

Analisi forense di modelli AI/ML: audit, rilevamento bias e manomissioni



La nuova frontiera dell'investigazione digitale nell'era dell'intelligenza artificiale

Di fronte alla proliferazione di sistemi di machine learning in settori critici, emerge una disciplina investigativa inedita: la forensics applicata ai modelli neurali. Un territorio inesplorato dove convergono cybersecurity, data science e compliance normativa.

Introduzione: quando l'algoritmo diventa scena del crimine

L'adozione massiva di sistemi di intelligenza artificiale in ambiti decisionali sensibili – dal *credit scoring* alla diagnostica medica, dalla selezione del personale ai sistemi di giustizia predittiva – ha generato una questione epistemologica e operativa senza precedenti: come si conduce un'indagine forense su un'entità matematica?

Il NIST AI Risk Management Framework, pubblicato il 26 gennaio 2023, ha formalizzato per la prima volta un approccio sistematico alla governance dei rischi AI, introducendo i concetti di *trustworthiness* e *accountability* come pilastri della gestione algoritmica. Tuttavia, il framework evidenzia una lacuna critica: mancano metodologie consolidate per l'investigazione post-incidente su modelli compromessi o malfunzionanti.

La forensics tradizionale opera su artefatti digitali statici – file system, log, memoria volatile. I modelli di *machine*

learning, invece, sono oggetti dinamici: matrici di pesi che codificano pattern appresi da dati che potrebbero non essere più disponibili, addestrati attraverso processi stocastici difficilmente riproducibili. Questa natura intrinseca pone sfide metodologiche che richiedono un ripensamento radicale degli strumenti investigativi.

Il Framework MITRE ATLAS: Tassonomia delle Minacce Adversariali

Il MITRE ATLAS (*Adversarial Threat Landscape for Artificial-Intelligence Systems*) rappresenta il tentativo più strutturato di mappare il panorama delle minacce specifiche per i sistemi AI. Modellato sul celebre ATT&CK framework per la cybersecurity tradizionale, ATLAS cataloga tattiche, tecniche e procedure (TTP) utilizzate da attori malevoli per compromettere sistemi di *machine learning*.

La tassonomia ATLAS identifica vettori d'attacco che spaziano dalla fase di raccolta dati (*data poisoning*) all'inferenza in produzione (*evasion attacks*), passando per tecniche di estrazione della proprietà intellettuale (*model stealing*) e violazione della privacy (*membership inference*). Ciascuna di queste categorie richiede approcci forensi distinti.

Il *data poisoning*, ad esempio, lascia tracce nei pattern statistici del dataset di training: distribuzioni anomale, cluster inattesi, correlazioni spurie. L'investigatore forense deve ricostruire la genealogia dei dati, verificando l'integrità della pipeline di acquisizione e *preprocessing*. Gli attacchi di *evasion*, invece, si manifestano esclusivamente in fase di inferenza, richiedendo l'analisi delle query ricevute dal modello e delle relative risposte per identificare pattern di *probing* sistematico.

Architettura di un audit forense: stratificazione metodologica

Un audit forense completo su un modello AI/ML richiede un'analisi stratificata che attraversa l'intero stack tecnologico e procedurale. L'OWASP Machine Learning Security Top 10 fornisce una base tassonomica per strutturare l'investigazione, identificando le vulnerabilità più critiche nei sistemi di apprendimento automatico.

Livello 1: Data Provenance e integrità

La catena di custodia dei dati di training costituisce il fondamento di qualsiasi investigazione. L'assenza di meccanismi robusti di *data lineage* rappresenta una delle debolezze sistemiche più diffuse. Tecnologie emergenti come i *data contracts* e i sistemi di versionamento basati su hash crittografici (implementati in strumenti come DVC – Data Version Control) permettono di ricostruire la storia evolutiva di un dataset.

L'investigatore deve verificare la provenienza delle fonti dati primarie, le trasformazioni applicate durante il *preprocessing*, l'eventuale presenza di dati sintetici o aumentati, e i meccanismi di anonimizzazione e pseudonimizzazione utilizzati.

Livello 2: training pipeline e riproducibilità

La riproducibilità degli esperimenti di *machine learning* è notoriamente problematica. Studi recenti indicano che solo una frazione delle pubblicazioni scientifiche in AI fornisce codice e dati sufficienti per replicare i risultati dichiarati. Un'analisi dei paper pubblicati a NeurIPS ha rilevato che solo il 42% includeva codice, e appena il 23% forniva link ai dataset utilizzati.

Dal punto di vista forense, questa opacità rappresenta un ostacolo significativo. L'investigatore deve raccogliere le configurazioni di *hyperparameter*, i *seed* dei generatori di numeri pseudocasuali, le versioni esatte delle librerie utilizzate, le specifiche hardware (particolarmente rilevanti per training distribuito su GPU/TPU) e i *checkpoint* intermedi del modello.

Livello 3: Model Introspection

L'analisi diretta dei parametri del modello richiede tecniche di *interpretability* ed *explainability*. Il framework Captum sviluppato da Meta AI offre implementazioni di metodi come *Integrated Gradients*, SHAP (*SHapley Additive exPlanations*) e *Layer-wise Relevance Propagation*.

Queste tecniche permettono di identificare feature anomale con peso eccessivo, rilevare neuroni o layer che codificano pattern inattesi, e mappare le regioni di *input space* dove il modello esibisce comportamenti aberranti.

Rilevamento del bias algoritmico: oltre la metrica

Il bias nei sistemi AI rappresenta una delle aree più studiate e al contempo più fraintese. La AI Fairness 360 toolkit sviluppata da IBM Research implementa oltre 70 metriche di *fairness*, rivelando una verità scomoda: diverse definizioni di equità sono matematicamente incompatibili tra loro.

I lavori indipendenti di Alexandra Chouldechova (2017) e Jon Kleinberg con colleghi (2016) hanno dimostrato formalmente che, salvo casi degeneri, non è possibile soddisfare simultaneamente *calibration*, *equalized odds* e *predictive parity*. Questa impossibilità matematica ha profonde implicazioni forensi: l'investigatore deve determinare quale definizione di *fairness* era appropriata per il contesto

applicativo e se il sistema rispettava tale standard.

Tipologie di bias investigabili

Il bias può infiltrarsi nel sistema a molteplici livelli.

Il **bias nei dati storici** riflette discriminazioni pregresse: un sistema di screening curricolare addestrato su decisioni di assunzione passate perpetuerà inevitabilmente i pregiudizi dei selezionatori umani.

Il **bias di rappresentazione** colpisce i sottogruppi demografici sottorappresentati nel training set, causando degradazione delle performance. Lo studio Gender Shades di Joy Buolamwini e Timnit Gebru, condotto presso il MIT Media Lab e pubblicato nel 2018, ha documentato disparità drammatiche nell'accuratezza di sistemi commerciali di classificazione di genere: il tasso di errore per donne con carnagione scura raggiungeva il 34,7%, contro lo 0,8% per uomini con carnagione chiara.

Il **bias di proxy** si verifica quando l'algoritmo apprende correlazioni spurie tra variabili apparentemente neutre e attributi protetti. Il codice postale, ad esempio, può fungere da proxy per etnia e status socioeconomico.

Il **bias di measurement** consiste in errori sistematici nella raccolta dati che colpiscono differenzialmente sottopopolazioni specifiche.

Manomissioni e backdoor: l'investigazione dell'intenzionalità

La distinzione tra malfunzionamento accidentale e manomissione deliberata rappresenta il cuore dell'investigazione forense criminale. I *backdoor attack* su modelli neurali costituiscono

una minaccia particolarmente insidiosa: l'attaccante inserisce durante il training un *trigger* nascosto che attiva comportamenti malevoli solo in presenza di input specifici.

Ricerche accademiche hanno dimostrato la possibilità di impiantare *backdoor* in modelli di visione artificiale che si attivano esclusivamente in presenza di pattern visivi impercettibili all'occhio umano. Il modello mantiene performance normali su input standard, superando qualsiasi test funzionale convenzionale.

Tecniche di rilevamento backdoor

La comunità scientifica ha sviluppato diverse metodologie per l'identificazione di *backdoor*.

Neural Cleanse analizza le perturbazioni minime necessarie per indurre misclassificazione verso ciascuna classe, identificando trigger potenziali come perturbazioni anomalmente piccole.

STRIP (*STRong Intentional Perturbation*) sfrutta la robustezza eccessiva degli input con *trigger* a perturbazioni casuali per identificare campioni sospetti.

Meta Neural Analysis addestra meta-classificatori per distinguere modelli puliti da modelli compromessi analizzando le attivazioni neurali.

Spectral Signature Analysis identifica anomalie nello spettro delle rappresentazioni latenti indotte dal training su dati avvelenati.

Il quadro normativo: EU AI Act e obblighi di audit

L'EU AI Act (Regolamento UE 2024/1689), entrato in vigore il 1° agosto 2024, introduce per la prima volta obblighi

vincolanti di valutazione della conformità per sistemi AI ad alto rischio. L'articolo 9 impone l'implementazione di sistemi di gestione del rischio lungo l'intero ciclo di vita del sistema AI.

I requisiti normativi con rilevanza forense includono:

Logging obbligatorio: i sistemi ad alto rischio devono implementare capacità di registrazione automatica che consentano la tracciabilità delle operazioni durante il ciclo di vita del sistema (Art. 12).

Documentazione tecnica: deve essere mantenuta documentazione aggiornata che dimostri la conformità ai requisiti del regolamento, incluse informazioni su dati di training, scelte progettuali e metriche di valutazione (Art. 11).

Supervisione umana: devono essere implementate interfacce che permettano l'*oversight* umano effettivo, inclusa la possibilità di interpretare gli output del sistema (Art. 14).

La violazione di questi obblighi può comportare sanzioni fino a 35 milioni di euro o il 7% del fatturato mondiale annuo, rendendo la capacità di dimostrare *compliance* attraverso audit forensi una necessità operativa.

Strumenti e metodologie operative

L'ecosistema degli strumenti forensi per AI è in rapida evoluzione. Alcune soluzioni emergenti meritano attenzione.

Model Cards e Data Sheets

Il concetto di Model Cards introdotto da Google Research nel 2019 propone un formato standardizzato per documentare le caratteristiche, limitazioni e contesti d'uso appropriati di un modello. Analogamente, i *Datasheets for Datasets* proposti da Timnit Gebru e colleghi forniscono un template per documentare provenienza, composizione e bias noti dei dataset.

Questi artefatti documentali rappresentano prove cruciali in un'investigazione forense, permettendo di valutare se il deployment del sistema rispettava i limiti dichiarati.

MLflow e Experiment Tracking

Piattaforme come MLflow permettono il tracking automatico di esperimenti, includendo parametri, metriche, artefatti e metadati. Dal punto di vista forense, i log di MLflow possono ricostruire la cronologia completa dello sviluppo di un modello.

Differential Privacy e Membership Inference

Le tecniche di *Differential Privacy* forniscono garanzie matematiche sulla protezione dei dati individuali nel training. Strumenti come Opacus di Meta implementano training differenzialmente privato per PyTorch.

L'investigatore forense può utilizzare attacchi di *membership inference* – tecniche che determinano se uno specifico record era presente nel training set – per verificare il rispetto delle politiche di privacy dichiarate.

Casi studio: lezioni dal campo

Caso COMPAS e bias razziale

Il sistema COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), utilizzato nel sistema giudiziario statunitense per valutare il rischio di recidiva, rappresenta un caso paradigmatico. L'inchiesta di ProPublica del maggio 2016 ha rivelato disparità significative: i defendant afroamericani avevano quasi il doppio delle probabilità di essere erroneamente classificati come ad alto rischio rispetto ai bianchi, pur non recidivando successivamente.

L'analisi forense ha coinvolto la ricostruzione del dataset di training, l'identificazione delle variabili proxy utilizzate e la quantificazione delle disparità di trattamento. Questo caso ha catalizzato l'intero campo della AI *fairness* e il successivo dibattito accademico sul teorema di impossibilità.

Amazon e il recruiting algoritmico

Nell'ottobre 2018, Reuters ha rivelato che Amazon aveva abbandonato un sistema di screening automatico dei curricula dopo aver scoperto che penalizzava sistematicamente le candidate donne. L'investigazione interna ha identificato come causa il training su dati storici di assunzione che riflettevano il bias di genere prevalente nel settore tech: il sistema aveva appreso a sfavorire curricula contenenti termini come «women's» o provenienti da college femminili.

Il caso illustra l'importanza dell'audit preventivo e della verifica continua dei sistemi in produzione.

Prospettive future: verso una forensics AI matura

L'evoluzione della disciplina richiede progressi su molteplici fronti.

Standardizzazione metodologica: l'ISO/IEC JTC 1/SC 42 sta sviluppando standard internazionali per l'AI, inclusi framework di testing e valutazione. L'adozione di metodologie standardizzate permetterà la comparabilità delle analisi forensi.

Certificazione professionale: emergerà la necessità di figure professionali certificate nell'audit di sistemi AI, con competenze ibride tra *data science*, *cybersecurity* e *compliance* normativa.

Automazione dell'audit: strumenti di *continuous monitoring*

integrati nelle pipeline MLOps permetteranno la rilevazione precoce di *drift*, anomalie e potenziali manomissioni.

Forensics federata: l'addestramento federato, dove i dati rimangono distribuiti presso le sorgenti, pone sfide forensi inedite. Come si investiga un modello addestrato su dati che non possono essere centralizzati per ragioni di privacy?

Conclusione: l'imperativo dell'*accountability* algoritmica

L'analisi forense dei sistemi AI non è un lusso accademico ma una necessità operativa urgente. Man mano che questi sistemi assumono ruoli sempre più critici nelle infrastrutture sociali ed economiche, la capacità di investigarne il comportamento, identificarne le anomalie e attribuire responsabilità diventa fondamentale per mantenere la fiducia pubblica.

La convergenza tra cybersecurity tradizionale, *data science* e *compliance* normativa sta generando una nuova disciplina professionale. I forensic analyst del futuro dovranno padroneggiare non solo le tecniche investigative classiche, ma anche la teoria dell'apprendimento automatico, la statistica bayesiana e i framework normativi emergenti.

L'EU AI Act rappresenta un punto di svolta: per la prima volta, la legge richiede esplicitamente capacità di audit e tracciabilità per i sistemi AI ad alto rischio. Le organizzazioni che non svilupperanno queste competenze si troveranno esposte a rischi legali, reputazionali e operativi significativi.

La forensics AI non è la risposta a tutti i problemi dell'intelligenza artificiale, ma è una componente essenziale di un ecosistema di governance più ampio. Solo attraverso la capacità di investigare, comprendere e attribuire responsabilità potremo costruire sistemi AI che meritino la

fiducia che sempre più spesso riponiamo in essi.

Condividi sui Social Network:

<https://www.ictsecuritymagazine.com/articoli/analisi-forense-ai/>