Valutare i rischi Cyber al tempo dell'Intelligenza Artificiale

A cura di: Andrea Pasquinucci 11 Ottobre 2024

I modelli di Intelligenza Artificiale (IA) e in particolare quelli di Machine Learning (ML), Predittivi, Generativi e "General Purpose" quali ad esempio i "Large Language Models" (LLM) come GPT, Bard/Gemini, LLaMA ecc., compaiono

sempre più spesso all'interno delle applicazioni aziendali come componenti di applicazioni o per uso diretto.

E' necessario quindi valutare i rischi di sicurezza informatica per le aziende, comunemente anche detti "Rischi Cyber", connessi all'utilizzo di queste tecnologie. In questo articolo si danno alcune indicazioni su come approcciare la valutazione dei Rischi Cyber in presenza di modelli IA, quali sono le principali novità e cosa invece dovrebbe rimanere immutato rispetto alla tradizionale valutazione dei Rischi Cyber aziendali.

Valutazione dei Rischi Cyber

Conviene iniziare riassumendo brevemente l'approccio di base alla valutazione dei rischi [1,2]. Il punto di partenza è il possibile avverarsi di un **Evento** che può portare danno all'azienda. L'Evento può essere non-intenzionale, come un terremoto, un guasto, un errore umano, o intenzionale, come un attacco di "Denial of Service" in Internet.

E' necessario quindi avere una tassonomia di Eventi che potrebbero essere nocivi e valutarne la **Probabilità di Accadimento** e l'eventuale **Impatto** (o danno).

Per ogni Evento, il **Rischio Inerente** è *formalmente* rappresentato dalla formula:

Rischio Inerente = (Probabilità di Accadimento) x Impatto

Per ridurre il Rischio Inerente si adottano delle **Misure di Mitigazione del Rischio**, ottenendo il **Rischio Residuo** associato ad ogni Evento:

Rischio Residuo = (Probabilità di Accadimento) x Impatto / Maturità delle Misure di Mitigazione

o anche

Rischio Residuo = Rischio Inerente / Maturità delle Misure di Mitigazione

Riassumendo, per ogni Evento potenzialmente nocivo va valutata la Probabilità di Accadimento, l'Impatto e la Maturità delle Misure di Mitigazione.

Rischi Cyber e Rischi Cyber dovuti all'IA

Le applicazioni che forniscono o contengono modelli di Intelligenza Artificiale sono comunque applicazioni informatiche e come tali soggette ai rischi comuni a tutte le applicazioni informatiche. Tra questi possono essere indicati ad esempio guasti all'hardware dei sistemi, guasti ai sistemi di alimentazione e di condizionamento delle sale macchine, o attacchi con malware/ ransomware, attacchi di phishing, attacchi ai sistemi di controllo accesso (quali quelli di forza bruta sulle password), eccetera.

La valutazione di tutti questi Rischi Cyber procede identica anche per le applicazioni che forniscono o contengono modelli IA, e non sono considerate in questo articolo. Nel prosieguo sono considerati solo Eventi e rischi specifici per i modelli di Intelligenza Artificiale.

Valutazione della Probabilità di Accadimento e Rischi nei Modelli di Intelligenza Artificiale: Sfide e Approcci Qualitativi

L'approccio standard alla valutazione della Probabilità di Accadimento di un Evento è di avere statistiche su periodi temporali appropriati sull'avverarsi dell'Evento. In informatica queste statistiche sono difficilmente disponibili o non esistono proprio. Infatti spesso la velocità degli sviluppi informatici è altamente superiore agli intervalli di tempo necessari per ottenere opportune statistiche, oltre alla difficoltà di raccogliere dati da molte diverse organizzazioni. Nel caso dei modelli di Intelligenza Artificiale, la novità di questa tecnologia e la rapidità del suo sviluppo implica una quasi totale assenza di statistiche su periodi di tempo sufficienti.

Come molto spesso per la valutazione dei rischi in informatica, è quindi necessario valutare qualitativamente la Probabilità di Accadimento di un Evento, adottando scale come la seguente:

- *molto bassa* (o quasi mai)
- *bassa* (al più una volta all'anno)
- *media* (al più quattro volte all'anno)
- *alta* (una o due volte al mese)
- *molto alta* (ogni giorno o quasi sempre).

La valutazione è fatta sulla base delle informazioni note sull'Evento e sul sistema informatico, nostro caso il modello IA in ambito, la sua configurazione, i suoi modi d'uso, i suoi utenti ecc.

Conviene fare un paio di esempi considerando come componente IA un Large Language Model, ovvero ad esempio un ChatBot basato su un LLM. Si consideri un'applicazione esposta in Internet con registrazione per l'accesso pressoché libera; a priori quindi chiunque può registrarsi e accedere all'applicazione basata sul LLM.

Si consideri come Evento un attacco di "Prompt Injection": solo sulla base delle notizie giornalistiche e degli articoli scientifici su questo tipo di attacchi, è possibile valutare qualitativamente che la Probabilità di Accadimento è Molto Alta.

Si consideri ora un altro modello IA addestrato con dati sintetici, generati unicamente all'interno dell'azienda, da personale interno, qualificato e molto limitato in numero. In questo caso un Evento di attacco di "Training Data Poisoning" sarà, sempre qualitativamente, molto poco probabile e quindi con Probabilità di Accadimento Molto Bassa.

Viceversa, se il modello IA è addestrato con dati scaricati con Bot da Internet e direttamente passati al modello per l'addestramento, la Probabilità di Accadimento di un Evento di attacco di "Training Data Poisoning" potrà anche essere Alta o anche Molto Alta.

Va ricordato che l'approccio adottato per la valutazione dei rischi suppone che la valutazione della Probabilità di Accadimento sia indipendente dalla presenza di Misure di Mitigazione. In altre parole, va fatta una per quanto possibile chiara distinzione tra Evento, la sua Probabilità di Accadimento, l'Impatto che l'Evento

ha in caso di suo accadimento e le Misure di Mitigazione del rischio già presenti, valutando per quanto possibile in maniera indipendente gli ultimi tre fattori.

In particolare è alle volte difficile valutare quali siano le Misure di Mitigazione del rischio da considerare. Conviene ragionare su un esempio semplice: si consideri un'applicazione non esposta in Internet.

Tipicamente l'applicazione è eseguita su un server (fisico o virtuale) sulla rete aziendale (on-premises o in-cloud) che è separata da Internet da uno o più firewall che non permettono l'accesso da Internet all'applicazione. Nella valutazione dei rischi in considerazione, la non-esposizione in Internet è una *caratteristica dell'applicazione*, e non (formalmente) una Misura di Mitigazione del rischio, anche se questa *caratteristica* è in pratica realizzata da uno o più firewall.

Nella valutazione del rischio si dà quindi per scontato che la non-esposizione in Internet porti un rischio già valutato e accettato. D'altra parte, se si considera un'applicazione esposta in Internet si può ridurre parte del rischio eliminando l'esposizione in Internet; così facendo alcuni Eventi non saranno più applicabili (o più precisamente, la loro probabilità di accadimento sarà nulla o quasi) e alcune specifiche Misure di Mitigazione non saranno più strettamente necessarie, quale ad esempio l'utilizzo di Web Application Firewall.

Questo si collega a un'altra considerazione che deve essere fatta quando si considera la probabilità di accadimento degli eventi, ovvero la valutazione di eventi che vanno sotto il nome di "Black Swan".

Come già indicato, l'avvio di una valutazione dei rischi richiede di identificare i possibili eventi nocivi, tra questi possono esserci eventi con bassissima Probabilità di Accadimento ma con altissimo Impatto potenziale.

Ad esempio, nel caso appena citato di un'applicazione non esposta in Internet, l'Evento *Black Swan* potrebbe essere la contemporanea compromissione di tutti i sistemi di segregazione della rete interna da Internet e la conseguente esposizione in Internet delle applicazioni aziendali.

O nel caso di un Datacenter e il suo Datacenter di Disaster Recovery, un guasto contemporaneo all'alimentazione o all'impianto di raffreddamento o alla rete di connettività che rende indisponibili entrambi allo stesso tempo.

In molti casi, come nei semplici esempi appena indicati, eventi *Black Swan* sono dati dall'avverarsi contemporaneamente di più Eventi semplici oggetto dell'analisi dei rischi e da questi ne può essere fatta un'analisi sia della Probabilità di Accadimento sia dei possibili Impatti.

Valutazione degli Impatti su Riservatezza, Integrità e Disponibilità nei Modelli di Intelligenza Artificiale: Classificazione e Rischi di Abuso

La valutazione degli Impatti che un Evento può avere deve essere sempre svolta considerando diversi aspetti o dimensioni di valutazione: quelli più comuni sono gli impatti economici, quelli d'immagine e quelli di conformità legislativa, a standard e certificazioni.

Per procedere alla valutazione dell'Impatto potenziale che un Evento può avere (in assenza di Misure di Mitigazione), è conveniente classificare gli Eventi secondo il tipo di impatto che potrebbero avere, ovvero se possono impattare Riservatezza, Integrità e Disponibilità (RID, in Inglese CIA) delle informazioni e/o del servizio informatico, o costituire un Abuso del sistema [Rif. 3].

Per una tradizionale applicazione informatica, un Abuso può verificarsi nel caso di utilizzo di funzionalità esistenti da parte di chi non ne è autorizzato (violazione dell'Integrità del sistema di autorizzazione dell'applicazione) o in violazione della licenza d'uso. Questo perché tutte le funzionalità e i possibili dati prodotti sono decisi, progettati e implementati dai programmatori; in altre parole, ogni dato prodotto dell'applicazione è previsto in fase di progettazione e quindi, almeno teoricamente, già noto.

Invece per i modelli IA Generativi, per loro stessa natura, vi è un ulteriore caso di

Abuso quando il modello IA è utilizzato per generare dati non previsti dai programmatori e/o in violazione degli scopi dell'addestramento del modello, senza però violarne le caratteristiche RID.

Seguendo anche [Rif. 4], vengono qui considerati Eventi specifici per i modelli IA/ML che possono impattare le informazioni e/o il servizio informatico che le gestisce, organizzati secondo la caratteristica che potrebbero violare, tra Disponibilità, Integrità, Abuso e Riservatezza.

Impatto degli Attacchi e Malfunzionamenti sulla Disponibilità dei Modelli IA/ML: Rischi e Vulnerabilità

I modelli IA/ML possono essere oggetti a violazioni della loro Disponibilità con specifici attacchi o malfunzionamenti. A parte la violazione del codice con l'inserimento di back/trap-door, specifici Eventi per i modelli IA/ML possono far decadere le loro prestazioni sino a violare i livelli di Disponibilità previsti.

Come primo esempio, si considerino modelli IA/ML utilizzati in un veicolo a guida autonoma, alcuni di questi devono valutare dati provenienti dai sensori esterni in brevissimo tempo e fornire indicazioni per la guida del veicolo. Ma è possibile che per particolari dati sui quali non sono stati specificamente addestrati, questi modelli possano impiegare più tempo del previsto per svolgere le analisi.

Va ricordato che i modelli IA/ML sono addestrati su un campione di dati e non su tutti i dati possibili per cui il programmatore non può valutare a priori i tempi di esecuzione dei programmi su tutti i possibili dati in ingresso. Quello che può accadere è che il modello IA/ML non fornisca i dati nei tempi previsti, ovvero che i dati non siano *disponibili* nelle tempistiche previste.

L'impatto di un ritardo nella generazione dei dati può dipendere dalla quantità del ritardo; si possono quindi considerare diversi Eventi a seconda della possibile durata del ritardo nella produzione dei dati.

Un altro Evento di cui valutare il possibile Impatto, è che il modello IA/ML non sia in grado di identificare e processare i dati in ingresso producendo un risultato

valido.

Ad esempio nel caso di un'immagine, il modello potrebbe non essere in grado di distinguere se è presente un cartello stradale, un'automobile o un albero, dando uguale probabilità a tutti i tre oggetti, anche se l'elaborazione è fatta nei tempi previsti.

Questo è un caso un po' particolare di mancanza di Disponibilità, in quanto l'applicazione è disponibile, mentre quello che non è disponibile è il risultato dell'elaborazione. Un evento paragonabile per un'applicazione non IA è il caso che un errore di programmazione blocchi l'esecuzione dell'applicazione producendo al più un risultato nullo o un messaggio di errore.

Spesso l'addestramento dei modelli IA/ML, in particolare quelli più avanzati e complessi, viene esteso e integrato con nuovi dati ad esempio per migliorarne il comportamento, eliminare "Adversarial Examples" e aggiungere nuovi dati e informazioni.

L'estensione e integrazione dell'addestramento di un modello IA/ML può anche far *dimenticare* informazioni precedentemente imparate, e in casi estremi anche peggiorare le prestazioni di elaborazione del modello IA/ML. Quindi anche un Evento di estensione dell'addestramento può essere causa di un Impatto sulla Disponibilità di un modello IA/ML.

I precedenti tipi di Eventi sono principalmente dovuti sia all'architettura del modello IA/ML sia al processo e ai dati utilizzati per il suo addestramento. Sono però possibili anche attacchi ai modelli IA/ML, in particolare a quelli Generativi e ai Large Language Model, allo scopo sia di rallentare l'esecuzione del modello sino a quasi fermarne completamente l'elaborazione, sia di aumentare i consumi energetici in quanto i server su cui il modello è eseguito utilizzano tantissime risorse per l'elaborazione di una singola richiesta.

Questo può essere ottenuto con la preparazione ed esecuzione di specifiche richieste ("Query") al modello costruite ad arte per poterne abusare. Questi tipi di attacco fanno anche parte degli attacchi di Abuso chiamati "Prompt Injection" o "Adversarial Prompting" [Rif. 5], ma sono qui considerati in quanto lo scopo di

questi attacchi è di violare la Disponibilità del modello IA/ML [Rif. 6], per questo sono anche chiamati attacchi di tipo "AI Model DoS".

Infine vanno anche considerati specifici Eventi relativi ai modelli IA/ML con funzionalità di tipo "Retrieval Augmented Generation (RAG)" [Rif. 7], ovvero che possono accedere a banche dati esterne, tra cui anche l'intero Internet, per integrare i dati con cui sono stati addestrati con informazioni puntuali, di dettaglio e aggiornate.

Tipicamente i dati provenienti dalle banche dati esterne sono pre-elaborati e aggiunti alla richiesta ("Query") dell'utente per essere utilizzati come informazioni di contesto dal modello IA/ML. L'utilizzo di dati esterni può anche diventare un canale di attacco, anche solo involontario, al modello ottenendo effetti simili a quelli appena descritti per gli Eventi di "AI Model DoS".

Violazioni dell'Integrità nei Modelli IA/ML: Attacchi Adversariali, Allucinazioni e Avvelenamento dei Dati

In una usuale applicazione informatica la violazione dell'Integrità dei dati (informazioni o codice) comporta la produzione di risultati errati sia come risultato di una elaborazione sia come semplice produzione dei dati archiviati.

Per i modelli IA/ML è possibile considerare come violazione della "Integrità" eventi in cui il risultato dell'elaborazione è errato. Quindi per i modelli IA/ML associamo al concetto di violazione dell'Integrità più che la violazione di dati quali informazioni archiviate o codice dell'applicazione, il risultato di questa violazione, ovvero il produrre risultati errati.

Gli Eventi, sia involontari che attacchi volontari, detti di Evasione e anche chiamati "Adversarial Attacks" o "Adversarial Examples", sono forse quelli più specifici dei modelli di Machine Learning. Questi sono principalmente dovuti al processo di apprendimento dei modelli stessi e al modo in cui le informazioni apprese vengono utilizzate dai modelli durante la loro esecuzione.

In pratica sono Eventi in cui il modello sbaglia completamente su dati di ingresso molto simili a quelli di addestramento [Rif. 8]. Esempi famosi sono i cartelli

stradali di stop con uno sticker giallo riconosciuti come cartelli di limite di velocità, o montature di occhiali, cappelli, trucco per il viso che portano il modello a riconoscere una persona per un'altra. Gli Impatti di un Evento di Evasione possono essere molto significativi e devono essere considerati attentamente nella valutazione dei rischi Cyber di un'applicazione IA/ML.

Altri Eventi specifici per i modelli Generativi / LLM sono chiamati Allucinazioni (o "Confabulation") sono dovuti al fatto che questi modelli possono produrre risultati che sembrano assolutamente reali o realistici (hanno un'alta probabilità di assomigliare alla *realtà* derivante dai dati di addestramento), ma in realtà sono del tutto falsi, inappropriati o pericolosi. Negli ultimi due anni, sono stati riportati dagli organi di informazioni molti casi, alcuni anche clamorosi, di Eventi di Allucinazione dei modelli Generativi / LLM più famosi, a partire da ChatGPT, Bard ecc.

Altri tipi di Eventi più tradizionali di violazione dell'Integrità ma sempre molto significativi per i modelli IA/ML sono quelli di "Avvelenamento" del codice o dei dati di addestramento. In entrambi i casi si tratta di Eventi di modifica volontaria (cioè attacchi) o involontaria del codice del modello IA/ML o dei suoi dati di addestramento.

In particolare, la maggior parte dei modelli IA/ML più avanzati ha la necessità di utilizzare grandi quantità di dati di addestramento di cui è quindi difficile verificare la qualità e sicurezza; un loro "Avvelenamento" può far sì che il modello produca, almeno in alcune situazioni, risultati diversi da quelli attesi e per i quali è stato programmato, violando appunto l'Integrità di quanto prodotto.

Abuso dei Modelli IA Generativi: Rischi di Utilizzo Improprio e Impatti su Sicurezza e Conformità

Come indicato precedentemente, un Evento di Abuso per i modelli IA Generativi consiste in un utilizzo (tipicamente volontario) del modello allo scopo di ottenere risultati non previsti dai programmatori e/o in violazione degli scopi dell'addestramento del modello.

Esempi semplici, se non banali, sono la produzione da parte del modello di codice di Malware, di testi di Phishing, di Fake News, di istruzioni per la preparazione di ordigni o per la realizzazione di eventi malevoli quali attentati, rapine o furti, ma anche la generazione di immagini, video, audio in violazione di Copyright ecc.

Questi Eventi possono produrre Impatti di immagine, di conformità ma anche economici nel caso di violazione di diritti altrui ed è quindi necessario valutarne sia la Probabilità di accadimento (potrebbe anche essere nulla per certi tipi di modelli IA/ML) e i possibili Impatti.

Si possono includere tra gli Abusi, anche se dovuti direttamente al modello IA/ML stesso e non ad un uso improprio da parte di un utente finale, Eventi in cui modelli con funzionalità di tipo RAG sono in grado di eseguire azioni ad esempio su altre applicazioni raggiungibili in Internet, non previste e non lecite, quali attacchi a siti web per estrarne informazioni non direttamente accessibili [Rif. 9].

Violazioni della Riservatezza nei Modelli IA/ML: Attacchi di Estrazione e Ricostruzione dei Dati di Addestramento

L'analisi degli Eventi che possono violare la riservatezza delle informazioni elaborate da un modello IA/ML è quella più simile al caso di una applicazione informatica non IA/ML. Vi sono alcune sorgenti di informazioni che possono essere riservate: i dati di addestramento, l'architettura, il disegno e il codice del modello, i dati degli utenti che utilizzano il modello ma che non sono utilizzati per addestrare il modello stesso.

Dal punto di vista di un'analisi dei rischi specifica per i modelli IA/ML, l'unico tipo di Eventi realmente rilevante è quello che riguarda i dati di addestramento, utilizzati sia nelle fasi di pre-training che nelle fasi di fine-tuning, elaborati e memorizzati dal modello stesso. I principali Eventi che possono violare la Riservatezza delle informazioni in questo caso sono gli attacchi di Estrazione e Ricostruzione.

Questi attacchi sfruttano delle debolezze o vulnerabilità dei modelli IA/ML per

estrarre dati riservati tramite l'utilizzo del modello stesso. Questo è possibile perché all'interno di un modello ML non è possibile, almeno ad oggi, implementare un sistema di controllo accessi alle informazioni, e quindi un attaccante può provare a formulare delle "Query" in modo tale da ottenere come risultato dell'elaborazione informazioni sui dati di addestramento o sulla struttura stessa del modello, e in alcuni casi di ottenere i dati stessi.

Con maggiori difficoltà, questo è possibile anche per i modelli di ML più semplici quali quelli di "Classificazione" in cui un attaccante può sottoporre al modello dei dati in ingresso così preparati da poter dedurre dalle risposte del modello se un certo dato è stato utilizzato per l'addestramento del modello stesso (questi attacchi sono chiamati di "Membership Inference").

Strategie di Mitigazione dei Rischi Cyber nei Modelli IA/ML: L'Importanza di Governance, Formazione e Controlli Operativi

In questo articolo non sono trattate in dettaglio possibili misure di mitigazione dei rischi specifici alle applicazioni IA/ML.

È utile comunque sottolineare che forse più delle misure tecniche di sicurezza applicabili ai modelli IA/ML stessi, oggi sono forse più importanti e più efficaci misure di sicurezza di governo, processo, utilizzo e gestione di queste tecnologie. Queste includono la formazione e informazione degli amministratori IT e degli utenti sulle reali caratteristiche, potenzialità, funzionalità e rischi di utilizzo di queste tecnologie.

In altre parole, il solo fatto che tipicamente i risultati prodotti da modelli IA/ML siano elaborazione *probabilistiche* sulla base dei soli dati di addestramento utilizzati e non risultati esatti, decisi da chi ha disegnato l'applicazione, richiede un diverso approccio degli utenti al loro utilizzo.

D'altra parte gli sviluppatori che le programmano e gli amministratori IT che le gestiscono devono introdurre misure di sicurezza spesso esterne ai modelli IA/ML quali ad esempio filtri sui dati in ingresso e verifiche di qualità sui dati

prodotti.

Il Ruolo della Supply Chain nella Valutazione dei Rischi Cyber per l'Utilizzo di Modelli IA/ML Avanzati: Sfide e Impatti delle Terze Parti

Per concludere questa breve presentazione di un possibile approccio alla valutazione dei "Rischi Cyber" dovuti a un eventuale utilizzo dei più recenti e avanzati modelli di Intelligenza Artificiale e Machine Learning, deve essere evidenziato e considerato il ruolo e il possibile impatto di terze parti e in generale della Supply Chain.

Per molti aspetti il paradigma informatico dei modelli IA/ML è molto diverso da quello di una tradizionale applicazione informatica e richiede personale con specifiche competenze per la loro creazione e gestione. Dato il rapidissimo sviluppo di queste tecnologie nell'ultimo paio di anni, è comune che all'interno delle aziende non vi siano ancora né specifiche competenze tecniche, né sensibilità del personale per comprendere le proprietà, utilizzare correttamente e valutare i rischi derivanti dall'uso di modelli IA/ML. Questo facilmente porta ad affidarsi a terze parti senza però essere in grado di valutare i possibili impatti e rischi che ne possono derivare.

Il problema è ancora più complesso in quanto ad oggi il numero di fornitori di queste tecnologie che sono all'avanguardia sono relativamente pochi, e può capitare che un servizio applicativo erogato in Internet (o "Cloud") basato o che include tecnologie IA/ML avanzate come quelle dei modelli Predittivi, Generativi, "General Purpose" e "Large Language Models" (LLM), in realtà abbia come terza parte che fornisce integralmente la tecnologia IA/ML uno dei maggiori fornitori quali Amazon, Google, Meta, Microsoft, OpenAI ecc. La catena di fornitura, o Supply Chain, è spesso lunga e le competenze tecniche in ambito AI/ML possono risiedere all'estremo opposto della catena rispetto all'utilizzatore, e quindi difficili da verificare e valutare.

L'esempio più semplice da considerare è quello della gestione dei dati di addestramento. Per valutare i rischi derivanti dall'utilizzo di un'applicazione di

Machine Learning erogata da un fornitore, bisogna comprendere su quali dati è addestrato il modello. Infatti un modello ML potrebbe essere addestrato utilizzando dati direttamente scaricati da Internet o, più probabilmente, forniti da una terza parte specializzata nella loro raccolta e preparazione.

Bisognerebbe comprendere quali procedure sono state adottate per garantire non solo la qualità dei dati, ma anche l'eticità, l'assenza di copyright o restrizioni sul loro uso, le misure di sicurezza adottate per evitare che un attaccante sia in grado di *avvelenare* i dati di addestramento per propri scopi che possono includere anche un attacco all'azienda che utilizza il servizio. Inoltre bisognerebbe sapere se tra i dati di addestramento sono utilizzate anche informazioni dei propri concorrenti, se sono utilizzati dati pubblici della propria azienda, se sono utilizzati i dati inviati dal personale aziendale all'applicazione nel corso del suo utilizzo (genericamente chiamati dati delle "Query"), ecc.

La complessità e al contempo novità di questa Supply Chain può quindi rendere ancora più complessa la valutazione dei Rischi Cyber connessi all'utilizzo dei più avanzati modelli di Intelligenza Artificiale e Machine Learning.

Riferimenti

[1] Quella presentata è solo una tra le più comuni metodologie di calcolo dei rischi, un altro approccio alle volte adottato per la valutazione dei rischi in informatica è quello proposto da OWASP con la "OWASP Risk Rating Methodology"

https://owasp.org/www-community/OWASP_Risk_Rating_Methodology

[2] si veda anche "Mapping LLM Security Landscapes: A Comprehensive Stakeholder Risk Assessment Proposal", https://arxiv.org/abs/2403.13309v1

[3] NIST AI 100-2 "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations", https://csrc.nist.gov/pubs/ai/100/2/e2023/final

[4] OWASP, "OWASP Top 10 for Large Language Models", https://genai.owasp.org/ ; si veda anche A. Pasquinucci, "Attacchi ai Modelli di

Intelligenza Artificiale", https://www.ictsecuritymagazine.com/articoli/attacchi-ai-modelli-di-intelligenza-artificiale/

[5] Per una breve introduzione si veda "Prompt Injection Attacks in Large Language Models", SecureFlag, https://blog.secureflag.com/2023/11/10/prompt-injection-attacks-in-large-language-models/

[6] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, R. Anderson , "Sponge Examples: Energy-Latency Attacks on Neural Networks", https://arxiv.org/abs/2006.03463

[7] Per una breve introduzione si veda ad esempio A. Kumar, "Retrieval Augmented Generation (RAG) & LLM", https://vitalflux.com/retrieval-augmented-generation-rag-llm-examples/

[8] A. Pasquinucci, "Adversarial Attacks a Modelli di Machine Learning", White Paper, ICT Security Magazine, https://www.ictsecuritymagazine.com/pubblicazioni/adversarial-attacks/

[9] R. Fang, R. Bindu, A. Gupta, Q. Zhan, D. Kang, "LLM Agents can Autonomously Hack Websites", https://arxiv.org/abs/2402.06664 ; R. Fang, R. Bindu, A. Gupta, D. Kang, "LLM Agents can Autonomously Exploit One-day Vulnerabilities", https://arxiv.org/abs/2404.08144

Articolo a cura di **Andrea Pasquinucci**

Profilo Autore

Andrea Pasquinucci

PhD 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.

Altri Articoli

- Aspetti tecnici degli Adversarial Examples
- Adversarial Machine Learning – Aspetti Scientifici
- Attacchi ai Modelli di Intelligenza Artificiale
- Adversarial Attacks a Modelli di Machine Learning

Condividi sui Social Network:

#adversarial attack #Adversarial Examples #Adversarial Prompting #Cyber Risk #Intelligenza Artificiale #Prompt Injection #risk assessment #Valutazione dei rischi IA

← PRECEDENTE

CYBEROO: Strategie Avanzate di Detection e Remediation per una Protezione Always-On

Articoli simili