# TASSONOMIA E TERMINOLOGIA DEGLI ATTACCHI E DELLE ATTENUAZIONI NELLA INTELLIGENZA ARTIFICIALE

Brevissimo riassunto tratto dal NIST AI 100-2e2023

*Autore: Aldo Pedico – Enterprise Cybersecurity*

*Contatto: pedicoaldo@gmail.com*

## *PREMESSA*

Per chi fosse interessato ad affrontare gli aspetti afferenti le minacce e le relative contromisure in ambito Machine Learning (AI), lo invito a studiare il manuale pubblicato da NIST, nel mese di marzo 2023, pubblicamente disponibile al sito *https://doi.org/10.6028/NIST.AI.100-2e2023.ipd*.

Il manuale ha il titolo: "NIST AI 100-2E2023 ipd – Adversarial Machine Learning – A Taxonomy and Terminology of Attacks and Mitigations".

Questo manuale lo ritengo ben fatto e ricco di spunti riguardanti i temi della cibersicurezza nel contesto della Intelligenza Artificiale.

Come mio esercizio, ho voluto riportare nelle pagine successive una sintesi sperando di fornire un quadro chiaro e succinto (gli aspetti trattati dalla fonte sono abbastanza sviluppati) delle tipologie di attacco effettuate da malintenzionati nei confronti di sistemi di IA e delle relative tecniche di mitigazione da adottare al fine di tutelare gli stessi sistemi.

Il manuale del NIST di fatto è un report all'interno del quale si sviluppano tassonomie di concetti e si definiscono terminologie nel campo dell'apprendimento automatico (ADVERSARIAL MACHINE LEARNING - AML).

Inoltre, il documento originale fornisce metodi corrispondenti per mitigare e gestire le conseguenze degli attacchi e sottolinea le sfide rilevanti da prendere in considerazione nel ciclo di vita dei sistemi di IA.

## INDICE DEGLI ARGOMENTI

Titolo Pag.

# INTRODUZIONE

*L'approccio basato sui dati del ML introduce ulteriori sfide di sicurezza e privacy nelle diverse fasi delle operazioni di ML oltre alle classiche minacce alla sicurezza e alla privacy affrontate dalla maggior parte dei sistemi operativi.*

*Queste sfide in materia di sicurezza e privacy includono il potenziale di manipolazione avversaria dei dati di addestramento, lo sfruttamento avversario delle vulnerabilità del modello per influire negativamente sulle prestazioni di classificazione e regressione ML e persino manipolazioni dannose, modifiche o semplici interazioni con modelli per esfiltrare informazioni sensibili sulle persone rappresentato nei dati o sul modello stesso.*

*Tali attacchi sono stati dimostrati in condizioni reali e la loro sofisticazione e il loro impatto potenziale sono aumentati costantemente.*

*I sistemi di IA sono su una traiettoria di espansione globale pluriennale in continua accelerazione.*

*Tuttavia, nonostante i significativi progressi compiuti dall'IA e dall'apprendimento automatico (ML) in diversi domini applicativi, queste tecnologie sono vulnerabili agli attacchi che possono causare fallimenti con conseguenze disastrose.*

*Ad esempio, nelle applicazioni di visione artificiale per la classificazione delle immagini, casi ben noti di perturbazioni avversarie delle immagini di input hanno causato la deviazione dei veicoli autonomi verso l'opposta corsia di direzione e l'errata classificazione dei segnali di stop come segnali di limite di velocità, la scomparsa di oggetti critici dalle immagini.*

*Allo stesso modo, nel campo medico, dove vengono implementati sempre più modelli ML per assistere i medici, esiste il potenziale per perdite di cartelle cliniche da modelli ML che possono esporre informazioni personali.*

*Gli aggressori possono anche manipolare i dati di addestramento degli algoritmi ML, rendendo così il sistema di IA, addestrato su di esso, vulnerabile agli attacchi.*

## PRINCIPALI TIPI DI ATTACCHI

*Il documento originale fornisce una categorizzazione degli attacchi e delle loro mitigazioni, a partire dai tre principali tipi di attacchi:*

> *1)* EVASIONE (EVASION),
>
> *2)* CONTAMINAZIONE *(*POISONING*) di dati e modelli e*
>
> *3)* PRIVACY *dei dati e dei modelli.*

*Inoltre, esamina gli attacchi contro tutti i metodi di apprendimento praticabili (ad esempio, apprendimento supervisionato, non supervisionato, semi-supervisionato, federato, apprendimento per rinforzo, ecc.) attraverso differenti modalità di gestione dei dati.*

*Fondamentalmente, la metodologia di apprendimento automatico, utilizzata nei moderni sistemi di IA, è suscettibile agli attacchi attraverso le API pubbliche contro il modello fornito e contro le piattaforme sulle quali sono distribuite.*

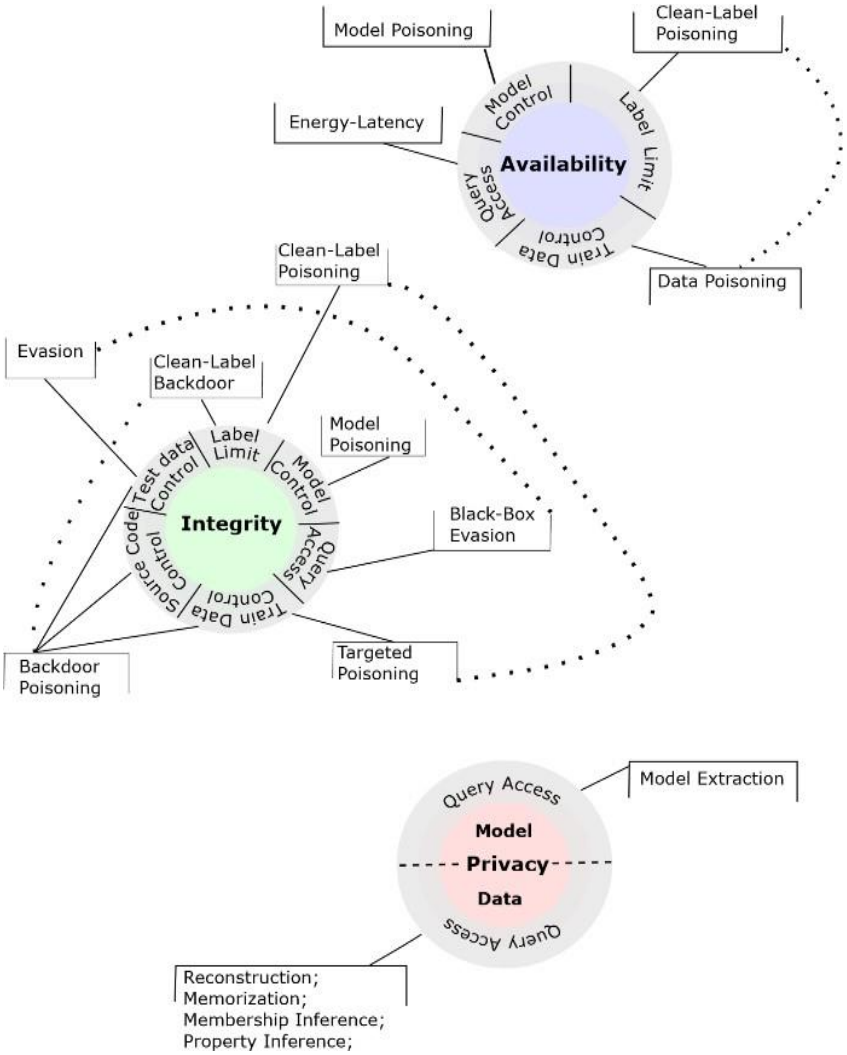
## 1 - CLASSIFICAZIONE DEGLI ATTACCHI

*La Figura 1 introduce una tassonomia degli attacchi nell'apprendimento automatico.*

*Gli obiettivi del malintenzionato sono mostrati come cerchi disgiunti; l'obiettivo del malintenzionato al centro di ogni cerchio:*

1. SUDDIVISIONE DELLA DISPONIBILITÀ *(*AVAILABILITY BREAKDOWN*),*
2. VIOLAZIONI DELL'INTEGRITÀ *(*INTEGRITY*)* e
3. COMPROMISSIONE DELLA PRIVACY *(*PRIVACY COMPROMISE*).*

FIG. 1. TASSONOMIA DEGLI ATTACCHI AI SISTEMI DI IA




*L'apprendimento automatico considera prevalentemente gli* ***attacchi avversari*** *contro i sistemi di IA che potrebbero verificarsi nella fase di addestramento o nella fase di distribuzione ML.*

*Durante la fase di addestramento ML, il malintenzionato potrebbe controllare parte dei dati di addestramento, le relative etichette, i parametri del modello o il codice degli algoritmi ML, svolgendo diversi tipi di* ***attacchi di contaminazione****.*

*Durante la fase di distribuzione ML, il modello ML è già addestrato e l'avversario potrebbe organizzare* ***attacchi di evasione*** *per creare violazioni dell'integrità e modificare il modello ML previsioni, nonché* ***attacchi alla privacy*** *per dedurre informazioni sensibili dai dati di training o sul modello ML.*

## 1.1 - CONOSCENZA DELL'ATTACCANTE

*Un'altra misura per la classificazione degli attacchi è la* quantità di conoscenza *che il malintenzionato ha sul sistema ML.*

### TIPI DI ATTACCHI

*Per questa dimensione, esistono tre tipi principali di attacchi:*

*1)* WHITE-BOX,

*2)* BLACK-BOX *e*

*3)* GRAY-BOX.

#### ATTACCHI WHITE-BOX

*Questi presuppongono che il malintenzionato operi con piena conoscenza del sistema ML, inclusi i dati di addestramento, l'architettura del modello e gli iperparametri del modello.*

*Mentre questi attacchi operano sotto presupposti molto forti, la ragione principale per analizzarli è testare la vulnerabilità di un sistema contro gli avversari e valutare le potenziali mitigazioni.*

#### ATTACCHI BLACK-BOX

*Questi attacchi presuppongono una conoscenza minima del sistema ML.*

*Un avversario potrebbe ottenere l'accesso alle query al modello, ma non dispone di altre informazioni su come viene addestrato il modello.*

*Questi attacchi sono i più pratici poiché presuppongono che l'attaccante non abbia alcuna conoscenza del sistema di IA e utilizzi interfacce di sistema prontamente disponibili per un uso normale.*

#### ATTACCHI GRAY-BOX

*Esistono una serie di attacchi* GRAY-BOX *che acquisiscono la conoscenza avversaria tra attacchi* BLACK-BOX *e* WHITE-BOX.

*Un malintenzionato potrebbe conoscere l'architettura del modello ma non i relativi parametri oppure il modello e i relativi parametri ma non i dati di addestramento.*

*Altri presupposti comuni per gli attacchi* GRAY-BOX *sono che il malintenzionato abbia accesso ai dati distribuiti in modo identico ai dati di addestramento e conosca la rappresentazione delle funzionalità.*

*Quest'ultimo presupposto è importante nelle applicazioni in cui l'estrazione di funzionalità è utilizzata prima di addestrare il modello ML, come la sicurezza informatica, la finanza e l'assistenza sanitaria.*

## 2 - ATTACCHI DI CONTAMINAZIONE E MITIGAZIONI

*Un'altra minaccia rilevante contro i sistemi di apprendimento automatico è il rischio che gli avversari organizzino* POISONING ATTACKS*, i quali sono ampiamente definiti come attacchi durante la fase di addestramento del ML algoritmo.*

*Gli attacchi di contaminazione sono stati ampiamente studiati in diversi domini applicativi:*

- ✓ *sicurezza informatica (per il rilevamento dello spam),*
- ✓ *rilevamento delle intrusioni di rete,*
- ✓ *previsione delle vulnerabilità,*
- ✓ *classificazione del malware,*
- ✓ *visione artificiale,*
- ✓ *elaborazione del linguaggio naturale e*
- ✓ *dati tabulari nei domini sanitari e finanziari.*

*Gli attacchi di contaminazione sfruttano un'ampia gamma di funzionalità avversarie, come:*

- ✓ *la contaminazione dei dati,*
- ✓ *la contaminazione del modello,*
- ✓ *il controllo delle etichette,*
- ✓ *il controllo del codice sorgente e*
- ✓ *il controllo dei dati di test,*

*con la ulteriore suddivisione di sottocategorie di attacchi di contaminazione.*

## 2.1 - ATTACCHI DI CONTAMINAZIONE DELLA DISPONIBILITÀ

*Gli attacchi di* AVAILABILITY POISONING *sono stati progettati anche per l'apprendimento non supervisionato contro il rilevamento delle anomalie basato sul punto focale e il clustering comportamentale per malware.*

*Nell'apprendimento, un avversario può creare un attacco di contaminazione del modello per indurre violazioni della disponibilità nel modello addestrato globalmente.*

### ATTENUAZIONI

*Tra le mitigazioni esistenti, alcune tecniche generalmente promettenti includono:*

- ✓ SANIFICAZIONE DEI DATI DI ADDESTRAMENTO*: questi metodi sfruttano l'intuizione che i campioni alterati sono in genere diversi dai normali campioni di addestramento non controllati dagli avversari.*
  *Pertanto, le tecniche di sanificazione dei dati sono progettate per pulire il set di addestramento e rimuovere i campioni alterati prima che venga eseguito l'addestramento di apprendimento automatico.*

- ✓ FORMAZIONE AFFIDABILE*: un approccio alternativo per mitigare gli attacchi di contaminazione della disponibilità consiste nel modificare l'algoritmo di addestramento ML ed eseguire un addestramento più affidabile.*

## 2.2 - ATTACCHI DI CONTAMINAZIONE MIRATO

*A differenza degli attacchi di disponibilità, gli attacchi di contaminazione mirati inducono un cambiamento nella previsione del modello ML su un numero limitato di campioni mirati.*

*Se l'avversario può controllare la funzione di etichettatura dei dati di addestramento, la contaminazione o alterazione dell'etichetta (LABEL FLIPPING) è un efficace attacco di contaminazione mirato.*

*L'avversario inserisce semplicemente diversi campioni alterati con l'etichetta di destinazione e il modello imparerà dall'etichetta sbagliata.*

*Pertanto, gli attacchi di contaminazione mirati sono studiati principalmente nell'impostazione dell'etichetta pulita in cui il malintenzionato non ha accesso alla funzione di etichettatura.*

## 2.3 - ATTACCHI DI CONTAMINAZIONE DA BACKDOOR

BACKDOOR GENERATING NETWORK (BAN) *è un attacco backdoor dinamico in cui la posizione del* TRIGGER *cambia nei campioni alterati affinché il modello apprenda* TRIGGER *in modo invariato di posizione.*

*I trigger funzionali sono incorporati in tutta l'immagine o cambiano in base all'input.*

*Ad esempio, sono stati alterati i sistemi di riconoscimento facciale usando oggetti fisici come fattori scatenanti, tipo occhiali da sole e orecchini.*

### ALTRE MODALITÀ

*Mentre la maggior parte degli attacchi di contaminazione backdoor sono progettati per applicazioni di visione artificiale, questo vettore di attacco è stato efficace in altri domini applicativi con diverse modalità di dati:*

1. *audio,*
2. *NLP e*
3. *impostazioni di sicurezza informatica.*

AUDIO*: nei sistemi audio, è stato mostrato come un avversario possa iniettare un trigger audio impercettibile nel discorso dal vivo, che è ottimizzato congiuntamente con il modello target durante l'addestramento.*

NLP*: nell'elaborazione del linguaggio naturale, la costruzione di campioni di contaminazione significativi è più impegnativa in quanto i dati di testo sono discreti e il significato semantico delle frasi sarebbe idealmente preservato affinché l'attacco rimanga impercettibile.*

SICUREZZA INFORMATICA*: i primi attacchi di contaminazione nella sicurezza informatica sono stati progettati contro la generazione di firme* WORM *nel 2006 e rilevatori di* SPAM *nel 2008, ben prima del crescente interesse per le* ML ADVERSARIAL.

### ATTENUAZIONI

*La letteratura sulla mitigazione degli attacchi backdoor è vasta rispetto ad altri attacchi di contaminazione.*

*Classi di difesa:*

1. LA SANIFICAZIONE DEI DATI,
2. LA RICOSTRUZIONE DEI TRIGGER,
3. L'ISPEZIONE E LA SANIFICAZIONE DEI MODELLI ED ANCHE
4. LE LORO LIMITAZIONI.

SANIFICAZIONE DEI DATI DI ADDESTRAMENTO: *analogamente agli attacchi alla disponibilità, la sanificazione dei dati può essere applicata al rilevamento di attacchi di contaminazione backdoor.*

RICOSTRUZIONE DEL TRIGGER: *questa classe di mitigazioni mira a ricostruire il* TRIGGER BACKDOOR*, supponendo che si trovi in una posizione fissa nei campioni di addestramento alterati.*

ISPEZIONE E SANIFICAZIONE DEL MODELLO: *l'ispezione del modello analizza il modello ML addestrato prima della sua distribuzione per determinare se è stato alterato.*

## 2.4 - ATTACCHI DI CONTAMINAZIONE DEL MODELLO

*Gli attacchi di contaminazione del modello tentano di modificare direttamente il modello ML addestrato per inserirgli funzionalità dannose.*

*Gli attacchi di contaminazione dei modelli possono causare sia la violazione della disponibilità che dell'integrità nei modelli federati:*

1. *Gli attacchi alla disponibilità che degradano l'accuratezza del modello globale sono stati efficaci, ma di solito richiedono che una grande percentuale di client sia sotto il controllo dell'avversario.*
2. *Gli attacchi mirati di contaminazione dei modelli inducono violazioni dell'integrità su un piccolo set di campioni al momento del test.*
   *Possono essere generati da un attacco di sostituzione del modello o di potenziamento del modello in cui il client compromesso sostituisce l'aggiornamento del modello locale in base all'obiettivo.*
3. *Gli attacchi di contaminazione del modello backdoor introducono un trigger tramite aggiornamenti client dannosi per indurre l'errata classificazione di tutti i campioni al momento del test.*

*Gli attacchi di contaminazione del modello sono possibili anche in scenari di* SUPPLY CHAIN *in cui i modelli o i componenti del modello forniti sono stati alterati.*

### ATTENUAZIONI

*La maggior parte delle attenuazioni tenta di identificare ed escludere gli aggiornamenti dannosi durante l'esecuzione dell'aggregazione sul server.*

*Il* GRADIENT CLIPPING *e la* PRIVACY DIFFERENZIALE *hanno il potenziale per mitigare gli attacchi di contaminazione del modello, ma di solito riducono l'accuratezza e non forniscono una completa mitigazione.*

## 3 - ATTACCHI ALLA PRIVACY

*L'obiettivo degli attacchi di* RECONSTRUCTION *è quello di decodificare le informazioni relative ad un utente o a dati sensibili dell'infrastruttura.*

*Un attacco alla privacy meno devastante è quello dell'inferenza dell'appartenenza* (MEMBERSHIP INFERENCE) *in cui un avversario può determinare se un particolare record è stato incluso nel set di dati utilizzato per il calcolo di informazioni statistiche o l'addestramento di un modello di ML.*

*Gli attacchi di inferenza dei membri* (MEMBERSHIP INFERENCE) *sono stati introdotti per la prima volta per i dati genomici.*

### 3.1 – ATTACCHI DI RICOSTRUZIONE DEI DATI

*Gli attacchi di* RICOSTRUZIONE *dei dati sono gli attacchi alla privacy più preoccupanti in quanto hanno la capacità di recuperare i dati di un individuo da informazioni statistiche aggregate rilasciate.*

### 3.2 - ATTACCHI ALLA MEMORIA

*Gli attacchi alla memoria sono una potente classe di tecniche che consentono a un avversario di estrarre dati di addestramento da modelli ML generativi, come i modelli linguistici.*

## 3.3 - ATTACCHI DI INFERENZA DELL'APPARTENENZA

*Nel* MEMBERSHIP INFERENCE*, l'obiettivo del malintenzionato è determinare se un particolare record o campione di dati fa parte del set di dati di addestramento utilizzato per l'algoritmo statistico o ML.*

*Un'altra tecnica è quella dei modelli ombra* (SHADOW MODELS)*, che addestra un meta-classificatore su esempi dentro e fuori il set di addestramento ottenuto dall'addestramento di migliaia di* SHADOW *dei modelli ML sulla stessa attività del modello originale.*

*Un metodo intermedio che sta attualmente raggiungendo prestazioni all'avanguardia in termini di metrica* AREA UNDER THE CURVE (AUC) *è l'attacco* LIRA*, che addestra un numero minore di modelli ombra per apprendere la distribuzione dei* LOGIT *del modello su esempi dentro e fuori il set di addestramento.*

## 3.4 - ATTACCHI DI ESTRAZIONE DAL MODELLO

*Diverse tecniche per montare attacchi di* MODEL EXTRACTION *sono state introdotte nella letteratura.*

1. *La prima tecnica è quella della* DIRECT EXTRACTION *basata sulla formulazione matematica delle operazioni eseguite nelle* DEEP NEURAL NETWORKS*, che consente all'avversario di calcolare il modello* WEIGHTS ALGEBRAICALLY.
2. *La seconda tecnica è quella di utilizzare metodi di* APPRENDIMENTO PER L'ESTRAZIONE.
3. *La terza tecnica è l'uso delle informazioni* SIDE CHANNEL *per l'estrazione del modello.*

## 3.5 - ATTACCHI DI INFERENZA DELLA PROPRIETÀ

*Negli attacchi di* PROPERTY INFERENCE*, il malintenzionato tenta di apprendere informazioni globali sulla distribuzione dei dati di training interagendo con un modello ML.*

## 3.6 - ATTENUAZIONI

*La* DIFFERENTIAL PRIVACY (DP) *è una definizione estremamente forte di privacy che garantisce un limite su quanto un malintenzionato con accesso all'output dell'algoritmo può imparare su ciascuno record individuale nel set di dati.*

*Per definizione,* DP *fornisce mitigazione contro gli attacchi di ricostruzione, la memorizzazione dei dati di addestramento e gli attacchi di inferenza dell'appartenenza.*

# 4 - DISCUSSIONE E SFIDE RIMANENTI

## 4.1 - TRADE-OFFS TRA GLI ATTRIBUTI DI UN'IA AFFIDABILE

L'AFFIDABILITÀ DI UN SISTEMA DI IA DIPENDE DA TUTTI GLI ATTRIBUTI CHE LO CARATTERIZZANO.

*Ad esempio, è improbabile che ci si attenda un sistema di IA accurato se facilmente suscettibile di exploit avversari. Allo stesso modo, è improbabile che un sistema di IA che produce risultati dannosi o ingiusti sia attendibile anche se è robusto.*

*Nei casi in cui l'equità è importante e la privacy è necessaria, è utile considerare il compromesso tra privacy ed equità.*

*Sfortunatamente, non è possibile massimizzare contemporaneamente le prestazioni del sistema di IA rispetto a questi attributi.*

*Ad esempio, i sistemi di IA ottimizzati per la sola precisione tendono a sotto-performare la robustezza e l'equità. Al contrario, un sistema di IA ottimizzato per la robustezza contro l'avversario può mostrare una precisione inferiore e risultati di equità deteriorati.*

## 4.2 - OLTRE I MODELLI E I DATI

*Tradizionali attacchi alle* CHATBOT *si sono concentrati sul sopraffare le* CHATBOT *con input tossici al fine di per alterare il suo comportamento.*

*Recentemente, attacchi specifici che utilizzano* "PROMPT INJECTIONS" *sono emersi come modi efficaci per innescare comportamenti scorretti nel* BOT.

*Inoltre, potenzialmente gli attacchi (i quali potrebbero compromettere la funzione del componente di dialogo e modificarne maliziosamente l'oggetto della conversazione per l'utente ignaro) possono portare la* CHATBOT *a offrire consigli fuorvianti o addirittura dannosi.*