Intelligenza Artificiale Spiegabile (XAI): quando la trasparenza diventa requisito di sicurezza informatica



L’**Intelligenza Artificiale Spiegabile (XAI)** rappresenta oggi una necessità strategica per la cybersecurity. La crescente adozione di sistemi AI in ambiti critici sta spostando il dibattito dalla pura efficienza prestazionale alla necessità di comprendere come e perché un algoritmo arrivi a una determinata decisione. Nel contesto della sicurezza informatica, questa evoluzione non rappresenta un vezzo accademico ma una necessità operativa e, sempre più spesso, un obbligo normativo.

Quando un sistema di rilevamento delle intrusioni basato su machine learning identifica un’anomalia classificandola come minaccia critica, gli analisti SOC non possono accontentarsi di un responso binario. Devono comprendere quali caratteristiche del traffico di rete hanno innescato l’allerta, quali pattern sono stati riconosciuti, quale livello di confidenza caratterizza quella specifica inferenza. Questa esigenza diventa ancora più stringente quando l’output dell’algoritmo guida decisioni che possono bloccare interi segmenti di rete, isolare sistemi produttivi o innescare procedure di incident response che coinvolgono risorse umane ed economiche significative.

Intelligenza Artificiale Spiegabile (XAI): colmare il divario tra

complessità algoritmica e comprensibilità umana

L'**Explainable AI** nasce per colmare il divario tra la complessità crescente dei modelli di apprendimento automatico e la necessità umana – e spesso legale – di comprenderne il funzionamento. Non si tratta semplicemente di "aprire la scatola nera", metafora ormai abusata ma ancora efficace: l'obiettivo è costruire sistemi che siano intrinsecamente interpretabili o dotarli di meccanismi che permettano di ricostruire a posteriori il ragionamento algoritmico.

Come evidenziato nell'articolo "Nuovi modelli di IA: Explainable AI e Hybrid AI", l'opacità dell'AI rappresenta un ostacolo fondamentale: "In ambito scientifico, l'opacità dell'AI si riferisce all'incapacità di comprendere il processo decisionale interno di un sistema di intelligenza artificiale, specialmente quelli basati su algoritmi complessi come il deep learning."

Nel panorama della **digital forensics**, questa capacità assume contorni ancora più definiti. Un'evidenza digitale prodotta da un sistema AI può essere contestata in sede giudiziaria se non è possibile dimostrare la tracciabilità del processo decisionale. Il perito deve poter spiegare non solo cosa ha rilevato l'algoritmo, ma attraverso quale logica inferenziale è giunto a quella conclusione. La catena di custodia dell'evidenza digitale si estende così al modello stesso e ai suoi parametri, richiedendo nuove competenze e protocolli di documentazione.

Il quadro normativo europeo: AI Act e NIS2

AI Act e requisiti di trasparenza

Dal punto di vista normativo, il GDPR aveva già introdotto il concetto di diritto alla spiegazione nelle decisioni automatizzate che producono effetti giuridici o incidono significativamente sull'interessato (articolo 22, artt. 13, 14, 15 e considerando 71). L'AI Act europeo (Regolamento UE 2024/1689), entrato in vigore il 1° agosto 2024 e pienamente applicabile dal 2 agosto 2026, rafforza ulteriormente questi requisiti classificando i sistemi AI in base al rischio e imponendo obblighi di trasparenza crescenti.

Per i sistemi ad alto rischio utilizzati in ambito di **sicurezza delle infrastrutture critiche**, identificazione biometrica o gestione di dati sensibili, la **spiegabilità** non è più opzionale: è un requisito di conformità. L'articolo 19 dell'AI Act (Automatically Generated Logs) impone agli operatori di sistemi ad alto rischio di mantenere log automatici che registrino le operazioni del sistema durante il suo ciclo di vita, permettendo la tracciabilità delle decisioni prese dal sistema AI, includendo input, output e timestamp.

Direttiva NIS2 e gestione del rischio cyber

La Direttiva NIS2, recepita in Italia con il Decreto Legislativo 138/2024 (entrato in vigore il 16 ottobre 2024), introduce un ulteriore livello di complessità. Gli enti essenziali e importanti sono chiamati a implementare misure tecniche e organizzative adeguate per gestire i rischi legati alla sicurezza di reti e sistemi informativi. Quando queste misure incorporano componenti di AI – dalla detection automatizzata delle minacce alla gestione dinamica degli accessi – la capacità di spiegare il comportamento di questi sistemi diventa parte integrante della dimostrazione di conformità verso le autorità competenti, in Italia rappresentate dall'ACN.

Per i **security operation center** che integrano strumenti AI,

questo significa ripensare l'intera architettura di logging e audit trail, garantendo che ogni decisione automatizzata – dal blocco di un IP alla quarantena di un file sospetto – sia documentata in modo da permettere una ricostruzione ex-post del ragionamento algoritmico.

Framework tecnici e standard emergenti per l'XAI

NIST AI Risk Management Framework

Sul piano operativo, l'industria sta convergendo verso framework standardizzati per l'implementazione dell'XAI. Il NIST AI Risk Management Framework, pubblicato il 26 gennaio 2023, dedica ampio spazio alla trasparenza e alla spiegabilità come elementi fondamentali della gestione del rischio AI. Il framework propone un approccio strutturato che include la documentazione del sistema, la validazione delle performance, il monitoraggio continuo e la gestione degli incidenti legati all'AI.

ISO/IEC 42001: standard per l'AI Management System

A livello europeo, lo standard ISO/IEC 42001 sui sistemi di gestione dell'intelligenza artificiale, pubblicato nel dicembre 2023, fornisce requisiti e linee guida per stabilire, implementare, mantenere e migliorare continuamente un AI Management System. Questo standard, che sarà probabilmente referenziato nei processi di certificazione previsti dall'AI Act, include requisiti specifici sulla tracciabilità e la spiegabilità dei sistemi AI, particolarmente per applicazioni in contesti critici.

MITRE ATT&CK e ATLAS

Nel contesto della **cybersecurity**, il framework MITRE ATT&CK ha iniziato a integrare considerazioni sui sistemi di AI utilizzati sia dagli attaccanti sia dai difensori. La matrice

ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) documenta tattiche e tecniche specifiche per compromettere sistemi di machine learning, includendo attacchi che sfruttano proprio la mancanza di trasparenza per iniettare bias, avvelenare i dataset di training o eludere i meccanismi di detection.

Come spiegato nell'articolo "Il problema del BIAS nei dataset e della XAI", "Una delle contromisure ed anche un sistema per capire meglio l'output di un algoritmo di IA è quello della XAI (eXplainable Artificial Intelligence), ossia l'intelligenza artificiale spiegabile. Gli algoritmi sono sostanzialmente delle black box, delle scatole nere nelle quali entrano i dati di input, che vengono elaborati per generare degli output, il motivo per cui sia stata generata una risposta rispetto ad un'altra non è facilmente comprensibile."

La supply chain dell'AI e il rischio di opacità stratificata

Una dimensione spesso trascurata riguarda la supply chain dell'intelligenza artificiale. Molte organizzazioni non sviluppano internamente i propri modelli di AI ma utilizzano API di terze parti, modelli pre-addestrati o soluzioni commerciali proprietarie. Questa dipendenza introduce quello che potremmo definire un "**rischio di opacità stratificata**": anche quando il fornitore offre interfacce di spiegabilità, queste possono operare a un livello di astrazione tale da mascherare problematiche più profonde nel modello sottostante.

Il caso dei **modelli foundation** illustra perfettamente questo dilemma. Quando un'organizzazione utilizza un LLM commerciale per automatizzare l'analisi di log di sicurezza o la correlazione di threat intelligence, si affida a un sistema la cui architettura interna, i dati di training e i processi di fine-tuning sono spesso protetti da segreto industriale. Come

possono i CISO garantire che questi sistemi non abbiano appreso bias dai dati di training? Come verificare che non siano suscettibili a prompt injection o data poisoning?

La risposta richiede un approccio contrattuale e tecnico integrato. I contratti di procurement devono includere clausole specifiche sui livelli di trasparenza garantiti, sulle modalità di audit algoritmico, sulla possibilità di testare il modello con dataset controllati. Sul piano tecnico, emergono metodologie di testing adversarial che permettono di sondare i confini comportamentali del modello, identificando potenziali vulnerabilità senza necessariamente accedere ai suoi parametri interni.

XAI e threat intelligence: il caso del ransomware-as-a-service

L'applicazione dell'XAI nella **threat intelligence** merita un approfondimento specifico. Consideriamo il fenomeno del ransomware-as-a-service, dove gruppi criminali operano con modelli di business strutturati e tattiche in continua evoluzione. I sistemi di AI vengono utilizzati per correlare indicatori di compromissione attraverso fonti eterogenee – dark web, telemetria di rete, rapporti di incident response – identificando pattern emergenti e prevedendo possibili vettori di attacco.

Un sistema XAI efficace in questo contesto non si limita a segnalare che un particolare hash di file è associato a una campagna ransomware. Dovrebbe spiegare perché quella associazione è significativa: quali caratteristiche comportamentali del malware sono state osservate, quale similarità strutturale esiste con campioni noti, quale correlazione temporale emerge con attività di reconnaissance su altri target. Questa ricchezza informativa trasforma il sistema AI da strumento di allerta a moltiplicatore di intelligence, permettendo agli analisti di comprendere non

solo la minaccia immediata ma il contesto strategico in cui si inserisce.

Alcuni fornitori di security intelligence stanno sviluppando dashboard di "**explainability**" che visualizzano graficamente il processo decisionale dell'AI. Grafici di feature importance mostrano quali attributi del malware hanno pesato maggiormente nella classificazione. Mappe di attivazione delle reti neurali evidenziano quali porzioni di codice hanno attirato l'attenzione del modello. Linee temporali correlano l'evento specifico con pattern storici osservati. Questi strumenti trasformano il responso algoritmico in un'esperienza investigativa, dove la macchina non sostituisce l'analista ma ne amplifica le capacità cognitive.

Come descritto nell'articolo "AI offensiva e difensiva: strategie di cybersecurity nel 2025", "L'AI spiegabile (XAI) offre una soluzione promettente fornendo interpretabilità e trasparenza, permettendo ai professionisti della sicurezza di comprendere meglio, fidarsi e ottimizzare i modelli IDS."

Il paradosso della spiegabilità nell'era del quantum computing

Un tema emergente, ancora poco esplorato ma destinato ad acquisire rilevanza, riguarda l'intersezione tra XAI e **quantum computing**. Man mano che algoritmi quantistici vengono integrati in sistemi di AI – sia per accelerare il training di modelli complessi sia per sviluppare nuove architetture di quantum machine learning – la sfida della spiegabilità assume dimensioni ulteriori.

I computer quantistici operano secondo principi che sfuggono all'intuizione classica: sovrapposizione, entanglement, interferenza. Quando questi fenomeni vengono sfruttati per costruire modelli di machine learning, il concetto stesso di "spiegazione" diventa problematico. Come si può spiegare una

decisione che emerge da uno stato quantistico collassato al momento della misurazione? Come documentare un processo inferenziale che sfrutta parallelismo quantistico attraverso spazi di Hilbert ad alta dimensionalità?

Paradossalmente, proprio mentre la comunità della cybersecurity si prepara alla transizione post-quantum per proteggere le comunicazioni dalla crittoanalisi quantistica, potremmo trovarci ad affrontare sistemi di **AI quantistica** intrinsecamente meno spiegabili dei loro predecessori classici. Questa prospettiva rende ancora più urgente lo sviluppo di framework concettuali e metodologie di XAI che possano adattarsi a paradigmi computazionali radicalmente diversi.

Certificazione e attestazione della spiegabilità

L'AI Act introduce il concetto di "**organismi notificati**" che avranno il compito di valutare la conformità dei sistemi AI ad alto rischio prima della loro immissione sul mercato. Questo meccanismo, simile a quanto già avviene per dispositivi medici o macchinari industriali, solleva la questione di come certificare la spiegabilità di un sistema AI.

A differenza di requisiti funzionali tradizionali, verificabili attraverso test deterministici, la spiegabilità ha una componente soggettiva: una spiegazione può essere tecnicamente accurata ma incomprensibile per l'utente finale, o viceversa intuitiva ma scientificamente imprecisa. Gli organismi di certificazione dovranno sviluppare metodologie di valutazione che bilancino accuratezza tecnica e comprensibilità contestuale, tenendo conto del profilo degli stakeholder che utilizzeranno le spiegazioni.

Alcune proposte in discussione prevedono l'adozione di "**benchmark di spiegabilità**" standardizzati, dove il sistema AI

viene testato su scenari di riferimento e le sue spiegazioni valutate secondo metriche oggettive: completezza, consistenza, contrastività (capacità di spiegare perché è stata presa una decisione invece di un'altra), continuità (stabilità delle spiegazioni rispetto a variazioni minime dell'input). Tuttavia, il dibattito rimane aperto su quali di queste proprietà dovrebbero essere prescrittive e quali semplicemente raccomandate.

Sovranità digitale e modelli europei di XAI

La dimensione geopolitica dell'AI non può essere ignorata quando si discute di spiegabilità. L'Europa, attraverso l'AI Act e iniziative come il progetto Gaia-X, sta tentando di costruire un ecosistema AI basato su valori di trasparenza, accountability e rispetto dei diritti fondamentali che si differenzia dall'approccio più liberista degli Stati Uniti e da quello più dirigista della Cina.

In questo contesto, l'XAI diventa elemento di **differenziazione competitiva**. Modelli europei che incorporano trasparenza by-design potrebbero trovare mercato presso organizzazioni e governi che privilegiano compliance normativa e riduzione del rischio reputazionale rispetto alla pura performance predittiva. Iniziative come i progetti Horizon Europe dedicati all'AI trustworthy stanno finanziando ricerca su architetture di XAI specificatamente progettate per soddisfare i requisiti normativi europei.

Per le organizzazioni italiane, questa dinamica crea opportunità ma anche responsabilità. Investire in competenze e soluzioni XAI europee significa contribuire alla costruzione di quella **sovranità tecnologica** che l'ACN ha identificato come priorità strategica. Significa anche prepararsi a un mercato dove la certificazione di conformità ai requisiti di spiegabilità potrebbe diventare prerequisito per accedere a

gare pubbliche o mercati regolamentati.

Verso un’etica operativa dell’AI in sicurezza

Oltre gli aspetti normativi e tecnici, l’XAI solleva questioni etiche che i professionisti della sicurezza non possono eludere. L’utilizzo di AI per decisioni che impattano la privacy, la libertà di movimento in ambienti fisici protetti, l’accesso a risorse critiche richiede non solo conformità legale ma anche legittimazione etica.

Un sistema di videosorveglianza intelligente che utilizza analisi comportamentale per identificare minacce potenziali opera in una zona grigia dove efficienza securitaria e diritti individuali si confrontano. La capacità di spiegare perché un comportamento è stato classificato come anomalo – e quindi potenzialmente minaccioso – non risolve il dilemma etico ma lo rende almeno trasparente, permettendo un dibattito informato e meccanismi di contestazione da parte degli individui interessati.

Alcune organizzazioni stanno adottando “**AI ethics boards**” interni che valutano l’introduzione di nuovi sistemi AI non solo sotto il profilo della compliance ma anche della sostenibilità etica. In questi comitati, la presenza di esperti di sicurezza è fondamentale per bilanciare l’imperativo della protezione con il rispetto dei diritti, utilizzando l’XAI come strumento di mediazione: sistemi più trasparenti sono anche sistemi più facilmente sottoponibili a scrutinio etico.

Implicazioni pratiche per CISO e

team di sicurezza

Per tradurre questi principi in azioni concrete, i responsabili della sicurezza devono integrare l'XAI nei processi decisionali quotidiani. Questo significa innanzitutto sviluppare o acquisire **competenze ibride**: data scientist che comprendano le minacce cyber, security analyst che sappiano interrogare modelli di machine learning, legal counsel che possano interpretare i requisiti di spiegabilità nel contesto specifico dell'organizzazione.

Sul piano procedurale, è opportuno istituire protocolli di "**AI incident response**" che definiscano come gestire situazioni in cui un sistema AI produce decisioni errate o inspiegabili. Chi ha l'autorità di disabilitare un sistema automatizzato in caso di comportamento anomalo? Come si documenta un falso positivo prodotto da AI per evitare recidive? Quali meccanismi di escalation si attivano quando le spiegazioni fornite dal sistema risultano insufficienti o contraddittorie?

La **formazione continua** diventa imperativa. I team SOC devono familiarizzare con i concetti fondamentali di machine learning – non per diventare data scientist, ma per sviluppare quella literacy che permette di distinguere una correlazione spuria da un pattern significativo, di riconoscere quando un modello sta operando fuori dal suo dominio di competenza, di valutare criticamente le confidence score associate alle predizioni.

Scenari futuri e direzioni di ricerca nell'XAI

Guardando avanti, alcune direzioni di sviluppo appaiono particolarmente promettenti. La ricerca su **modelli "intrinsecamente interpretabili"** – architetture di neural network progettate per essere trasparenti per costruzione anziché richiedere tecniche post-hoc di spiegazione – potrebbe

ridurre il trade-off tra performance e spiegabilità. Approcci come le Neural Additive Models o le Concept Bottleneck Networks mostrano che è possibile costruire sistemi potenti mantenendo un livello di interpretabilità umana.

L'integrazione tra XAI e sistemi di **knowledge graph** rappresenta un'altra frontiera interessante. Mappando le decisioni algoritmiche su grafi di conoscenza esplicita – ontologie di minacce, tassonomie di vulnerabilità, modelli di comportamento normale – si può costruire un ponte tra il ragionamento sub-simbolico delle reti neurali e la logica esplicita comprensibile agli umani.

Infine, l'utilizzo dell'AI per spiegare l'AI – modelli linguistici che traducono output tecnici in linguaggio naturale, sistemi che generano contro-fattuali per illustrare decisioni ("il sistema ha bloccato questa transazione perché, se il parametro X fosse stato diverso, sarebbe stato classificato come legittimo") – apre possibilità di rendere la spiegabilità accessibile anche a stakeholder non tecnici.

Conclusione: trasparenza come investimento strategico

In definitiva, l'**Intelligenza Artificiale Spiegabile** non configura una singola tecnologia ma un paradigma progettuale che deve permeare l'intero ciclo di vita dei sistemi AI utilizzati in contesti di sicurezza. Dalla raccolta e selezione dei dati di training, alla scelta dell'architettura del modello, fino al deployment e al monitoraggio continuo, ogni fase deve incorporare considerazioni di interpretabilità.

Per i CISO e i responsabili della sicurezza, questo significa adottare framework di governance che includano requisiti espliciti di spiegabilità nei processi di procurement e sviluppo di soluzioni AI, prevedere meccanismi di audit algoritmico periodico e investire nella formazione dei team

affinché sviluppino quella literacy necessaria per interrogare criticamente gli output dei sistemi intelligenti.

La posta in gioco supera la mera conformità normativa: riguarda la costruzione di una **fiducia giustificata** nei sistemi automatizzati che sempre più frequentemente prendono decisioni che impattano la sicurezza delle organizzazioni e, in ultima analisi, la protezione dei dati personali e dei diritti fondamentali degli individui. Un sistema opaco può funzionare perfettamente fino al momento in cui fallisce, e senza spiegabilità quel fallimento rimane inspiegabile, irreplicabile, e quindi impossibile da prevenire in futuro.

La trasparenza algoritmica diventa così non un vincolo allo sviluppo tecnologico, ma una condizione abilitante per un'innovazione sostenibile e responsabile. Nel contesto italiano ed europeo, dove la cybersecurity si intreccia sempre più strettamente con considerazioni normative, etiche e geopolitiche, l'investimento in XAI rappresenta non un costo ma un **vantaggio competitivo**: la capacità di dimostrare non solo che i propri sistemi funzionano, ma anche di spiegare come e perché lo fanno.

Condividi sui Social Network:
https://www.ictsecuritymagazine.com/articoli/intelligenza-artificiale-spiegabile-xai/