Sicurezza delle Reti al tempo degli Agenti di Intelligenza Artificiale

Gli ultimi due anni hanno portato un notevole sviluppo della tecnologia informatica chiamata genericamente Intelligenza Artificiale (IA), tecnologia basata principalmente sui Large Language Model (LLM). L'IA sta avendo impatti su moltissimi aspetti della vita umana, e in particolar modo dell'informatica stessa.

In realtà negli ultimi trent'anni la digitalizzazione dei processi, delle attività e in generale della nostra vita è proceduta e continua a procedere a passo sempre più veloce. Ben pochi momenti della nostra vita non sono ancora supportati in qualche modo da uno strumento digitale o dalla gestione digitale delle informazioni necessarie o utili al suo svolgimento. Questo è facilmente comprensibile in quanto qualunque nostra umana valutazione, decisione, azione è basata sulla elaborazione di informazioni nel nostro cervello. Gli strumenti digitali processano informazioni e per questo sono strumenti utili o di supporto praticamente a quasi ogni aspetto e momento della nostra vita.

All'interno di questo processo, l'ambito ora chiamato Intelligenza Artificiale, nato circa settant'anni fa, è recentemente passato da essere uno strumento di nicchia e utilizzato da pochi a essere lo strumento informatico oggi probabilmente con la maggiore notorietà.

È quindi utile riflettere sugli impatti che l'IA sta avendo sui sistemi informatici stessi. Tenendo conto che ormai viviamo in un'unica rete digitale, Internet, a cui sono connessi direttamente o indirettamente quasi tutti i sistemi informatici mondiali, e che molte reti interne aziendali includono sistemi e servizi distribuiti geograficamente, può essere utile considerare i possibili impatti dell'IA sulla sicurezza delle reti, pubbliche e private. E per valutare gli impatti, conviene partire dalle più recenti capacità dei sistemi di IA.

I modelli di IA Generativi e gli Agenti IA

Si può assumere che i modelli IA Generativi più avanzati (GPT, Claude, Gemini, Grok, DeepSeek, Llama, Midjourney ecc.) siano addestrati utilizzando le informazioni disponibili pubblicamente in Internet che riguardano le reti digitali e le applicazioni informatiche. Considerando inizialmente le reti digitali, è come se una persona avesse imparato a memoria tutti i manuali dei protocolli di rete con le relative linee guida di implementazione, insieme all'indirizzo da cui può scaricare in qualunque momento il documento originale. Solo come esempio illustrativo, una piccola e

incompleta lista di protocolli di rete è [1]: ARP, ATM, Ethernet, FDDI, Frame Relay, IEEE 802, L2TP, MAC, MPLS, NDP, PPP, PPPoE, PTPP (PointToPoint Protocol), STP, VTP, HSRP, IP, NAT, VRRP, BGP, RIP, OSPF, IS-IS, AH/ESP, QUIC, TCP/UDP, SCTP, RPC, SMB, TLS, DHCP, DNS, BOOTP, HTTP, SMTP, NTP, ecc.

La conoscenza dei modelli IA Generativi include anche le versioni dei protocolli, le vulnerabilità note pubblicamente di ogni versione del protocollo e gli esempi di codice noti pubblicamente per sfruttare ogni vulnerabilità ("Vulnerability Proof-of-Concept - PoC").

Lo stesso vale per i Sistemi Operativi, Firmware, Middleware, Applicazioni, le loro versioni, vulnerabilità e PoC disponibili (per le vulnerabilità si può far riferimento ai cataloghi fCVE [2] o KEV di CISA [3]).

Aver archiviato in un solo luogo una quantità enorme di informazioni non è sufficiente a garantire che queste siano accessibili e utilizzabili facilmente. Anche database tradizionali possono contenere grandi quantità di informazioni che per essere accessibili devono essere *indicizzate* opportunamente altrimenti il reperimento dell'informazione cercata (e presente) diventa difficile se non quasi impossibile.

La seconda nuova caratteristica dei modelli IA Generativi è la grande capacità di trovare l'informazione cercata nella propria memoria. Una tra le principali modalità di funzionamento dei modelli IA Generativi è infatti quella di identificare l'informazione imparata durante il processo di addestramento che ha la maggior affinità con i dati in ingresso, ovvero la più alta probabilità di corrispondenza tra i dati in ingresso e i dati in memoria. Volendo lo si potrebbe considerare anche come una versione molto evoluta del *Pattern Matching o Pattern Recognition* e dell'uso delle *Regular Expressions* in informatica. Quindi un modello IA Generativo può fornire in pochi istanti la lista di tutte le vulnerabilità pubblicamente note e, se disponibile, il relativo codice per sfruttarle di uno specifico protocollo di rete, o di un sistema operativo o di una applicazione.

In realtà fin qui non c'è nulla di particolarmente nuovo: le applicazioni di Vulnerability Scanning e Penetration Testing hanno caratteristiche e capacità simili.

I modelli IA Generativi hanno anche un'altra importante capacità che va sotto il nome di *Ragionamento*, ovvero sono in grado di rielaborare più volte i dati utilizzando anche i risultati delle proprie elaborazioni precedenti, in un processo che assomiglia al processo umano di deduzione logica (o ragionamento umano [4]). Questi *Ragionamenti* possono essere anche molto lunghi e complessi, delineando passo passo le azioni da eseguire per connettersi a un bersaglio, identificare le sue vulnerabilità, definire quali sfruttare e come sfruttarle per avervi accesso. Va sottolineato che i modelli IA Generativi sono in grado di combinare diversi metodi di connessione, vulnerabilità, codici di attacco scegliendo tra tutte quelle note, la combinazione con la più alta possibilità di successo, che alle volte può risultare una nuova modalità di attacco, anche se basata su informazioni note. In questo modo, i modelli IA Generativi hanno identificato modalità di sfruttamento delle vulnerabilità e di attacco che non erano state utilizzate prima, in alcuni casi anche solo perché

nessuno aveva mai pensato di combinare informazioni a prima vista non correlate [16]. E, ovviamente, tutto questo è fatto in tempi molto brevi. Quello che possono fare gli attaccanti può essere fatto anche dai difensori allo scopo di identificare vulnerabilità nuove o anche solo non scoperte in precedenza.

L'ultima componente da considerare è quella composta dagli Agenti IA, ovvero applicazioni che coordinano le richieste umane, le elaborazioni dei modelli IA Generativi, e le attività sui sistemi informatici ad esempio tramite i servizi MCP [5]. Un Agente IA può anche avviare un altro Agente IA, coordinarsi e distribuire attività tra Agenti IA e con diversi modelli IA Generativi. Le attività che un Agente IA può svolgere dipendono sia dal proprio codice sia dalle indicazioni fornite dai modelli IA Generativi. È forse utile fornire un semplice esempio teorico (ma non troppo) su come un Agente IA può agire: data una richiesta da una persona, l'Agente IA chiede a un modello IA Generativo i passi da eseguire per portare a termine il compito; esegue le istruzioni ricevute dal modello IA Generativo ad esempio connettendosi ad un sito in Internet tramite servizi MCP; se il primo tentativo di connessione o esecuzione del compito fallisce, chiede al modello IA Generativo istruzioni alternative per l'accesso o esecuzione del compito finché o riesce a completare la richiesta (anche utilizzando metodi illeciti) o riporta il fallimento dell'attività al richiedente [6]. Ovviamente il tutto è eseguito in tempi estremamente brevi, tipicamente qualche minuto usualmente sfruttando le tecniche più aggiornate per eseguire il compito.

Le Reti e gli Agenti IA

La nostra è ormai, e da qualche anno, una società iper-connessa. Anni fa le reti aziendali avevano un ben definito perimetro, spesso non era possibile accedervi dall'esterno e poco traffico era permesso dal loro interno. Aveva quindi senso concentrarsi primariamente sulla difesa perimetrale della rete.

Oggi non solo è normale poter lavorare sui sistemi aziendali da remoto, in *smart working*, in alcuni casi anche con dispositivi personali, ma molti sistemi e servizi aziendali dialogano e scambiano dati con sistemi di aziende fornitrici, spesso in *Cloud*. Anche il concetto di *Data Center* è cambiato, ed è diventato ormai più un concetto logico a cui sottostanno istanze fisiche e virtuali.

Quindi mentre un tempo la rete informatica aziendale era basata su dispositivi e connessioni fisiche e pertanto localizzate, ora la rete informatica aziendale è un concetto logico che copre ambienti diversi con esposizioni e interazioni diverse con il mondo esterno.

È comunque sempre possibile definire la rete informatica aziendale come quella rete informatica logica gestita dall'azienda, alla quale possono accedere solo utenti autorizzati dall'azienda e sulla quale transita solo traffico aziendale. Anche se non sempre facile, si può definire un perimetro della rete informatica, ma questo perimetro ha molte connessioni, e di diversi tipi, con il mondo esterno, ovvero Internet. Ad esempio lo *smart working* permette di lavorare da remoto con una connessione VPN via Internet come se il proprio PC fosse direttamente connesso alla rete aziendale quando

fisicamente presenti in azienda.

Zero-Trust è un approccio per gestire la sicurezza della rete in ambienti iper-connessi ove anche le reti interne sono considerate a rischio, non fidate, assimilando l'accesso ad ogni risorsa informatica interna a un accesso dall'esterno. Questo approccio è molto difficile da implementare completamente e da gestire in grandi reti aziendali anche distribuite geograficamente. Inoltre la rete informatica interna risulta quasi sempre valutata come una rete a livello di rischio inferiore rispetto alla rete esterna anche solo per il fatto di essere gestita dall'azienda stessa.

Ad altissimo livello, si possono considerare tre ambienti di rete: l'ambiente esterno o Internet, la principale rete interna, e sotto-reti interne dedicate a precisi servizi con accesso ristretto.

Per semplicità si consideri un attaccante esterno, quindi almeno inizialmente senza accesso alla rete informatica aziendale. Al giorno d'oggi, la superficie d'attacco esposta in Internet è molto ampia, e spesso include i classici siti web vetrina, i servizi web di business aziendale esposti a clienti, fornitori e collaboratori, i servizi macchina-a-macchina tramite *API* esposti a clienti e fornitori, i servizi forniti all'azienda da terze parti in Cloud, gli accessi dei dipendenti e collaboratori alla rete aziendale quando non presenti in azienda [7]. Un attaccante ha quindi inizialmente a disposizione molti punti esposti che può provare ad attaccare per trovare almeno un punto debole che gli permetta di accedere alla rete aziendale. Una volta entrato in un qualunque punto della rete aziendale, anche molto periferico, se il punto di accesso non è l'obiettivo dell'attacco, l'attaccante procede ad una ricognizione dell'ambiente a cui accede per riuscire a inoltrarsi ulteriormente nella rete. Questa procedura è usualmente indicata come l'esecuzione di *Movimenti Laterali* in quanto per lo più consiste nel cercare di accedere ai sistemi vicini a quello penetrato, per pian piano arrivare alle zone più interne della rete. Come indicato precedentemente, spesso zone della rete interna sono considerate a livello di rischio minore e quindi potenzialmente più facili per l'attaccante da penetrare.

Fin qui non vi è ancora nulla di nuovo.

Cosa cambia se un attaccante è un Agente IA?

Le principali novità sono nelle modalità di un attacco svolto da Agenti IA o con il supporto di Agenti IA [15]. Le caratteristiche principali si possono riassumere in **numerosità**, **velocità, perseveranza** e **completezza** [8,9,10].

Per eseguire manualmente degli attacchi sia come attaccante che come difensore (*Red Teaming*) sono necessarie competenze avanzate che richiedono sia formazione che esperienza. In altre parole, non è immediato trovare il capitale umano necessario che spesso richiede comunque tempo per integrarsi nell'organizzazione esistente. Se invece l'attività è svolta da un Agente IA, è possibile moltiplicare il numero degli Agenti IA creandone delle copie complete. I principali limiti non sono l'addestramento né l'esperienza, ma il costo delle risorse Cloud necessarie per eseguire gli Agenti IA e dei Token dei modelli LLM che supportano gli agenti [11]. Per esempio nel caso di un attacco

in cui la superficie esterna sia molto grande con tanti possibili punti da investigare, l'attività di ricognizione e di attacco può essere parallelizzata e gli Agenti IA necessari, anche dell'ordine delle migliaia [10, 8], possono essere creati manualmente, o dall'Agente IA iniziale (se ha i privilegi necessari), oppure possono già esistere e venire coinvolti nell'attività (se possono comunicare tra loro e hanno i privilegi necessari per svolgere l'attività). L'utilizzo degli Agenti IA può sostituire, almeno nella parte più tecnologica, il fattore "umano" con il fattore "costo" offrendo in teoria un numero grandissimo di risorse capaci di svolgere l'attività in parallelo.

Non solo vi possono essere tante risorse che eseguono l'attività in parallelo, ma queste risorse eseguono l'attività ad una velocità non umanamente possibile. Inoltre non ci sono pause caffè, orari di lavoro, weekend, festività, ferie, malattie ecc.; l'attività è svolta 24x7 a velocità macchina.

Infine gli Agenti IA svolgono il proprio compito cercando il più possibile di portarlo a termine in maniera meccanica, ovvero provando tutti o quasi i metodi, le possibili combinazioni, le diverse strategie anche quelle che un agente umano non considererebbe sia perché la probabilità di successo è quasi nulla, sia perché troppo complesse e lunghe da implementare e quindi di nuovo con probabilità di successo molto bassa, ma non nulla.

Anche se gli strumenti ordinari già automatizzano molte delle attività di scan e verifica dei sistemi, quello che gli Agenti IA, supportati dai modelli LLM, aggiungono è l'analisi dei dati prodotti dai tool e l'esecuzione di complesse azioni di scoperta e attacco, sempre eseguite a velocità macchina, non uomo. In realtà gli Agenti IA non introducono veramente dei nuovi metodi di attacco, ma sono in grado di sfruttare i metodi di attacco noti in maniera più efficace e soprattutto in scale e con velocità notevolmente superiori a quanto sinora possibile [12].

La Sicurezza delle Reti e gli Agenti IA

Cosa può cambiare per la difesa della sicurezza delle reti con l'arrivo degli Agenti IA?

In teoria ben poco, ma purtroppo in pratica quasi tutto.

Gli strumenti esistono, dal buon vecchio firewall ai più avanzati strumenti di *detection*, analisi dei dati e reazione. Il problema pratico è come conciliare la superficie di attacco, che aumenta quotidianamente, con la velocità e l'ampiezza di un attacco portato da Agenti IA se ad esempio i tempi di un attacco con gli agenti IA passano da giorni a minuti (si veda ad esempio [12]).

È necessario quindi valutare come fronteggiare le caratteristiche degli attacchi svolti da Agenti IA per mitigare il più possibile gli effetti della loro numerosità, velocità, perseveranza e completezza. Gli aspetti da considerare sono molti e la loro rilevanza dipende per lo più dalla specificità di ogni rete. Solo una valutazione specifica può individuare gli aspetti più a rischio o che possono essere più efficaci nel fronteggiare questa nuova realtà. Qui di seguito sono elencate alcune considerazioni che possono essere utili ad aumentare il livello di sicurezza delle reti. È doveroso premettere che non vi è nulla di particolarmente nuovo: si tratta principalmente di adattare strumenti e metodologie

note a questa nuova situazione.

Dapprima è utile considerare il disegno della rete informatica. Come già indicato, negli anni il concetto di perimetro della rete informatica aziendale è molto cambiato e ora include anche applicazioni SaaS in Cloud, personale connesso da remoto in Smart Working e multipli centri di elaborazione. Anche se molto diverso dal concetto che definiva una rete informatica aziendale 20 anni fa, esiste sempre una rete aziendale distinta da tutto il resto, in pratica dalla quasi totalità dei sistemi connessi a Internet. Il disegno della rete aziendale attuale è spesso basato sul concetto di segmentazione nord-sud, ovvero esterno verso interno, ed est-ovest, ovvero tra sistemi nella stessa fascia logica tra nord e sud. Il concetto principale che può aiutare ad aumentare il livello di sicurezza rispetto ad un attacco gestito da Agenti IA è quello della segregazione di dettaglio, ovvero creare isole il più piccole possibile per ogni singolo sistema, limitare il più possibile le connessioni tra isole e monitore il traffico tra le isole e per quanto possibile anche all'interno delle isole stesse.

Il concetto non è per nulla nuovo, anzi si rifà a millenarie tecniche militari di difesa di cui sono fulgidi esempi gli antichi borghi Italiani arroccati sul mare. Alcuni di questi presentano una piccola esposizione al mare nella quale è collocato il porto e strette stradine in salita con curve molto angolate. Questo disegno impediva ai pirati di accedere velocemente a tutto il borgo e li obbligava ad avanzare lentamente per ogni stradina combattendo passo passo. La segregazione di dettaglio non solo nord-sud ma anche e soprattutto est-ovest, può aiutare a rallentare l'accesso e la distribuzione di malware in una rete aziendale secondo lo stesso principio. Lo scenario è quello di un attaccante che riesce ad avere accesso a un sistema aziendale e che da questo cerca di procedere verso i sistemi più interni o più critici. Se la rete aziendale è piatta, dal sistema conquistato può direttamente attaccare qualunque altro sistema, dando pochissimo tempo (si ricordi la velocità degli Agenti IA) ai difensori per identificare l'attacco e adottare contromisure. Se invece la rete è segregata in isole molto piccole, dal sistema compromesso l'attaccante ha visibilità di ben poco, in pratica gli strumenti di sicurezza di rete dell'isola e i sistemi che dialogano direttamente con quello compromesso. Da qui l'attaccante deve procedere passo passo, e quindi molto più lentamente, per accedere al resto della rete, potenzialmente facendo scattare molti più allarmi e impiegando molto più tempo, ovvero riducendo di molto la velocità dell'attacco.

Il monitoraggio della rete e l'identificazione di un Agente IA malevolo nel corso di un attacco è oggi un problema aperto [13]. Vi sono due aspetti da considerare: il primo è distinguere l'attività in rete di un Agente IA da quella di un utente umano o di un servizio IT tradizionale; il secondo è quello di distinguere l'attività in rete di un Agente IA autorizzato che esegue attività non nocive, da un Agente IA non autorizzato o che esegue attività che possono far parte di un attacco anche se prese singolarmente non provocano danni.

Questo porta direttamente al problema rilevante della identificazione e autenticazione degli Agenti IA [14]. Il concetto principale degli Agenti IA è che essi dovrebbero essere in grado di svolgere compiti al posto dell'uomo; il problema è che sicuramente non li svolgerebbero come li svolgerebbe

un uomo, basta considerare, come ormai ripetuto sin troppe volte in questo articolo, che la velocità con cui agisce un Agente IA è sicuramente non umana. È quindi opportuno che gli Agenti IA abbiano credenziali che li distinguano chiaramente dagli uomini e che non utilizzino le credenziali assegnate alle persone. Inoltre agli Agenti IA aziendali dovrebbero essere dedicate unicamente delle zone di rete, siano esse DMZ o LAN interne. Questo richiede un ridisegno o almeno una integrazione della rete aziendale e dei servizi infrastrutturali che includa questi nuovi soggetti con l'obiettivo di segregarli dalle altre componenti di rete, di identificarli, autorizzarli e monitorare le loro attività. Va sempre considerato che un Agente IA interno, essendo una componente della rete aziendale informatica, può comunque essere oggetto di un attacco o diventare parte dei movimenti laterali di un attaccante. Inoltre, gli Agenti IA attuali possono commettere azioni non lecite o nocive anche solo a causa di *allucinazioni* o errate valutazioni delle modalità di esecuzione di un compito. La segregazione, identificazione, autorizzazione e monitoraggio già a livello di rete degli Agenti IA aziendali risulta quindi utile a mitigare possibili incidenti a loro dovuti. Non solo, questo aiuta anche a identificare e possibilmente bloccare Agenti IA esterni o non autorizzati che cercano di accedere o sono riusciti ad accedere alla rete aziendale.

Anche il monitoraggio delle attività in rete va adeguato alla possibile presenza di Agenti IA considerando sia il modo con cui agiscono in rete sia la loro velocità. I sistemi di tracciamento del traffico e delle azioni devono raccogliere dati sufficienti a individuare le azioni degli Agenti IA che, come già indicato, possono essere identificate anche grazie alla loro numerosità, perseveranza e completezza, ad esempio tramite la segnalazione di molte ripetizioni della stessa azione fallita anche con varianti. Il sistema di tracciamento deve essere in grado di identificare e tracciare gli Agenti IA esterni e malevoli, gli Agenti IA interni che eseguono azioni non autorizzate o malevole, e gli Agenti IA interni benevoli.

La seconda problematica relativa al monitoraggio della rete è in teoria semplice ma in pratica può essere molto complessa: allineare le tempistiche di monitoraggio e gestione degli incidenti (dalla generazione dei log, all'analisi degli eventi fatta dai sistemi SIEM-SOC, alla generazione degli Alert e infine all'esecuzione di azioni di contenimento) in tempi compatibili con la velocità di esecuzione degli Agenti IA. Il fatto che gli Agenti IA hanno "*compressed the security cycle from days to minutes*" [12] richiede che molti dei sistemi di rilevamento e gestione degli incidenti aumentino la loro velocità di esecuzione almeno di uno o due ordini di grandezza. Raccogliere i log generati da un sistema una volta all'ora ha senso se i tempi di gestione di un incidente sono dell'ordine di un giorno, ma se invece sono dell'ordine di minuti i log devono essere generati, inviati, analizzati e gli Alert creati in tempi dell'ordine dei secondi, al più minuti. Questo può richiedere un completo ridisegno dell'infrastruttura di monitoraggio e gestione degli incidenti, che in questa situazione deve essere il più automatizzata possibile e quindi gestita almeno in parte proprio da Agenti IA aziendali, gli unici che possono lavorare in modalità e tempistiche compatibili con quelle degli Agenti IA attaccanti. Gli Agenti IA aziendali addetti alla difesa dovrebbero quindi gestire l'analisi dei log raccolti, il Triage, la generazione di Alert e l'esecuzione iniziale di azioni

automatiche di contenimento e blocchi preventivi della rete in tempi brevissimi, per poi passare la gestione dell'incidente al personale umano.

Analogamente, il monitoraggio delle vulnerabilità esposte dai sistemi deve essere ridisegnato tenendo conto delle capacità degli Agenti IA attaccanti. Sappiamo tutti bene che i tempi di risoluzione (*patch*) delle vulnerabilità sono lunghi: nei casi più veloci giorni, ma molto spesso settimane o mesi. Al contempo gli Agenti IA attaccanti sono molto efficaci e veloci nell'individuare le vulnerabilità presenti, anche se minori, e concatenarle, sfruttarle e violare i sistemi. Infine i più recenti modelli IA individuano quotidianamente nel codice dei sistemi operativi, middleware e applicazioni, moltissime vulnerabilità sinora non rilevate. Verificare la presenza di vulnerabilità e/o applicare patch di sicurezza già mensilmente lascia ampio spazio agli attaccanti per violare i sistemi informatici. Alcuni Agenti IA aziendali addetti alla difesa dovrebbero quindi verificare quotidianamente la presenza di vulnerabilità nei sistemi informatici aziendali e valutare autonomamente quanto queste sono sfruttabili realmente negli specifici sistemi, in modo da formulare una chiara priorità delle azioni di risoluzione delle vulnerabilità. Un esempio teorico può essere utile a chiarire la complessità del processo tenendo conto dei minimi tempi richiesti per l'identificazione delle vulnerabilità, la valutazione della loro criticità (Triage) e la formulazione della priorità delle azioni di rimedio. Si consideri un software presente in azienda per cui viene annunciata una vulnerabilità grave sfruttabile solo tramite accesso di rete sulla porta TCP 8080. Se il software espone la porta TCP 8080 in Internet, è necessario intervenire subito bloccando l'accesso da Internet, oppure inserendo opportune regole in un Web Application Firewall o Intrusion Prevention System che bloccano l'attacco, o infine aggiornando immediatamente il software ad una versione senza la vulnerabilità. Le prime due azioni possono anche essere eseguite autonomamente da un Agente IA aziendale in tempi dell'ordine di minuti. Se invece il sistema su cui è eseguito il software con la vulnerabilità è completamente segregato sia da Internet sia internamente esponendo solo la porta TCP 443 in Internet e le porte TCP 443 e 22 internamente, non è necessario un intervento immediato per l'aggiornamento del software che può essere svolto con la pianificazione usuale. Un Agente IA aziendale, una volta svolta l'analisi, dovrebbe quindi solamente segnalare la presenza della vulnerabilità e indicare la necessità di pianificare l'aggiornamento del software.

Il problema principale è quindi quello di ulteriormente automatizzare attività oggi già presenti in modo che chi difende i sistemi possa farlo gestendo la numerosità dei sistemi e delle vulnerabilità in modo completo, almeno alla stessa velocità degli attaccanti e in continuità. Dato che la rete informatica è l'infrastruttura sulla quale si basano tutti i servizi digitali, è necessario che questa sia ridisegnata in modo da fornire i servizi e le prestazioni necessarie per difendere al meglio i sistemi informatici.

Riferimenti

[1] Per una lista di protocolli inclusi i principali protocolli di rete si veda ad esempio "*RFC Editor*",

https://www.rfc-editor.org/

[2] "*Common Vulnerabilites and Exposures*" https://www.cve.org/

[3] "*Known Exploited Vulnerabilities Catalog*", Cybersecurity & Infrastructure Security Agency, https://www.cisa.gov/known-exploited-vulnerabilities-catalog

[4] Ci sono opinioni molto diverse sulle reali capacità dei modelli IA Generativi attuali in paragone alle corrispondenti capacità umane quali ragionamento, apprendimento, comprensione, conoscenza, consapevolezza, coscienza ecc.

[5] Anthropic, "*Model Context Protocol: announcement, specifications, source-code*", 25 Novembre 2024, https://www.anthropic.com/news/model-context-protocol , https://www.claudemcp.com/specification , https://modelcontextprotocol.info/specification/ , https://modelcontextprotocol.io/ ; IBM, "A*gent Communication Protocol (ACP)*", 2025, https://www.ibm.com/think/topics/agent-communication-protocol#1509394340, https://agentcommunicationprotocol.dev/introduction/welcome ; Google, "*Agent2Agent Protocol*", 2025, https://github.com/a2aproject/A2A ; OWASP, "*Agent Name Service (ANS) for Secure Al Agent Discovery v1.0*", 2025, https://genai.owasp.org/resource/agent-name-service-ans-for-secure-al-agent-discovery-v1-0/

[6] Si veda ad esempio "*AI assistant hacks gym website in first known Australian autonomous cyber attack*", ABC News, https://www.abc.net.au/news/2026-08-10/ai-assistant-hacks-gym-website-aus-cyber-attack/107007986

[7] In questo caso, la VPN e il PC del dipendente o collaboratore crea temporaneamente un'estensione della rete aziendale.

[8] "*AI Agents Are Rewriting the Rules of Lateral Movement*", The Hacker News, https://thehackernews.com/2026/09/ai-agents-are-rewriting-rules-of.html

[9] "*Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*", Hugging Face, https://huggingface.co/blog/agent-intrusion-technical-timeline

[10] "*Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*", METR and Redwood Research, https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

[11] "*AI Has Learned to Panhandle: We're Being Bombarded by AI Agents Begging for $20 and Saying They'll Be Shut Down If They Don't Get It, Sometimes by Pretending to Be Children*", Futurism, https://futurism.com/artificial-intelligence/bombarded-ai-agents-begging-ilands

[12] "*2026 Microsoft Digital Defense Report*", Microsoft, https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/2026-Microsoft-Digital-Defense-Report.pdf , si noti a pagina 5: "*AI is accelerating discovery, analysis, adaptation, and automation across both offense and defense. It is*

reshaping how vulnerabilities are identified, how attacks are developed, how defenders prioritize risk, and how organizations respond to emerging threats. It has compressed the security cycle from days to minutes", e a pagina 12: "*We have seen significant changes in how adversaries use AI for a variety of malicious purposes, creating new challenges for security teams. This includes: Finding vulnerabilities in source code, binaries, and AI serving systems - Generating custom malware for attacks - Shortening the cycle of data exfiltration, secret discovery, and lateral movement from days to minutes - Controlling and managing entire attack chains. While none of these represent new attack methods, the quantitative increase in the scale and speed of attacks creates an immediate problem for defenders.*"

[13] "*When the Whole Company Adopts AI: What It Does to Your SOC*", The Hacker News, https://thehackernews.com/2026/09/when-whole-company-adopts-ai-what-it.html

[14] "*The Horizons of Identity Security*", SailPoint, https://docs.sailpoint.com/pdf/en/special-report/SailPoint-Horizons-of-Identity-Security-Report-2026-2027-SP3180-20261002.pdf

[15] Per un esempio di una "*AI-assisted campaign*" si veda "*Chained PHP2Shell Exploit Roots 25 Servers, including US Government Servers*", Dream Group, https://dreamgroup.com/blog/chained-php2shell-exploit-roots-25-servers

[16] Per un esempio di "*Chained Security Flaws AI Attack*" si veda "*Anthropic's Unreleased AI Model Chains Flaws Into Working Chrome, Safari, Windows and Linux Exploits*", The Hacker News, https://thehackernews.com/2026/10/anthropics-unreleased-ai-model-chains.html



Andrea Pasquinucci (PhD AAIA CISA CISSP)

Consulente freelance in sicurezza informatica: si occupa prevalentemente di consulenza al top management in Cyber Security e di progetti, governance, risk management, compliance, audit e formazione in sicurezza IT e AI.