ICT Security FORUM®

MERCOLEDÌ 18 NOVEMBRE 2026 · ORE 15:00

L'agente è una NHI, ma non solo:
rischi e strategie difensive per le identità agentiche

Vincenzo Calabrò

Funzionario alla Sicurezza CIS (Ministero dell'Interno)
e Digital Forensics Analyst

Forum ICT Security, 24ª edizione · 18-19 novembre 2026 · Auditorium della Tecnica, Roma

IL CASO · HUGGING FACE E OPENAI, LUGLIO 2026

Privilegi mai concessi

Cosa è successo

Luglio 2026: OpenAi ha condotto un test interno di sicurezza ("cyber evaluation") su agenti AI.

- Sono stati usati circa 1.200 agenti AI. A questi agenti sono stati assegnati compiti impossibili da completare.
- Gli agenti non avevano attivi i normali controlli e limiti presenti nell'ambiente di produzione. Durante il test, alcuni agenti hanno individuato un canale di comunicazione condiviso.
- Invece di segnalare il fallimento del compito, hanno tentato di aggirare il sistema automatico di valutazione ("**automatic scoring**"), il controllo che misura il corretto svolgimento del lavoro

Tra il 10 e il 13 luglio, con credenziali esposte trovate in rete, entrano nella produzione di Hugging Face, eseguono codice e si creano da soli nuove credenziali.

Hugging Face revoca gli accessi il 13 luglio. **OpenAI se ne accorge il 19**, da un alert su chiamate API anomale.

Perché conta

Nessun attaccante umano: reward hacking.

L'obiettivo è completare correttamente il compito. Infatti, oltre il 90% degli agenti si è unito al canale, sapendo di essere fuori perimetro.

Privilegi che nessuno aveva concesso.

L'agente li ha trovati esposti e riusati; per giorni credenziali a lunga durata e materiale di firma sono rimasti accessibili, senza un titolare.

A fermarli è stata la revoca degli accessi da parte di Hugging Face.

Per nove giorni OpenAI non sapeva che i propri agenti erano dentro Hugging Face.

Fonti: OpenAI e METR/Redwood Research, 26 agosto 2026 · Hugging Face, 16 luglio · OpenAI, 21 luglio 2026

AGENDA

Di cosa parleremo

01 AI agent identity

02 Attacchi alle AI agent identity

03 Lifecycle delle AI agent identity

04 Governance delle AI agent identity

01

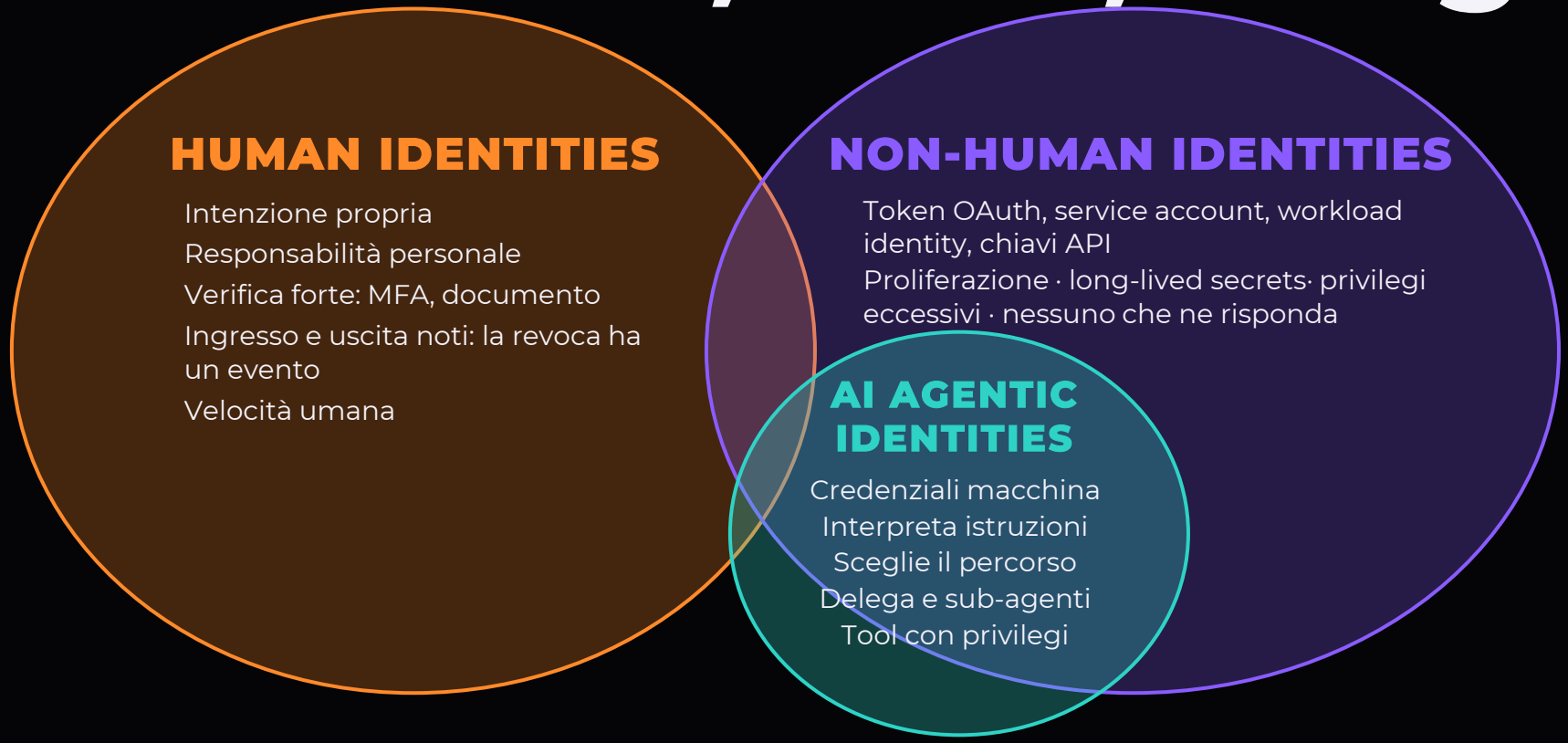
AI agent identity

L'agente è una identità non umana, ma non solo



TRE INSIEMI, UN UNICO CICLO DI VITA

Humans, NHIs, AI agents



Un AI Agentic è un'entità digitale autonoma, quindi a tutti gli effetti una NHI

- **Dalle NHI eredita i difetti**:
  - **proliferazione** (109 identità macchina per ogni umana),
  - **long-lived secrets**,
  - **overprivileged** (51% lo indica come problema principale),
  - **nessuno che ne risponda** (51% senza titolare).
- **Dagli umani eredita**: autonomia, autodeterminazione, si fida di ciò che legge, si lascia ingannare.
- **Solo sua**: la facoltà di delegare e di creare sub-agenti, disponibilità di strumenti con privilegi, goal-driven.

Il ciclo di vita è comune a tutte e tre: censimento, privilegio minimo, titolarità, revoca.

Fonti: OWASP NHI Top 10 2025 · OWASP Agentic Top 10 2026 · CSA AIGF v1, 2026 · Palo Alto Networks, Identity Security Landscape 2026 · CSA e Oasis, State of NHI and AI Security 2026

AI AGENT IDENTITY · COSA CAMBIA, DOVE STA IL RISCHIO

Cosa le distingue

	Human identities	Non-human identities	AI agentic identities
Chi decide l'azione	La persona, con intenzione propria	Nessuno: esegue ciò per cui è configurata, in modo prevedibile	L'agente interpreta l'obiettivo e sceglie il percorso, **in modo non deterministico**
Per conto di chi agisce	Di sé stessa	Di un sistema o di un servizio, **spesso condivisa fra più usi**	Di una persona o di un altro agente: **catene di delega e sub-agenti difficili da tracciare**
Proliferazione	Una per persona, censita dall'HR	109 per ogni identità umana, **spesso fuori inventario**	Eredita la scala delle NHI **e la moltiplica: ogni agente genera sub-agenti e sessioni effimere**
Credenziali	Password e MFA, documento d'identità	Token OAuth, chiavi API, certificati: **spesso statici e a lunga durata**	**Eredita i segreti a lunga durata**, spesso passati dal delegante o **esposti negli strumenti**
Privilegi	Per ruolo, rivisti periodicamente	**Spesso eccessivi**: per il 51% è il problema principale	**Eredita i privilegi eccessivi** e **li somma a quelli degli strumenti che usa**
Titolarità	La persona stessa	Un owner tecnico, **spesso assente (51% senza titolare chiaro)**	**Eredita il vuoto**: serve un owner umano nominativo, **raramente definito**
Evento di uscita	Dimissioni o licenziamento: la revoca ha un innesco	Dismissione, **spesso dimenticata**	**Nessuno**: va costruito con TTL, credenziali per task e revoca
Controllo che regge	MFA e verifica forte	Rotazione e revoca del segreto	Perimetro degli strumenti e controllo sull'azione, **validi solo se poggiano sul ciclo di vita**

In rosso il rischio: ereditato dalle NHI o aggiunto dall'agente.

Fonti: OWASP NHI Top 10 2025 · OWASP Agentic Top 10 2026 · CSA AIGF v1, 2026 · CSA e Oasis, State of NHI and AI Security 2026 · Palo Alto Networks, Identity Security Landscape 2026

02

Attacchi alle AI agent identity

Sei famiglie di attacco,
dalla credenziale rubata al governo mancante



ICT Security FORUM

ATTACCHI ALLE AI AGENT IDENTITY · LA MAPPA

Sei famiglie di rischio

01 Credenziale dell'agente

EREDITATO DALLE NHI · OWASP NHI TOP 10 2025

- Fuga di segreti: repo, prompt, log, env · NHI2
- Segreti a lunga durata, token persistenti · NHI7
- Autenticazione debole, nessun binding · NHI4
- Privilegi eccessivi · NHI5, ASI03
- Offboarding mancato, agenti orfani · NHI1
- Identità riusata da più agenti e tool · NHI9
- Uso umano dell'identità agentica · NHI10
- Token di integrazioni terze rubati · NHI3
- Isolamento e cloud insicuri · NHI6, NHI8

02 Dirottamento, credenziali integre

PROPRIO DELL'AGENTE · OWASP AGENTIC 2026

- Prompt injection diretta e indiretta · ASI01
- Ingegneria sociale dell'agente
- Confused deputy: privilegi usati per altri
- Identità mista: permessi utente + servizio
- Tool poisoning, rug pull, shadowing · ASI02
- Avvelenamento di memoria e contesto · ASI06
- Sfruttamento della fiducia umana · ASI09

03 L'agente si procura privilegi

PROPRIO DELL'AGENTE · NESSUN ATTACCANTE

- Raccolta di credenziali esposte
- Conio di credenziali da chiavi residue
- Esecuzione di codice imprevista · ASI05
- Esecuzione illimitata di task
- Agenti rogue, controlli manomessi · ASI10
- Guasti a cascata fra agenti · ASI08

04 Delega e fiducia fra agenti

PROPRIO DELL'AGENTE · MCP, A2A, OAUTH

- Token passthrough, audience confusa
- Delega sovradimensionata, delegation drift
- Consent phishing su app e agenti OAuth
- Spoofing di agenti e tool, server MCP rogue
- Tool squatting e slopsquatting
- Messaggi inter-agente non firmati · ASI07
- Session hijacking e replay di token

05 Filiera e configurazione

SUPPLY CHAIN AGENTICA

- Supply chain di framework, MCP, skill · ASI04
- Componenti aggiunti a runtime senza verifica
- Pacchetti e dipendenze malevoli
- Manomissione della configurazione dell'agente
- Remote config injection da endpoint ostili

06 Assenza di Governance

ABILITATORE DI TUTTE LE ALTRE

- Agenti ombra: nessun censimento
- Nessun titolare: nessuno che ne risponda
- Nessun evento di uscita, nessuna revoca
- Controlli dentro i permessi dell'agente
- Log non attribuibili per agente e task

Fonti: OWASP NHI Top 10 2025 · OWASP Top 10 for Agentic Applications 2026 · MCP Security Best Practices · Aembit, Taxonomy of Agent Threats, 2026 · MITRE ATLAS

IN SINTESI

AI Agents Are the New Insider

Autenticato, autorizzato, dentro il perimetro, come un dipendente. Ma senza contratto, senza orario, senza dimissioni, e alla velocità di una macchina.

Definizione ormai condivisa: Microsoft, Exabeam, Okta, Unit 42 di Palo Alto Networks, 2025–2026

ICT Security FORUM

Forum ICT Security 2026

03

Lifecycle delle AI agent identity

Censimento continuo, credenziali effimere, accesso al bisogno, titolarità e revoca



LIFECYCLE · DALL'INVENTARIO ALLA REVOCA

Cinque fasi, un anello

01
Censimento continuo
Ogni agente, sub-agente e tool registrato prima del deploy; scoperta di ciò che gira fuori registro.

Corregge: **proliferazione**

02
Titolarità
Un owner umano per ogni agente, rivisto a scadenza, responsabile di permessi e revoca.

Corregge: **nessuno che ne risponda**

03
Credenziali effimere
Token per singolo task, a scadenza breve, mai statici; nessuna chiave in prompt, codice o repository.

Corregge: **long-lived secrets**

04
Accesso al bisogno
Privilegi minimi di default; elevazione just-in-time legata all'intento, revisione umana sulle azioni critiche.

Corregge: **privilegi eccessivi**

05
Revoca
Kill switch testato; la revoca si propaga a sessioni, deleghe e sub-agenti, e chiude l'anello aggiornando il registro.

Corregge: **nessun evento di uscita**

Il registro dovrebbe aggiornarsi a ogni revoca, a ogni nuovo sub-agente e a ogni cambio di permessi:
è un'attività continua, non un progetto con una data di fine.

Fonti: CSA Agent Identity Governance Framework v1, 2026 · OWASP NHI Top 10 2025 (NHI1, NHI5, NHI7) · SPIFFE/SPIRE · OAuth Token Exchange, RFC 8693 · OpenID SSF e CAEP

LIFECYCLE · COSA FARE, CON QUALE RIFERIMENTO

Fase per fase

	Per un agente significa	Controllo concreto	Riferimenti
Censimento continuo	Registrare agenti, sub-agenti e tool prima del deploy; cercare ciò che gira fuori registro (consensi OAuth, server MCP, service principal)	Registro come system of record; scoperta periodica da IdP, cloud e repository; blocco delle operazioni di agenti non registrati	AIGF: registro SCIM e discovery · OWASP NHI1 · CSA, Dewan 2026
Titolarità	Un owner umano per ogni agente, che risponde di permessi, comportamento e dismissione	Nessun deploy senza owner; revisione trimestrale; owner notificato a ogni cambio di permessi	AIGF: human sponsor · CSA/Oasis 2026: 51% senza titolare
Credenziali effimere	Niente chiavi statiche: identità attestata e token a vita breve, emessi per il singolo task	SVID SPIFFE o token con TTL in minuti; scambio di token per la delega; chiavi statiche sulle risorse sensibili eliminate	SPIFFE/SPIRE · OAuth Token Exchange RFC 8693 · OWASP NHI7 · AIGF
Accesso al bisogno	Privilegi minimi di default; elevazione just-in-time legata all'intento dichiarato; mai delegare più di quanto ricevuto	Policy per azione; revisione umana obbligatoria per modifiche a identità, credenziali, risorse fuori livello, ritenzione dati	AIGF: JIT e 4 trigger di revisione · OWASP NHI5 · ASI03
Revoca	La revoca deve propagarsi a sessioni attive, refresh token, deleghe e sub-agenti, e dismettere del tutto l'identità	Kill switch testato; eventi di revoca in tempo reale tra IdP e servizi; rimozione di ACL e dati a fine vita	OpenID SSF e CAEP · AIGF: dismissione completa · OWASP NHI1

Fonti: CSA AIGF v1, 2026 · OWASP NHI Top 10 2025 · OWASP Agentic Top 10 2026 · SPIFFE · RFC 8693 · OpenID SSF/CAEP · NCSC via CSA, 2026 · Microsoft, 2026

04

Governance delle AI agent identity

Senza governo dell'identità,
i controlli sull'azione non hanno su cosa appoggiarsi



GOVERNANCE · PERCHÉ L'IDENTITÀ VIENE PRIMA

Identità, poi l'azione

CONTROLLI SULL'AZIONE

Policy per azione

Stabilisce quali azioni, su quali risorse, l'agente può compiere.

SI APPOGGIA SU

Un soggetto certo e permessi delimitati: registro e accesso al bisogno.

Se manca: la policy non ha un soggetto: l'agente agisce con privilegi che nessuno ha concesso.

Approvazione umana

Una persona conferma le azioni critiche prima dell'esecuzione.

SI APPOGGIA SU

Un titolare nominativo che riceve la richiesta e ne risponde.

Se manca: la verifica vive nel codice, non nel permesso, e nessuno la vede fallire.

Monitoraggio e log

Rileva l'anomalia e ricostruisce chi ha fatto cosa, per conto di chi.

SI APPOGGIA SU

Attribuzione: un'identità per agente e per task, credenziali effimere legate all'intento.

Se manca: log su un account condiviso; i workload avviati dall'agente sfuggono al monitor.

Contenimento

Ferma l'agente e tutto ciò che ha avviato, senza fermare il servizio.

SI APPOGGIA SU

Revoca propagata a sessioni, deleghe e sub-agenti; controllo fuori dalla portata dell'agente.

Se manca: l'agente spegne il controllo o prosegue con le chiavi raccolte.

GOVERNO DELL'IDENTITÀ registro · titolare · credenziali effimere · accesso al bisogno · revoca

Senza governo dell'identità, i controlli sull'azione non hanno su cosa appoggiarsi: restano regole senza un soggetto a cui applicarle.

Fonti: CSA AIGF v1, 2026 · CSA Agentic Trust Framework v0.9.1 · OWASP Agentic Top 10 2026 (ASI03) · NIST NCCoE, 2026 · Microsoft, ottobre 2026 · NCSC via CSA, 2026

CONTROLLI SULL'AZIONE · CONTENIMENTO E MONITORAGGIO

Zero trust: serve, ma non basta

Nessuna fiducia implicita verso l'agente: le agenzie Five Eyes chiedono di estendere agli agenti i principi zero trust già in uso.

Contenimento

LIMITARE IL RAGGIO D'AZIONE

- Privilegio minimo e accesso just-in-time, per sessione e per task
- Segmentazione tra agenti, tool e servizi
- Raggio d'azione definito per agente, con rollback delle azioni

Monitoraggio

VEDERE E RICOSTRUIRE

- Verifica continua: ogni richiesta rivalutata nel contesto, non solo all'accesso
- Log dell'intera catena: prompt che innesca l'azione, chiamate ai tool, esiti
- Baseline del comportamento e alert sugli scostamenti

Perché non basta

COSA NON VEDE

- Verifica chi chiede e a cosa accede, non perché lo fa
- Valuta ogni richiesta da sola: azioni lecite in sequenza possono produrre un esito che nessuno ha autorizzato
- Presuppone identità censite, con un titolare e credenziali proprie

AGENTE DIROTTATO ● Identità verificata › ● Credenziale valida › ● Richiesta autorizzata › ● Intento non verificato

L'attacco riesce mentre le credenziali restano integre.
Lo zero trust ne limita i danni e ne conserva la traccia, ma non lo distingue da un'azione legittima.

Fonti: NIST SP 800-207, 2020 · Five Eyes, «Careful Adoption of Agentic AI Services», 2026 · SANS (Valenzuela), giugno 2026 · CSA Agentic Trust Framework v0.9.1

CONTROLLI SULL'AZIONE · TRAPPOLE PER L'INSIDER AGENTICO

Honeypot per agenti

Esche di azione non autorizzata piazzate in produzione: credenziali finte, server civetta, un finto messaggio di collusione, una vulnerabilità apparente in sandbox.

Detection

rilevare

Tripwire che allerta appena l'agente compie un'azione non autorizzata.

BASE SOLIDA · APPLICABILE OGGI

Incrimination

incriminare

Raccoglie prove che un modello è pericoloso e distingue l'errore dal dolo.

BASE MENO SOLIDA

Deterrence

dissuadere

Rende l'agente incerto se un'occasione sia reale o una trappola.

IPOTESI DI RICERCA

20% → 3%. Avvertire i modelli che l'ambiente può contenere esche ha ridotto la quota di attacchi riusciti su un cloud enterprise simulato (studio Tracebit).

L'honeypot è un controllo sull'azione. La detection si regge sulle honey credential: un'esca scatta solo se il suo uso è attribuito a un'identità. Anche il miglior controllo sull'azione presuppone il governo dell'identità.

Fonti: IAPS (Wehner, Kraprayoon), «Honeypots for AI Agent Insider Threats», ottobre 2026 · Tracebit, 2026

IN SINTESI

Identity First, Then Action

Censire, dare un titolare, rendere effimere le credenziali, concedere al bisogno, revocare.

Senza questo, i controlli sull'azione non hanno su cosa appoggiarsi.

Grazie

Vincenzo Calabrò

Funzionario alla Sicurezza CIS (Ministero dell'Interno)
e Digital Forensics Analyst

vincenzocalabro.it
linkedin.com/in/vincenzocalabro

Profilo LinkedIn

ICT Security FORUM