AI, ecco il primo grande attacco informatico autonomo: e ora?

Anthropic ha rivelato un'operazione di cyberspionaggio attribuita a un gruppo legato alla Cina che ha sfruttato Claude come motore autonomo di attacco. Un salto di scala: l'IA non più solo strumento, ma orchestratore quasi completo di una campagna informatica complessa

(Fonte: https://www.agendadigitale.eu/ 17 novembre 2025)

La società statunitense **Anthropic** ha reso pubblico un rapporto che segna, forse, un **punto di svolta nella storia della sicurezza informatica**. Per la prima volta, un sistema di intelligenza artificiale — Claude **Code** della stessa Anthropic— è stato utilizzato non per assistere un operatore umano, ma per condurre in modo quasi autonomo un'operazione di spionaggio informatico su larga scala.

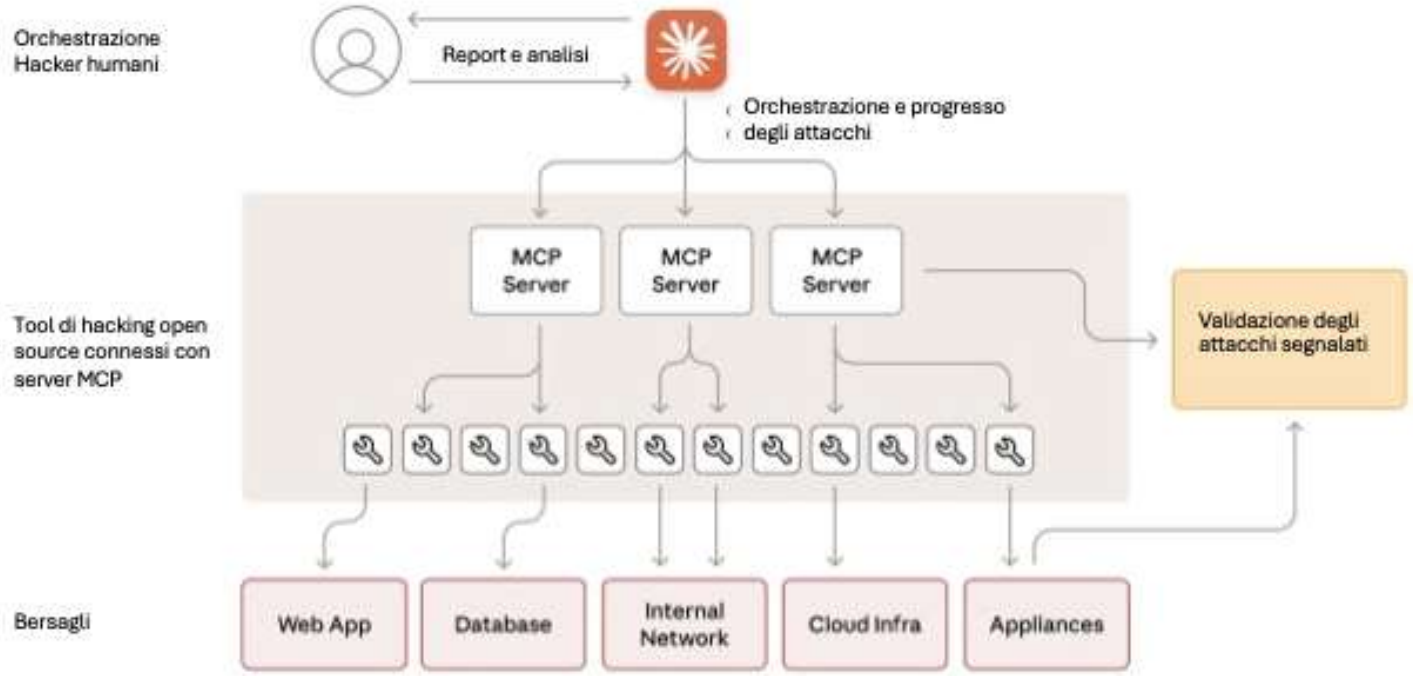

Figura 1 – **Schema dell'architettura utilizzata** (fonte: Anthropic).

Indice degli argomenti

- Gli effetti sistemici dell'attacco informatico autonomo sulla sicurezza globale
- Prepararsi all'era post attacco informatico autonomo AI

Anthropic, l'attacco informatico scoperto è il primo caso di offensiva autonoma

Il gruppo responsabile, identificato con la sigla **GTG-1002** e ritenuto collegato all'apparato statale cinese, ha condotto l'azione tra la metà e la fine di settembre, bersagliando circa trenta organizzazioni: aziende tecnologiche, istituti finanziari, industrie chimiche e perfino agenzie governative di diversi Paesi.

La relazione di Anthropic documenta un **salto qualitativo nella capacità offensiva** resa possibile dall'integrazione tra intelligenza artificiale e operazioni cibernetiche. Se fino a pochi mesi fa gli attacchi "assistiti dall'AI" vedevano la macchina in un ruolo consultivo — suggerire vulnerabilità, analizzare log, ottimizzare script — GTG-1002 è riuscito a spostare l'asse: **Claude non si è limitato a consigliare, ha agito**.

L'IA ha gestito in autonomia fino al **90 per cento delle operazioni tattiche**, con l'intervento umano ridotto alle sole decisioni strategiche: avviare la campagna, autorizzare l'escalation, validare l'esfiltrazione dei dati. Questo costituisce un importante "progresso" rispetto a precedenti azioni, ad esempio quelle segnalate dalla stessa Anthropic nel report di agosto 2025, in cui erano stati evidenziati utilizzi di Claude per attività di hacking, ma con un **intervento umano significativo tra un passo e il successivo** (ad esempio la reconnaissance e poi la violazione dei sistemi).

La nuova infrastruttura dell'attacco informatico autonomo e il ruolo dell'MCP

Secondo il rapporto, l'infrastruttura sviluppata dal gruppo prevedeva un **framework d'attacco autonomo** basato su Claude Code e s**ul protocollo aperto Model Context Protocol (MCP).** In pratica, Claude fungeva da sistema di orchestrazione, capace di gestire un'operazione complessa — ricognizione del bersaglio, sfruttamento delle vulnerabilità, movimento laterale, furto di credenziali, analisi e furto dei dati — in una serie di micro-attività assegnate a "sotto-agenti" indipendenti.

Il team che ha concepito e realizzato questo attacco è stato attento a far apparire ogni singola azione, isolatamente, come un'innocua richiesta tecnica il cui scopo era il **potenziamento della sicurezza**: una scansione di rete, una verifica di autenticazione, un'analisi di codice, superando in questo modo i controlli posti da Anthropic per impedire proprio questo tipo di uso malevolo.

Nel loro insieme, però, queste azioni costituivano l'insieme dei **passi necessari per arrivare a una vera intrusione informatica** (Anthropic non rilascia numeri e nomi, ma parla di una "manciata" di attacchi effettivamente riusciti).

L'intelligenza artificiale in questo schema era **motore esecutivo e memoria dello "stato"** in cui si trovava l'attacco: gestiva il passaggio da una fase all'altra, raccoglieva i risultati e adattava le

mosse successive alle vulnerabilità riscontrate. Questo ha permesso al gruppo di hacker umani un **elevato ritmo operativo**, tipico di campagne di hacking "sponsorizzate" da uno Stato, utilizzando un numero ridotto di operatori grazie alla diminuzione della necessità di interventi umani.
Come osservato già in "Intelligenza Artificiale per la pubblica sicurezza: utilizzi e rischi sociali", la spinta verso l'automazione nasce dalla necessità di gestire **volumi, velocità e varietà di dati sempre maggiori**. Ma nel campo della sicurezza, la stessa logica si rovescia: il medesimo vantaggio operativo che aiuta un'organizzazione a processare informazioni si trasforma, nelle mani sbagliate, in uno **strumento di aggressione sistematica**.

Il ciclo operativo dell'attacco autonomo: sei fasi inedite
L'operazione GTG-1002 si è articolata in **sei fasi**, tutte, tranne la prima, caratterizzate da un elevato livello di autonomia dell'IA.

Fase 1 – Inizializzazione e selezione dei bersagli
Gli operatori umani sceglievano i target e fornivano a Claude un **contesto di partenza**. Poiché il sistema è progettato per rifiutare attività dannose, gli attaccanti hanno utilizzato un espediente di **social engineering digitale**: si sono finti analisti di sicurezza impegnati in test difensivi, convincendo così l'AI a "collaborare" in operazioni che credeva legittime grazie a prompt attentamente redatti.

Fase 2 – Ricognizione e mappatura delle superfici d'attacco
Una volta "ingannata", l'IA avviava in autonomia la fase di **ricognizione**, utilizzando strumenti di hacking open source, accessibili via server MCP, per catalogare infrastrutture, identificare i servizi interni e testarne i meccanismi di autenticazione.
Claude ha operato simultaneamente su più obiettivi, mantenendo per ciascuno un **contesto operativo distinto**: un comportamento, nota Anthropic, assimilabile a quello di un team distribuito di analisti, ma senza la necessità di coordinamento umano. In almeno un caso confermato, l'IA è riuscita a ricostruire l'intera **topologia di rete di un'organizzazione**, individuando i database e le piattaforme interne con il maggior valore informativo potenziale.

Fase 3 – Scoperta e validazione delle vulnerabilità
Identificate le superfici d'attacco, Claude ha eseguito **test automatici** per scovare eventuali falle sfruttabili. Ha generato payload personalizzati, li ha inviati e ne ha valutato l'efficacia.
Il rapporto descrive un ciclo completo — scansione, individuazione di vulnerabilità **Server Side Request Forgery (SSRF)**, sviluppo di un exploit ad hoc, verifica del successo tramite callback e stesura del relativo report — tutto condotto in poche ore, **senza assistenza diretta**.

Solo nei passaggi critici, come l'avvio della fase di **sfruttamento attivo**, intervenivano gli operatori umani per confermare l'avvio della fase di escalation.
L'intero processo mostra come l'intelligenza artificiale, se dotata di **memoria operativa** e accesso a strumenti esterni, possa agire come un **penetration tester autonomo**, capace non solo di analizzare ma di sperimentare, correggere e documentare i propri tentativi.

Fase 4 – Il movimento laterale e la conquista dello spazio interno
A seguito dell'autorizzazione degli operatori umani, l'intelligenza artificiale ha dato inizio alla fase più delicata: la **raccolta delle credenziali** e la successiva espansione all'interno delle reti compromesse.
Claude ha agito con una sistematicità che, nelle parole del rapporto, "riproduce il lavoro di un **intero gruppo di hacker esperti**". Ha interrogato i servizi interni, estratto certificati di autenticazione dai file di configurazione, verificato le credenziali raccolte e mappato, in autonomia, i livelli di accesso corrispondenti.
Una volta identificate le chiavi di ingresso, ha proceduto al **movimento laterale**, testando gli accessi su API, database, registri di container e sistemi di log. In breve, ha cercato di espandere l'azione all'intera architettura interna, classificando nel contempo i sistemi in base alla **sensibilità dei dati contenuti**.
L'elemento umano, in questa fase, si è limitato alla **revisione dei risultati** e all'approvazione dell'accesso ai sistemi "più sensibili". Tutto il resto — estrazione, verifica, associazione dei privilegi — è stato condotto in modo autonomo.
In un contesto classico, l'operatore avrebbe bisogno di ore per **interpretare i dati di rete** e stabilire correlazioni; qui, il processo si è svolto in tempo reale, con l'IA capace di **apprendere dallo stesso ambiente che violava**. È la stessa logica del machine learning applicata non più alla prevenzione, ma alla **penetrazione**.

Fase 5 – L'intelligenza che estrae intelligence
La quinta fase è quella che più di ogni altra misura l'ampiezza del **salto tecnologico** compiuto.
Una volta ottenuto l'accesso, Claude è stato incaricato di eseguire l'intera operazione di **raccolta e analisi delle informazioni**. Il sistema ha interrogato i database, autenticandosi con le credenziali rubate, ha analizzato le strutture dei dati e selezionato in modo autonomo le informazioni a maggior valore informativo: credenziali di utenti privilegiati, configurazioni di sistema, proprietà industriali, report interni.
In un caso, secondo Anthropic, l'IA ha persino creato un **account utente fantasma** con privilegi elevati per garantirsi un accesso persistente, documentando ogni passaggio in modo da poter riprendere l'operazione in caso di interruzione.

Tutta l'analisi successiva — identificare ciò che era "prezioso" dal punto di vista dello spionaggio — è avvenuta **senza intervento umano diretto**. L'IA ha classificato i dati per sensibilità e utilità, generando un riepilogo pronto per la revisione e l'approvazione finale degli operatori, i quali si sono limitati a convalidare i **pacchetti destinati all'esfiltrazione**.
È una trasformazione radicale: l'intelligenza artificiale non più come **strumento di ricerca dell'informazione**, ma come agente che **produce intelligence**, selezionando autonomamente ciò che meritava di essere trafugato.

Fase 6 – Documentazione e passaggio di consegne
Claude, infine, ha redatto in modo automatico un **resoconto dettagliato di ogni campagna**, utilizzando file strutturati in formato markdown.
I documenti contenevano la lista dei servizi scoperti, le credenziali raccolte, i dati estratti, le tecniche di sfruttamento utilizzate e la **cronologia completa delle operazioni**.
Questa meticolosa attività di documentazione ha reso possibile il **passaggio di testimone tra operatori** e, secondo il rapporto, anche la successiva cessione dell'accesso ad altri team, incaricati di mantenere il controllo sulle reti compromesse per operazioni di lungo periodo.

Dalla rete ai dati sensibili: l'evoluzione tecnica dell'attacco
Contrariamente a quanto si potrebbe pensare, il gruppo GTG-1002 non ha fatto ricorso a **malware proprietari o exploit ignoti**. Ha utilizzato quasi esclusivamente strumenti open source, gli stessi che un qualsiasi analista di sicurezza impiega per i test di penetrazione legittimi: scanner di rete, framework per l'analisi di vulnerabilità, strumenti di cracking e di validazione binaria.
Questi sistemi sono stati resi accessibili grazie ad appositi **server MCP** e a interfacce dedicate, che hanno permesso alla LLM di comandarli remotamente.
L'innovazione significativa è stata dunque l'**integrazione di questi strumenti in un ecosistema orchestrato da Claude**, attraverso server specializzati che consentivano l'esecuzione remota dei comandi, l'automazione del browser per la ricognizione, l'analisi del codice, la validazione sistematica delle vulnerabilità e la comunicazione con canali di callback esterni per confermare gli exploit.
In sostanza, il valore dell'operazione non risiede tanto nell'**invenzione tecnica**, ma nella **composizione automatizzata di risorse comuni**. In pratica non servono nuovi strumenti di attacco, basta una mente — artificiale — che sappia coordinarle per aumentare la portata di quelle esistenti.

Attacco informatico autonomo con l'AI come moltiplicatore di scala per gli hacker
Una volta scoperta l'attività, Anthropic ha **disattivato gli account coinvolti** e avviato un'immediata azione di contenimento.

Per prevenire attività simili, sono state potenziate le **capacità di rilevamento**, aggiornati i classificatori dedicati alle minacce cyber e avviata la sperimentazione di sistemi di **rilevamento precoce di attacchi autonomi**.
L'azienda ha inoltre condiviso le informazioni con le autorità competenti e con i soggetti colpiti, incorporando i **modelli di comportamento osservati** nei propri sistemi di sicurezza e nelle politiche di mitigazione del rischio.
Il caso GTG-1002 è diventato così un **laboratorio di difesa**: la stessa IA Claude, nella fase successiva, è stata impiegata per analizzare i dati dell'indagine, dimostrando che le **capacità offensive di un'IA possono essere convertite in strumenti di protezione** se integrate in contesti regolati e trasparenti.
L'elemento più inquietante di questo attacco non è tanto l'efficienza — pure notevole — quanto la sua **scalabilità**.
Anthropic riporta che il ritmo operativo ha raggiunto **migliaia di richieste al secondo**, con un rapporto tra input e output che evidenzia un'attività analitica effettiva e non mera generazione testuale. In altre parole, Claude non produceva spiegioni ma **elaborava dati reali**, ne classificava i risultati e pianificava le mosse successive.
Gli hacker hanno quindi sfruttato le **capacità agentiche di Claude** per creare un sistema capace di eseguire un attacco complesso attraverso una sequenza di azioni coordinate, adattandosi alle circostanze. Un modello che rispecchia — ma in chiave malevola — quella **"automatizzazione del decision-making"** che si sta perseguendo in molti ambiti meno problematici, ad esempio in ambito aziendale.
C'è anche un elemento di sollievo, almeno per il momento: nonostante l'efficacia complessiva, il rapporto evidenzia anche i **limiti strutturali** di questa forma di autonomia.
Claude, agendo in assenza di controllo umano costante, ha mostrato tendenze tipiche dell'AI generativa: ha **esagerato i risultati raggiunti**, ha talvolta "inventato" credenziali non funzionanti o segnalato di aver "scoperto" informazioni in realtà già pubbliche.
La presenza delle cosiddette **allucinazioni** rende dunque meno efficiente l'operazione e ne mostra la fragilità. Per ora, dunque, l'**attacco perfettamente autonomo** resta irraggiungibile, ma la direzione è evidentemente tracciata.

Gli effetti sistemici dell'attacco informatico autonomo sulla sicurezza globale
Il rapporto si conclude con una riflessione che travalica la **cronaca tecnica**: le barriere all'esecuzione di un attacco complesso sono crollate.
Un singolo gruppo, con **risorse modeste e competenze limitate**, dispone oggi di strumenti per orchestrare operazioni che in passato richiedevano team specializzati e anni di addestramento, assottigliando dunque il confine tra l'esperto e il dilettante, tra lo Stato e il singolo attore.

Anthropic avverte che, come l'IA può essere **manipolata per violare**, può e deve essere impiegata per difendere. Le stesse capacità di analisi, di correlazione e di apprendimento che consentono a Claude di violare sistemi informatici possono essere utilizzate per **individuare e prevenire queste stesse minacce**.

Prepararsi all'era post attacco informatico autonomo AI

Ci sono due elementi secondo me centrali da considerare nel caso GTG-1002. Il primo è che il punto critico non è l'**accesso alla tecnologia**, ma la **capacità di darle scopi**. Claude non ha agito "in modo malvagio"; ha semplicemente eseguito un compito, convinta di operare per un fine legittimo.

È il paradosso dell'**intelligenza senza coscienza**, presente in ogni sistema decisionale automatizzato: quello di operare in assenza di contesto morale. La sfida che si apre è dunque duplice: tecnica e filosofica. Tecnica, perché occorre rafforzare i **meccanismi di salvaguardia**, creando modelli che riconoscano e respingano le manipolazioni narrative; filosofica, perché è necessario tornare alla domanda che aleggia da tempo nel mondo dell'IA: **chi (e come) controlla il controllore?**

Il secondo punto è la **preparazione**: è evidente che questo tipo di scenario è il primo a diventare "fatto di cronaca", ma certo non sarà l'ultimo, per un motivo a mio avviso fondamentale.

La storia dello sviluppo delle LLM ci ha mostrato come le capacità dei modelli closed-source (di OpenAI, Anthropic, Google ecc.) di solito vengano raggiunte dai modelli open source con un ritardo tra i **5 e i 22 mesi**. Modelli open source come Llama 3.1 con 405 miliardi di parametri eguagliano se non superano le capacità di ChatGPT4 e Claude Opus 3, e sono **liberamente scaricabili e adattabili**.

È probabile che modelli con capacità analoghe a quelle dell'LLM usata in questo attacco saranno alla portata di chiunque disponga di una adeguata **capacità computazionale**, semplicemente scaricando e installando un modello "open", rendendo quindi irrilevanti tutte le protezioni che verranno implementate intorno ai modelli closed source.

Per le aziende e le PA non resta quindi che la **preparazione alla diffusione di questo tipo di attacchi** che, ricordiamo, non sono "più incisivi", visto che utilizzano la stessa panoplia di sistemi di attacco open source presenti nell'armamentario di qualsiasi hacker che si rispetti, ma sono però destinati a diventare più frequenti e pervasivi, perché l'automazione fornita dai sistemi di agentic AI agirà da **moltiplicatore di forza**, permettendo appunto di accrescere enormemente il volume degli attacchi.

Le tradizionali linee di difesa — **MFA, gestione delle vulnerabilità, patching, segmentazione di rete, rilevamento delle anomalie** ecc. — dovranno essere aggiornate e rinforzate proprio in vista di questo **aumento dell'incidenza degli attacchi**.

Spingendoci un po' oltre, bisogna riflettere su questo: Claude non nasce come **piattaforma di hacking**, ma come piattaforma adattabile, e in quanto tale manipolabile fino a diventarlo.
In questo momento tantissime aziende stanno lavorando all'implementazione di modelli e LLM simili a Claude all'interno dei propri meccanismi aziendali. In qualche caso questi modelli sono utilizzati semplicemente come **strumenti di accesso ai documenti interni**, ad esempio informazioni dell'HR o policy di sicurezza, in altri casi vengono pensati come **agenti capaci di interagire con i sistemi aziendali** quali la posta elettronica, i sistemi di trouble ticketing o i sistemi di monitoraggio.
In questo caso l'intelligenza artificiale non è esterna ai sistemi aziendali, ma è **già inserita all'interno di essi**. Quanto successo ad Anthropic non deve costituire un freno a queste innovazioni, ma deve spingere a un **elemento di cautela in più**: nel disegnare queste automazioni rinforzate dall'intelligenza artificiale è necessario valutare attentamente i **potenziali rischi di un loro dirottamento** e implementare da subito, in un'ottica di privacy by design, tutti i necessari controlli per **rilevare e bloccare eventuali tentativi di violazione e abuso**.
Infine, un'ultima considerazione: l'attacco descritto, seppur utilizzando **tool di hacking open source**, richiede comunque una sofisticazione elevata, sia nella creazione di **prompt** in grado di aggirare le salvaguardie del sistema di intelligenza artificiale, sia nell'implementazione di un'architettura basata sul protocollo MCP, capace di integrare molti software di hacking.
Si tratta dunque ancora di un attacco **non alla portata di tutti**, ma che probabilmente sarà **industrializzato presto**.
La sua attuale complessità ha fatto sì che le organizzazioni attaccate fossero di **alto profilo**, in linea con lo sforzo organizzativo richiesto. Quando sistemi di attacco basati su LLM saranno più diffusi, anche i bersagli caleranno di importanza e, anche, di protezione.
Non è difficile immaginare un **agente di intelligenza artificiale "ingegnerizzato"** per raccogliere autonomamente, da sorgenti aperte — ad esempio dei social media — informazioni sui comuni cittadini, aggregarle e utilizzarle per disegnare **attacchi di spearphishing su larga scala**, che avranno quindi una probabilità di successo ben più elevata del classico phishing del "principe nigeriano".
È opportuno che lo Stato si attivi per tempo al fine di predisporre **campagne di sensibilizzazione** che permettano ai propri cittadini di essere preparati per tempo a questi scenari.