Digeat N.7 - 19 Settembre 2025

Avvelenare l’IA: la nuova arma degli hacker

Di Giusy Distratis

Abstract

Il contributo analizza il cambiamento di paradigma nel campo della cybersecurity determinato dall’integrazione pervasiva dei sistemi di intelligenza artificiale. L’introduzione di queste tecnologie comporta nuove vulnerabilità specifiche, come gli adversarial attack (attacchi avversariali) e il data poisoning (avvelenamento dei dati), una tecnica che altera il dataset di addestramento di un modello con l’obiettivo di indurre errori, introdurre pregiudizi o creare backdoor nascoste. Questa minaccia, spesso difficile da individuare e contrastare, evidenzia la necessità di un rinnovato approccio normativo e tecnico. Il Regolamento europeo sull’Intelligenza Artificiale affronta tali sfide imponendo rigorosi requisiti di robustezza e accuratezza per i sistemi ad alto rischio, rendendo imprescindibile l’adozione di un approccio security by design per garantire l’integrità dei dati e la resilienza dei modelli contro manipolazioni malevoli.

Indice

Quando i dati tradiscono: il data poisoning come minaccia all’intelligenza artificiale. Chi ne risponde davvero?

L’intelligenza artificiale rappresenta, ormai, un elemento determinante nei processi decisionali strategici in ambiti quali la finanza, la sanità e la sicurezza, nei quali l’affidabilità dei sistemi costituisce un presupposto imprescindibile ai fini della legittimità e dell’efficacia delle determinazioni adottate. Ma mentre ne celebriamo la potenza, spesso ignoriamo la sua più profonda vulnerabilità: l’integrità dei dati da cui apprende. Un avversario può deliberatamente “avvelenare” queste informazioni, trasformando un sistema fidato in un’arma imprevedibile o in un consulente fallace.

Due scenari, una sola minaccia

Primo scenario: compromissione della fiducia

Si consideri il sistema di punta di un **e-commerce**: un sofisticato algoritmo di IA che personalizza le raccomandazioni per milioni di utenti, generando una parte significativa del fatturato. **Il sistema inizia a promuovere con insistenza un prodotto di bassa qualità, persino difettoso**. Le vendite di quel prodotto aumentano esponenzialmente, ma vengono presto sommerse da reclami, resi e recensioni negative. La fiducia dei clienti, costruita in anni, si erode in poche settimane.

> L'incidente è stato causato da un "avvelenamento" dei dati di addestramento, che ha indotto il sistema ad associare quel prodotto a un alto indice di gradimento. Il danno non è solo economico, per aver promosso un articolo non conforme, ma anche reputazionale.

Secondo scenario: la falla invisibile

Si analizzi ora il caso di un **sistema di IA che protegge** un'istituzione finanziaria **dalle frodi**, esaminando milioni di transazioni al giorno. Il modello appare robusto, preciso e affidabile. **Un avversario inserisce nei dati di addestramento un "cavallo di Troia": informazioni manipolate che creano una backdoor nascosta. Il modello impara a ignorare uno schema di frode molto specifico, noto solo all'attaccante.** Per mesi, piccole somme di denaro vengono sottratte senza essere rilevate, fino a causare una perdita milionaria.

> Le perdite economiche sono dirette, ma la vera crisi è di fiducia e di responsabilità legale: chi risponde di una falla che non risiede nel codice, ma nella "conoscenza" stessa della macchina?

Questi non sono scenari ipotetici, ma esempi concreti di una minaccia silenziosa e pervasiva: il data poisoning.

Il data poisoning: una minaccia alla "catena di approvvigionamento" dei dati

Il data poisoning (avvelenamento dei dati) è una forma sofisticata di attacco informatico che colpisce i modelli di intelligenza artificiale alle loro fondamenta.

> A differenza degli attacchi che mirano a un modello già addestrato, il data poisoning corrompe il processo di apprendimento stesso, introducendo dati manipolati nel set di addestramento. L'obiettivo non è semplicemente ingannare il sistema, ma alterarne la "base di conoscenza", compromettendone i parametri e la logica decisionale.

Un sistema di riconoscimento facciale, ad esempio, addestrato con immagini etichettate erroneamente, potrebbe fallire nel riconoscere i soggetti o, peggio, identificarli in modo errato, con conseguenze gravi.

Se gli **attacchi di evasione (Evasion Attacks)**, possono essere paragonati a un raggiro del controllo qualità a fine produzione, e gli **attacchi al modello (Model Attacks)** a un sabotaggio dei macchinari, il **data poisoning** equivale a contaminare le materie prime ancora prima che entrino in fabbrica.

> Se i dati sono compromessi all'origine, il prodotto finale (il modello di IA) sarà inevitabilmente difettoso, indipendentemente dalla qualità del processo di produzione.

Questo rende il rilevamento e la correzione degli effetti dell'avvelenamento estremamente più complessi.

Comprendere come vengono eseguiti questi attacchi è fondamentale per sviluppare difese efficaci. Gli avversari sfruttano le vulnerabilità presenti nella catena di approvvigionamento dei dati, che spesso si basa su set di dati open-source, scraping dal web o fornitori terzi.

> I dati avvelenati vengono elaborati per mimetizzarsi con i dati legittimi, eludendo i tradizionali meccanismi di rilevamento delle anomalie.

Le alterazioni possono provenire da **diverse fonti**:

- **Minacce Interne (Insider Attack):** Individui con accesso legittimo ai dati possono introdurre campioni falsi.
- **Compromissione della Supply Chain:** L'uso di fonti di dati esterne già "avvelenati" contamina a cascata tutti i modelli che le utilizzano.
- **Accesso Non Autorizzato:** Ottenuto tramite phishing o altre intrusioni di rete.

Gli attacchi di data poisoning possono essere classificati in base al loro obiettivo. Si potranno avere **attacchi non mirati (Degradazione)**, nei quali l'obiettivo è minare le prestazioni generali del modello, riducendone l'accuratezza complessiva. Ad esempio, aggiungere dati etichettati in modo casualmente errato a un filtro antispam potrebbe portare alla classificazione errata di e-mail importanti. Oppure trovarci di fronte ad **attacchi mirati (Manipolazione)** i quali sono più insidiosi, mirano a corrompere l'output del modello solo in risposta a input specifici (trigger), lasciandolo funzionare normalmente negli altri casi. Un esempio è quello di addestrare un filtro a considerare sempre sicuri gli URL provenienti da un dominio malevolo specifico, pur continuando a bloccare tutto il resto dello spam.

Per raggiungere questi scopi, gli attaccanti impiegano varie tecniche di manipolazione dei dati:

1. **La rietichettatura (Label Flipping)** che consiste nell'alterare deliberatamente le etichette di campioni di dati legittimi. Ad esempio, etichettare foto di cibo per cani come cibo per gatti per confondere un sistema di riconoscimento di immagini e indurlo a classificare erroneamente i prodotti;
2. **L'Iniezione di dati sintetici:** vengono creati e inseriti dati completamente nuovi, progettati per alterare il comportamento del modello;
3. **Gli attacchi backdoor**: vengono inseriti dati che creano una vulnerabilità nascosta attivabile solo da uno specifico trigger. Il modello si comporta normalmente finché non incontra quel trigger.

> Quando un modello di IA è integrato in sistemi critici (diagnostica medica, veicoli autonomi) le sue prestazioni degradate si traducono direttamente in rischi operativi e di sicurezza.

Vulnerabilità, responsabilità e strategie di difesa

I **modelli di Deep Learning** richiedono set di dati massicci, spesso aggregati da fonti eterogenee, rendendo impraticabile un'ispezione manuale e aumentando il rischio di introdurre dati malevoli.

Inoltre, la loro elevata capacità permette di "memorizzare" campioni avvelenati senza che le prestazioni generali degradino visibilmente, consentendo a un comportamento malevolo di rimanere dormiente.

Un ulteriore fattore di vulnerabilità è rappresentato dagli ambienti di addestramento distribuiti, come il **federated learning**, che coinvolgono più partecipanti nel contribuire con dati e risorse computazionali.

> Attori malevoli possono iniettare dati avvelenati senza avere un accesso diretto al modello centrale, complicando notevolmente il rilevamento. Le stesse qualità che rendono il deep learning potente – la sua capacità di apprendere da vasti set di dati e la sua elevata capacità rappresentativa – sono precisamente ciò che lo rende vulnerabile. **L'avvento di modelli generativi come ChatGPT ha radicalmente espanso la superficie di attacco**.

La cybersecurity ora **deve considerare** non solo le minacce al codice o all'infrastruttura di rete, ma **anche attacchi diretti ai dati e ai modelli stessi**. I sistemi di IA attuali sono spesso dinamici, opachi e "vulnerabili by design", essendo stati commercializzati prima di raggiungere una piena robustezza. Addestrati su enormi quantità di dati non sempre controllati, reagiscono a input esterni in modi imprevedibili e non del tutto spiegabili, creando un ambiente ideale per attori malevoli.

Per affrontare queste sfide, **l'AI Act introduce un quadro normativo rigoroso**. L'Articolo 15, in particolare, stabilisce obblighi fondamentali per i sistemi di IA "ad alto rischio":

- **Progettazione sicura (Security by Design)**: i sistemi devono resistere a input maligni e attacchi avversariali fin dall'inizio.;
- **Accuratezza e tracciabilità**: il sistema deve garantire un funzionamento coerente, ripetibile e interamente auditabile. La capacità di tracciare ogni azione è fondamentale per identificare manipolazioni o difetti strutturali;
- **Resistenza agli attacchi**: il sistema deve essere progettato per resistere a manipolazioni. Ciò richiede attività continue di penetration test, simulazioni di attacco e auditing da parte di enti esterni;
- **Documentazione e valutazione del rischio**: chi sviluppa, distribuisce o implementa un sistema di IA deve documentare le misure di robustezza adottate, gli scenari di attacco considerati e i rischi residui accettati.

Tuttavia, la normativa sull'IA non opera in isolamento, ma è essenziale considerare la convergenza normativa con altre norme come il Cyber Resilience Act, la Direttiva NIS 2 e il GDPR. Un sistema di IA utilizzato in un'infrastruttura critica dovrà conformarsi anche ai requisiti della NIS 2.

L'AI Act distribuisce le responsabilità lungo l'intera catena del valore. Per i sistemi ad alto rischio, gli obblighi operativi sono suddivisi tra diversi attori:

- **Produttori (Provider)**: sono i primi responsabili della progettazione e dello sviluppo sicuro del sistema;
- **Distributori e Importatori**: hanno il dovere di verificare la conformità del sistema alle norme prima di immetterlo sul mercato;
- **Utilizzatori (Deployer)**: chi implementa e utilizza il sistema deve garantirne l'uso in un ambiente sicuro e protetto.

Immaginiamo che una grande azienda sanitaria utilizzi un modello di Natural Language Processing (NLP) per il triage medico. I pazienti descrivono i sintomi via chat e il sistema li classifica per priorità (es. urgente, differibile, non critico), secondo una logica analoga a quella dei codici a colori utilizzati nel triage dei pronto soccorso. Un attore malevolo avvelena il dataset di addestramento open source, usato dal modello, associando sintomi gravi a una classificazione non urgente. Il sistema, dopo il retraining, inizia a classificare erroneamente casi potenzialmente letali come non critici.

La responsabilità sarà condivisa:

- **Il Fornitore (Provider)** è responsabile ai sensi dell'art. 15 AI Act per non aver adottato adeguate misure di verifica e data auditing sul dataset open source;
- **l'Utilizzatore (Deployer)**, l'ospedale, è responsabile per aver implementato il sistema senza prevedere un adeguato controllo umano per supervisionare i casi a rischio, violando gli obblighi di implementazione sicura.

> Qualora l'attacco fosse esterno e non prevedibile, la responsabilità del fornitore potrebbe essere esclusa solo a condizione che questi dimostri di aver adottato *ex ante* tutte le misure ragionevoli e tecnicamente possibili per prevenirlo, rispettando gli standard e le best practices di settore. **La responsabilità, quindi, dipende dal livello di diligenza tecnica dimostrabile**.

L'**implementazione di adeguate strategie di difesa** costituisce un obbligo giuridico. Tali strategie si articolano **su tre livelli**:

1. **Misure preventive e governance dei dati (ex ante)**: adottare protocolli rigorosi e auditabili per la validazione dei dati di addestramento e implementare un controllo granulare degli accessi. Il controllo degli accessi e la segregazione dei dati dovranno garantire che l'accesso ai dati sensibili di addestramento sia limitato al personale autorizzato e debitamente tracciato tramite log immodificabili, validi a fini probatori;
2. **Meccanismi di rilevamento e risposta agli Incidenti**: monitorare le prestazioni in tempo reale per rilevare anomalie e attivare piani di risposta agli incidenti predefiniti (*Incident Response Plan*) che includa procedure per il contenimento del danno, l'analisi forense e la notifica alle autorità competenti, ove previsto dalla normativa (es. Direttiva NIS 2);
3. **Resilienza del modello e conformità by design**: impiegare tecniche di addestramento robusto, come l'*adversarial training*, e adottare difese la cui efficacia sia dimostrabile al fine di costituire una prova della diligenza del produttore nel mitigare i rischi di manipolazione prevedibili.

Conclusioni

> La minaccia del data poisoning impone un ripensamento fondamentale della cibersicurezza, spostando il focus dalla protezione delle infrastrutture alla salvaguardia dell'intera "catena di approvvigionamento cognitivo" dei modelli di IA.

Le conseguenze di un attacco minano direttamente le performance, l'etica e la reputazione dei sistemi, erodendo la fiducia, che è prerequisito per la loro adozione diffusa. La normativa, con l'AI Act in primo piano, sancisce una responsabilità condivisa lungo l'intera filiera, imponendo un approccio proattivo alla sicurezza**. La capacità di neutralizzare tali minacce determinerà non solo la conformità normativa, ma anche la fiducia del mercato e la sostenibilità a lungo termine dell'adozione dell'IA, rendendo la sicurezza il fondamento stesso del suo valore.**