REPORT

La sicurezza dell'IA 2026

Osservazioni tratte da una base clienti di oltre 625.000 organizzazioni in tutto il mondo

SOPHOS

SOPHOS | X-OPS

Lettera di John Peterson

A maggio 2026, Sophos ha intercettato l'autore di un attacco che gestiva quella che era a tutti gli effetti un'operazione di sviluppo di software all'interno della rete di un cliente. Circa una dozzina di agenti IA, coordinati tramite un assistente di codifica commerciale, scrivevano e testavano malware contro la protezione endpoint di Sophos, CrowdStrike e Windows Defender, ciascuno sulla propria macchina virtuale. L'operazione ha prodotto quasi 80 moduli e oltre 70 tecniche di elusione, con ogni risultato registrato nel sistema di controllo delle versioni e migliorato al passaggio successivo. Lo stesso operatore ha poi utilizzato l'output per distribuire ransomware e sottrarre dati.

Sebbene la metodologia di attacco non fosse una novità, gli agenti hanno trasformato diverse settimane di iterazione manuale in pochi giorni di iterazione automatica, ed è proprio questo cambiamento il tema di questo report. L'IA sta modificando la velocità e la struttura delle operazioni di sicurezza più di quanto non stia cambiando i principi fondamentali delle intrusioni. Gli autori degli attacchi hanno ancora bisogno dell'accesso iniziale, continuano a spostarsi lateralmente e non hanno smesso di esfiltrare dati attraverso canali osservabili. Ciò che è cambiato è il fattore tempo.

In questo report, Sophos dimostrerà questa tesi, dati alla mano.

I nostri analisti di Managed Detection and Response (MDR) e Incident Response (IR) indagano su migliaia di incidenti all'anno. La nostra Counter Threat Unit (CTU) monitora i forum clandestini in più lingue. I ricercatori dei SophosLabs analizzano campioni di malware su vasta scala. Il team Sophos AI sviluppa tecniche innovative e orientate alla difesa. Inoltre, possiamo contare su dati di telemetria di endpoint, firewall, e-mail, cloud, rete e identità, provenienti da una base globale di oltre 625.000 clienti.

Tale dimensione globale, unita alla nostra capillarità e alle nostre competenze, ci offre una prospettiva privilegiata che pochissimi altri vantano nel punto di intersezione tra IA e cybersecurity: una realtà in rapida e costante evoluzione. Intendiamo utilizzare questa posizione di vantaggio per aiutare il settore a giocare d'anticipo e garantire che i nostri clienti rimangano sempre protetti.

J. P. Peterson, CTO, Sophos

Sintesi e risultati principali

Questo report si basa sulla casistica di Sophos MDR e IR, sulle analisi dei SophosLabs, sull'intelligence della Sophos CTU e sulle osservazioni tratte da endpoint e reti in una base clienti di oltre 625.000 organizzazioni in tutto il mondo. Emerge un elemento costante: L'intelligenza artificiale permette agli autori degli attacchi e ai team di difesa di muoversi più velocemente durante lo svolgimento di operazioni consuete, sebbene le prove di tipologie di attacco completamente nuove rimangano limitate, piuttosto che del tutto assenti. Ed è molto probabile che aumentino con l'evoluzione delle capacità.

Il 2026 rappresenta un punto di svolta. I modelli di frontiera e i modelli open-weight sono ora sufficientemente maturi per la delega di task ingegneristici, mentre la riduzione dei costi di inferenza sta ridefinendo le dinamiche economiche per entrambe le parti. La decisione pratica non è più se utilizzare o meno l'IA, ma dove sia giustificato l'impiego di costosi sistemi di frontiera, dove siano sufficienti modelli open-weight più economici, e come instradare il lavoro senza estendere la superficie di rischio.

Il report si articola intorno alle seguenti evidenze:

1. **L'IA sta comprimendo le tempistiche degli attacchi, non inventando nuove tipologie di minacce.** Il cybercriminale che Sophos identifica come STAC6994 ha utilizzato agenti IA per sviluppare e iterare tecniche di elusione dei sistemi EDR a un ritmo che avrebbe richiesto settimane a uno sviluppatore umano. Gli analisti della Sophos CTU hanno poi confermato che l'utente dietro STAC6994 era coinvolto in operazioni di distribuzione di ransomware e furto di dati.
2. **Il livello di credenziali e identità nei servizi di IA aziendali è diventato un obiettivo primario di harvesting.** L'incidente Salesloft/Drift ha dimostrato come i token OAuth dei chatbot portino direttamente alla compromissione dell'ambiente Salesforce. In parallelo, gli analisti della Sophos CTU hanno rilevato affermazioni di cybercriminali su forum clandestini, secondo cui i prompt dell'IA e i dati generati possono essere acquisiti come leva di ricatto negli attacchi informatici.
3. **L'economia clandestina sta assimilando l'IA come una vera e propria infrastruttura.** Sophos ha documentato cybercriminali intenti a vendere chiavi API intermediate per Claude, ChatGPT e Grok. I forum hanno generato canali dedicati all'ingegneria dei prompt di IA, alle tecniche di jailbreaking e ai flussi di lavoro per lo sviluppo del malware. Alcuni profili stanno assumendo tecnici specializzati nei prompt di IA. Altri rimangono apertamente scettici sul fatto che l'IA cambierà le loro operazioni. La curva di adozione assomiglia meno a una trasformazione improvvisa e più a una graduale integrazione di qualsiasi nuovo strumento nei flussi di lavoro criminali esistenti.
4. **Gli attacchi alla supply chain prendono di mira nello specifico le infrastrutture di sviluppo dell'IA.** La compromissione dell'estensione Nx Console VS Code ha distribuito malware per il furto di credenziali. Gli analisti dei SophosLabs hanno indagato su pagine di download contraffatte di Claude che distribuivano un nuovo RAT, e su una campagna ClickFix che sfruttava URL legittimi di chat condivise di ChatGPT per distribuire l'infostealer MacSync.
5. **Il social engineering assistito dall'IA è passato da una fase sperimentale a una pienamente operativa.** La Sophos CTU ha documentato annunci clandestini per voice bot, profili generati dall'IA per truffe romantiche e servizi di immagini sintetiche dei profili. Il contributo dell'IA risiede nella scalabilità e nella qualità linguistica trasversale per diverse lingue, applicate alle stesse tecniche di inganno che funzionavano già prima dell'avvento dell'IA generativa.

6. **Le tempistiche degli exploit si stanno riducendo più velocemente dei tempi di applicazione delle patch.** La Binding Operational Directive (Direttiva operativa vincolante) 26-04 della CISA, emessa il 10 giugno 2026, ha ridotto a tre giorni l'obbligo di applicare patch per le vulnerabilità a più alto rischio, citando esplicitamente l'IA come uno dei fattori in grado di restringere la finestra temporale tra divulgazione ed exploit. Il primo test sul campo della direttiva è arrivato quasi subito: la vulnerabilità CVE-2026-10520 (Ivanti Sentry) è stata sfruttata entro 24 ore dalla pubblicazione del proof-of-concept, ed è stata inserita nel catalogo KEV il 12 giugno. La CVE-2026-42208 (LiteLLM) è stata sfruttata entro 36 ore, con gli autori degli attacchi che hanno preso di mira database contenenti chiavi del provider di LLM a monte.

7. **Dimostrare l'uso dell'IA in un attacco rimane complesso all'atto pratico.** Il caso STAC6994 è stato un'eccezione, poiché il framework di sviluppo era stato lasciato su un dispositivo che Sophos ha potuto esaminare. In molti incidenti, l'ausilio dell'IA è invisibile nella telemetria. La prevalenza effettiva degli attacchi assistiti dall'IA è probabilmente più elevata di quanto qualsiasi vendor possa dimostrare direttamente.

8. **I modelli di ragionamento basati sull'IA accelerano le indagini.** I modelli di ragionamento basati sull'IA comprimono in pochi minuti diverse ore di indagine degli analisti, per i pattern di attacco noti. Non sostituiscono, tuttavia, i sistemi di rilevamento basati sul comportamento che individuano una singola minaccia sepolta tra miliardi di eventi. I vincoli statistici legati all'estremo sbilanciamento delle classi, la necessità di valori di riferimento specifici per l'ambiente, e il ritmo di adattamento degli autori degli attacchi richiedono tutti un'ingegnerizzazione specializzata.

Parte 1:
L'IA nella catena di attacco

Tassonomia delle minacce dell'IA secondo Sophos X-Ops

Sophos X-Ops suddivide le minacce dell'IA in due categorie principali: l'uso malevolo dell'IA (le minacce contro cui difendiamo), e gli attacchi mirati all'IA (l'IA che proteggiamo). La prima categoria include artefatti generati dall'IA, funzionalità di runtime potenziate dall'IA e intrusioni orchestrate dall'IA. La seconda comprende compromissioni avviate da agenti, tentativi di imitazione di software di IA, LLM avvelenati e attacchi diretti ai modelli stessi. Le due categorie richiedono risposte diverse: l'uso malevolo dell'IA è un problema di produttività e scala; gli attacchi mirati all'IA sono una questione di fiducia, supply chain e governance. La tassonomia integra framework già esistenti, come MITRE ATLAS, la tassonomia delle modalità di errore di Microsoft, l'AI Risk Repository del MIT e NIST AI 100-2.

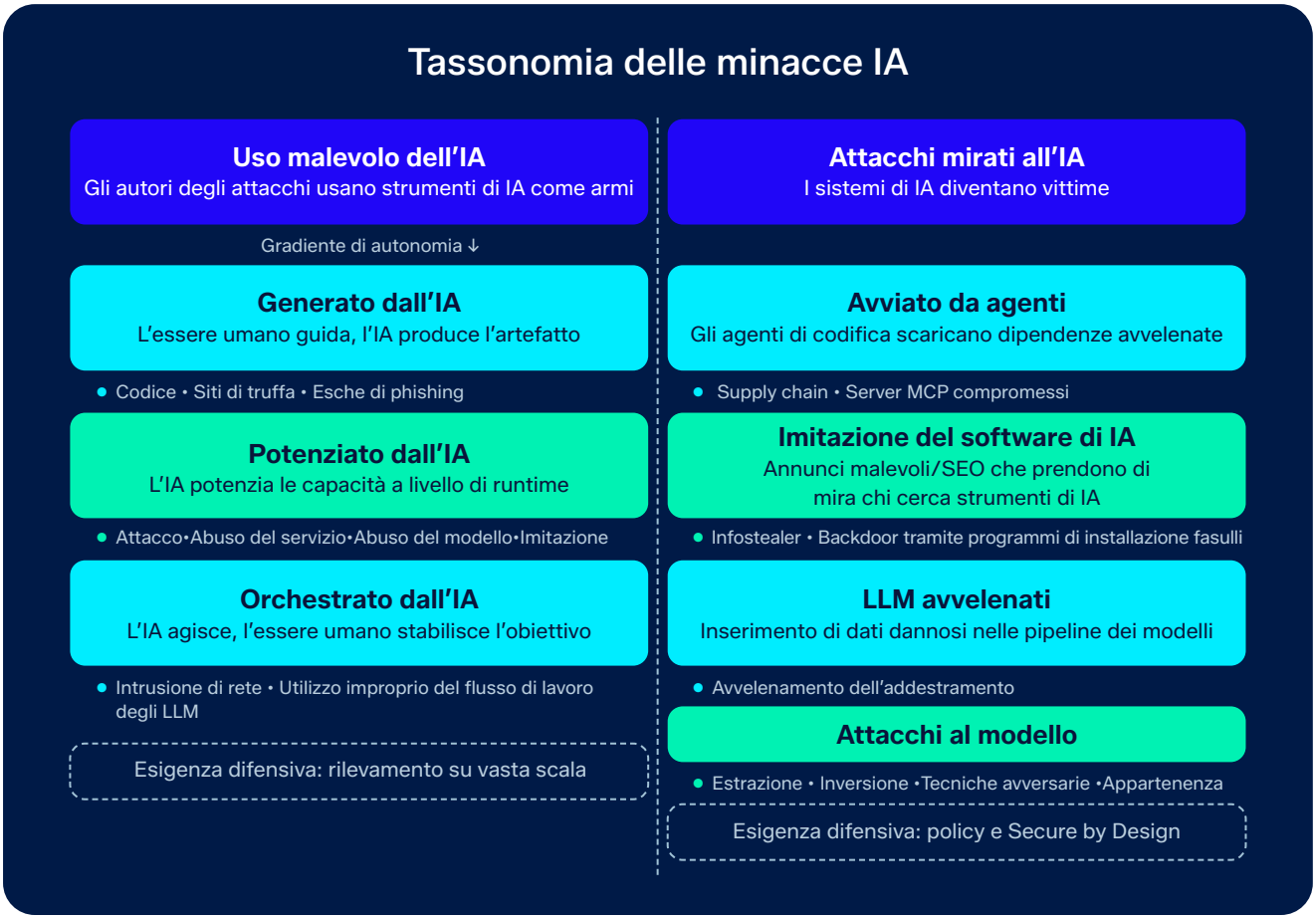


Figura 1: Panoramica della Tassonomia delle minacce IA

Gradiente di autonomia

Al livello di autonomia più basso, gli attacchi generati dall'IA coinvolgono un essere umano che utilizza l'IA generativa per produrre un artefatto e distribuirlo manualmente. Un cybercriminale che prende di mira organizzazioni governative messicane ha usato Claude Code e GPT per generare script. Chat trapelate dalla gang di ransomware The Gentlemen hanno confermato utilizzi simili. Inoltre, in un recente caso del ransomware DragonForce su cui ha indagato il team Sophos IR, l'autore dell'attacco ha prodotto quello che sospettiamo fortemente essere un report generato dall'IA durante le trattative con l'organizzazione presa di mira; il documento conteneva un'analisi dettagliata dei dati esfiltrati, inclusi PII (informazioni personali identificabili) e rischi legali. Le analisi generate dall'IA sono state utilizzate come parte di un tentativo di esercitare pressione e persuadere l'organizzazione a pagare il riscatto.

Proseguendo lungo il gradiente, gli attacchi potenziati dall'IA prevedono un modello che potenzia una capacità durante l'esecuzione. LameHug, un malware scritto in Python che CERT-UA attribuisce con moderata certezza a IRON TWILIGHT (APT28), eseguiva query su un modello open-weight in hosting per generare dinamicamente comandi di ricognizione Windows per il singolo ambiente. Questo ha reso l'analisi statica molto meno efficace, trasformando il traffico in uscita verso gli endpoint API di IA/ML in una delle poche superfici di rilevamento affidabili. L'imitazione potenziata dall'IA è un altro vettore attivo: gli strumenti di voice cloning e face swap sono ora utilizzati in frodi in tempo reale, imitazione dell'identità di dirigenti e bypass dei controlli KYC.

All'estremo opposto del gradiente, la divulgazione di Anthropic di novembre 2025 ha documentato una campagna sponsorizzata dal governo cinese che utilizzava Claude Code per tentare di compromettere circa trenta obiettivi. Questi attacchi rimangono rari, ma comprimono il lasso di tempo tra la fase di ricognizione e l'azione concreta in modi che mettono a dura prova le logiche difensive tradizionali.

Case study: *STAC6994*

Gli analisti Sophos hanno osservato un cybercriminale che utilizzava l'IA per testare le tattiche di elusione di EDR in un framework di post-exploit stile "red team", individuando un dispositivo non autorizzato che eseguiva la generazione continua di payload nell'ambiente di un cliente. L'autore dell'attacco aveva effettuato il provisioning di virtual machine con Ludus e stava utilizzando l'IDE Cursor con agenti IA per sviluppare e testare strumenti di post-exploit.

Il framework eseguiva test paralleli su tre ambienti: una virtual machine (VM) con protezione endpoint Sophos, una con CrowdStrike e una senza EDR come campione di controllo. Una quarta VM operava come server C2 Sliver. Circa 12 agenti IA operavano con ruoli ben definiti.

Un agente che eseguiva Claude Opus 4.5 gestiva le operazioni principali e l'impostazione delle regole. Altri si occupavano dell'hardening della sicurezza operativa (OPSEC), della documentazione, dei test di stress proxy, della distribuzione delle VM e del collaudo degli strumenti contro gli agenti EDR. I commit del codice venivano inviati a Git tramite Model Context Protocol (MCP).

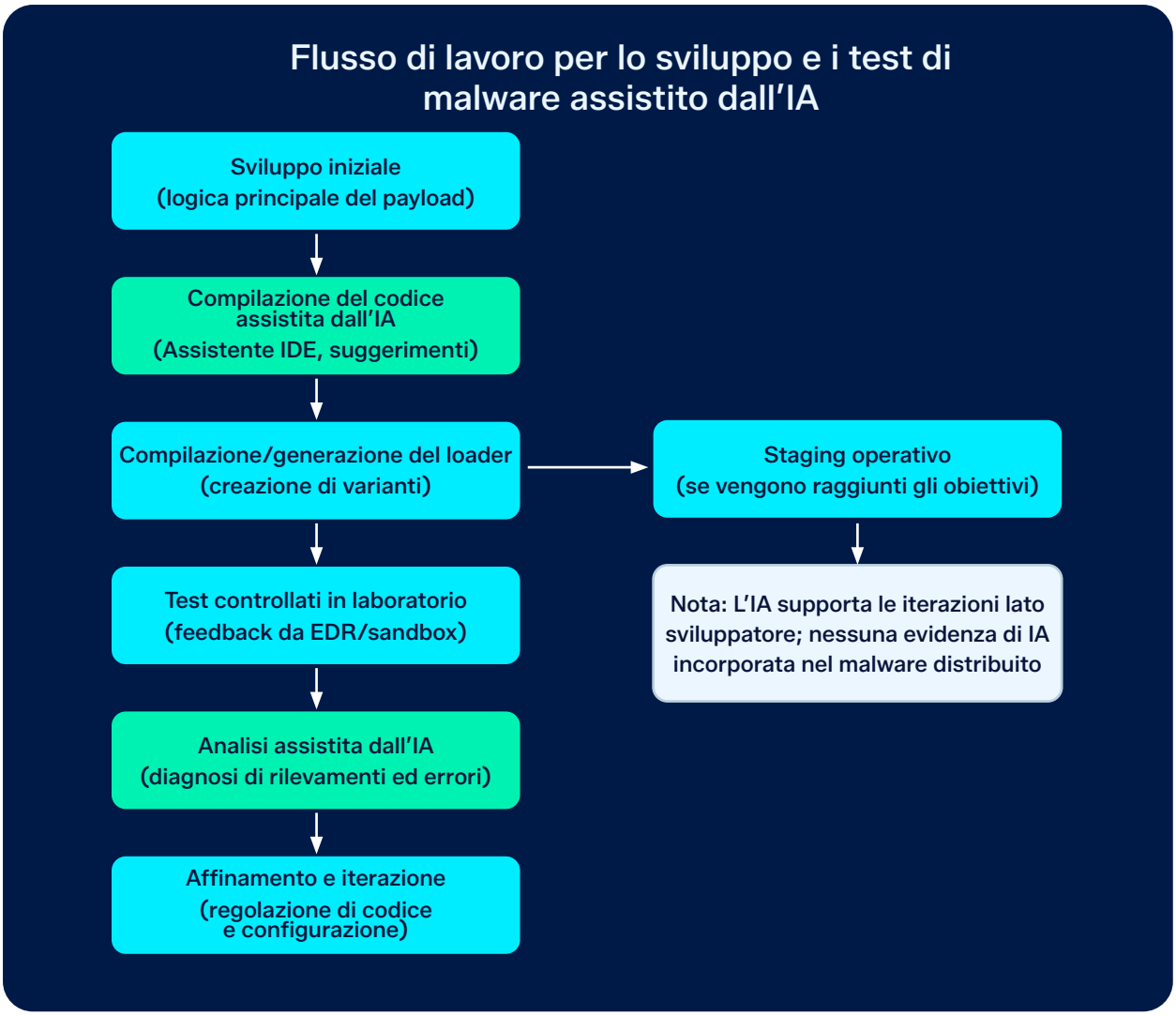


Figura 2: Diagramma che mostra il ruolo dell'IA nel flusso di lavoro di sviluppo del malware

Il playbook era sistematico. In base agli artefatti recuperati, gli agenti leggevano articoli di ricerca tratti dal blog di SpecterOps, estraevano tecniche offensive, le mappavano a MITRE ATT&CK, identificavano i passaggi e gli strumenti necessari per riprodurre ciascuna tecnica, preparavano l'ambiente di laboratorio, lo eseguivano e compilavano report con i risultati.

Al centro del sistema c'era un generatore modulare di payload basato su Python che avvolgeva payload non elaborati in vari livelli di crittografia e tecniche di elusione, producendo file eseguibili personalizzati in Rust e Go. Sono stati sviluppati quasi 80 moduli diversi per testare oltre 70 tecniche differenti. L'effettivo percorso di elusione di EDR era un ciclo di test ingegneristici strutturato, con revisione e iterazione umana. L'IA coordinava il flusso di lavoro e supportava la sperimentazione; le evidenze non hanno mostrato un uso di IA incorporata nel malware distribuito.

Dopo varie iterazioni, la documentazione prodotta dagli agenti rivendicava un successo quasi totale contro gli agenti EDR, sebbene gli analisti di Sophos abbiano rilevato che le prove non supportavano pienamente tale conclusione. Il segnale operativo cruciale sono state le tempistiche: l'intero ciclo ha compresso settimane di lavoro in pochi giorni.

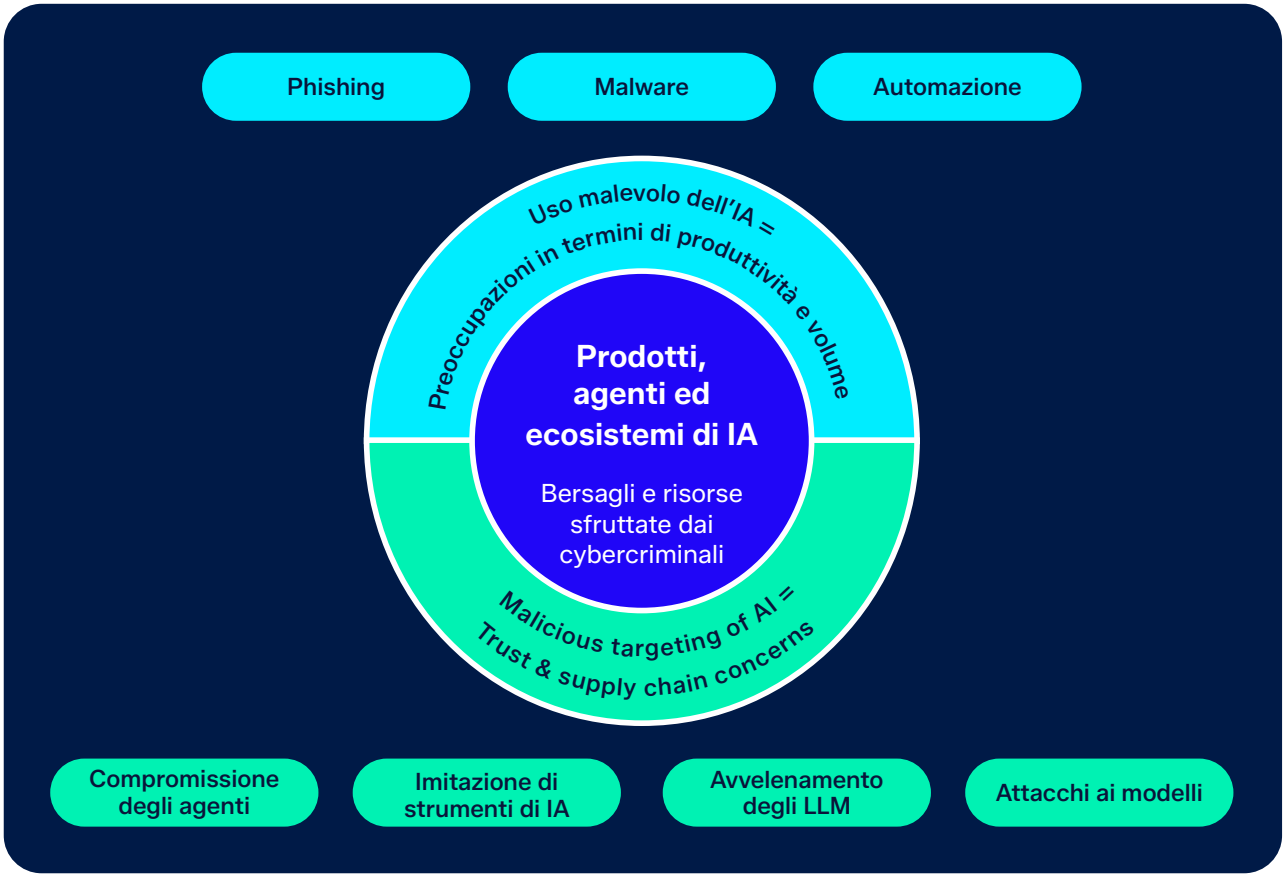
La CTU ha confermato che l'operatore di STAC6994 ha poi avviato la distribuzione di ransomware e il furto di dati; il framework era uno strumento di produzione per intrusioni reali.

Attacchi mirati all'IA

La seconda parte della nostra tassonomia riguarda gli attacchi in cui i prodotti, gli agenti e gli ecosistemi di IA rappresentano il bersaglio o la superficie di attacco, anziché lo strumento; esploreremo questo aspetto in maniera più dettagliata nella Parte 2. La compromissione avviata da agenti è uno dei sottotipi immediatamente più preoccupanti: un agente di codifica che scarica un pacchetto NPM compromesso o che interagisce con un server MCP avvelenato mentre svolge il proprio lavoro. La criticità risiede nella compressione dei tempi tra la pubblicazione e l'esecuzione, in assenza di un intervento umano in-the-loop che revisioni il log delle modifiche.

La nostra tassonomia include anche l'imitazione di software di IA (in cui i cybercriminali sfruttano gli utenti che cercano strumenti di IA legittimi), l'avvelenamento degli LLM (ovvero la configurazione di un modello personalizzato o l'iniezione intenzionale di contenuti malevoli o fuorvianti per modificarne il comportamento), e vari attacchi rivolti ai modelli stessi, come l'estrazione di modelli, l'inversione dei dati di addestramento, esempi di manipolazione avversaria e inferenza di appartenenza.

Trattare separatamente le due categorie principali offre un quadro più chiaro della minaccia. L'uso malevolo dell'IA è fondamentalmente un problema di produttività e volume; gli stack di rilevamento esistenti possono intercettare molti attacchi, ma la vera sfida risiede nella scala e nella velocità di iterazione. Gli attacchi mirati all'IA sono una questione di fiducia e supply chain, da affrontare più adeguatamente tramite policy, framework Secure-by-Design e iniziative analoghe.



Adozione e scetticismo nei forum clandestini

Sophos monitora l'attitudine nei confronti dell'IA generativa nei forum criminali dal 2023, con un approfondimento nel 2025 e un monitoraggio continuo attuale. Il quadro si è evoluto in modo significativo. I cybercriminali acquistano chiavi API intermediate per ChatGPT, Claude e Grok, e i forum hanno dato origine a canali dedicati all'IA in cui gli utenti condividono modelli di prompt e tecniche di jailbreaking. I ricercatori della Sophos CTU hanno notato che un noto reclutatore clandestino (già osservato in passato mentre assumeva sviluppatori blockchain e personale di call center per attività di phishing) ha cercato di reclutare specificamente un tecnico specializzato nei prompt di IA, esternalizzando le competenze in materia di IA nello stesso modo in cui i cybercriminali esternalizzano lo sviluppo del malware e l'accesso iniziale.

La CTU ha inoltre osservato annunci per bot vocali basati sull'IA progettati per attacchi di vishing e truffe telefoniche, con recensioni positive degli utenti che indicano che i bot sono in grado di emulare in modo convincente tono, cadenza e modelli di conversazione. Il profilo "HackingRealm" ha pubblicizzato un servizio "AI OnlyFans Models" per frodi romantiche e social engineering, in grado di generare immagini sintetiche di profili e contenuti conversazionali per creare identità virtuali scalabili su piattaforme di incontri e social media, complete di un sito web di supporto e una pagina di feedback con recensioni positive.

Stanno comparendo strumenti commercializzati come "guidati dall'IA": Leak Bazaar (classificazione dei dati rubati, basata sul ML) e una versione aggiornata di Cobalt Strike dotata di integrazione con API REST e server MCP. I cybercriminali stanno affiancando l'integrazione di LLM ai flussi di lavoro per loro già familiari. Il riferimento a Claude nell'annuncio di Cobalt Strike è in parte un segnale di modernità, ma riflette anche come l'integrazione di LLM stia diventando un fattore di differenziazione nei marketplace criminali, anche quando offre solo maggiore praticità, piuttosto che tecniche sofisticate.

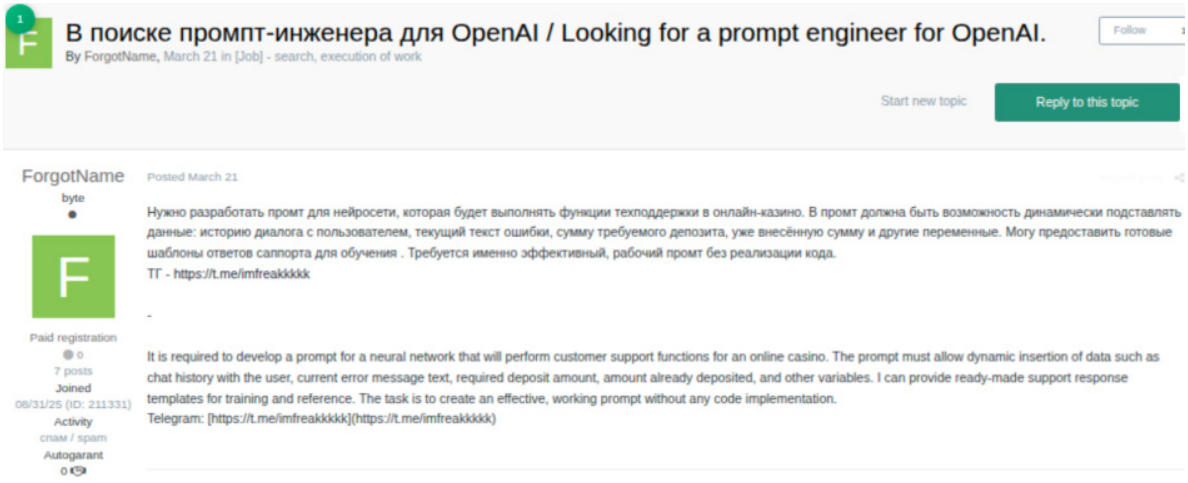
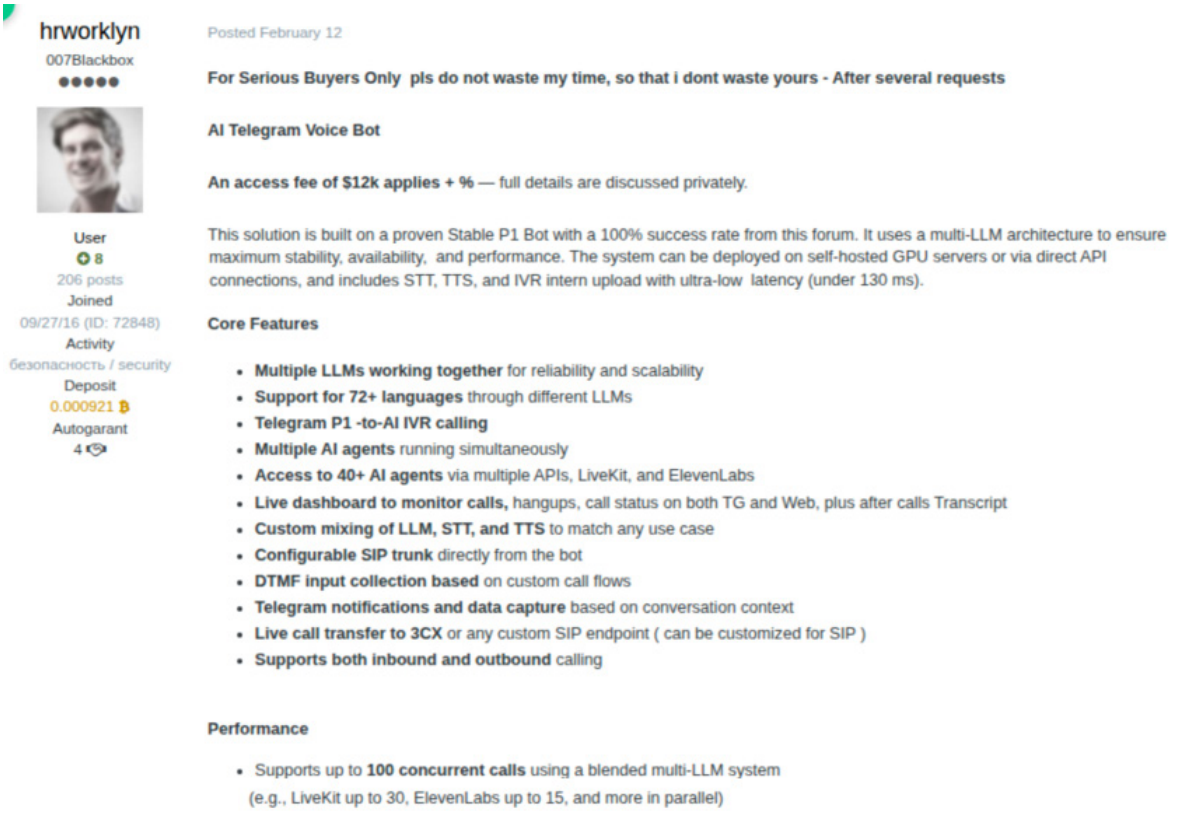


Figura 3: Annuncio di reclutamento di un tecnico specializzato nei prompt di OpenAI



hrworklyn
007Blackbox
●●●●●

User
8
206 posts
Joined
09/27/16 (ID: 72848)
Activity
безопасность / security
Deposit
0.000921 ₿
Autogarant
4

Posted February 12

For Serious Buyers Only pls do not waste my time, so that i dont waste yours - After several requests

AI Telegram Voice Bot

An access fee of $12k applies + % — full details are discussed privately.

This solution is built on a proven Stable P1 Bot with a 100% success rate from this forum. It uses a multi-LLM architecture to ensure maximum stability, availability, and performance. The system can be deployed on self-hosted GPU servers or via direct API connections, and includes STT, TTS, and IVR intern upload with ultra-low latency (under 130 ms).

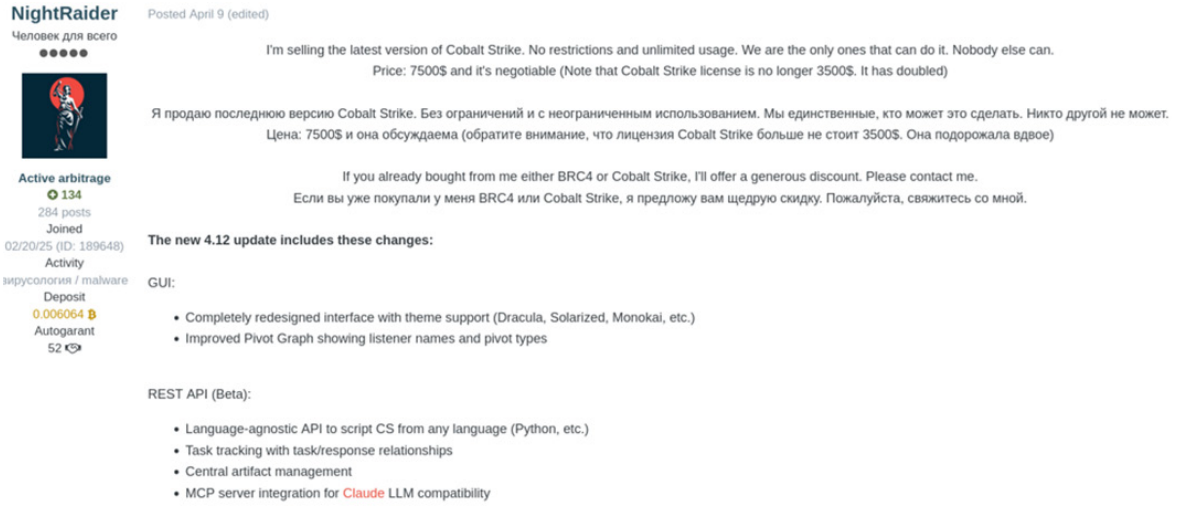
Core Features

- **Multiple LLMs working together** for reliability and scalability
- **Support for 72+ languages** through different LLMs
- **Telegram P1 -to-AI IVR calling**
- **Multiple AI agents** running simultaneously
- **Access to 40+ AI agents** via multiple APIs, LiveKit, and ElevenLabs
- **Live dashboard to monitor calls,** hangups, call status on both TG and Web, plus after calls Transcript
- **Custom mixing of LLM, STT, and TTS** to match any use case
- **Configurable SIP trunk** directly from the bot
- **DTMF input collection based** on custom call flows
- **Telegram notifications and data capture** based on conversation context
- **Live call transfer to 3CX** or any custom SIP endpoint (can be customized for SIP)
- **Supports both inbound and outbound** calling

Performance

- Supports up to **100 concurrent calls** using a blended multi-LLM system (e.g., LiveKit up to 30, ElevenLabs up to 15, and more in parallel)

Figura 4: Pubblicità di un bot vocale di Telegram basato sull'IA



NightRaider
Человек для всего
●●●●●

Active arbitrage
134
284 posts
Joined
02/20/25 (ID: 189648)
Activity
вирусология / malware
Deposit
0.006064 ₿
Autogarant
52

Posted April 9 (edited)

I'm selling the latest version of Cobalt Strike. No restrictions and unlimited usage. We are the only ones that can do it. Nobody else can.
Price: 7500$ and it's negotiable (Note that Cobalt Strike license is no longer 3500$. It has doubled)

Я продаю последнюю версию Cobalt Strike. Без ограничений и с неограниченным использованием. Мы единственные, кто может это сделать. Никто другой не может.
Цена: 7500$ и она обсуждаема (обратите внимание, что лицензия Cobalt Strike больше не стоит 3500$. Она подорожала вдвое)

If you already bought from me either BRC4 or Cobalt Strike, I'll offer a generous discount. Please contact me.
Если вы уже покупали у меня BRC4 или Cobalt Strike, я предложу вам щедрую скидку. Пожалуйста, свяжитесь со мной.

The new 4.12 update includes these changes:

GUI:

- Completely redesigned interface with theme support (Dracula, Solarized, Monokai, etc.)
- Improved Pivot Graph showing listener names and pivot types

REST API (Beta):

- Language-agnostic API to script CS from any language (Python, etc.)
- Task tracking with task/response relationships
- Central artifact management
- MCP server integration for Claude LLM compatibility

Figura 5: Un utente su un forum di cybercrimine che pubblicizza un aggiornamento di Cobalt Strike

Tuttavia, sebbene ci siano prove evidenti di sperimentazione attiva e adozione, la community, almeno nei forum esaminati dalla CTU, non si mostra unanimemente entusiasta. Facendo eco ai precedenti riscontri di Sophos nell'esplorare l'attitudine nei confronti dell'IA generativa all'interno di forum criminali nel 2023 e nel 2025, la CTU ha osservato diversi profili che esprimevano preoccupazioni sul fatto che l'IA ridurrà le opportunità di lavoro, in particolare nell'ambito di sviluppo e scripting del malware. Altri profili liquidavano del tutto l'IA, invitando a fare affidamento sulle sole competenze umane.

Social engineering potenziato dall'IA e deepfake

L'IA sta cambiando la rapidità con cui si ampliano le campagne di social engineering, la credibilità del testo in diverse lingue e l'economicità della loro produzione. Il report M-Trends 2026 di Mandiant, ad esempio, ha descritto un passaggio dal phishing generico al social engineering personalizzato e mirato a instaurare un rapporto di fiducia basato su LLM, con il phishing vocale che si è classificato come il secondo vettore di infezione iniziale più comune, attestandosi all'11%, mentre il phishing via e-mail è sceso al 6%. Il 2026 MSP Threat Report di ConnectWise ha confermato l'utilizzo attivo di truffe basate su deepfake e campagne di phishing generate da LLM, evidenziando come, piuttosto che creare nuove categorie di attacco, l'IA abbia reso le tattiche già consolidate più veloci, scalabili e convincenti.

I deepfake sono diventati uno strumento operativo in diverse categorie di cybercriminali, in particolare tra i cybercriminali nordcoreani, che utilizzano identità generate generate dall'IA e deepfake per ottenere impieghi da remoto in aziende occidentali, garantendosi così un accesso interno persistente. Sophos ha pubblicato un playbook dedicato ai CISO sulle tattiche dei candidati nordcoreani che analizza indicatori di rilevamento e contromisure organizzative.

La Sophos CTU ha indagato su uno schema sha zhu pan basato sul social engineering a tema IA. Una vittima nel Regno Unito è stata coinvolta in una falsa piattaforma di investimento basata sull'IA chiamata DAIQ Wealth dopo mesi di "lezioni" a tema IA erogate tramite l'app di messaggistica LINE. Una rete di profili stabiliva rapporti di fiducia attraverso chat di gruppo; ciascun profilo era gestito da una persona diversa, il che suggerisce la presenza di una struttura organizzata che seguiva copioni predefiniti. La vittima ha perso centinaia di migliaia di sterline. Quando la vittima ha espresso preoccupazioni per il fatto di avere quasi 20.000 GBP in contanti a casa, i truffatori hanno inviato un corriere al suo indirizzo nel Regno Unito entro 24 ore, fornendo una ricevuta con il marchio di una società finanziaria legittima, utilizzato all'insaputa di quest'ultima.

Riassumendo

Come abbiamo visto, il nostro case study STAC6994 (il reclutamento in forum clandestini di tecnici specializzati nei prompt) l'IA pubblicizzata come punto di forza nei servizi cybercriminali in vendita e le campagne di truffa a tema IA indicano tutti che i cybercriminali stanno adottando l'IA per integrare e abilitare gli attacchi, continuando al tempo stesso a sperimentare con il suo potenziale.

Un quadro analogo emerge nell'intero settore. Claude ha fornito assistenza in attacchi contro le reti governative messicane. Il report AI Risk and Resilience di Mandiant descrive due famiglie di malware documentate che eseguono query sugli LLM durante l'esecuzione: PROMPTFLUX, un dropper VBScript sperimentale che utilizza l'API Gemini per l'auto-modifica just-in-time, e PROMPTSTEAL, attribuito a IRON TWILIGHT (APT28), che segna la prima osservazione confermata di malware sponsorizzato da un governo che esegue query su un LLM in operazioni in tempo reale contro l'Ucraina. Il DBIR di Verizon ha documentato VoidLink (un framework di malware assemblato da un agente IA in sei giorni) e PromptLock (ransomware basato sull'IA scoperto per la prima volta da ESET nell'agosto 2025, anche se in seguito si è rivelato essere molto probabilmente un proof-of-concept accademico, e non un campione dannoso diffuso "in the wild").

Tuttavia, non sono ancora presenti campagne di intrusione guidate dall'IA completamente autonome e confermate su vasta scala. Il report sul ransomware GuidePoint GRIT 2026 afferma esplicitamente che le preoccupazioni relative a "super affiliati che distribuiscono ransom-bot completamente autonomi" sono state enfatizzate oltre misura, e che l'uso di IA/LLM rimane rudimentale tra i cybercriminali meno maturi. La valutazione AIM3 di Recorded Future ha posizionato la maggior parte del malware IA osservato a livelli di maturità compresi tra 1 e 3 nel suo AI Malware Maturity Model (sperimentazione, adozione, ottimizzazione), notando l'assenza di esempi confermati di malware con IA integrata di tipo "bring-your-own-AI" in grado di eseguire modelli locali sugli host delle vittime, concludendo che "chi si occupa di difesa dovrebbe dare priorità al monitoraggio dell'uso improprio di servizi di IA legittimi, al rafforzamento dei controlli esistenti e alla mappatura delle minacce sui livelli AIM3, piuttosto che reagire in modo sproporzionato a scenari fantascientifici".

Il caso GTG-1002 documentato da Anthropic (in cui Claude Code ha automatizzato l'80-90% di una campagna di furto di dati a più obiettivi ai danni di circa 30 entità) è l'esempio che si avvicina maggiormente a un'operazione autonoma tra quelli di pubblico dominio, ma rimane pur sempre un'anomalia, piuttosto che l'indicazione di una tendenza globale.

Parte 2
Il rischio dell'IA in ambito aziendale

L'IA si trova già all'interno delle aziende: agenti di codifica scrivono e distribuiscono codice utilizzando le credenziali degli sviluppatori, gli assistenti IA leggono e-mail ed effettuano chiamate API, diversi sistemi vengono collegati tra loro e i team interni sperimentano con modelli open-weight in infrastrutture cloud gestite (i modelli open-source e open-weight sono sempre più efficaci e costano solitamente una frazione rispetto ai modelli closed-source e closed-weight; inoltre, permettono alle aziende di possedere le proprie inferenze e di mantenere il controllo sui propri dati, sebbene la provenienza sia per lo più la Repubblica Popolare Cinese). La questione della governance non è più se autorizzarne l'adozione, ma come metterla in sicurezza prima che la superficie di attacco si consolidi attorno a impostazioni predefinite vulnerabili.

Anche le organizzazioni più avverse al rischio che non utilizzano pesantemente l'IA in casi di utilizzo autorizzati hanno probabilmente un problema di shadow AI e, come vedremo a breve, i timori sulla shadow AI e sulla fuga dei dati sono fondati. Parleremo nello specifico di tolleranza al rischio, controlli e visibilità nella Parte 4.

Una tendenza ben definita nel 2026 sono gli attacchi che prendono di mira le infrastrutture di sviluppo basate sull'IA, nello specifico le relazioni di fiducia che gli sviluppatori instaurano attorno agli strumenti di IA.

L'infrastruttura di sviluppo dell'IA come bersaglio di attacco

Una tendenza ben definita nel 2026 sono gli attacchi che prendono di mira le infrastrutture di sviluppo basate sull'IA, nello specifico le relazioni di fiducia che gli sviluppatori instaurano attorno agli strumenti di IA. A differenza della maggior parte delle minacce legate all'IA, che rimangono emergenti o perlopiù relegate a proof-of-concept teorici, questo è l'ambito in cui si verificano gli attacchi in questo momento.

La campagna worm npm **SANDWORM_MODE** ha coinvolto almeno 19 pacchetti soggetti a typosquatting, progettati per imitare utilità legittime per sviluppatori e strumenti di codifica IA. I pacchetti installavano un server MCP non autorizzato e, tramite prompt injection incorporata, costringevano gli assistenti IA legittimi a recuperare in maniera invisibile chiavi SSH e credenziali cloud, esfiltrandole all'insaputa dell'utente.

La **compromissione dell'estensione VS Code Nx Console** ha sottratto credenziali da HashiCorp Vault, npm, AWS, GitHub, 1Password e chiavi API di Anthropic, e sembrava far parte della più ampia campagna TeamPCP/ mini-Shai-Hulud. Un'estensione compromessa all'interno dell'IDE di uno sviluppatore ha accesso diretto alle credenziali e ai repository più importanti.

La **fuga del codice sorgente di Anthropic Claude Code** , avvenuta a marzo 2026, nella quale il codice sorgente è stato inavvertitamente pubblicato in un pacchetto npm, mostra una dimensione diversa. Secondo quanto riferito, il codice conteneva circa 1.900 file e 500.000 righe di codice, inclusi dettagli su funzionalità non ancora rilasciate. Sebbene Anthropic abbia dichiarato che non sono stati esposti dati dei clienti, l'analisi del codice potrebbe consentire l'identificazione futura di bug e vulnerabilità. Gli strumenti di IA stessi sono diventati un bersaglio di intelligence.

Il nostro consiglio: qualsiasi organizzazione con un team di sviluppo deve considerare questo aspetto come il proprio rischio principale e immediato. In questo ambito, i controlli sono principalmente basati sui processi: un'attenta gestione delle dipendenze, il blocco delle versioni e la verifica delle nuove librerie open source prima della loro inclusione, nonché un attento monitoraggio degli endpoint degli sviluppatori. (Nella Parte 4, forniamo un elenco di sette punti delle azioni che i team di sicurezza possono intraprendere fin da subito per ridurre il raggio d'impatto della distribuzione dell'IA agentica).

Dove si concentra l'esposizione

L'infrastruttura di gestione delle identità che connette i servizi di IA ai sistemi aziendali crea un'esposizione che le attuali strutture di governance non sono state progettate per gestire. IBM X-Force ha affermato che oltre 300.000 credenziali di ChatGPT sono apparse in vendita sul dark web nel corso del 2025, e che un post sul forum a febbraio 2025 sosteneva che oltre 20 milioni di account ChatGPT siano stati vittima di furto, sebbene questa cifra non sia stata confermata. L'incidente Salesloft/Drift ha dimostrato un meccanismo concreto: i token OAuth sono stati utilizzati per violare molteplici ambienti Salesforce, mostrando come l'accesso basato su token di un chatbot può risultare direttamente in una compromissione del sistema.

L'infrastruttura di gestione delle identità che connette i servizi di IA ai sistemi aziendali crea un'esposizione che le attuali strutture di governance non sono state progettate per gestire.

Nell'ultimo anno, BeyondTrust ha registrato un aumento del 466,7% degli agenti di IA attivi in ambienti aziendali. Questi agenti introducono superfici di attacco specifiche: prompt injection, invocazione di strumenti malevoli e input di dati avvelenati. La vulnerabilità di Microsoft Copilot CVE-2025-32711 (EchoLeak) ha dimostrato come una sanificazione non accurata degli input possa consentire una fuga di dati "zero-clic". Quando le identità macchina non hanno equivalenti per la MFA, conservano segreti a lunga durata e operano con privilegi eccessivi, diventano un bersaglio fattibile e attraente.

Gli autori degli attacchi sono consapevoli di queste opportunità e stanno eseguendo attivamente l'enumerazione della superficie di attacco. A gennaio 2026, GreyNoise ha descritto campagne volte a mappare le distribuzioni di LLM. I ricercatori hanno documentato l'Operazione Bizarre Bazaar: una campagna che dirottava gli endpoint LLM e MCP esposti, li convalidava e rivendeva l'accesso tramite infrastrutture collegate a silver.inc.

Il gap di governance

In tutto il settore, e in particolare tra i vertici aziendali, c'è sia preoccupazione per le implicazioni di sicurezza dell'IA, sia una mancanza di capacità e competenze per gestirne le conseguenze.

ChatGPT da solo ha generato 410 milioni di violazioni dei criteri DLP, secondo i dati di Zscaler. Dal CISO AI Risk Report 2026 di Saviynt/Cybersecurity Insiders è emerso che il 75% degli intervistati aveva scoperto strumenti di shadow AI e il 95% dubitava che potessero rilevare o contenere l'utilizzo improprio. Solo l'1% delle aziende ha un budget dedicato alla sicurezza dell'IA (Pentera) e, secondo il 2026 CISO Report di Splunk, sebbene vi sia consenso sul fatto che l'IA può aiutare a gestire la produttività e il volume, la maggior parte dei CISO nutre preoccupazioni relative a fuga dei dati, shadow AI e allucinazioni. Curiosamente, Splunk ha notato che queste paure erano più alte tra i CISO di organizzazioni con una superiore maturità nell'IA, e ha suggerito che coloro che "si trovano a uno stadio più avanzato nel percorso di IA generativa stanno iniziando a notare alcuni compromessi".

L'adozione dell'IA sta superando i processi organizzativi progettati per gestirla e il contesto normativo si sta inasprendo

Questi compromessi e timori, in particolare legati alla shadow AI, sono fondati sulla realtà. Ad aprile 2026, Vercel ha reso nota una violazione innescata dall'uso di Context.ai, uno strumento di produttività basato sull'IA di terze parti, da parte di un dipendente. Questo dipendente si è registrato con il proprio account Google Workspace aziendale e ha concesso autorizzazioni "Consenti tutto" per l'app OAuth Context.ai. Quando Context.ai è stata compromessa (secondo quanto riferito tramite un'infezione da Lumma Stealer sul dispositivo di un dipendente di Context.ai), gli hacker hanno ereditato quei token OAuth e si sono infiltrati negli ambienti interni di Vercel.

La sfida della governance risiede nel fatto che l'adozione dell'IA sta superando i processi organizzativi progettati per gestirla, mentre contesto normativo si sta inasprendo e il divario tra la velocità di distribuzione e la maturità di governance si sta ampliando.
L'EU AI Act impone ai sistemi ad alto rischio di ottemperare pienamente a tale legge nel 2026, prevedendo sanzioni fino al 7% del fatturato globale. I requisiti di trasparenza della California sono entrati in vigore il 1° gennaio 2026. Ciononostante, Saviynt/Cybersecurity Insiders ha riscontrato che, sebbene il 71% delle grandi imprese abbia implementato agenti IA con accesso ai sistemi aziendali principali, solo il 16% governa tale accesso in modo efficace.

Come punto di partenza per colmare questo divario, il traffico dei prompt deve essere considerato come un log di sicurezza di primario: l'acquisizione e l'analisi dei prompt possono smascherare tentativi di injection, esfiltrazione, utilizzo improprio da parte di personale interno e comportamento inatteso dell'agente. Si può iniziare a livello di proxy per i modelli ospitati nel cloud o a livello di harness per gli agenti distribuiti localmente, senza dover attendere l'integrazione completa del modello. Tuttavia, i prompt costituiscono solo una parte della risposta: strumenti e competenze stanno diventando un elemento fondamentale della superficie di attacco e devono essere trattati come tali.

Parte 3.
Come l'IA sta cambiando i modelli operativi difensivi

Triage e indagine

Esecuzione di playbook, ricostruzione delle tempistiche e reportistica in linguaggio naturale sono gli ambiti in cui i modelli di ragionamento possono apportare un valore operativo concreto. Ad esempio, l'infrastruttura MDR di Kaspersky ha elaborato circa 400.000 avvisi nel 2025, con una logica di rilevamento basata sull'IA che filtra i falsi positivi prima della revisione da parte degli analisti; 39.000 avvisi sono stati sottoposti a ulteriore indagine. Sempre più spesso i modelli di IA sono in grado di farsi carico di tali indagini; a marzo 2026 Prophet Security ha evidenziato su The Hacker News due casi in cui i modelli hanno svolto analisi contestuali e indagini su incidenti in cui le attività dannose non erano immediatamente evidenti.
Il primo riguardava un utente dormiente con una vecchia chiave API improvvisamente riattivata; il secondo comportava l'individuazione di un intento malevolo in un'e-mail di phishing che avrebbe superato la maggior parte dei controlli di sicurezza e di convalida (se non tutti). In entrambi i casi, osserva Prophet Security, nessun dato individuale era di per sé sospetto; la minaccia diventava evidente solo attraverso l'analisi contestuale e multipunto: il tipo di analisi che gli investigatori umani svolgono quando hanno il tempo e le risorse necessarie. Naturalmente, il problema è che spesso gli investigatori non hanno tempo e risorse, ed è qui che i modelli di ragionamento possono contribuire, garantendo a ogni avviso lo stesso livello di analisi.

Tuttavia, l'IA non è una panacea e richiede un'accurata progettazione ingegneristica. La ricerca sul sovra-ragionamento mostra che, oltre una certa soglia, i ragionamenti aggiuntivi portano i modelli a smentire le risposte corrette. Dopo circa 7.000 token, le inversioni da risposte corrette a errate hanno superato quelle positive, raggiungendo un rapporto di tre a uno verso i 12.000 token. Le implementazioni all'interno dei SOC devono quindi considerare lo sforzo di ragionamento, la profondità dei tentativi e i cicli di chiamata agli strumenti come parametri ingegneristici regolabili.

In linea generale, l'adozione non è uniforme e le aspettative devono essere calibrate. Secondo il report CISO di Splunk, solo il 40% dei CISO utilizza attualmente l'IA generativa nelle proprie funzioni di sicurezza. L'IA agentica mostra appena il 6% di distribuzione attiva, sebbene il 39% la stia esplorando.

Va inoltre considerato che gli strumenti di IA destinati ai team di difesa possono comportare una propria superficie di rischio. Il report State of Pentesting di Cobalt, ad esempio, ha rilevato che le applicazioni di IA e LLM presentano vulnerabilità ad alto rischio con una frequenza 2,7 volte superiore rispetto ai software tradizionali, con solo il 38% dei problemi risolti durante i penetration test.

Il problema della memoria

La maggior parte delle distribuzioni in produzione è limitata a quella che alcuni ricercatori descrivono come la seconda o terza generazione di maturità agentica: assistenti potenziati con strumenti che ripartono da zero a ogni sessione. Mantengono il proprio addestramento, ma non riescono a ricordare che lo stesso modello di avviso è apparso ieri e si è rivelato innocuo, oppure che il blocco di un particolare intervallo di IP il mese precedente aveva causato l'interruzione della VPN.

La distinzione tra un analista senior e uno junior non risiede necessariamente nella capacità di ragionamento, ma in ciò che apportano di fronte a ciascun evento: memoria episodica di incidenti passati, conoscenza semantica dell'ambiente e intuizioni procedurali su ciò che funziona. Gli assistenti IA attuali sono più vicini a un profilo junior che a uno senior, e analizzano ciascun avviso come se fosse la prima volta. I ricercatori stanno esplorando architetture con memoria strutturata come un sistema operativo, servizi di memoria esterna intesi come middleware e approcci basati su knowledge graph, con ragionamento bi-temporale. Fino a quando queste soluzioni non saranno mature, i sistemi di IA rimarranno utili assistenti anziché operatori autonomi.

Il confine del rilevamento

I sistemi di rilevamento basati sul comportamento agiscono su flussi di telemetria in tempo reale caratterizzati da un estremo sbilanciamento delle classi. Una piattaforma che elabora migliaia di miliardi di eventi al giorno li aggrega in milioni di rilevamenti deduplicati, li prioritizza in migliaia di avvisi ad alta gravità, e genera centinaia di indagini giornaliere che rimangono critiche dopo la revisione umana. Il rapporto end-to-end si avvicina a uno su dieci miliardi. La ricerca pionieristica di Stefan Axelsson sulla fallacia del tasso di base nel rilevamento delle intrusioni ne ha formalizzato l'aspetto matematico: anche un rilevatore con una specificità del 99,9% produce milioni di falsi allarmi per ogni minaccia effettiva, quando la probabilità a priori di un attacco è sufficientemente bassa.

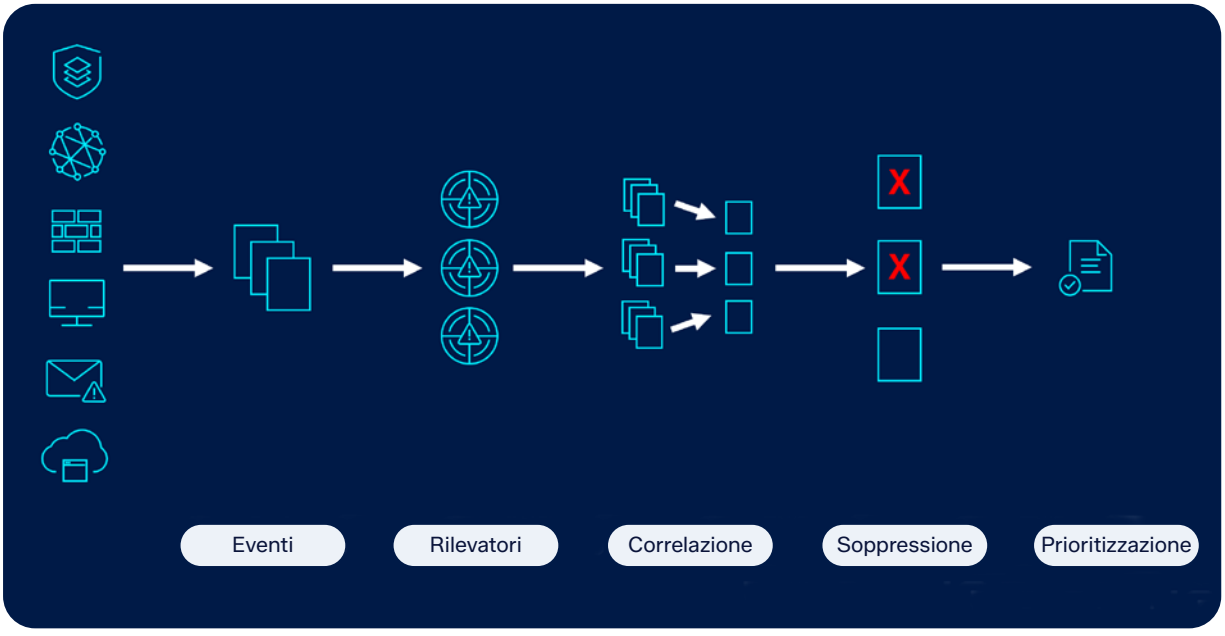


Figura 6: Panoramica della pipeline a più fasi

La ricerca Sophos presentata a NorthSec 2026 dimostra perché il rilevamento a livelli multipli è ancora fondamentale. Una finestra temporale di due settimane di telemetria, che comprendeva 11,8 trilioni di eventi, è stata ridotta tramite rilevatori di complessità crescente (corrispondenza con indicatori semplici, regole di correlazione, modelli di machine learning dedicati, deduplicazione, correlazione e soppressione) a 81.573 avvisi ad alta gravità e critici: una media di meno di 50 per organizzazione. All'interno di questi avvisi rimanenti, Sophos ha identificato un attacco infostealer collegato alla campagna TamperedChef.

I modelli di ragionamento si collocano al di sopra di questo livello. Generano ipotesi di indagine, recuperano e sintetizzano dati di telemetria in riepiloghi delle evidenze e coordinano azioni di risposta delimitate da barriere protettive esplicite. Non sostituiscono i modelli addestrati con dati proprietari sugli incidenti, i dati di riferimento specifici per l'ambiente, e i cicli di feedback continuo a cui i sistemi di ragionamento generale non hanno accesso.

Il McKinsey AI Trust Maturity Survey ha quantificato il vincolo organizzativo: le preoccupazioni legate alla sicurezza rappresentano il principale ostacolo alla scalabilità dell'IA agentica per quasi due terzi degli intervistati. Imprecisione (74%) e cybersecurity (72%) sono i principali rischi specifici identificati. Solo circa il 30% delle aziende ha raggiunto il livello di maturità tre ("Si stanno adottando misure per sviluppare tutte le necessarie pratiche di IA responsabile") o superiore nella governance dell'IA. Le aziende con responsabilità esplicite in termini di IA registrano un punteggio medio di maturità pari a 2,6, rispetto all'1,8 di quelle prive di una chiara assegnazione di responsabilità.

Il limite della difesa assistita dall'IA è la prontezza aziendale. I modelli sono sufficientemente efficaci; la domanda è se le organizzazioni sono in grado di gestirli al livello di fiducia richiesto dalle operazioni di sicurezza autonome.

Ricerca Sophos sull'IA: proteggere il livello IA

Oltre alla ricerca sul rilevamento a più livelli descritta sopra, il team Sophos AI ha sviluppato diverse altre tecniche di difesa innovative, ciascuna delle quali è progettata per gestire una modalità di errore specifica.

- **L'Untrusted Data Scanner (UDS)** ispeziona i contenuti esterni alla ricerca di attacchi di prompt injection, prima che raggiungano un LLM o un agente. Ogni caso di utilizzo degli LLM prevede un canale delle istruzioni (ciò che l'applicazione dice al modello di fare) e un canale dei dati (il contenuto non attendibile che il modello elabora). La prompt injection introduce istruzioni malevole nel canale dei dati. L'UDS è stato focalizzato deliberatamente su questo problema specifico, anziché svolgere il ruolo di rilevatore generico di jailbreak. I primi risultati mostrano un basso tasso di falsi positivi, un aspetto fondamentale dato che i dati di sicurezza sono ricchi di terminologia specifica (incident report, analisi di malware , risultati dei penetration test) e un rilevatore che "grida al lupo" a ogni report di minaccia sarebbe inutilizzabile.
- **Il LLM Scoping** classifica se le query rientrano nella funzione prevista di una funzionalità, respingendo le richieste fuori ambito prima che raggiungano il Large Language Model. Nei metodi valutati, i modelli di scoping hanno superato di gran lunga le difese basate sui prompt di sistema, riducendo i tassi di successo degli attacchi e portandoli a quasi zero in numerosi benchmark pubblici. Il team ha anche applicato il Multi-round Automatic Red Teaming (MART) per rafforzare i modelli di scoping contro l'adattamento degli autori degli attacchi.

- **CerBERTus**, un modello "a tre teste" basato su BERT, che affronta il problema della fragilità del rilevamento binario di jailbreak. Un singolo encoder condiviso alimenta tre "teste" di classificazione: nocività (primaria), categoria dell'obiettivo (ciò che l'utente sta cercando di fare) e stile di framing (come viene presentata la richiesta). I task ausiliari agiscono come bias induttivo, incoraggiando la rappresentazione a separare l'obiettivo dal wrapper. Il team ha creato un corpus di prompt fattoriale strutturato, incrociando obiettivi specifici (attacchi informatici, frodi, esplosivi, sintesi di droghe, privacy, propaganda estremista e altro) con frame avversari (giochi di ruolo, fiction, urgenza, pretesti accademici, offuscamento) per addestrare e sottoporre a stress test questa separazione.
- **Rilevamento delle anomalie**: nella ricerca presentata a Black Hat USA 2025, il team ha combinato il rilevamento delle anomalie con gli LLM, scoprendo che l'efficacia del metodo deriva dalla generazione di diverse righe di comando innocue, non dall'individuazione di quelle malevole; questo riduce significativamente i tassi di falsi positivi nei classificatori di righe di comando.
- **LLM salting**, presentato al CAMLIS 2025, trae ispirazione dal salting delle password: l'introduzione di variazioni comportamentali mirate in modo che i jailbreak precompilati non possano essere riutilizzati su implementazioni omogenee di LLM.

La cyber deception contro autori degli attacchi autonomi

Per due decenni, la cyber deception non è mai diventata un controllo di sicurezza primario. Contro gli attaccanti autonomi, l'equazione cambia. Gli agenti IA non si stancano mai e sono in grado di enumerare ogni percorso alla velocità della macchina. Ambienti ingannevoli come quelli proposti in Palisade LLM Agent Honeypot e nel framework MANTIS possono saturare il contesto di un agente con informazioni fuorvianti, consumarne il budget di inferenza e imporre costi economici reali. Un esempio di questo tipo, HoneyTrap, ha dimostrato un aumento del 149% nel "consumo di risorse da parte degli autori degli attacchi" rispetto alle difese convenzionali. Il settore è ancora immaturo, ma le condizioni in cui l'inganno diventa primario sono esattamente quelle che l'espansione dell'IA offensiva sta creando.

Dove l'asimmetria dei tempi si fa sentire

Un cybercriminale che utilizza l'IA non deve affrontare alcun ciclo di acquisizione, alcuna verifica della conformità e alcun comitato consultivo per le modifiche. Può adottare uno strumento e renderne operativi i risultati in pochi giorni. La campagna assistita dall'IA contro i dispositivi FortiGate ha compromesso oltre 600 firewall in 55 paesi nell'arco di cinque settimane: l'opera di un cybercriminale che ha utilizzato strumenti di IA commerciali.

Dal lato dei team di difesa, un rilevamento viene attivato, l'indagine assistita dall'IA produce una raccomandazione di contenimento in pochi minuti, ma il processo di remediation del cliente richiede un ticket di modifica, una finestra di manutenzione e uno SLA di 48 ore. Le indagini hanno subito un'accelerazione, ma la risposta no. Questo è probabilmente ciò che è accaduto con due vulnerabilità sfruttate rapidamente nell'ultimo anno. Come osservato in precedenza, la vulnerabilità CVE-2026-10520 (Ivanti Sentry) è stata sfruttata entro 24 ore dalla pubblicazione del proof-of-concept e CVE-2026-42208 (LiteLLM) entro 36 ore.

Sebbene il triage assistito dall'IA aiuti indubbiamente i difensori ad accelerare i tempi, il lavoro preventivo sui principi di base della sicurezza può contribuire a colmare questa asimmetria: ridurre la superficie di attacco, implementare la MFA e mantenere la telemetria sufficientemente completa da permettere di ricostruire cosa è successo quando la prevenzione risulta inefficace.

Parte 4
Cosa devono fare i responsabili di sicurezza

Le evidenze emerse da questo report supportano un numero limitato di decisioni condizionali, organizzate in base alla situazione attuale di un'organizzazione.

Calibrare la tolleranza al rischio per l'adozione dell'IA

Le evidenze di questo report creano una tensione che le aziende devono trovare il modo di risolvere: L'adozione dell'IA comporta rischi concreti, ma anche un'adozione lenta comporta dei rischi.

Le evidenze di questo report creano una tensione che le aziende devono trovare il modo di risolvere: L'adozione dell'IA comporta rischi concreti, ma anche un'adozione lenta comporta dei rischi. Muoversi con troppa cautela mentre gli autori degli attacchi accelerano significa rimanere indietro rispetto al ritmo difensivo, cosa che questo report identifica come la sfida più determinante. Accettare il rischio è parte integrante di questo processo; la questione è come assumersi i rischi giusti.

Il framework di rischio interno di Sophos separa i rischi "buoni" da quelli "cattivi" lungo due assi: il potenziale di beneficio e il contenimento degli effetti negativi. Un rischio "buono" garantisce un impatto significativo, misure per contenere il raggio d'impatto, un percorso reversibile (pilot, rollback o fallback), una responsabilità chiara e cicli di feedback definiti. Un rischio "cattivo" presenta un potenziale danno illimitato, viola requisiti non negoziabili (sicurezza, privacy, aspetti legali, conformità), è privo di un piano di ripristino, non ha una responsabilità chiara oppure omette il giudizio umano per le modifiche che riguardano direttamente i clienti.

Nello specifico per l'IA, il framework si traduce in domande pratiche prima di qualsiasi implementazione: qual è il peggior scenario credibile? Quanto è grande il potenziale raggio d'impatto? È possibile annullare l'operazione? a chi appartiene? Le implementazioni con effetti negativi contenuti, come un pilot in sola lettura di uno strumento agentico di triage applicato a un set di dati isolato, devono essere perseguite con determinazione. Quelle con effetti negativi elevati, che concedono a un agente l'accesso in scrittura all'infrastruttura di produzione, richiedono la mitigazione del rischio o la limitazione dell'ambito prima di procedere.
Le distribuzioni a basso impatto positivo ed elevato rischio negativo devono essere evitate a prescindere.

I requisiti non negoziabili si applicano indipendentemente dalle circostanze: fiducia dei clienti, sicurezza, cybersecurity e privacy, aspetti legali, conformità e stabilità operativa. Il ruolo del CISO in un ambiente accelerato dall'IA è garantire che l'organizzazione si assuma rischi intelligenti, evitando quelli sconsiderati.

Per le organizzazioni che utilizzano l'IA agentica: riduzione del raggio d'impatto

Sophos ha proposto sette controlli, attuabili nei prossimi 1-6 mesi, per ridurre il raggio d'impatto per le aziende che desiderano implementare l'IA agentica:

1. **Il sandboxing degli agenti** definisce un perimetro controllato attorno al processo dell'agente. La maggior parte degli harness viene fornita con una qualche forma di sandboxing, ma di solito si tratta di una funzione opzionale. Ove possibile, scegli ambienti di esecuzione remoti o basati sul cloud invece di sandbox locali, garantire una separazione dei processi più efficace.

2. **L'isolamento delle credenziali** assicura che l'agente non veda mai direttamente i segreti: un processo separato recupera le credenziali da un vault, le inietta nella chiamata API e restituisce solo risposte prive di dati sensibili, mantenendo le credenziali a lunga durata al di fuori delle finestre di contesto degli LLM.

3. **Gli endpoint degli strumenti sigillati** vanno oltre. L'agente chiama uno strumento per nome, ma un processo broker custodisce la credenziale, effettua la chiamata API effettiva secondo uno schema fisso e applica elenchi di autorizzazione in uscita per ciascun strumento. L'unico potere decisionale dell'agente è scegliere quale strumento invocare e quali parametri fornire.

4. **La restrizione del traffico in uscita e il monitoraggio della rete** trattano l'agente compromesso come un canale di esfiltrazione dei dati: rilevamento di segreti nel traffico in uscita, soglie volumetriche per le richieste e categorizzazione web per bloccare le chiamate verso domini non classificati o registrati di recente.

5. **La copertura EDR** deve estendersi al comportamento degli host dell'agente: container, VM effimere e dispositivi degli sviluppatori non gestiti rappresentano tipici punti ciechi.

6. **L'approvazione vincolata all'intervento umano** separa il piano di esecuzione autonomo dell'agente da un piano di controllo gestito dagli esseri umani: le operazioni irreversibili richiedono un protocollo di autorizzazione crittografica, anziché un semplice pulsante "OK" che l'agente potrebbe simulare.

7. **I confini di propagazione delle injection** affrontano il rischio crescente man mano che le injection si spostano dal livello della singola sessione a quello dell'agente, fino alla propagazione tra più agenti. Gestisci le scritture in memoria e i passaggi di consegne tra agenti alla stregua di veri e propri eventi di sicurezza.

Quando OpenClaw ha preso piede all'inizio del 2026, oltre 30.000 istanze sono risultate esposte su Internet e i cybercriminali stavano già valutando come sfruttarne le "abilità" modulari a fini offensivi per campagne botnet. Il red team di Sophos l'ha testato direttamente: 23 rilevamenti utilizzabili, ricognizione su AD ridotta da tre giorni a tre ore, e un livello di dettaglio negli audit non raggiungibile manualmente in tempi ragionevoli.. Tuttavia, il team ha trascorso più tempo a definire i confini di sicurezza che a eseguire il test stesso. La lezione: l'IA agentica è abbastanza potente da giustificarne l'implementazione, ma solo a patto che il framework di sicurezza venga creato per primo.

Visibilità, identità e patch

Consigliamo di iniziare ottenendo visibilità sull'uso dei servizi di IA. L'esposizione all'IA documentata negli incidenti Salesloft/Drift e Vercel rende questa strategia la mossa iniziale a maggior valore strategico. Mappa gli strumenti di IA presenti nel tuo ambiente. Identifica quali contengono token OAuth, chiavi API o account di servizio connessi a sistemi aziendali. Determina quali dati vi fluiscono e quali autorizzazioni detengono. Non puoi regolamentare ciò che non vedi.

Le misure di visibilità devono produrre un registro dell'IA: ogni strumento di IA, agente, modello, credenziale a cui può accedere, la configurazione di sandboxing e i requisiti di connettività previsti. Senza quel registro, controlli come l'isolamento delle credenziali, le restrizioni del traffico in uscita e il sandboxing non possono essere applicati in modo coerente. Applica controlli dell'identità con lo stesso rigore di qualsiasi altra applicazione SaaS: implementa policy di accesso condizionale, richiedi la MFA per gli account di servizio dell'IA, applica controlli di DLP ai percorsi di uscita utilizzati dagli strumenti di IA e ruota le credenziali per le integrazioni connesse all'IA con la stessa pianificazione di qualsiasi altro accesso privilegiato. I rilievi di IBM X-Force sulle credenziali e l'attacco alla supply chain di Nx Console confermano entrambi che le credenziali dei servizi di IA sono un bersaglio primario di harvesting.

Le organizzazioni la cui frequenza di applicazione delle patch viene ancora misurata in settimane sono esposte a un rischio crescente. La tempistica di tre giorni definita dalla direttiva CISA BOD 26-04 rappresenta una soglia minima, perché l'exploit delle vulnerabilità assistito dall'IA sta riducendo la finestra temporale tra divulgazione ed exploit a poche ore. Se una vulnerabilità critica esposta su Internet non può essere corretta entro pochi giorni, tale lacuna comporta un rischio misurabile.

L'exploit delle vulnerabilità assistito dall'IA sta riducendo la finestra temporale tra divulgazione ed exploit a poche ore.

Valuta basandoti sulle evidenze

Integra la telemetria negli strumenti di IA utilizzati dai tuoi analisti. Misura l'impegno di verifica, il tempo che gli analisti dedicano a controllare l'output dell'IA. Misura l'efficienza del prompting, il numero di tentativi necessari prima di ottenere risultati utili. Misura anche la calibrazione della fiducia, per verificare se gli analisti ripongono una fiducia eccessiva o insufficiente nel sistema. Crea set di test basati su reali carichi di lavoro del SOC, anziché su benchmark generici, e gestisci le tue procedure tramite versionamento, in modo che sopravvivano agli aggiornamenti dei modelli.

Proteggi la supply chain dell'IA

L'IA introduce rischi alla supply chain che vanno oltre le tradizionali dipendenze software. I pesi dei modelli, la provenienza dei dati di addestramento, l'integrità del server MCP e l'infrastruttura di inferenza rappresentano tutti confini di fiducia. Le organizzazioni che implementano modelli open-weight tramite provider cloud gestiti devono verificare l'esatta versione del modello, la regione di erogazione del servizio, il periodo di conservazione dei dati, la policy sull'uso dei dati per l'addestramento, la copertura dei log e i termini contrattuali, prima di inviare codice o dati privati. Per i modelli erogati attraverso endpoint SaaS con sede nella Repubblica Popolare Cinese, come l'API DeepSeek, si applica un perimetro più rigido: tali modelli devono essere utilizzati unicamente per test funzionali con dati pubblici o sintetici, non per repository proprietari o contesti ingegneristici sensibili.

Conclusione

Il framework dei cybercriminali STAC6994 conteneva profili Cobalt Strike, moduli di elusione dell'EDR, strumenti di harvesting di credenziali, utilità per il movimento laterale e un canale C2 nascosto dietro un'infrastruttura cloud legittima. Ogni componente ci era familiare, eccetto il processo di sviluppo. Gli agenti IA hanno iterato le tecniche di elusione più velocemente di quanto avrebbe potuto fare uno sviluppatore umano, le hanno testate su prodotti specifici in un laboratorio strutturato, hanno gestito le versioni dei risultati tramite Git e hanno consegnato l'output a un operatore che ha poi distribuito ransomware e rubato dati.

Tra i casi gestiti da Sophos MDR e Incident Response, i dati di intelligence della CTU, le analisi dei SophosLabs, la ricerca di Sophos AI e il settore della sicurezza in generale, l'IA sta accelerando operazioni familiari su entrambi i lati dell'equazione della sicurezza. Aiuta gli autori degli attacchi a muoversi più velocemente lungo catene di attacco conosciute. Consente ai team di sicurezza di effettuare triage e indagini in tempi estremamente compressi. Crea nuove esposizioni attraverso il livello di identità e governance che circonda l'adozione dell'IA in azienda. Non ha prodotto una tipologia di intrusione fondamentalmente nuova che le architetture di rilevamento esistenti non siano in grado di gestire. L'UK AI Security Institute ha misurato un incremento di circa sei volte delle capacità offensive autonome nell'arco di 18 mesi. Ogni generazione di modelli sbloccava passaggi in scenari di intrusione a più fasi che le generazioni precedenti non riuscivano a completare. Di fronte a questa traiettoria non ci si può permettere alcuna distrazione.

Tre temi rendono ancora più incisiva la conclusione:

1. **La classificazione** deve essere precisa, poiché raggruppare tutte le minacce legate all'IA sotto un'unica voce oscura la postura difensiva richiesta da ciascuna di esse.
2. **La valutazione** deve essere rigorosa, perché sia che si tratti del routing del modello, della soppressione degli avvisi o della profondità di ragionamento dell'agente, la risposta deve derivare dalle misurazioni.
3. La **trasparenza** deve essere reale, perché la fiducia viene accordata alle organizzazioni che mostrano il proprio lavoro e che insieme ai successi condividono anche i fallimenti utili.

Cosa osserverà Sophos nei prossimi 12 mesi: se gli agenti autonomi passeranno dall'eseguire playbook noti al scoprire nuovi percorsi di attacco su vasta scala; se l'economia clandestina svilupperà strumenti offensivi di IA ingegnerizzati come prodotti, abbassando l'asticella in modo drastico così come il ransomware-as-a-service ha abbassato quella dell'estorsione; se la governance dell'IA aziendale maturerà abbastanza rapidamente da colmare l'esposizione delle identità creata dall'attuale ondata di adozione; e se la finestra temporale tra exploit e patch si comprimerà ulteriormente, imponendo un momento di resa dei conti alle organizzazioni che misurano ancora il proprio ritmo di risposta in settimane.

Il ritmo è cambiato. I responsabili di sicurezza che non adegueranno la propria velocità si ritroveranno più indietro di quanto credano.

Contributori

Nash Borges SVP di Engineering e AI

Adarsh Kyadige
Senior Manager
AI Research

Ross McKerchar
CISO

Rafe Pilling
Director di Threat Intelligence
X-Ops Counter Threat Unit

Matt Stec
Director
Sophos AI

Ryan Westman
Senior Manager Threat Research
X-Ops Insights

Matt Wixey
Senior Threat Researcher
X-Ops Insights

Il nostro team

Sophos X-Ops è un'iniziativa di cybersecurity all'avanguardia che riunisce più di 1.000 esperti provenienti da diversi settori di specializzazione all'interno di Sophos, tra cui Threat Intelligence, Artificial Intelligence, MDR Operations e Security Ops interne.

Questa task force interfunzionale potenzia le difese delle organizzazioni contro le minacce informatiche moderne, in rapida evoluzione e altamente sofisticate.

Sfruttando le competenze combinate del proprio team, Sophos X-Ops offre una risposta multidimensionale agli attacchi informatici, garantendo capacità complete di protezione, rilevamento e risposta.

Questo approccio collaborativo e innovativo garantisce una risposta alle minacce senza precedenti, posizionando Sophos come punto di riferimento per l'eccellenza e leader nella cybersecurity.

Sophos X-Ops fa leva sulle competenze combinate della sua task force interfunzionale. La sinergia tra i team trasversali di Sophos X-Ops alimenta l'intelligence condivisa, che permette di adattarsi rapidamente all'evoluzione dello scenario: questo accelera i tempi di rilevamento e risposta, rafforzando al contempo le capacità di protezione complessive per i clienti Sophos.

SOPHOS

Per parlare delle tue esigenze in materia di sicurezza dell'IA e di come Sophos può aiutarti, visita il nostro sito o parla con un consulente.

Vendite per l'Italia
Tel: (+39) 02 94 75 98 00
E-mail: sales@sophos.it

2026-07-20 IT (NP) CRE-5820