

AI offensiva e difensiva: strategie di cybersecurity nel 2025



L'intelligenza artificiale sta trasformando radicalmente il panorama della cybersecurity, operando simultaneamente come catalizzatore per sofisticate minacce offensive e come pilastro fondamentale per la difesa proattiva.

Il presente articolo esamina l'evoluzione duale dell'AI nella sicurezza informatica, analizzando in particolare l'emergenza di campagne di Business Email Compromise (BEC) iper-personalizzate abilitate da modelli linguistici avanzati e l'implementazione di sistemi di threat detection basati su machine learning per l'identificazione di anomalie in infrastrutture di rete complesse. L'analisi evidenzia come l'anno 2025 rappresenti un punto di svolta critico nell'utilizzo dell'AI per scopi malevoli, con attacchi sempre più sofisticati che sfruttano capacità generative avanzate, mentre parallelamente si assiste all'evoluzione di framework difensivi intelligenti che integrano tecniche di behavioral analytics e anomaly detection per la mitigazione delle minacce in tempo reale.

Introduzione

Il paradigma della sicurezza informatica sta attraversando una trasformazione epocale guidata dall'integrazione pervasiva dell'intelligenza artificiale. L'anno 2025 è previsto essere caratterizzato da un incremento massiccio nell'utilizzo malevolo dell'AI, con esperti che stimano un cambiamento fondamentale nel modo in cui l'intelligenza artificiale influenza la cybersecurity. Questa dualità tecnologica si manifesta attraverso due vettori principali: l'AI offensiva,

che alimenta attacchi di precisione scalabili e adaptive, e l'AI difensiva, che abilita capacità di detection e response automatizzate per contrastare minacce emergenti.

Gli attacchi di Business Email Compromise continuano a rappresentare una delle minacce più persistenti e distruttive, con perdite che superano i 2,95 miliardi di dollari, mentre l'integrazione di tecnologie generative AI amplifica significativamente il potenziale di danno di tali campagne. Parallelamente, i sistemi di cybersecurity basati su AI e machine learning offrono la capacità di rilevare minacce precedentemente sconosciute attraverso tecniche avanzate di anomaly detection e pattern recognition.

L'evoluzione dell'AI offensiva

Campagne BEC iper-personalizzate

L'evoluzione dei modelli linguistici di grandi dimensioni (LLM) ha rivoluzionato le capacità di ingegneria sociale dei threat actor. L'AI generativa ha aumentato significativamente il rischio di BEC, e le difese tradizionali di cybersecurity faticano a tenere il passo con la crescente velocità, scala e sofisticazione degli attacchi. I cybercriminali sfruttano ora capacità linguistiche avanzate per creare contenuti di phishing che bypassano i filtri antispam tradizionali attraverso diverse strategie tecniche:

Generazione di contenuti adattivi: I sistemi AI-powered possono analizzare profili pubblici, cronologie di comunicazione e pattern linguistici per creare messaggi che mimano perfettamente lo stile comunicativo di individui specifici. Le campagne di phishing AI-driven generano messaggi perfettamente personalizzati in tempo reale, imitando i pattern del linguaggio umano e personalizzando il contenuto basandosi su dati pubblicamente disponibili.

Evasione automatizzata dei filtri: Gli algoritmi di machine

learning possono essere addestrati per identificare e aggirare le signature di detection degli security gateway, modificando dinamicamente il contenuto per evitare la classificazione come spam o phishing.

Scalabilità operativa: I chatbot alimentati da LLM possono rimuovere una delle limitazioni storiche dell'ingegneria sociale generica, permettendo ai cybercriminali tecnologicamente avanzati di migliorare la credibilità degli schemi di advance fee.

Malware autoevolutivo

Gli attaccanti stanno implementando malware auto-apprendente che si adatta agli ambienti, evita la detection e altera il suo comportamento dinamicamente per bypassare le misure di sicurezza tradizionali. Questa categoria di minacce presenta caratteristiche tecniche avanzate:

Adaptive behavior modification: I payload malevoli integrano algoritmi di reinforcement learning per modificare le proprie tattiche di persistenza e lateral movement basandosi sull'ambiente target.

Morphic code generation: Utilizzo di tecniche generative per produrre varianti di codice funzionalmente equivalenti ma morfologicamente diverse, compromettendo l'efficacia della signature-based detection.

Environment-aware execution: Implementazione di logiche di decisione che valutano l'architettura del sistema target, la presenza di strumenti di security e i pattern di utilizzo per ottimizzare le strategie di attack.

Deepfake e manipolazione mediatica

I deepfake rappresenteranno una minaccia reale nel 2024-2025, con implicazioni critiche per l'autenticazione e la verifica

dell'identità. Le tecnologie di sintesi vocale e video raggiungono livelli di fedeltà che rendono difficile la distinzione tra contenuti autentici e artificiali, abilitando:

- **Voice cloning attacks:** Replicazione precisa di caratteristiche vocali per compromise di sistemi di autenticazione biometrica
- **Video impersonation:** Creazione di contenuti video falsi per campagne di disinformazione o social engineering avanzato
- **Real-time generation:** Capacità di generazione di contenuti deepfake in tempo reale per interazioni dirette

Strategie di AI difensiva

Behavioral analytics e anomaly detection

I sistemi basati su AI e ML hanno la capacità significativa di rilevare minacce precedentemente sconosciute (attacchi zero-day) attraverso anomaly detection e pattern recognition. L'implementazione di queste tecnologie in ambienti enterprise richiede architetture sofisticate:

User and Entity Behavior Analytics (UEBA): Sviluppo di baseline comportamentali per utenti, dispositivi e applicazioni attraverso l'analisi di pattern di accesso, timing e sequence di operazioni. Il machine learning è utilizzato per stabilire baseline di comportamento normale per utenti e sistemi, successivamente segnalando attività inusuali that potrebbero indicare minacce da parte di threat actor esterni o insider malevoli.

Network traffic analysis: L'anomaly detection è fondamentale nel traffic di rete per identificare attacchi DDoS, tentativi di intrusione di rete e altri pattern sospetti che indicano accessi non autorizzati. Gli algoritmi utilizzati includono:

- **Isolation Forest:** Algoritmo specializzato per identificare outlier in spazi ad alta dimensionalità
- **One-Class SVM:** Apprendimento di confini intorno ai dati normali per identificare anomalie
- **DBSCAN clustering:** Identificazione di regioni sparse come anomalie in dataset con densità variabile

Real-time correlation: Il machine learning incorpora la correlazione di eventi, l'anomaly detection e la riduzione di falsi positivi attraverso l'apprendimento da dati storici di eventi di sicurezza.

Threat intelligence automatizzata

L'AI trasforma il threat intelligence da processo reattivo a capacità predittiva. Gli Indicators of Attack (IOA) basati su AI rilevano proattivamente minacce emergenti indipendentemente dal malware o dagli strumenti utilizzati, operando in modo asincrono rispetto ai modelli on-sensor. Le implementazioni avanzate includono:

Predictive threat modeling: Utilizzo di algoritmi di machine learning per analizzare pattern di attacco storici e predire tattiche, tecniche e procedure future dei threat actor.

Automated indicator generation: Sistemi che generano automaticamente IoC (Indicators of Compromise) e IoA attraverso l'analisi di campioni di malware e behavioral patterns.

Attribution analysis: Tecniche di clustering e classification per correlazioni tra campagne di attacco e threat group specifici.

Incident response automatizzata

I sistemi di sicurezza AI-driven possono rispondere autonomamente agli attacchi, contenendo le breach più

velocemente di quanto potrebbero fare team umani. L'orchestrazione automatizzata include:

Dynamic containment: Isolamento automatico di asset compromessi basato su risk scoring e impact assessment in tempo reale.

Adaptive countermeasures: Implementazione di controlli di sicurezza personalizzati basati sulle caratteristiche specifiche dell'attacco rilevato.

Remediation workflows: Esecuzione automatizzata di procedure di ripristino e hardening post-incident.

Il problema della trasparenza algoritmica

La sfida del black box

I sistemi di Intrusion Detection tradizionali spesso si basano su algoritmi complessi di machine learning che mancano di trasparenza nonostante la loro alta accuratezza, creando un effetto "black box" che può ostacolare la comprensione dei processi decisionali da parte degli analisti. Questa opacità presenta sfide critiche:

Trust and accountability: Se il processo decisionale di un sistema AI non è trasparente, diventa difficile fidarsi dei suoi output, specialmente in applicazioni critiche come healthcare, finance e cybersecurity.

Regulatory compliance: Alcune normative, come l'EU AI Act e il California Consumer Privacy Act (CCPA), stabiliscono regole su come le organizzazioni possono utilizzare dati personali sensibili in strumenti decisionali basati su AI.

Vulnerability identification: L'opacità dei modelli black box può nascondere vulnerabilità di cybersecurity, bias,

violazioni della privacy e altri problemi.

Approcci all'explainable AI (XAI)

L'AI spiegabile (XAI) offre una soluzione promettente fornendo interpretabilità e trasparenza, permettendo ai professionisti della sicurezza di comprendere meglio, fidarsi e ottimizzare i modelli IDS. Le tecniche implementate includono:

Model-agnostic explanations:

- **LIME (Local Interpretable Model-agnostic Explanations):** LIME sonda essenzialmente il modello con input casuali per vedere come ogni piccolo cambiamento influisce sull'output complessivo, rivelando i key pattern e le relazioni "osservate" dall'algoritmo
- **SHAP (SHapley Additive exPlanations):** Framework per spiegare l'output di qualsiasi modello di machine learning

Counterfactual reasoning: L'analisi controfattuale presenta una via per generare scenari "what if" e possibili mondi alternativi che producono risultati diversi.

Visualization techniques: Sviluppo di heatmap, decision-path chart e altre rappresentazioni grafiche per rendere comprensibili i processi decisionali degli algoritmi.

Hybrid interpretability models

I modelli ibridi che combinano algoritmi interpretabili, come gli alberi decisionali, con sistemi più complessi come le deep neural network offrono un equilibrio promettente tra accuratezza e trasparenza. Questi approcci permettono:

- **Layered explanation:** Diversi livelli di spiegazione per differenti stakeholder (tecnici vs management)
- **Context-aware interpretation:** Spiegazioni adattate al

contesto operativo specifico

- **Performance optimization:** Mantenimento delle performance predittive con miglioramento dell'interpretabilità

Framework normativi e standard emergenti

NIST Cybersecurity and AI Framework

NIST sta sviluppando il Cyber AI Profile basato sul suo landmark Cybersecurity Framework, con un rilascio previsto entro i prossimi sei mesi. Questo framework mira a:

Risk taxonomy development: Sviluppo di una tassonomia chiara tra la community AI e i professionisti della cybersecurity per facilitare la discussione sull'intersezione dei rischi.

Control overlay implementation: NIST intende sfruttare completamente i Framework di cybersecurity esistenti e le linee guida tecniche per sviluppare una serie di overlay di controlli di cybersecurity focalizzati sui casi d'uso e informati sulle minacce.

Community profile development: I Community Profile forniscono un modo per le comunità di descrivere un punto di vista di consenso sulla gestione del rischio di cybersecurity.

AI Risk Management Framework

Il NIST AI Risk Management Framework (AI RMF) è inteso per uso volontario e per migliorare la capacità di incorporare considerazioni di trustworthiness nella progettazione, sviluppo, uso e valutazione di prodotti, servizi e sistemi AI. Il framework enfatizza:

Trustworthy AI characteristics: Validità, affidabilità, sicurezza, security e resilience come caratteristiche fondamentali.

Socio-technical approach: NIST riconosce che i rischi AI si estendono oltre le considerazioni tecniche per comprendere complesse implicazioni sociali, legali ed etiche.

Adaptive implementation: Il framework è progettato per essere adattabile, applicabile sia a piccole startup che a grandi multinazionali, a sistemi AI a basso rischio o ad alto rischio.

Evoluzione della workforce e competenze

NIST sta incorporando l'AI nel NICE Framework attraverso due modi principali: una nuova Competency Area per la Sicurezza dell'Intelligenza Artificiale e l'aggiornamento dei Work Role esistenti per includere considerazioni AI. Questo include:

AI Security Competency Area: Descrizione delle conoscenze e competenze fondamentali necessarie per comprendere l'intersezione tra AI e cybersecurity.

Enhanced work roles: Integrazione di statement di Task, Knowledge e Skill (TKS) relativi all'AI nei ruoli esistenti e nuovi.

Professional development: Necessità di sviluppare nuovi ruoli professionali come "AI security ethicist" e "machine learning defense specialist".

Implicazioni strategiche e raccomandazioni

Architetture di difesa stratificata

La complessità delle minacce AI-enabled richiede approcci di difesa multilivello:

Zero Trust Architecture: L'adozione di architetture zero trust creerà un cambiamento sismico nella cultura delle organizzazioni, richiedendo stretta collaborazione tra IT, team di sicurezza e business unit.

AI-augmented SIEM: L'AI deve essere al centro dello stack di cybersecurity, non semplicemente un overlay di agenti aggiuntivi sui endpoint.

Continuous adaptive assessment: Implementazione di sistemi di valutazione continua che si adattano alle tattiche emergenti dei threat actor.

Investment priorities

L'industria healthcare dovrebbe spendere 125 miliardi di dollari in cybersecurity dal 2020 al 2025, con previsioni di crescita della spesa in tutti i settori industriali critici. Le priorità di investimento dovrebbero concentrarsi su:

Behavioral analytics platforms: Sistemi capaci di apprendimento continuo e adattamento ai pattern comportamentali emergenti.

Automated response capabilities: Workflow automatizzati e detection AI-enabled di deviazioni dalle installazioni baseline.

Threat intelligence automation: Piattaforme che integrano multiple fonti di intelligence per generazione automatizzata di IoC e IoA.

Preparazione organizzativa

Un terzo delle organizzazioni non ha una strategia documentata per difendersi dalle minacce AI e generative AI, con l'89% dei CISO e senior IT leader che credono che le minacce AI-powered stiano appena iniziando. Le organizzazioni devono:

Develop AI governance frameworks: Stabilire policy e procedure

per l'uso responsabile dell'AI in contesti di sicurezza.

Enhance skill development: Investire in training per team di sicurezza che possano ottimizzare processi attraverso predictive analytics.

Implement continuous monitoring: Sviluppare capacità di monitoraggio 24/7 per minacce AI-enabled in evoluzione rapida.

Conclusioni e prospettive future

L'intelligenza artificiale rappresenta un paradigm shift fondamentale nella cybersecurity, operando simultaneamente come acceleratore di minacce sofisticate e come enabler di difese adaptive. L'analisi delle tendenze emergenti evidenzia come il 2025 segni un punto di non ritorno nell'utilizzo dell'AI per scopi offensivi, con previsioni che indicano il 40% di tutti gli attacchi informatici saranno AI-driven.

Le organizzazioni che adotteranno proattivamente strategie di AI difensiva, integrandole con framework di explainable AI per mantenere trasparenza e accountability, saranno meglio posizionate per contrastare la crescente sofisticazione delle minacce. Solo strategie multilayered, defense-in-depth possono contrastare la minaccia BEC potenziata dall'AI.

La convergenza di tecnologie emergenti come quantum computing, edge AI e 5G introdurrà ulteriori vettori di attacco e opportunità difensive, richiedendo un approccio sistematico all'evoluzione delle capacità di cybersecurity. L'investimento in competenze specializzate, l'adozione di standard normativi emergenti e l'implementazione di architetture di sicurezza AI-native rappresentano imperativi strategici per mantenere posture di sicurezza efficaci nell'era dell'AI.

La sfida principale rimane il bilanciamento tra automazione e controllo umano, garantendo che i sistemi AI-powered mantengano livelli appropriati di trasparenza e accountability

mentre operano alla velocità e scala richieste per contrastare minacce che evolvono rapidamente.

Fonti

Cyber Strategy Institute. (2025). *A Deep Dive into 2024 and a Glimpse into 2025 BEC and Phishing Attacks*.
<https://cyberstrategyinstitute.com/a-deep-dive-into-2024-and-a-glimpse-into-2025-bec-and-phishing-attacks-the-relentless-evolution-of-business-email-compromise-bec-and-phishing/>

SecurityWeek. (2025). *Cyber Insights 2025: Artificial Intelligence*.
<https://www.securityweek.com/cyber-insights-2025-artificial-intelligence/>

Darktrace. (2024). *Business Email Compromise (BEC) in the Age of AI*.
<https://www.darktrace.com/blog/business-email-compromise-bec-in-the-age-of-ai>

Mohamed, N. (2025). *Cutting-edge advances in AI and ML for cybersecurity: a comprehensive review of emerging trends and future directions*. Knowledge and Information Systems.
<https://www.tandfonline.com/doi/full/10.1080/23311975.2025.2518496>

Capitol Technology University. (2025). *Emerging Threats to Critical Infrastructure: AI Driven Cybersecurity Trends for 2025*.
<https://www.captechu.edu/blog/ai-driven-cybersecurity-trends-2025>

Bitdefender. (2024). *2024 Cybersecurity Predictions for AI: A Technical Deep Dive*.
<https://www.bitdefender.com/en-us/blog/businessinsights/2024-cybersecurity-predictions-for-ai-a-technical-deep-dive>

SC Media. (2025). *2025 Forecast: AI to supercharge attacks,*

quantum threats grow, SaaS security woes.
<https://www.scworld.com/feature/cybersecurity-threats-continue-to-evolve-in-2025-driven-by-ai>

Abusix. (2025). *AI-Powered Cyber Threats in 2025: How Attackers Use Machine Learning.*
<https://abusix.com/blog/the-rise-of-ai-powered-cyber-threats-in-2025-how-attackers-are-weaponizing-machine-learning/>

VentureBeat. (2025). *Outsmarting AI-powered cyber attacks: A 2025 playbook for real-time endpoint defense.*
<https://venturebeat.com/security/outsmarting-ai-powered-cyber-attacks-endpoint-defense-2025/>

Cobalt. (2025). *Top Cybersecurity Statistics for 2025.*
<https://www.cobalt.io/blog/top-cybersecurity-statistics-2025>

Springer. (2025). *Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms.*
<https://link.springer.com/article/10.1007/s10115-025-02429-y>

SentinelOne. (2025). *AI Threat Detection: Leverage AI to Detect Security Threats.*
<https://www.sentinelone.com/cybersecurity-101/data-and-ai/ai-threat-detection/>

Fidelis Security. (2025). *Anomaly Detection: How Algorithms Spot Cyber Threats.*
<https://fidelissecurity.com/threatgeek/network-security/anomaly-detection-algorithms/>

SailPoint. *Machine learning (ML) in cybersecurity.*
<https://www.sailpoint.com/identity-library/how-ai-and-machine-learning-are-improving-cybersecurity>

ResearchGate. (2024). *Machine Learning for Cybersecurity: Threat Detection, Prevention, and Response.*
https://www.researchgate.net/publication/377990654_Machine_Lea

rning_for_Cybersecurity_Threat_Detection_Prevention_and_Response

ScienceDirect. *Network anomaly detection and security defense technology based on machine learning: A review*.
<https://www.sciencedirect.com/science/article/abs/pii/S0045790624005081>

BMC Medical Informatics and Decision Making. *Anomaly-based threat detection in smart health using machine learning*.
<https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02760-4>

CrowdStrike. (2025). *Machine Learning (ML) in Cybersecurity: Use Cases*.
<https://www.crowdstrike.com/en-us/cybersecurity-101/artificial-intelligence/machine-learning/>

Nextgov/FCW. (2025). *Forthcoming NIST profile to address growing AI-cyber challenges*.
<https://www.nextgov.com/cybersecurity/2025/04/forthcoming-nist-profile-address-growing-ai-cyber-challenges/404887/>

NIST. (2025). *The Impact of Artificial Intelligence on the Cybersecurity Workforce*.
<https://www.nist.gov/blogs/cybersecurity-insights/impact-artificial-intelligence-cybersecurity-workforce>

NIST. (2025). *Cybersecurity and AI: Integrating and Building on Existing NIST Guidelines*.
<https://www.nist.gov/blogs/cybersecurity-insights/cybersecurity-and-ai-integrating-and-building-existing-nist-guidelines>

NIST. *AI Risk Management Framework*.
<https://www.nist.gov/itl/ai-risk-management-framework>

NIST. (2024). *Cybersecurity, Privacy, and AI*.
<https://www.nist.gov/itl/applied-cybersecurity/cybersecurity-privacy-and-ai>

Executive Gov. (2025). *NIST Working on Cyber AI Profile Framework*.

<https://executivegov.com/2025/04/nist-develops-cyber-ai-profile/>

NIST. *Artificial intelligence*.

<https://www.nist.gov/artificial-intelligence>

The White House. (2025). *Sustaining Select Efforts to Strengthen the Nation's Cybersecurity and Amending Executive Order 13694 and Executive Order 14144*.

<https://www.whitehouse.gov/presidential-actions/2025/06/sustaining-select-efforts-to-strengthen-the-nations-cybersecurity-and-amending-executive-order-13694-and-executive-order-14144/>

Palo Alto Networks. *NIST AI Risk Management Framework (AI RMF)*.

<https://www.paloaltonetworks.com/cyberpedia/nist-ai-risk-management-framework>

Schellman. (2024). *NIST's AI Risk Management Framework Explained*.

<https://www.schellman.com/blog/cybersecurity/nist-ai-risk-management-framework-explained>

Frontiers in Artificial Intelligence. (2025). *A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity*.

<https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1526221/full>

Gibraltar Solutions. (2025). *Navigating the AI Black Box Problem*.

<https://gibraltarsolutions.com/blog/navigating-the-ai-black-box-problem/>

Brighterion AI. (2024). *Explainable AI: From Black Box to Transparency*.

<https://b2b.mastercard.com/news-and-insights/blog/explainable-ai-from-black-box-to-transparency/>

Invoca. *What Is Black Box AI?*
<https://www.invoca.com/blog/what-is-black-box-ai>

CommTel Networks. *Illuminating the Black Box: Explainable AI and Algorithmic Transparency.*
<https://commtelnetworks.com/illuminating-the-black-box-explainable-ai-and-algorithmic-transparency/>

Egyptian Journal of Radiology and Nuclear Medicine. *Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology.*
<https://ejrnm.springeropen.com/articles/10.1186/s43055-024-01356-2>

IBM. *What Is Black Box AI and How Does It Work?*
<https://www.ibm.com/think/topics/black-box-ai>

Springer. *Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence.*
<https://link.springer.com/article/10.1007/s12559-023-10179-8>

TechTarget. *What is Black Box AI? Definition from TechTarget.*
<https://www.techtarget.com/whatis/definition/black-box-AI>

Arthur AI. *From Black Box to Glass Box: Transparency in XAI.*
<https://www.arthur.ai/blog/from-black-box-to-glass-box-transparency-in-xai>

Condividi sui Social Network:
<https://www.ictsecuritymagazine.com/articoli/ai-offensiva/>