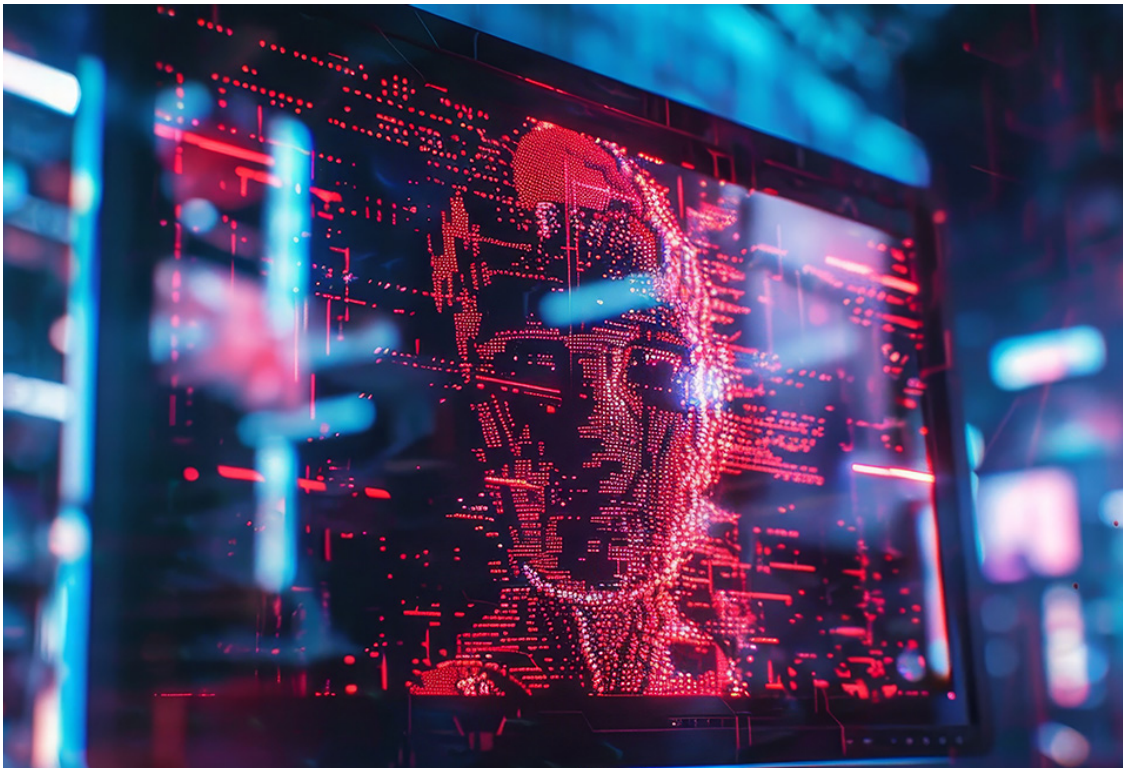

Attacchi ai Modelli di Intelligenza Artificiale

A cura di: Andrea Pasquinucci 7 Maggio 2024

I modelli di Intelligenza Artificiale (AI) stanno assumendo molto velocemente un grande ruolo nella vita di tutti noi. Le previsioni sono che nei prossimi anni utilizzeremo quotidianamente modelli AI sia in ambito lavorativo sia personale per moltissimi scopi. Ad esempio IDC [Rif. 1] prevede che entro il 2027 il 60% dei PC venduti avrà componenti Hardware ("system-on-a-chip – SoC", diversi dalle GPU) dedicate all'elaborazione locale di modelli AI (si veda anche [Rif. 2]). Ci si aspetta un simile sviluppo anche per i dispositivi mobili quali smartphone, tablet ecc. oltre, ovviamente, ai server.

Oltre alle possibili debolezze comuni a qualsiasi programma informatico, i modelli AI hanno delle debolezze specifiche e in alcuni casi nuove, contro le quali non siamo abituati a confrontarci quasi quotidianamente. La ricerca in questo campo è molto attiva e nuovi risultati sono annunciati ogni giorno. È comunque possibile fare già ora una rassegna dei principali tipi di debolezze dei modelli AI che possono essere sfruttate da un attaccante.

Approcci allo Studio dei Tipi di Attacchi ai Modelli AI

In generale lo studio di un attacco ad un sistema informatico può essere analizzato considerando le **vulnerabilità** che l'attaccante intende sfruttare, le **tecniche** per implementare l'attacco e l'**effetto** dell'attacco sul sistema attaccato. Vi sono vari approcci allo studio degli attacchi ai sistemi informatici. Fra i più noti è quello che fa riferimento alla "Kill Chain", ovvero identificare nell'ordine le azioni che un attaccante porta a termine per realizzare l'attacco per trovare i punti ove poterlo fermare. La metodologia più utilizzata per i modelli AI è quella sviluppata da MITRE con ATLAS™ [Rif. 3]. Questo approccio, focalizzandosi sulle azioni dell'attaccante, è sicuramente quello più appropriato per la gestione degli incidenti in tempo reale perché aiuta il difensore a identificare la migliore maniera per gestire l'incidente.

Un altro approccio è più preventivo e consiste nel valutare principalmente le vulnerabilità e i possibili effetti di un eventuale loro sfruttamento da parte di un attaccante per poter implementare misure di sicurezza sulla base di una valutazione dei rischi. In questo caso si ricorre spesso alla formulazione di tassonomie quale quella di NIST [Rif. 4] o di Microsoft [Rif. 5].

In questa sede non si intende descrivere in dettaglio una "Kill Chain" o proporre una tassonomia, ma fare una rassegna ad alto livello di quanto ad oggi si è capito cercando di organizzare le informazioni in uno schema sufficientemente semplice. Per questo l'approccio adottato è più simile a quello di una tassonomia mancandone però il rigore e la completezza.

Per questo articolo sono state considerate le debolezze o vulnerabilità dei modelli AI e le possibili conseguenze o effetti del loro sfruttamento. Questi sono stati raggruppati in quelli che genericamente saranno chiamati "attacchi". Seguendo [Rif. 4] conviene distinguere due tipi di attacchi: quelli comuni a molti tipi di modelli AI, che sono indicati come modelli **AI Predittivi**, e quelli specifici ai modelli **Generativi** e "General Purpose" quali ad esempio i "Large Language Models" (LLM) come GPT, Bard/Gemini, LLaMA ecc.

In sicurezza informatica, gli effetti dello sfruttamento di una debolezza o vulnerabilità sono usualmente classificati dalla violazione di una o più caratteristiche di sicurezza tra Riservatezza, Integrità e Disponibilità (RID, in Inglese CIA). Nella letteratura già richiamata, questa classificazione è utilizzata anche per i modelli AI con due piccole varianti: si considerano Riservatezza insieme a Privacy, Integrità, Disponibilità e per i modelli Generativi anche Abuso. Con Abuso si intende proprio l'utilizzo di un modello AI Generativo per generare dati in violazione degli scopi e/o dei modi di utilizzo del modello stesso. Si noti che l'abuso di un modello AI Generativo qui considerato è diverso dall'abuso di una tradizionale applicazione informatica. Tutte le funzionalità e possibili dati prodotti da una usuale applicazione informatica sono disegnati e implementati dai programmatori; in questa situazione un abuso si può verificare 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. Come sarà descritto in seguito, per i modelli Generativi un ulteriore caso di abuso è l'utilizzo del modello per generare dati non previsti dai programmatori o in violazione degli scopi dell'addestramento del modello, senza però violarne le caratteristiche RID.

Attacchi ai Modelli AI Predittivi

In generale gli attacchi ai modelli AI possono essere classificati in tre classi:

- "Poisoning" (Avvelenamento)
- "Evasion / Adversarial Attacks" (Evasione)
- "Extraction / Reconstruction" (Estrazione / Ricostruzione).

Avvelenamento

Gli attacchi di avvelenamento sono quelli più simili ai comuni attacchi alla sicurezza dei sistemi informatici e avvengono usualmente nella fase di addestramento dei modelli AI, qui da intendersi generalmente come le attività tramite le quali un modello AI *impara* modificando i propri dati interni. L'addestramento può essere fatto in molte modalità: dalla modalità più semplice di una volta sola prima della messa in esercizio del modello AI, a modelli AI che imparano continuamente dai feedback che ricevono dai dati prodotti.

Oggetto dell'attacco possono essere sia i dati utilizzati per l'addestramento del modello AI sia il modello stesso, ovvero il codice che lo implementa.

Avvelenamento del codice

L'attacco forse più tradizionale è quello per la modifica del codice di un modello AI: l'attaccante accedendo in maniera illecita ai sistemi IT che gestiscono il codice del sistema AI, o a quelli di terze parti che sviluppano il codice o anche componenti di supporto come possono essere librerie crittografiche o numeriche (in molti casi *open-source*), o ancora impersonando uno sviluppatore di modelli AI e abusando di codice *open-source*, può modificare il codice aggiungendo ad esempio una *backdoor* o modificando in maniera opportuna ai propri scopi il codice del modello. È stato mostrato che è possibile introdurre modifiche al codice del modello estremamente difficili da identificare nel codice stesso, se non tracciando l'accesso illecito ai sistemi IT. L'effetto di queste modifiche comporta usualmente una violazione dell'Integrità del modello AI, ovvero dell'integrità sia del codice del modello sia del comportamento e dei dati prodotti dal modello. Un attaccante può quindi introdurre nel modello dei comportamenti noti soltanto a lui, ad esempio che a certi specifici dati in ingresso corrispondono certi tipi di dati in uscita. Oppure le modifiche possono alterare le capacità del modello impedendogli di gestire correttamente certi dati in ingresso o limitandone le prestazioni. In alcuni casi le modifiche al codice possono portare alla violazione della Disponibilità del modello AI, ad esempio il modello può utilizzare troppe risorse di calcolo per l'elaborazione di alcuni tipi di dati in ingresso e non è più in grado di completare le elaborazioni nei tempi attesi.

Avvelenamento dei dati di addestramento

I modelli AI di Machine Learning, tra cui la maggior parte dei modelli AI al momento più conosciuti, hanno come caratteristica la necessità di utilizzare grandi quantità di dati per l'addestramento. Ovviamente questo fornisce un nuovo e molto interessante punto di attacco: modificando i dati di addestramento, un attaccante può modificare il comportamento di un modello AI, e quindi i dati da questo prodotti, per raggiungere i propri scopi. Il primo esempio è quello di un modello AI utilizzato per individuare malware o spam: un attaccante può cercare di inserire tra i dati di addestramento il proprio malware o i propri messaggi di spam indicandoli come dati benevoli in modo che, quando operativo, il modello AI non li segnali come malware o spam. In questo caso, effetto dell'attacco è di violare

l'Integrità del modello tramite la modifica dei dati di addestramento.

La necessità di ottenere grandi quantità di dati di addestramento aumenta le possibili modalità di attacco e rende più difficile adottare misure di difesa. Spesso i dati di addestramento sono pubblici, cercati e scaricati da Internet, ma questo rende molto difficile garantire che la sorgente sia qualificata, lecita e non a sua volta oggetto di attacco o direttamente creata dall'attaccante. Ma anche se i dati di addestramento inizialmente sono corretti, spesso vi sono varie terze parti che li raccolgono, li analizzano, formattano ecc. in modo da poter essere utilizzati come dati di addestramento nei modelli AI. Anche tutte queste terze parti possono diventare oggetto di attacco o essere gestite direttamente da un attaccante. Inoltre, tutti coloro coinvolti nella gestione dei dati di addestramento devono garantire la sicurezza di veramente grandi quantità di dati, il che può portare difficoltà nel controllo accessi o anche solo nella protezione dei dati sia in transito sia quando memorizzati (a riposo). Infine, non è facile per chi utilizza i dati per addestrare i propri modelli AI, verificarne l'integrità, qualità e l'assenza di dati maligni.

Per questi motivi è utile considerare la possibilità di utilizzare dati di addestramento sintetici, ovvero creati appositamente per l'addestramento. Oltre a permettere una maggiore protezione dei dati contro attacchi, l'utilizzo di dati sintetici evita anche il rischio di infrangere diritti d'autore, licenze e la riservatezza o privacy delle informazioni. D'altra parte, generare dei sintetici e realistici non è per nulla facile, e non è applicabile a tutti i possibili utilizzi dei modelli AI.

Evasione

Gli attacchi detti di Evasione, anche chiamati "Adversarial Attacks" o "Adversarial Examples", sono forse quelli più specifici dei modelli di Machine Learning perché per lo più connessi al processo di apprendimento dei modelli stessi e al modo in cui le informazioni apprese vengono utilizzate dai modelli durante la loro esecuzione. Il nome di questi tipi di attacchi viene dal fatto che i modelli di Machine Learning possono compiere errori che dal punto di vista umano sono macroscopici, violando completamente il comportamento atteso e oggetto dell'addestramento. In pratica si tratta di dati di ingresso al modello molto simili a quelli di addestramento sui quali il modello sbaglia completamente [Rif. 6], come i famosi esempi dei 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. Questi errori dei modelli di Machine Learning sono delle vulnerabilità che possono essere sfruttate da un attaccante. Se l'attaccante conosce alcune di queste vulnerabilità di un modello, può sfruttarle a proprio favore e quindi "evadere" dal comportamento del modello sfruttando degli "Adversarial Examples". L'esempio più semplice è ancora quello di un sistema AI per il riconoscimento dei volti ad esempio utilizzato in un luogo pubblico. Se l'attaccante è a conoscenza che il modello AI compie un errore nel riconoscimento dei volti se sul volto sono presenti particolari montature di occhiali, le può utilizzare per non essere individuato.

Vi sono molteplici modi in cui un attaccante può individuare queste vulnerabilità: ad esempio se l'attaccante ha accesso completo o ha una copia completa del modello AI addestrato, lo può studiare per individuare gli "Adversarial Examples"; o può studiare modelli simili adattando poi i risultati al modello da attaccare; o se può accedere ai risultati dell'elaborazione dei dati del modello ma non al modello stesso, può adottare un attacco di forza bruta nel quale modifica i dati in ingresso al modello sino a individuare l'errore di elaborazione cercato.

Questi attacchi portano ad una violazione dell'Integrità dei dati prodotti dal un modello AI e sono particolarmente pericolosi perché permettono all'attaccante di sovvertire completamente il comportamento del modello AI con dati che sono molto simili ai dati di addestramento, per cui a priori del tutto leciti. È quindi molto difficile individuare questi attacchi solo analizzando i dati di ingresso al modello AI. Invece per individuare queste vulnerabilità è necessaria un'attività simile a quella ben nota dei Read Team, ovvero un gruppo di esperti di sicurezza che attacca il modello AI per individuarne le vulnerabilità come se fosse un vero attaccante.

Estrazione / Ricostruzione

Questo tipo di attacchi sfrutta delle debolezze o vulnerabilità dei modelli AI per violare la Riservatezza e/o Privacy dei dati di addestramento o della struttura del modello, tramite l'utilizzo del modello stesso. È importante ricordare che in molti modelli AI i dati di addestramento non sono archiviati in modo simbolico (l'archiviazione simbolica permette di individuare la locazione di memoria nella quale è archiviato un singolo dato), ma sub-simbolico e sparso, ovvero un singolo dato è destrutturato e poi distribuito e aggregato con altri dati nella memoria del modello AI. Quindi, almeno ad oggi, non è possibile implementare all'interno di un modello AI un sistema di controllo accessi alle informazioni utilizzate per l'addestramento e ai dati della struttura stessa del modello AI.

Gli attacchi di estrazione e ricostruzione permettono all'attaccante di sottoporre al modello AI degli input formulati in modo tale da ottenere in output informazioni sui dati di addestramento o della struttura stessa del modello AI, e in alcuni casi di ottenere i dati stessi.

Se per i modelli Generativi è abbastanza semplice immaginare come a una ben formulata richiesta il modello possa rispondere fornendo in output come risposta un dato di addestramento, per altri tipi di modelli l'estrazione o ricostruzione delle informazioni presenti nel modello è più complessa, ma spesso possibile. Ad esempio, un attaccante avendo idea del tipo di dati utilizzati per addestrare un modello di Machine Learning di classificazione ma non sapendo esattamente quali dati sono stati utilizzati, 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").

Un approccio per difendersi da questi attacchi utilizza il concetto di "Differential Privacy" che misura la quantità massima di informazioni sui dati di addestramento che un attaccante può estrarre dal modello AI. Sono state proposte varie tecniche che implementano la "Differential Privacy" utilizzando specifiche procedure di preparazione dei dati di addestramento e di addestramento dei modelli AI ma che non sono efficaci rispetto a tutti i possibili tipi di attacchi di Estrazione o Ricostruzione. Altri metodi di difesa utilizzano il monitoraggio e l'applicazione di filtri ai dati in ingresso e a quelli prodotti dai modelli AI.

Attacchi ai Modelli AI Generativi

Oltre agli attacchi ai Modelli AI appena descritti, i modelli AI Generativi e "General Purpose" quali ad esempio i "Large Language Models" (LLM) come GPT, Bard/Gemini, LLaMA ecc., possono essere oggetto di un'altra classe di attacchi chiamata

- "Abuse" (Abuso)

Abuso

Con Abuso si intende proprio l'utilizzo di un modello AI Generativo in modo difforme da quanto previsto o consentito. I modelli AI Generativi possono essere abusati per scopi diversi tra i quali:

- La generazione di malware (generazione di codice maligno);
- La generazione di campagne di Phishing;
- La generazione di informazioni (testi, audio, immagini, video) false ma con grande rassomiglianza a informazioni vere ("Fake News"), inclusa la modifica di dati storici in modo credibile e la generazione di campagne per influenzare il pubblico su temi politici, di discriminazione, di costume ecc.;
- La generazione di dati a supporto di frodi, truffe, azioni di delinquenza, attacchi informatici ecc.
- La generazione di testi (ad esempio per un Chatbot) che convincono un utente a cliccare su un link che porta a un malware, o a divulgare informazioni riservate.

In alcuni casi i modelli AI Generativi possono essere abusati tramite l'uso delle loro interfacce macchina (API) che producono in maniera automatica i dati utilizzati in tempo reale per l'attività illegale. Un esempio di questo può essere la generazione automatica di testi utilizzati da un Chatbot o per uno scambio di messaggi email per portare a termine una truffa.

Questi tipi di attacchi sono per lo più compiuti utilizzando l'interfaccia interattiva, o tramite API, del modello AI Generativo. Queste interfacce, in particolare per i Large Language Models, accettano testi scritti o parlati (o immagini) in linguaggio corrente e non programmatico. L'utente quindi invia una richiesta ("Query") al modello AI che la interpreta, elabora e fornisce il risultato. Come indicato precedentemente, l'elaborazione può includere anche la ricerca di informazioni presso altre fonti, incluso Internet. L'attaccante sfrutta quindi possibili debolezze del modello AI nell'accettazione e interpretazione della richiesta per ottenere dati a priori non permessi. Questi tipi di attacchi vanno usualmente sotto il nome di "Prompt Injection" o "Adversarial Prompting" in quanto l'attaccante inserisce al "Prompt" del modello AI una richiesta il cui scopo è di abusare del modello stesso. A livello teorico, questo tipo di attacchi assomiglia molto agli attacchi di SQL Injection nei quali si modifica la Query SQL in modo che il database esegua un comando non permesso.

Sono stati proposti e utilizzati molti diversi metodi per costruire richieste che portano ad un abuso di un modello AI Generativo tra cui "Prefix injection", "Refusal suppression", "Style injection", "Role-play", "Ignore previous instructions", "Special encoding", "Character transformation", "Word transformation", "Prompt-level obfuscation" ecc. [Rif. 7], e la ricerca in questo ambito è in continuo sviluppo.

Vi sono diverse tecniche per difendersi da attacchi di questo tipo. Una tecnica adottata nel processo di addestramento di questi modelli è chiamata "Reinforcement Learning from Human Feedback (RLHF)" e consiste in una fase ulteriore di addestramento svolto tramite l'intervento diretto di persone specialmente addestrate a insegnare al modello a non generare risposte offensive, dannose o non lecite. Un altra tecnica consiste nell'affiancare al modello AI Generativo un modello AI Predittivo per la classificazione di testi (o di immagini ecc.) che è addestrato a identificare dati malevoli, offensivi o non leciti eventualmente inviati al, o generati dal modello AI Generativo. Il modello AI Predittivo esamina quindi i dati in ingresso e i dati prodotti dal modello AI Generativo e calcola la probabilità che un dato prodotto si configuri come risultato di un abuso del modello AI Generativo. Infine si utilizzano anche

filtri programmati specificatamente per individuare abusi nei dati in ingresso e prodotti dal modello.

Infine va citata anche la vulnerabilità dei modelli AI Generativi indicata genericamente come "Allucinazione" ovvero la generazione di dati che non hanno senso, sono irrealistici o falsi ma che sono presentati come reali e appaiono molto convincenti. Questa vulnerabilità può essere sfruttata da un attaccante per creare campagne di disinformazione, di phishing o truffe e in pratica può risultare molto efficace a questi scopi.

AI Generativa come attaccante autonomo

Sinora i modelli AI sono stati considerati come l'oggetto di un attacco, in realtà modelli AI possono anche essere utilizzati come strumento d'attacco. Questo non è strano in quanto le informazioni di sicurezza informatica in generale e sui modelli AI stessi, possono essere utilizzate sia dai difensori sia dagli attaccanti. Ad esempio, un modello AI può essere utilizzato per studiare i punti deboli di un sistema informatico o di un altro modello AI, e queste informazioni possono essere sfruttate sia da un difensore sia da un attaccante. Un attaccante può quindi adottare modelli AI tra i propri strumenti per eseguire attacchi a sistemi informatici, e un difensore può fare lo stesso. Si ha quindi una situazione in cui i modelli AI svolgono allo stesso tempo il ruolo di attaccante e difensore: in altre parole si ha uno scenario di "guerra" tra i modelli AI utilizzati da attaccanti e difensori.

E, almeno a livello di laboratorio, è già possibile considerare la creazione di "worm" che autonomamente diffondono malware tra modelli AI Generativi [Rif. 8].

Di maggiore interesse è considerare la possibilità che un modello AI Generativo esegua **autonomamente** attacchi ad altri sistemi informatici. Questo scenario richiama immediatamente alla mente molti romanzi e film di fantascienza, spesso con conseguenze tragiche (si pensi ad esempio a *Skynet* della serie *Terminator*). La possibilità teorica esiste, come ad esempio descritto in [Rif. 9]. Una modalità teorica in cui già oggi potrebbe essere possibile un attacco eseguito autonomamente da un modello AI Generativo è la seguente. Un modello AI Generativo quale un Large Language Model (LLM) è addestrato con una enorme quantità di informazioni che tipicamente includono anche informazioni di sicurezza informatica quali le modalità di eseguire attacchi a un sito Web come SQL Injection o Cross Site Scripting (XSS). Inoltre già oggi i modelli più avanzati hanno la capacità di eseguire ricerche in Internet per ottenere informazioni in tempo reale, e il risultato della ricerca può essere utilizzato dal modello AI come nuovo dato di ingresso in un ciclo complesso di elaborazione. In altre parole è tecnologicamente possibile che un modello AI Generativo possa ad esempio utilizzare un attacco SQL Injection per ottenere delle informazioni da un sito Web, ritenendo che questa tecnica sia la più efficace per accedere alle informazioni, senza che l'utente sia a conoscenza né del fatto che il modello sta cercando di estrarre informazioni da uno specifico sito Web né che la tecnica utilizzata per ottenere le informazioni sia un attacco informatico.

Riferimenti

[1] IDC, "IDC Forecasts Artificial Intelligence PCs to Account for Nearly 60% of All PC Shipments by 2027", 7/2/2024, https://www.idc.com/getdoc.jsp?containerId=prUS51851424

[2] La Repubblica "Quel microchip fa Nvidia a tutti. L'Eldorado produttivo promesso dall'AI trascina le

Borse mondiali", 23 febbraio 2024; MarketWatch "Nvidia makes Wall Street history as stock surge adds $277 billion in market cap", https://www.marketwatch.com/story/nvidias-stock-surge-could-add-200-billion-in-market-cap-with-mammoth-growth-on-tap-8d9472d2

[3] MITRE ATLAS™ "Adversarial Threat Landscape for Artificial-Intelligence Systems", https://atlas.mitre.org/

[4] 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

[5] Microsoft "Failure Modes in Machine Learning", https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

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

[7] 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/

[8] S. Cohen, R. Bitton, B. Nassi, "Here Comes the AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications", https://sites.google.com/view/compromptmized

[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

- Adversarial Attacks a Modelli di Machine Learning
- Considerazioni su Modelli di Intelligenza Artificiale Generativa
- Machine Learning, "Deep Fake" ed i rischi in un mondo iperconnesso
- Intelligenza Artificiale / Machine Learning: tra Complessità e Sicurezza

Condividi sui Social Network:

#Adversarial Examples #Intelligenza Artificiale

← PRECEDENTE

Forum Cyber 4.0, a Roma il 3 e 4 Giugno 2024

Articoli simili