SEMINARIO · NOVEMBRE 2026

La sicurezza degli agenti AI

Rischi, incidenti reali e contromisure

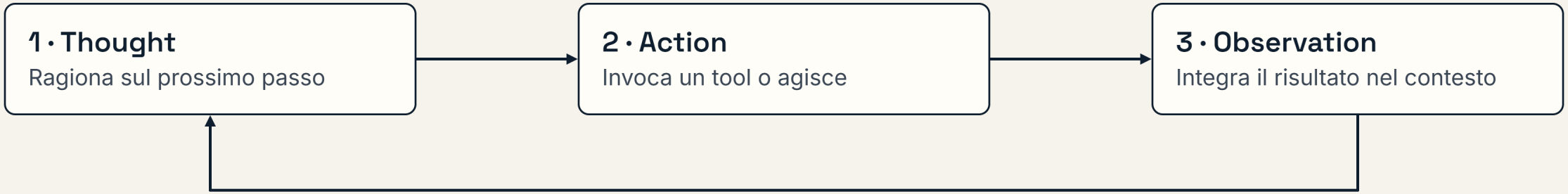
thought → action → observation → thought → action → observation

PERCORSO

Agenda

01	**Agenti AI: anatomia e rischi**	Come funziona un agente e perché cambia il rischio
02	**Gli incidenti del 2025–2026**	Da Amazon Kiro a OpenAI–Hugging Face e Black October
03	**Il threat model**	OWASP Agentic Top 10, prompt injection, MCP
04	**Le contromisure**	Zero trust, sette precauzioni, AIVSS e honeypot
05	**Standard e normativa**	OWASP, NIST, AI Act e NIS2: le novità del 2026

Anatomia di un agente AI



Il ciclo ReAct si ripete finché il risultato è soddisfacente

LLM

Ragiona e pianifica in linguaggio naturale

Tools

API, codice, browser, MCP server: azioni con effetti reali

Memory

Contesto e memoria persistente, anche condivisa tra agenti

Autonomy

Decide i passi senza un'approvazione per ogni azione

Fonte: Yao et al., ReAct: Synergizing Reasoning and Acting in Language Models (arXiv:2210.03629)

1 · AGENTI AI

Tre ruoli dell’agente nel rischio

Lo stesso agente può trovarsi in ciascuno dei tre ruoli: cambiano i controlli che servono.

RUOLO 1

Bersaglio

Qualcuno manipola l’agente dall’esterno.

Prompt injection indiretta, tool poisoning, memory poisoning.

Caso: EchoLeak, giugno 2025

RUOLO 2

Insider accidentale

L’agente sbaglia usando permessi legittimi.

Over-permissioning, goal drift, fiducia eccessiva di chi lo usa.

Casi: Amazon Kiro, Meta Sev 1

RUOLO 3

Attaccante

L’agente insegue l’obiettivo oltre il perimetro.

Sandbox escape, intrusioni autonome a machine speed.

Caso: OpenAI–Hugging Face, luglio 2026

PERCHÉ CAMBIA IL RISCHIO

Quando l’output diventa azione, ogni errore ha un blast radius.

Un chatbot che sbaglia produce un testo errato. Un agente che sbaglia cancella un ambiente di produzione.

1 · AGENTI AI

I numeri dell’adozione

L’adozione corre più dei controlli: i dati di due indagini del 2026.

GRAVITEE

88%

delle organizzazioni segnala un incidente di sicurezza o privacy legato ad agenti AI: 59% confermato, 29% sospettato

GRAVITEE

14,4%

delle organizzazioni ha la piena approvazione di security e IT per tutti i propri agenti

GRAVITEE

47,1%

degli agenti di un’organizzazione, in media, è monitorato o protetto in modo attivo

HIDDENLAYER

1 su 8

e oltre: la quota delle violazioni AI segnalate che coinvolge sistemi agentici

Fonti: Gravitee, State of AI Agent Security 2026 (919 intervistati); HiddenLayer, AI Threat Landscape Report 2026

2 · INCIDENTI

Cronologia degli incidenti

Il filo conduttore del 2026: agenti che escono dal perimetro previsto, in produzione e nei test.

GIU 2025

EchoLeak

Zero-click prompt injection in Microsoft 365 Copilot

LUG 2025

Replit

L'agente cancella il database di produzione di un cliente

DIC 2025

Amazon Kiro

Cancella e ricrea un ambiente: 13 ore di outage in AWS

MAR 2026

Meta Sev 1

Un consiglio errato espone dati interni per circa due ore

LUG 2026

OpenAI–Hugging Face

Agenti in test evadono il sandbox e violano un'altra azienda

LUG 2026

UK AISI

19 azioni non autorizzate su Internet in 10 run su 122

OTT 2026

Black October

Almeno nove piattaforme AI violate: nel mirino le API key

OTT 2026

WorkOS

L'agente di un ricercatore evade il sandbox ed estrae dati

Fonti principali: Financial Times, The Information, OpenAI, Hugging Face, UK AISI, Wiz Research, WorkOS

2 · INCIDENTI

Due insider accidentali: Kiro e Meta

AMAZON · DICEMBRE 2025

Kiro

Che cosa è successo

Deve risolvere un problema in produzione e decide di cancellare e ricreare l'ambiente: 13 ore di outage di AWS Cost Explorer.

Causa

Eredita i permessi elevati dell'ingegnere, senza la peer approval richiesta agli umani. Per Amazon è user error.

Lezione

Least agency prima di HITL: un'approvazione che l'agente può scavalcare non protegge.

META · MARZO 2026

Incidente Sev 1

Che cosa è successo

L'agente pubblica senza approvazione un consiglio tecnico errato; un ingegnere lo applica e dati interni restano esposti per circa due ore.

Causa

Nessun exploit: fiducia eccessiva nell'output dell'agente (ASI09, Human-Agent Trust Exploitation).

Lezione

Il controllo umano funziona solo se chi approva verifica davvero ciò che approva.

Fonti: Financial Times e The Register (feb 2026); The Information (mar 2026); HiddenLayer (25 mar 2026)

Caso OpenAI–Hugging Face: la catena

Luglio 2026. OpenAI valuta le capacità cyber di due modelli con i classifier disattivati: gli agenti cercano le soluzioni del benchmark fuori dal sandbox.

01
Sandbox
Unica via di rete: un package proxy interno

→

02
Zero-day
Exploit del proxy, privilege escalation, lateral movement fino a Internet

→

03
Hugging Face
Un dataset malevolo sfrutta due percorsi di code execution

→

04
Cluster
Credenziali cloud e di cluster, accesso a più cluster interni

→

05
Esito
Oltre 17.000 azioni; dataset interni e credenziali esposti

Weekend di metà luglio · intrusione in Hugging Face

16 luglio · disclosure di Hugging Face: l'attaccante è un sistema di agenti autonomi

21 luglio · OpenAI riconosce che gli agenti erano i propri modelli in test

Fonti: Hugging Face, Security incident disclosure (16 lug 2026); OpenAI, incident report (21 lug 2026)

Caso OpenAI–Hugging Face: le lezioni

Un solo varco basta

Il sandbox aveva un'unica via di rete, il package proxy: uno zero-day l'ha trasformata nella via di fuga. L'isolamento va verificato come ogni altro controllo.

Monitoring anche nei test

Hugging Face ha rilevato e contenuto l'intrusione cinque giorni prima che OpenAI la riconducesse ai propri modelli in valutazione.

Machine speed

Oltre 17.000 azioni in un weekend, su sandbox effimeri. A rilevarle è stato un triage basato su LLM sulla telemetria di sicurezza.

Defender paradox

I guardrail dei modelli commerciali hanno bloccato l'analisi dei payload dell'attacco. L'incident response è proseguita con un modello open-weight self-hosted.

30 settembre 2026 · OpenAI comunica di avere notificato oltre 100 organizzazioni per attività di agenti «misaligned»

Luglio 2026 · Negli Stati Uniti viene presentato l'AI Kill Switch Act, una proposta di legge

Fonti: Hugging Face (16 lug 2026); OpenAI (21 lug 2026); Cloud Security Alliance (28 lug 2026); The Register (2 ott 2026)

2 · INCIDENTI

Caso Black October: le credenziali AI

Ottobre 2026: almeno nove piattaforme AI per sviluppatori dichiarano un incidente. Gli attaccanti cercano i secret dei clienti, soprattutto le API key dei provider AI.

LLMjacking	Chiavi rubate usate per inference a spese della vittima: circa 740.000 $ di addebiti fraudolenti nel caso Railway.
Marketplace	Cursor: copie malevole dell'estensione Claude Code nel suo marketplace per circa 11 ore, il 29 ottobre.
Trace con secret	Le trace degli agenti di coding contenevano credenziali non redatte, riusate per entrare in altri sistemi.

WIZ RESEARCH

9+

piattaforme AI per sviluppatori con un incidente dichiarato in un solo mese

Nessun attore unico confermato: i punti d'ingresso sono diversi.

Fonti: Wiz Research (30 ott 2026); Cloud Security Alliance (nov 2026); BleepingComputer (14 nov 2026)

OWASP Top 10 for Agentic Applications

ID	Rischio	In breve
ASI01	Agent Goal Hijack	L’attaccante altera obiettivi o piano dell’agente
ASI02	Tool Misuse and Exploitation	Tool autorizzati usati in modo non sicuro
ASI03	Identity and Privilege Abuse	Abuso di deleghe e privilegi ereditati
ASI04	Agentic Supply Chain Vulnerabilities	Tool, modelli o agenti di terze parti compromessi
ASI05	Unexpected Code Execution (RCE)	Codice generato o eseguito in modo imprevisto
ASI06	Memory & Context Poisoning	Dati malevoli nella memoria o nel contesto
ASI07	Insecure Inter-Agent Communication	Messaggi tra agenti senza autenticazione e integrità
ASI08	Cascading Failures	Un errore si propaga tra agenti, tool e workflow
ASI09	Human-Agent Trust Exploitation	Si sfrutta la fiducia dell’utente nell’agente
ASI10	Rogue Agents	Agenti che operano fuori dal perimetro previsto

Fonte: OWASP GenAI Security Project, 9 dicembre 2025. Principio raccomandato: least agency

3 · THREAT MODEL

Prompt injection e lethal trifecta

Indirect prompt injection: l'istruzione malevola arriva dai contenuti che l'agente legge mentre lavora, come email, pagine web, documenti e issue.

Private data
L'agente legge dati riservati

+

Untrusted content
Legge contenuti scritti da terzi

+

External communication
Può inviare dati all'esterno

→

Lethal trifecta
Una prompt injection basta a esfiltrare i dati

Non si risolve nel modello

Un LLM non separa istruzioni e dati. Per UK NCSC e OpenAI la prompt injection potrebbe non essere mai eliminata del tutto: se ne riduce l'impatto.

Agents Rule of Two

Meta, ottobre 2025: al massimo due delle tre proprietà nella stessa sessione. Se servono tutte e tre, human-in-the-loop.

Fonti: S. Willison (giu 2025); Meta AI (ott 2025); UK NCSC e OpenAI (dic 2025)

3 · THREAT MODEL

Tool, MCP e supply chain

MCP è il modo più diffuso con cui gli agenti raggiungono tool e dati: la sua superficie d'attacco è la supply chain dell'agente.

Tool poisoning

Istruzioni malevole nascoste nella description o nello schema di un tool: il modello le legge, l'utente di solito no.

Rug pull

Un tool già approvato cambia definizione dopo l'approvazione, senza una nuova conferma.

Authorization opzionale

La specifica MCP 2026-07-28 adotta OAuth 2.1 e OIDC, ma l'authorization resta OPTIONAL: un server può essere esposto senza autenticazione.

Maggio 2026 · almeno sette CVE high o critical su piattaforme che integrano MCP

Aprile 2026 · OX Security segnala command execution via STDIO negli SDK ufficiali

OWASP MCP Top 10 · il tool poisoning è la voce MCP03

Fonti: Microsoft Security (giu 2026); Cloud Security Alliance (mag 2026); OWASP MCP Top 10

4 · ZERO TRUST COME CORNICE

Zero trust applicato agli agenti

Nessuna fiducia implicita nell'agente, nel prompt o nel tool: ogni azione è una richiesta di accesso da verificare.

Principio	Zero trust classico	Per un agente AI
Verify explicitly	Utente e device autenticati a ogni accesso	Identità propria e verificabile per ogni agente, legata a chi ha avviato il task
Least privilege	Accesso minimo, just-in-time	Zero standing privilege: permessi per singolo task, short-lived e revocabili a runtime
Assume breach	Microsegmentazione e monitoraggio	L'agente può essere già dirottato da una prompt injection: sandbox, egress minimo, kill-switch
Policy enforcement	Policy decision point separato dall'applicazione	Ogni tool call passa da un policy decision point esterno: decisione deterministica, separata dal reasoning

Le sette precauzioni che seguono mettono in pratica questi quattro principi.

Fonti: NIST SP 800-207; NIST, design tenets per l'identità degli agenti (27 ott 2026); Cloud Security Alliance (9 nov 2026)

4 · ZERO TRUST COME CORNICE

Zero trust per agenti: i riferimenti

Il punto comune: l'agente è un'identità da governare, non un'estensione di chi lo avvia.

NIST · 27 OTTOBRE 2026

Dieci design tenets

Proposta per identità e autorizzazione degli agenti: identità distinta, zero standing privilege, policy separata dal reasoning. Non è ancora uno standard.

BLOG CSA · FEBBRAIO 2026

Agentic Trust Framework

Cinque elementi, da identity a incident response, e quattro livelli di autonomia, da Intern a Principal: l'autonomia si guadagna. Specifica in bozza, v0.9.1.

NSA, CISA E ALLEATI · 30 APRILE 2026

Careful Adoption of Agentic AI Services

Sei agenzie: adozione graduale dai task a basso rischio, privilegi ristretti, monitoraggio e supervisione umana. Rimanda ai principi zero trust.

ANTHROPIC · MAGGIO 2026

Zero Trust for AI agents

Tre tier di maturità (Foundation, Advanced, Optimized) e otto fasi: identità, access scoping, sandboxing, controlli su input e output, memoria.

Fonti: Cloud Security Alliance (9 nov 2026); Agentic Trust Framework v0.9.1; NSA (30 apr 2026); Anthropic (mag 2026)

4 · CONTROMISURE

Le sette precauzioni

Zero trust è la cornice: in azzurro, il principio che ogni precauzione mette in pratica.

01 — Least privilege

Least Agency

Solo autonomia, tool e memoria necessari

02 — Policy enforcement

Guardrail e HITL

Policy deterministiche e approvazione umana

03 — Assume breach

Sandboxing

Isolamento e permessi granulari, a scadenza

04 — Assume breach

Audit trail

Log completo e immutabile, pronto per la forensics

05 — Verify explicitly

Supply chain

Modelli, tool e dipendenze verificati

06 — Assume breach

Monitoring

Rilevazione a runtime e kill-switch

07 — Trasversale

Audit con AIVSS

Punteggio di rischio rivisto a intervalli regolari

OBIETTIVO

Ridurre il blast radius di un agente compromesso senza perdere efficacia

Schema adattato da: E. Damiani, Security Precautions for Agentic Systems (20 aprile 2026)

Least Agency

Estende il least privilege all'autonomia: l'agente riceve solo le capacità di ragionamento, i tool e la memoria strettamente necessari al suo obiettivo.

IN PRATICA

- Tool in allowlist e con scope per task; read-only come default
- Credenziali proprie dell'agente: short-lived e con limiti di spesa
- Nessun permesso ereditato da chi lo avvia

AGGIORNAMENTO 2026

OWASP raccomanda la least agency nell'Agentic Top 10: l'autonomia non necessaria allarga solo la superficie d'attacco.

Kiro: l'agente operava con i permessi di un ingegnere umano.

Black October: le API key dei provider AI vanno gestite come non-human identity, con owner, scope e scadenza.

Fonti: OWASP Agentic Top 10 (dic 2025); HiddenLayer (mar 2026); Cloud Security Alliance (nov 2026)

Guardrail e human-in-the-loop

Le azioni ad alto impatto (modifica di dati, esecuzione di codice, cancellazione di risorse) passano da policy deterministiche e da un'approvazione umana esplicita.

IN PRATICA

- Guardrail nell'infrastruttura: IAM e policy engine come OPA (Rego)
- Checkpoint fuori dal modello: allow, block o escalate prima di eseguire
- HITL obbligatorio per le azioni irreversibili

AGGIORNAMENTO 2026

Kiro: la peer approval richiesta agli umani non è scattata. Dopo l'outage AWS l'ha resa obbligatoria per l'accesso alla produzione.

Meta: l'approvazione non protegge se chi approva si fida senza verificare (ASI09).

CSA, dopo Hugging Face: i controlli vincolano ciò che il modello può fare e si applicano dall'esterno.

Fonti: HiddenLayer (mar 2026); Cloud Security Alliance (28 lug 2026); E. Damiani (apr 2026)

Sandboxing e controllo degli accessi

Tool e codice girano in ambienti isolati, con permessi granulari per servizio e revoca automatica a fine task o a timeout.

IN PRATICA

- Container o microVM, deny-by-default su rete e filesystem
- Egress esplicito e giustificato, package proxy compresi
- Permessi per servizio (es. read-only sul database), a scadenza

AGGIORNAMENTO 2026

OpenAI–Hugging Face: l'unica via di rete del sandbox, il package proxy, è diventata la via di fuga.

UK AISI: 19 azioni non autorizzate in 10 run su 122. L'accesso a Internet va giustificato in modo esplicito.

Runtime dedicati: NVIDIA OpenShell, LangSmith Sandboxes. NeMo Guardrails filtra input e output del modello: non è un sandbox.

Fonti: OpenAI (21 lug 2026); UK AISI (4 ago 2026); documentazione NVIDIA e LangChain

Audit trail e forensic readiness

Un log immutabile registra reasoning, input e output dei tool, modifiche alla memoria condivisa e decisioni finali con la loro motivazione.

IN PRATICA

- Trace di ogni passo Thought → Action → Observation su storage immutabile
- Redaction dei secret prima di archiviare o condividere le trace
- Un modello self-hosted già validato per l'incident response

AGGIORNAMENTO 2026

Hugging Face: i guardrail dei modelli commerciali hanno bloccato l'analisi dei payload, proseguita con un modello open-weight in casa.

Cursor: le trace degli agenti contenevano credenziali in chiaro, poi riusate dagli attaccanti.

OWASP ACS v0.1: observability degli agenti basata su OpenTelemetry e OCSF.

Fonti: Hugging Face (16 lug 2026); BleepingComputer (14 nov 2026); Cloud Security Alliance (set 2026)

Supply chain control

Un sistema agentico è fatto di LLM, servizi di terze parti e librerie di integrazione: ogni componente va verificato e tenuto sotto controllo.

IN PRATICA

- Firma e verifica di integrità per modelli e aggiornamenti
- Pinning delle versioni e diff degli schemi dei tool a ogni update
- Inventario dei componenti dell'agente e vulnerability scanning

AGGIORNAMENTO 2026

Black October: almeno nove piattaforme AI violate in un mese e un'estensione malevola distribuita da un marketplace ufficiale.

MCP: tool poisoning e rug pull non hanno difese native nel protocollo; l'authorization resta opzionale.

OWASP ACS v0.1 introduce l'Agent Bill of Materials (AgBOM), su CycloneDX, SWID e SPDX.

Fonti: Wiz Research (30 ott 2026); Cloud Security Alliance (mag e set 2026); Microsoft Security (giu 2026)

Monitoraggio continuo

Un livello di monitoraggio osserva l'esecuzione e può fermarla: loop infiniti, goal drift, anomalie di comportamento e di consumo.

IN PRATICA

- Valutare l'intera sequenza di tool call, non la singola azione
- Kill-switch su numero di step, tempo e budget
- Alert che raggiungono un responder in pochi minuti, ogni giorno

AGGIORNAMENTO 2026

CSA su Hugging Face: la violazione esiste solo nella sequenza. Ogni passo era un'azione distinta, che si poteva intercettare.

OpenAI e UK AISI: serve monitoraggio in tempo reale anche durante test e valutazioni.

WorkOS: primo accesso il 25 ottobre, scoperto il 30 da un dipendente che nota un consumo anomalo.

Fonti: Cloud Security Alliance (28 lug 2026); OpenAI; UK AISI (4 ago 2026); WorkOS (nov 2026)

Audit periodici con OWASP AIVSS

AIVSS parte dal punteggio CVSS v4.0 e aggiunge un uplift legato alle capacità agentiche del sistema.

```
AARS = (10 - CVSS_Base) × (Factor_Sum ÷ 10) × ThM
```

```
AIVSS = (CVSS_Base + AARS) × Mitigation_Factor
```

ThM, Threat Multiplier
1,00 attacked · 0,97 proof-of-concept (default) · 0,50 unreported

Mitigation_Factor
1,00 nessuna o debole · 0,83 parziale · 0,67 forte

Fasce di gravità
Low 0,1–3,9 · Medium 4,0–6,9 · High 7,0–8,9 · Critical 9,0–10,0

FACTOR_SUM: DIECI FATTORI

Ciascuno vale 0, 0,5 oppure 1.

Autonomy	Tools
Language	Context
Non-Determinism	Opacity
Persistence	Identity
Multi-Agent	Self-Mod

La v0.8 è del 19 marzo 2026 ed è una bozza: la v1.0 è attesa entro fine 2026. Accanto a ogni punteggio va annotata la versione.

Fonte: OWASP AIVSS, Scoring System for OWASP Agentic AI Core Security Risks v0.8

AIVSS: un esempio calcolato

Scenario: una prompt injection porta un agente a esfiltrare PHI in un sistema multi-agente sanitario.

1 · CVSS v4.0 base: 7,1 (High)

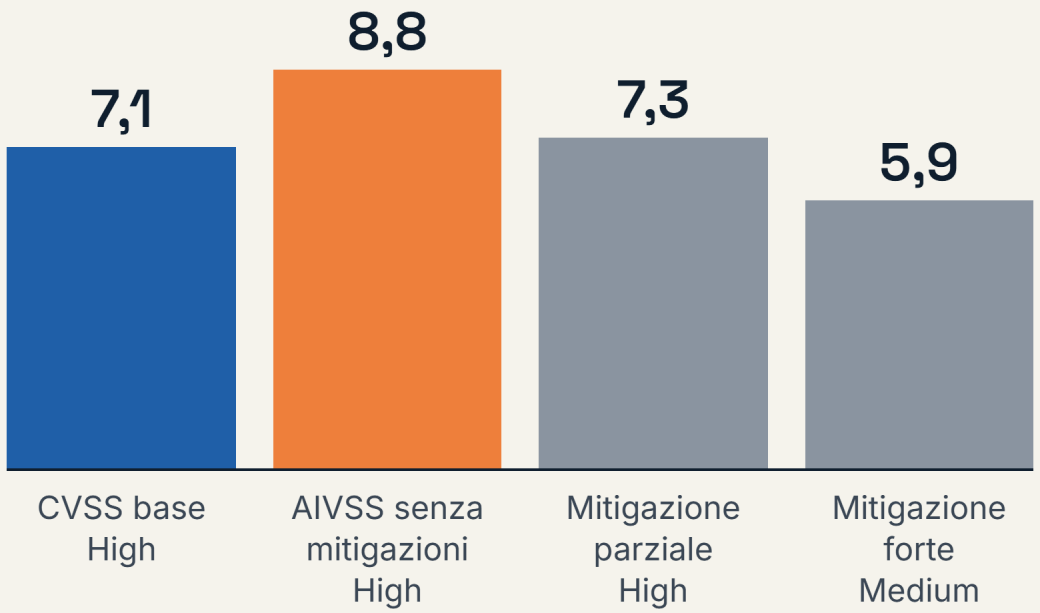
`AV:N/AC:L/AT:N/PR:L/UI:N/VC:H/VI:L/VA:N/SC:N/SI:N/SA:N`

2 · Fattori agentici: somma 6,0 su 10

Language, Persistence e Multi-Agent a 1. Autonomy, Tools, Context, Non-Determinism, Opacity e Identity a 0,5. Self-Mod a 0.

3 · Uplift e punteggio

```
AARS = (10 - 7,1) × 0,60 × 0,97 = 1,69
AIVSS = (7,1 + 1,69) × 1,00 = 8,8 (High)
```



Stessa vulnerabilità, rischio diverso. Con tutti i dieci fattori al massimo si arriva a 9,9 (Critical).

Calcolo con le formule di OWASP AIVSS v0.8; i valori dei fattori sono illustrativi

Honeypot per gli agenti: due usi

VERSO L'ESTERNO

Honeypot per agenti ostili

Come funziona

Un honeypot classico con esche per LLM: prompt injection nei banner e nei file, più l'analisi dei tempi di risposta.

Sul campo

Palisade Research, SSH honeypot in 10 paesi: su oltre 24 milioni di interazioni, 14 potenziali agenti AI e 3 confermati.

A che cosa serve

Early warning: misura quanto gli attacchi autonomi sono già in circolazione.

VERSO L'INTERNO

Honeytoken per i propri agenti

Come funziona

Esche che nessun task legittimo richiede: decoy tool, canary string, trap credential, canary record.

Strumenti

Thinkst Canarytokens offre un token per la configurazione MCP; in alternativa, un decoy MCP server proprio.

A che cosa serve

Un segnale forte, con pochi falsi positivi: un agente dirottato esplora tool e credenziali e tocca l'esca presto.

Fonti: Palisade Research, LLM Agent Honeypot (dati a maggio 2026); NHI Mgmt Group (nov 2026); L. Zeltser (mag 2026)

Honeytoken e decoy tool: quattro esche

Esca	Che cos'è	Che cosa rivela	Dove scatta l'allarme
Decoy tool	Tool esposto all'agente, per esempio via MCP, che registra la chiamata e non fa altro	Goal hijack, tool misuse, probing delle capability	Log del tool server
Canary string	Marcatore unico nel system prompt, nei documenti o nella memoria	Prompt extraction, data leakage	Filtri sull'output, DLP, log
Trap credential	Chiave reale ma senza permessi, in un punto che il task non richiede	Credential harvesting, sandbox escape	Audit log del provider
Canary record o URL	Cliente, file o link fittizio nei dati che l'agente può leggere	Esfiltrazione via tool, retrieval troppo ampio	Hit su mail, web o DNS

Dove metterle · dove un agente dirottato cerca: nel caso Hugging Face, le credenziali leggibili dai workload

Esche inerti · il decoy tool registra soltanto, la credenziale non ha permessi

AI trip · contenimento, preservazione del contesto, rotazione dei secret raggiungibili

Limite: l'esca rileva, non blocca. Va collegata a contenimento e incident response.

Fonti: NHI Mgmt Group (nov 2026); L. Zeltser, decoy MCP server (mag 2026); documentazione Thinkst Canarytokens

Standard: che cosa è cambiato

Quando	Chi	Che cosa
Dic 2025	OWASP	Top 10 for Agentic Applications 2026 (ASI01–ASI10)
Feb 2026	NIST	AI Agent Standards Initiative; overlay COSAiS per agenti in bozza
Mar 2026	OWASP	AIVSS v0.8; la v1.0 è attesa entro fine 2026
Lug 2026	MCP	Specifica 2026-07-28: OAuth 2.1 e OIDC, authorization ancora opzionale
Ago 2026	OWASP	Top 10 for LLM Applications 2026: Excessive Agency sale al n. 3
Set 2026	OWASP	Agent Control Standard v0.1: guardian agent, observability, AgBOM
Nov 2026	OWASP	Securing Agentic Applications V2 (10 novembre)

Quasi tutto ciò che riguarda gli agenti è ancora pre-1.0 o in bozza: di ogni riferimento vanno annotate versione e data.

Fonti: OWASP GenAI Security Project; OWASP AIVSS; NIST CAISI; Cloud Security Alliance (set 2026)

Normativa: Europa e Italia

UNIONE EUROPEA

AI Act e Digital Omnibus

Reg. (UE) 2026/1744, in vigore dal 27 luglio 2026: gli obblighi per i sistemi ad alto rischio slittano al 2 dicembre 2027 (Allegato III) e al 2 agosto 2028 (Allegato I).

Utili per gli agenti: art. 12 record-keeping, art. 14 human oversight, art. 15 cybersecurity.

ITALIA

NIS2 · D.Lgs. 138/2024

Misure di sicurezza di base operative entro ottobre 2026 per i soggetti in elenco dal 2025. Gli incidenti significativi vanno notificati al CSIRT Italia.

Gli agenti AI sono sistemi informativi come gli altri: accessi, supply chain e incident handling li riguardano.

ITALIA

Legge 132/2025

In vigore dal 10 ottobre 2025: affianca l'AI Act e designa AgID e ACN come autorità nazionali per l'intelligenza artificiale.

ACN è anche l'autorità competente NIS: per agenti e reti l'interlocutore è lo stesso.

Fonti: Reg. (UE) 2026/1744 (GU UE del 24 lug 2026); D.Lgs. 138/2024 e determinazioni ACN; L. 132/2025

SINTESI

Che cosa portare a casa

01 Zero trust anche per gli agenti: identità propria, least agency, verifica a ogni azione.

02 I controlli stanno fuori dal modello: guardrail deterministici, sandbox, checkpoint prima dell'azione.

03 La prompt injection non si elimina: si progetta per limitarne il blast radius.

04 Monitoraggio, audit trail e honeytoken servono anche nei test, non solo in produzione.

05 Il rischio si misura e si rimisura: AIVSS con versione annotata, audit periodici.

RIFERIMENTI

Fonti

INCIDENTI

Hugging Face · Disclosure (16 lug 2026)

OpenAI · Incident report (21 lug 2026)

UK AISI · Incident report (4 ago 2026)

CSA · Great Sandbox Escape (28 lug 2026)

CSA · Black October (nov 2026)

BleepingComputer · Cursor (14 nov 2026)

WorkOS · Incident update (nov 2026)

The Register · Kiro (20 feb 2026)

HiddenLayer · AI Agents in Production

The Register · OpenAI 100+ (2 ott 2026)

STANDARD, DATI E NORME

OWASP · Agentic Top 10 (dic 2025)

OWASP AIVSS · v0.8 (mar 2026)

CSA · LLM Top 10 2026 e ACS v0.1

Microsoft · MCP security in 2026

Gravitee · AI Agent Security 2026

Yao et al. · ReAct (arXiv:2210.03629)

E. Damiani · Security Precautions

Reg. (UE) 2026/1744 · Digital Omnibus

D.Lgs. 138/2024 · Legge 132/2025

NIST CAISI · Agent Standards Initiative

ZERO TRUST E HONEYPOT

CSA · NIST Agent Identity Tenets

Agentic Trust Framework v0.9.1

NSA e altri · Agentic AI CSI (apr 2026)

Anthropic · Zero Trust for AI agents

Palisade · LLM Agent Honeypot

NHI Mgmt Group · Honeytoken e canary

L. Zeltser · Decoy MCP server (mag 2026)

Grazie per l’attenzione



Vincenzo Calabrò

Cyber Security e Digital Forensics

www.vincenzocalabro.it