Red Teaming e Penetration Testing con intelligenza artificiale: la nuova frontiera della sicurezza informatica

L’evoluzione del paradigma offensivo nella cybersecurity AI-driven

L’integrazione dell’intelligenza artificiale nelle metodologie di valutazione della sicurezza rappresenta una trasformazione epocale nel panorama della cybersecurity contemporanea. Il red teaming e il penetration testing, discipline tradizionalmente dominate dall’expertise umana, stanno attraversando una metamorfosi profonda che ridefinisce le dinamiche tra attaccanti e difensori nel cyberspazio.

L’Executive Order 14110 firmato dal Presidente Biden il 30 ottobre 2023 ha cristallizzato questa evoluzione definendo l’AI red teaming come “uno sforzo di testing strutturato per individuare falle e vulnerabilità in un sistema AI, spesso in ambiente controllato e in collaborazione con gli sviluppatori di AI. Il red teaming dell’intelligenza artificiale è eseguito tipicamente da ‘red team’ dedicati che adottano metodi avversariali per identificare falle e vulnerabilità, come output dannosi o discriminatori da un sistema AI, comportamenti di sistema indesiderati o imprevisti, limitazioni o potenziali rischi associati all’uso improprio del sistema”.

Tale definizione istituzionale ha segnato il passaggio da una pratica emergente a un requisito normativo. È importante notare che l’EO 14110 è stato revocato dall’amministrazione

Trump il 20 gennaio 2025; tuttavia, le linee guida tecniche e i framework sviluppati in sua attuazione dal NIST e altre agenzie federali continuano a rappresentare riferimenti metodologici fondamentali per l'industria.

Fondamenti teorici del red teaming applicato ai sistemi AI

La convergenza tra sicurezza tradizionale e AI security

Il National Institute of Standards and Technology ha delineato attraverso il framework NIST AI 600-1, rilasciato nella versione finale il 26 luglio 2024, le coordinate metodologiche per la valutazione dei rischi associati all'intelligenza artificiale generativa. Questo profilo, complementare all'AI Risk Management Framework, introduce una tassonomia di 12 categorie di rischio specifiche che sono "nuove o esacerbate dall'uso dell'AI generativa", distinguendo le vulnerabilità intrinseche dei modelli di machine learning da quelle ereditate dall'infrastruttura tecnologica sottostante.

Il documento identifica rischi quali: CBRN Information (accesso facilitato a informazioni su armi chimiche, biologiche, radiologiche o nucleari), Confabulation (generazione di contenuti falsi presentati come fattuali), Dangerous or Violent Recommendations, Data Privacy, Environmental impacts, Harmful Bias, Homogenization, Information Integrity, Information Security, Intellectual Property, Obscene/Degrading Content, e Value Chain/Component Integration.

La Cybersecurity and Infrastructure Security Agency (CISA) ha ulteriormente consolidato questo approccio integrando i servizi di red teaming AI nell'ambito del più ampio framework TEVV (Testing, Evaluation, Verification and Validation).

Secondo CISA, le valutazioni di sicurezza dei sistemi AI devono essere concepite come sottoinsiemi delle metodologie consolidate per il software tradizionale, beneficiando così di "decenni di approcci comprovati e testati per assicurare che il software sia adatto allo scopo". CISA ha già iniziato a ricevere richieste da partner per condurre penetration testing e testing tecnico di sicurezza su Large Language Model (LLM) e prevede che la domanda per questi servizi crescerà con l'adozione crescente di strumenti AI.

Il framework MITRE ATLAS: le minacce AI

Il progetto ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) sviluppato da MITRE Corporation costituisce la base di conoscenza più completa sulle tattiche e tecniche avversariali specifiche per i sistemi di intelligenza artificiale. Modellato sul celebre framework ATT&CK, ATLAS cataloga attualmente 15 tattiche distinte, 66 tecniche e 46 sotto-tecniche derivate da osservazioni di attacchi reali e dimostrazioni condotte da team di sicurezza specializzati, come documentato nel changelog ufficiale aggiornato a ottobre 2025.

Le tattiche peculiari di ATLAS che non trovano corrispondenza in ATT&CK includono:

- **ML Model Access (AML.TA0004)**: descrive le modalità attraverso cui gli avversari ottengono accesso ai modelli target tramite API di inferenza o accesso diretto agli artefatti
- **ML Attack Staging (AML.TA0012)**: copre le fasi preparatorie degli attacchi, inclusi l'avvelenamento dei dati di training e l'inserimento di backdoor nei modelli

Il framework ATLAS eredita 13 tattiche da ATT&CK, tra cui Reconnaissance, Initial Access, Execution ed Exfiltration, applicandole specificamente ai contesti AI. I dati ATLAS sono disponibili in formato STIX 2.1, consentendo l'integrazione

machine-readable con strumenti di sicurezza e piattaforme SIEM. Il framework si connette anche a MITRE D3FEND, che fornisce contromisure difensive mappabili contro le tecniche ATLAS.

L’OWASP Top 10 per applicazioni LLM: vulnerabilità emergenti nel 2025

Rischi per i Large Language Model

L’edizione 2025 dell’OWASP Top 10 for LLM Applications, rilasciata a novembre 2024 come parte dell’OWASP GenAI Security Project, riflette l’evoluzione accelerata del panorama delle minacce, incorporando le lezioni apprese da incidenti reali e le trasformazioni nelle pratiche di deployment. Come evidenziato nella documentazione ufficiale: “Le organizzazioni stanno entrando in territorio inesplorato nella sicurezza e supervisione delle soluzioni GenAI. Il rapido avanzamento della GenAI apre anche le porte agli avversari per migliorare le loro strategie di attacco.”

La classificazione aggiornata per il 2025 comprende:

1. **LLM01: Prompt Injection** – Manipolazione dei LLM tramite input costruiti che portano ad accesso non autorizzato, violazioni di dati e compromissione del processo decisionale
2. **LLM02: Sensitive Information Disclosure** – La mancata protezione contro la divulgazione di informazioni sensibili negli output LLM può comportare conseguenze legali o perdita di vantaggio competitivo
3. **LLM03: Supply Chain** – Vulnerabilità introdotte quando componenti di terze parti, dati o modelli vengono utilizzati senza adeguata verifica
4. **LLM04: Data and Model Poisoning** – Dati di training

manipolati che possono compromettere sicurezza, accuratezza o comportamento etico dei modelli
5. **LLM05: Improper Output Handling** – Output non validati che possono generare contenuti dannosi e introdurre rischi di sicurezza aggiuntivi
6. **LLM06: Excessive Agency** – LLM a cui viene concessa troppa autonomia per compiere azioni, con conseguenze non intenzionali che compromettono affidabilità, privacy e fiducia
7. **LLM07: System Prompt Leakage** – Rischio di esposizione delle istruzioni di sistema che possono rivelare informazioni sulla logica applicativa
8. **LLM08: Vector and Embedding Weaknesses** – Vulnerabilità nei sistemi RAG e nelle pipeline di embedding
9. **LLM09: Misinformation** – Output LLM che generano disinformazione, compromettendo il processo decisionale
10. **LLM10: Unbounded Consumption** – Consumo eccessivo di risorse che può portare a denial of service o sfruttamento finanziario

Vulnerabilità critiche nella supply chain AI

La categoria Supply Chain ha acquisito rilevanza crescente in risposta alla proliferazione di componenti AI di terze parti. Come riportato nei dati di settore, il 53% delle aziende opta per non effettuare fine-tuning dei propri modelli, affidandosi invece a pipeline RAG e agentiche, il che rende le vulnerabilità relative a vettori e embedding particolarmente critiche.

L’avvelenamento dei dati di training (Data Poisoning) rappresenta una minaccia particolarmente subdola, capace di alterare il comportamento del modello in modi difficilmente rilevabili attraverso testing funzionale convenzionale. Le organizzazioni che adottano modelli pre-addestrati, dataset esterni o librerie di machine learning ereditano potenziali

vettori di compromissione che sfuggono ai tradizionali controlli di sicurezza applicativa.

Strumenti e metodologie per il penetration testing AI-native

L’automazione del red teaming attraverso agenti avversariali

La ricerca accademica ha prodotto framework sofisticati per l’automazione delle attività di red teaming. PentestGPT, presentato alla conferenza USENIX Security 2024 e vincitore del Distinguished Artifact Award, rappresenta un paradigma emergente di agente autonomo per penetration testing alimentato da Large Language Model.

La ricerca, condotta da Gelei Deng e colleghi della Nanyang Technological University, dimostra che “PentestGPT non solo supera gli LLM standard con un incremento del completamento dei task del 228.6% rispetto al modello GPT-3.5 tra i target benchmark, ma si dimostra efficace anche nell’affrontare target di penetration testing reali e challenge CTF”. L’architettura incorpora tre moduli auto-interagenti, persistenza delle sessioni, containerizzazione per ambienti riproducibili e capacità multi-categoria che spaziano dal web exploitation alla forensics, crypto, reversing e privilege escalation.

Il framework AutoPentester propone un’evoluzione ulteriore attraverso un sistema multi-agente che replica l’approccio cognitivo dei penetration tester umani. Dato un indirizzo IP target, il sistema esegue autonomamente ricognizione, scanning, valutazione delle vulnerabilità e sfruttamento attraverso strumenti di sicurezza consolidati, adattando dinamicamente la strategia di attacco sulla base degli output delle iterazioni precedenti. AutoPentester utilizza cinque

agenti basati su LLM: Summarizer, Strategy Analyzer, Generator, Results Verifier e Report Generator.

Tecniche avanzate di jailbreaking multi-turn

La ricerca recente ha evidenziato l'importanza degli attacchi multi-turn rispetto alle strategie single-turn tradizionali. Come documentato nel paper GOAT (Generative Offensive Agent Tester), gli utenti interagiscono con i chatbot spesso "avversarialmente" su più turni per un dato argomento. Il framework GOAT raggiunge un Attack Success Rate (ASR@10) del 97% contro Llama 3.1 e dell'88% contro GPT-4-Turbo sul dataset JailbreakBench.

Anthropic ha introdotto nel 2024 il concetto di "Many-Shot" jailbreaking, dove l'inserimento di un numero crescente di esempi avversariali nel prompt template aumenta progressivamente la probabilità di successo dell'attacco. Questa tecnica sfrutta la tendenza dei modelli a seguire pattern dimostrati attraverso esempi in-context.

Framework open source per AI security testing

La comunità open source ha sviluppato numerosi strumenti per il red teaming e il penetration testing di sistemi AI, offrendo alle organizzazioni la possibilità di condurre valutazioni di sicurezza indipendenti e personalizzabili:

- DeepTeam – Framework per red teaming LLM con oltre 40 vulnerabilità pre-configurate e 10 metodologie di attacco per scenari sia single-turn che conversazionali multi-turn. Supporta conformità nativa a standard OWASP Top 10 e NIST AI RMF
- Garak – Framework di attacco LLM per vulnerability scanning e generazione di probe avversariali

- Promptfoo – Strumento per testing di regressione con integrazione MITRE ATLAS, progettato per l'integrazione in pipeline CI/CD
- PyRIT (Python Risk Identification Tool) – Framework per red teaming automatizzato di sistemi AI generativa, rilasciato con licenza MIT

Questi strumenti condividono alcune caratteristiche metodologiche comuni: la capacità di generare input avversariali sistematici, la valutazione automatizzata degli output tramite classificatori di sicurezza, e il supporto per benchmark standardizzati che consentono la comparazione dei risultati nel tempo e tra diversi modelli.

Benchmark e dataset per la valutazione

La comunità di ricerca ha sviluppato benchmark standardizzati per valutare le tecniche di red teaming:

- **AdvBench**: Introdotto da Zou et al. (2023), contiene 58 comportamenti dannosi specifici per testare la suscettibilità degli LLM agli attacchi avversariali
- **HarmBench**: Proposto da Mazeika et al. (2024), è un benchmark più completo con 510 comportamenti unici attraverso varie categorie
- **ALERT**: Sviluppato da Tedeschi et al. (2024), contiene circa 15.000 prompt di red-teaming standard categorizzati secondo una tassonomia fine-grained con 6 categorie generali e 32 sottocategorie

Il quadro normativo: EU AI Act e requisiti di sicurezza

Obblighi di testing avversariale per

sistemi ad alto rischio

Il Regolamento UE 2024/1689, noto come EU AI Act, è entrato in vigore il 1° agosto 2024 come primo framework legale completo sull'AI a livello mondiale. Il regolamento introduce requisiti vincolanti di valutazione e testing avversariale per specifiche categorie di sistemi AI.

I provider di modelli General Purpose AI (GPAI) che presentano rischio sistemico sono tenuti a:

- Condurre valutazioni dei modelli e **testing avversariale strutturato**
- Tracciare e reportare gli incidenti gravi
- Implementare protezioni di cybersecurity adeguate

L'articolo 15 dell'AI Act prescrive per i sistemi ad alto rischio la dimostrazione di **accuratezza, robustezza e sicurezza informatica**, con misure tecniche proporzionate al contesto operativo e ai rischi associati. I provider devono implementare misure contro minacce quali:

- Data poisoning (avvelenamento dei dati)
- Model evasion (evasione del modello)
- Attacchi avversariali
- Vulnerabilità emergenti

Timeline di implementazione e implicazioni operative

L'entrata in vigore progressiva dell'EU AI Act stabilisce milestone critiche per le organizzazioni:

In base alle scadenze normative, i team di sicurezza hanno ora da 6 a 36 mesi per raggiungere la conformità alle diverse fasi dell'EU AI Act. Le organizzazioni devono stabilire sistemi di risk management per governare l'intero ciclo di vita delle

applicazioni ad alto rischio e condurre data governance per assicurare che i dataset utilizzati per training, validazione e testing siano rilevanti, rappresentativi, accurati e completi.

Prospettive evolutive: dall'adversarial testing all'AI governance

La convergenza multi-stakeholder nel red teaming pubblico

Iniziative pioneristiche come l'esercizio ARIA (Assessing Risks and Impacts of AI) condotto dal NIST in collaborazione con Humane Intelligence hanno dimostrato la fattibilità di modelli di red teaming pubblico cooperativo. L'esercizio del 2024 ha coinvolto 457 partecipanti in una competizione capture-the-flag che ha testato software di produttività office alimentati da AI generativa.

L'operazione, svolta durante la conferenza CAMLIS (Conference on Applied Machine Learning for Information Security) nell'ottobre 2024, ha utilizzato il framework NIST AI 600-1 come rubrica operativa per valutare sei categorie di rischio. I modelli testati includevano applicazioni di Meta, Robust Intelligence, Synthesia e Anote, con capacità che spaziavano dalla generazione text-to-video al ragionamento su documenti via RAG.

La metodologia multi-fase ha combinato qualificazioni virtuali con eventi intensivi in presenza, selezionando red teamer da background diversificati che includevano ricercatori AI, professionisti di cybersecurity, data scientist, esperti legali e policy maker. I risultati hanno evidenziato pattern di exploit ricorrenti e informato lo sviluppo di metriche standardizzate per la valutazione della sicurezza AI.

Insight dal NVIDIA AI Red Team

Il NVIDIA AI Red Team (AIRT) ha condiviso insight pratici derivanti dalla valutazione di numerosi sistemi AI-enabled per potenziali vulnerabilità. Il team, composto da ricercatori con background in penetration testing, adversarial machine learning e AI security, sottolinea l'importanza di trattare i sistemi LLM come input non fidati e di implementare architetture difensive stratificate.

Kai Greshake, ricercatore del team e autore di ricerche ampiamente citate sulla prompt injection indiretta, evidenzia come i sistemi complessi tendano a fallire in modi inaspettati, rendendo essenziale un approccio proattivo alla sicurezza.

Verso un paradigma di continuous adversarial validation

L'evoluzione delle pratiche di red teaming AI converge verso modelli di validazione avversariale continua integrata nei cicli DevSecOps. Come riportato da VentureBeat, i risultati del red teaming espongono una verità dura sull'arms race della sicurezza AI: "Ogni modello frontier fallisce sotto pressione sostenuta."

I dati pubblicati dalla piattaforma Gray Swan Shade su modelli Claude illustrano questa dinamica:

- Claude Opus 4.5 in ambienti di coding: 4.7% ASR a 1 tentativo, 33.6% a 10 tentativi, 63.0% a 100 tentativi
- In computer use con extended thinking: 0% ASR anche dopo 200 tentativi (primo modello a saturare il benchmark)

Questi risultati evidenziano il gap crescente tra capacità offensive e prontezza difensiva. Come osservato da Elia Zaitsev, CTO di CrowdStrike: "Se hai avversari che effettuano breakout in due minuti, e ti serve un giorno per ingestire

dati e un altro giorno per eseguire una ricerca, come puoi sperare di tenere il passo?”

L’automazione attraverso agenti LLM consente scaling del testing che sarebbe impraticabile con approcci esclusivamente manuali, mentre la supervisione umana rimane essenziale per interpretazione contestuale e discovery di vulnerabilità emergenti che sfuggono a pattern noti. L’OWASP Gen AI Red Teaming Guide fornisce metodologie strutturate per identificare vulnerabilità a livello di modello e di sistema, raccomandando testing avversariale trimestrale come pratica standard per qualsiasi team che rilascia funzionalità LLM-powered.

Considerazioni conclusive

Il red teaming e il penetration testing dei sistemi AI rappresentano discipline in rapida maturazione che richiedono competenze ibride all’intersezione tra cybersecurity tradizionale, machine learning e comprensione dei rischi specifici dell’intelligenza artificiale. L’adozione di framework strutturati come MITRE ATLAS e OWASP Top 10 for LLM Applications fornisce vocabolari condivisi e metodologie riproducibili, mentre l’evoluzione normativa con l’EU AI Act impone standard minimi di due diligence per organizzazioni che deployano sistemi AI in contesti ad alto rischio.

La traiettoria evolutiva indica una progressiva automazione supportata da agenti AI specializzati, ma la dimensione umana del red teaming, caratterizzata da pensiero laterale, comprensione contestuale e giudizio etico, rimane il differenziatore critico tra esercizi di compliance formale e genuine attività di assicurazione della sicurezza. Le organizzazioni che investono in capability mature di AI security testing si posizionano vantaggiosamente in un panorama dove le vulnerabilità dei sistemi intelligenti costituiscono vettori di rischio strategico con potenziali

impatti su sicurezza, reputazione e conformità normativa.

Condividi sui Social Network:
https://www.ictsecuritymagazine.com/articoli/red-teaming/