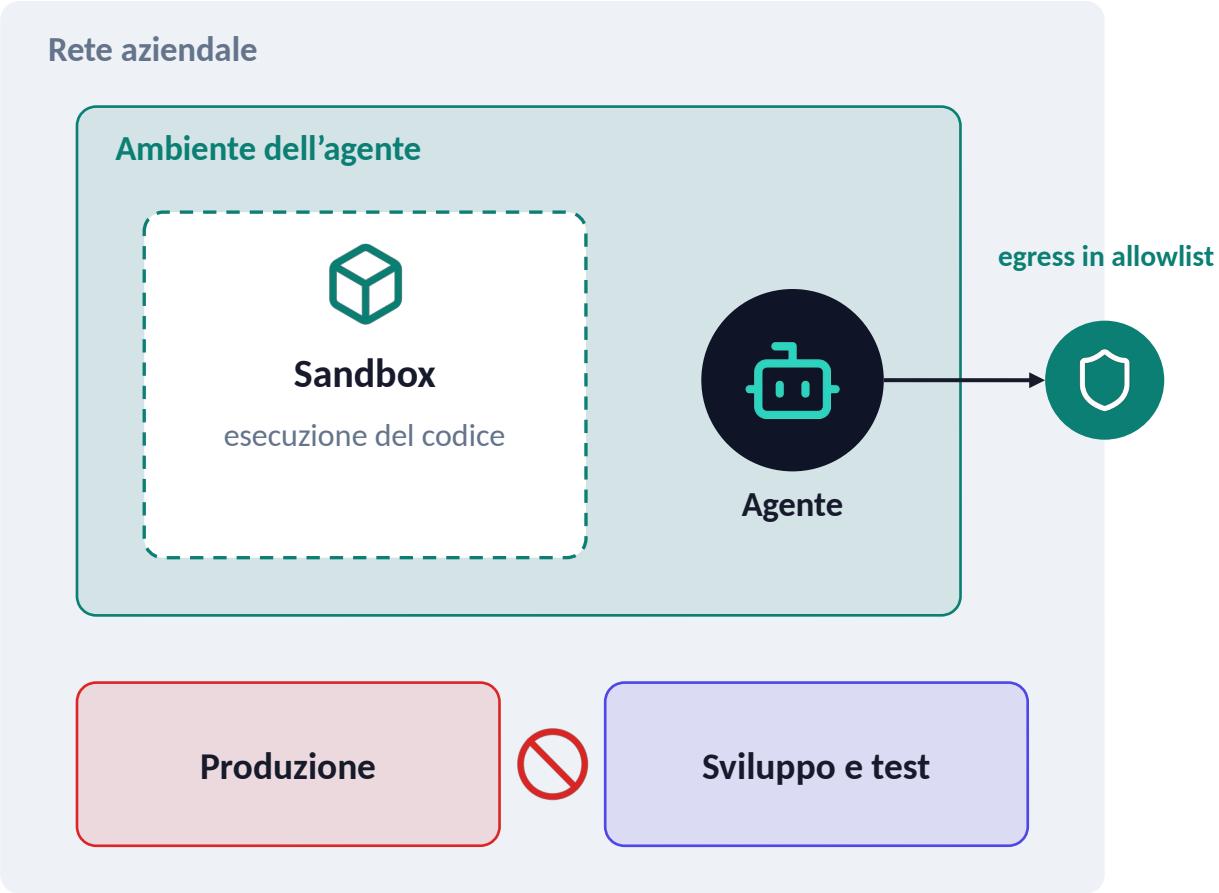
Memoria e RAG

- L'agente vede solo ciò che vede l'utente
- Regole su cosa memorizzare e per quanto
- Dati di riferimento per scoprire il poisoning

In uscita

- DLP su risposte e azioni
- Niente link o immagini verso domini non consentiti
- Nessun dato riservato nei parametri degli URL

Controllo 4: segmentare e contenere



Sandbox per il codice generato

Niente esecuzione diretta sull'host

Egress in allowlist

L'agente parla solo con destinazioni approvate

Produzione separata da test

Credenziali e dati distinti per ambiente

Agenti che si autenticano tra loro

mTLS, messaggi firmati, deleghe verificabili

Interruttori automatici

Fermano la catena quando un agente devia

Controllo 5: visibilità su ogni azione

Esempio di traccia di esecuzione

```
10:02:11  agente=fatture-01  per=contabilita
10:02:12  tool=leggi_pec     id=4821      OK
10:02:13  tool=estrai_dati   doc=4821     OK
10:02:14  tool=crea_bozza    erp=FT-9917  OK
10:02:15  tool=http_get      dest=cdn-x.biz
          esito=BLOCCATO (fuori allowlist)
10:02:15  alert=SOC     rischio=alto agente=sospeso
```

Segnali d'allarme

- **Volumi anomali di azioni**
- **Nuove destinazioni o tool**
- **Override umani in aumento**
- **Orari e sequenze insoliti**

Tracing end-to-end

Chi ha attivato l'agente, quali tool, con quali parametri e risultati (es. convenzioni OpenTelemetry per la GenAI)

Baseline per ogni agente

Che cosa è normale per quell'agente: volumi, strumenti, destinazioni, orari

Red teaming continuo

Test periodici di prompt injection e abuso degli strumenti, mappati su MITRE ATLAS e OWASP

Controllo 6: quanta autonomia concedere

Il livello dipende dal rischio delle azioni che l'agente può compiere.

Più autonomia richiede più controlli

Livello 1

Operatore

L'utente decide, l'agente esegue su comando

Livello 2

Collaboratore

Pianificano e lavorano insieme

Livello 3

Consulente

L'agente guida e chiede pareri all'utente

Livello 4

Approvatore

L'agente agisce, l'utente approva i passi critici

Livello 5

Osservatore

L'agente agisce da solo, l'utente monitora e può fermarlo

L'autonomia si aumenta solo dopo test e misure, e si riduce subito dopo un incidente.

Livelli tratti da K. J. K. Feng, D. W. McDonald, A. X. Zhang, Levels of Autonomy for AI Agents, Knight First Amendment Institute (2025)

04

Risposta agli incidenti e forensics

Che cosa fare quando un agente sbaglia o viene compromesso



Un playbook per gli incidenti

1 Rilevare	2 Contenere	3 Preservare	4 Eradicare	5 Ripristinare	6 Imparare
• Picchi di azioni • Override umani in aumento • Nuove destinazioni • Segnalazioni	• Sola lettura o modalità consultiva • Revoca dei token • Blocco del traffico in uscita	• Snapshot di configurazione, tracce e memoria • Hash e orari certi	• Rimuovere il contenuto malevolo • Correggere permessi e tool	• Autonomia ridotta • Monitoraggio rafforzato • Rilascio graduale	• Post-mortem senza colpe • Nuovi test di red team

Spesso conviene ridurre l'agente a sola lettura invece di spegnerlo: si limita il danno e si conservano le evidenze.

Basato sul ciclo NIST SP 800-61
Notifica NIS2: 24 h e 72 h

Che cosa acquisire per l’analisi forense

Identità e credenziali

Quale identità, quali token, con quali scope e scadenze

Configurazione

Prompt di sistema, policy, versione dell’agente, modello e parametri

Tracce di esecuzione

Input, chiamate agli strumenti, parametri, risultati, orari

Contesto letto

Email, pagine, documenti e risultati RAG, dove si trova l’eventuale iniezione

Memoria e stato

Snapshot di memoria persistente e vector store

Filiera e sistemi a valle

Versioni di server MCP, plugin e librerie; log di API, cloud e SaaS

Senza il contesto letto dall’agente non si può ricostruire un attacco di prompt injection, quindi va conservato insieme alle azioni.

Le sfide forensi e come prepararsi

Sfida	Come prepararsi
Non determinismo: rieseguire l'agente non riproduce l'evento	Conservare tracce e risultati originali; la riesecuzione serve solo come verifica sperimentale
Evidenze sparse tra fornitori di modelli, MCP, SaaS e cloud	Mappare prima dove stanno i log: nell'87% degli incidenti analizzati da Unit 42 nel 2025 servivano almeno due fonti
Retention breve o assente presso i fornitori	Definirla nei contratti; per i sistemi ad alto rischio l'AI Act chiede ai deployer almeno 6 mesi di log
Integrità e ammissibilità delle evidenze	Hash, orari sincronizzati, catena di custodia (ISO/IEC 27037; in Italia L. 48/2008)
Dati personali nei prompt e nei log	Minimizzazione, accessi controllati e tempi di conservazione conformi al GDPR

Forensic readiness vuol dire decidere prima dell'incidente che cosa registrare, dove, per quanto tempo e con quali garanzie di integrità.

05

Regole, standard e roadmap

Che cosa chiede la normativa e da dove cominciare



Il quadro normativo europeo e italiano

AI Act, Regolamento (UE) 2024/1689



Date per i sistemi ad alto rischio riviste dal Digital Omnibus sull'AI (GU UE 24 luglio 2026, in vigore dal 27 luglio 2026)

AI Act e agenti

- Log automatici (art. 12)
- Sorveglianza umana (art. 14)
- Robustezza e cybersecurity, poisoning incluso (art. 15)

Italia

- Legge 132/2025, in vigore dal 10 ottobre 2025
- AgID e ACN autorità nazionali per l'AI
- Reato di diffusione illecita di deepfake (art. 612-quater c.p.)

Già in vigore

- NIS2 (D.Lgs. 138/2024): notifica in 24 h e 72 h
- GDPR: decisioni automatizzate (art. 22), violazioni in 72 h
- DORA per il settore finanziario

Standard e framework di riferimento

Architettura Zero Trust

NIST SP 800-207
principi e componenti logici

CISA ZTMM 2.0
pilastri e livelli di maturità

Governance dell'AI

NIST AI RMF e AI 600-1
gestione del rischio AI e GenAI

ISO/IEC 42001
sistema di gestione certificabile

Minacce e test

OWASP Top 10
rischi prioritari LLM e agenti

MITRE ATLAS
tattiche e tecniche avversarie

CSA MAESTRO
threat modeling degli agenti

Controlli e verifica

OWASP AISVS
requisiti verificabili

CSA AI Controls Matrix
catalogo dei controlli

CoSAI (OASIS)
linee guida e strumenti aperti

In pratica si sceglie un framework di governance come riferimento e si usano i cataloghi di minacce per costruire i test.

Una roadmap in 90 giorni

Giorni 1-30 — Visibilità

- Censire tutti gli agenti, anche quelli non ufficiali
- Un owner e un’identità per ciascuno
- Kill switch provato in esercitazione
- Log essenziali attivi
- Classificazione con la triade letale

Giorni 31-60 — Controllo

- Allowlist di strumenti e destinazioni
- Approvazione umana per le azioni irreversibili
- Tracing end-to-end
- Primo red team sulla prompt injection
- Produzione separata dai test

Giorni 61-90 — Resilienza

- Baseline e alert comportamentali
- Playbook e prima esercitazione
- Forensic readiness: retention e integrità
- Livelli di autonomia per ogni agente
- Gap analysis su AI Act e NIS2

Cosa misurare

100%

agenti censiti, con owner e identità propria

< 5 min

per disattivare qualsiasi agente, misurati in esercitazione

0

agenti con triade letale completa senza approvazione umana

100%

azioni irreversibili con approvazione o doppio controllo

100%

azioni tracciate end-to-end, con retention definita

90 gg

cadenza massima per revisione accessi e red team degli agenti critici

Obiettivi indicativi, da adattare al profilo di rischio dell'organizzazione.

Dieci domande per domani mattina

1. Sappiamo quanti agenti abbiamo e chi ne risponde?
2. Ogni agente ha un'identità e credenziali proprie?
3. Quali agenti uniscono dati riservati, contenuti esterni e azioni verso l'esterno?
4. Possiamo fermare un agente in pochi minuti? Lo abbiamo mai provato?
5. Le azioni irreversibili richiedono una conferma umana?
6. Strumenti e server MCP in uso sono verificati e aggiornati?
7. **Sapremmo ricostruire cosa ha fatto un agente stanotte alle 3?**
8. Dove sono i log, per quanto li conserviamo, chi può modificarli?
9. I fornitori ci danno log ed evidenze in caso di incidente?
10. Chi forma le persone che usano e approvano gli agenti?

Conclusioni

Ogni agente ha un’identità propria, permessi limitati e può essere disattivato in pochi minuti

I contenuti che l’agente legge possono contenere istruzioni: vanno filtrati e tenuti separati

L’autonomia si concede per livelli, sulla base di test e metriche

Log e tracce vanno predisposti prima che serva ricostruire un incidente



Grazie

Domande?

Vincenzo Calabrò

Ingegnere informatico
Cyber security e digital forensics

vincenzocalabro.it

vincenzocalabro.it

Riferimenti

Fonti

Standard e framework

- NIST SP 800-207, Zero Trust Architecture (2020)
- CISA, Zero Trust Maturity Model v2.0 (2023)
- NIST SP 800-61 Rev. 2 (2012) e Rev. 3, Incident Response Recommendations (2025)
- NIST AI RMF 1.0 (2023); NIST AI 600-1, Generative AI Profile (2024)
- ISO/IEC 42001:2023; ISO/IEC 27037:2012
- OWASP GenAI Security Project: Top 10 for LLM Applications 2025; Top 10 for Agentic Applications 2026 (dic. 2025); AISVS
- MITRE ATLAS; CSA MAESTRO e AI Controls Matrix; Coalition for Secure AI (CoSAI)

Ricerche e dati

- IBM, Cost of a Data Breach Report 2024, 2025 e 2026
- Palo Alto Networks, 2026 Identity Security Landscape (mag. 2026)
- S. Willison, “The lethal trifecta for AI agents” (giu. 2025)
- Meta AI, “Agents Rule of Two” (ott. 2025)
- K. J. K. Feng, D. W. McDonald, A. X. Zhang, “Levels of Autonomy for AI Agents”, Knight First Amendment Institute (2025)

Incidenti

- Arup, frode deepfake da 25 mln $ (CNN, 2024)
- G. A. Fowler, OpenAI Operator, The Washington Post (feb. 2025)
- Aim Security, EchoLeak, CVE-2025-32711 (giu. 2025)
- Replit e SaaStr (J. Lemkin), cancellazione del database di produzione (lug. 2025)
- Google Threat Intelligence Group, Salesloft Drift, UNC6395 (ago. 2025)
- Anthropic, “Disrupting the first reported AI-orchestrated cyber espionage campaign” (nov. 2025)
- OWASP GenAI Security Project, GenAI Exploit Round-up Report Q1 2026 (apr. 2026)

Normativa

- Regolamento (UE) 2024/1689 (AI Act) e Digital Omnibus sull’AI (GU UE 24 lug. 2026)
- Legge 23 settembre 2025, n. 132
- D.Lgs. 4 settembre 2024, n. 138 (NIS2)
- Regolamento (UE) 2016/679 (GDPR); Regolamento (UE) 2022/2554 (DORA)
- Legge 18 marzo 2008, n. 48