MITRE ATLAS e l’Adversarial Machine Learning

Un’analisi strutturale e operativa per la difesa dei sistemi di Intelligenza Artificiale

Il Cambio di Paradigma: Dal Software Deterministico all'AI

Software Deterministico

- Basato su logica e codice esplicito.
- Difesa incentrata su reti, endpoint e infrastrutture IT tradizionali.

AI Probabilistica

- Sistemi guidati dai dati e processi di apprendimento continui.
- Vulnerabilità uniche: dipendenza dai dati di addestramento, manipolazione degli input (Evasion, Poisoning).

L’integrazione dell’AI richiede un passaggio dalla protezione dell’infrastruttura alla protezione del ciclo di vita del modello.

Cos’è MITRE ATLAS?

ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) è un database globale e in continua evoluzione delle tattiche e tecniche avversarie, basato su osservazioni reali di attacchi ai sistemi AI.

Genesi (Giugno 2021)

Colma il divario tra la cybersecurity tradizionale e la data science.

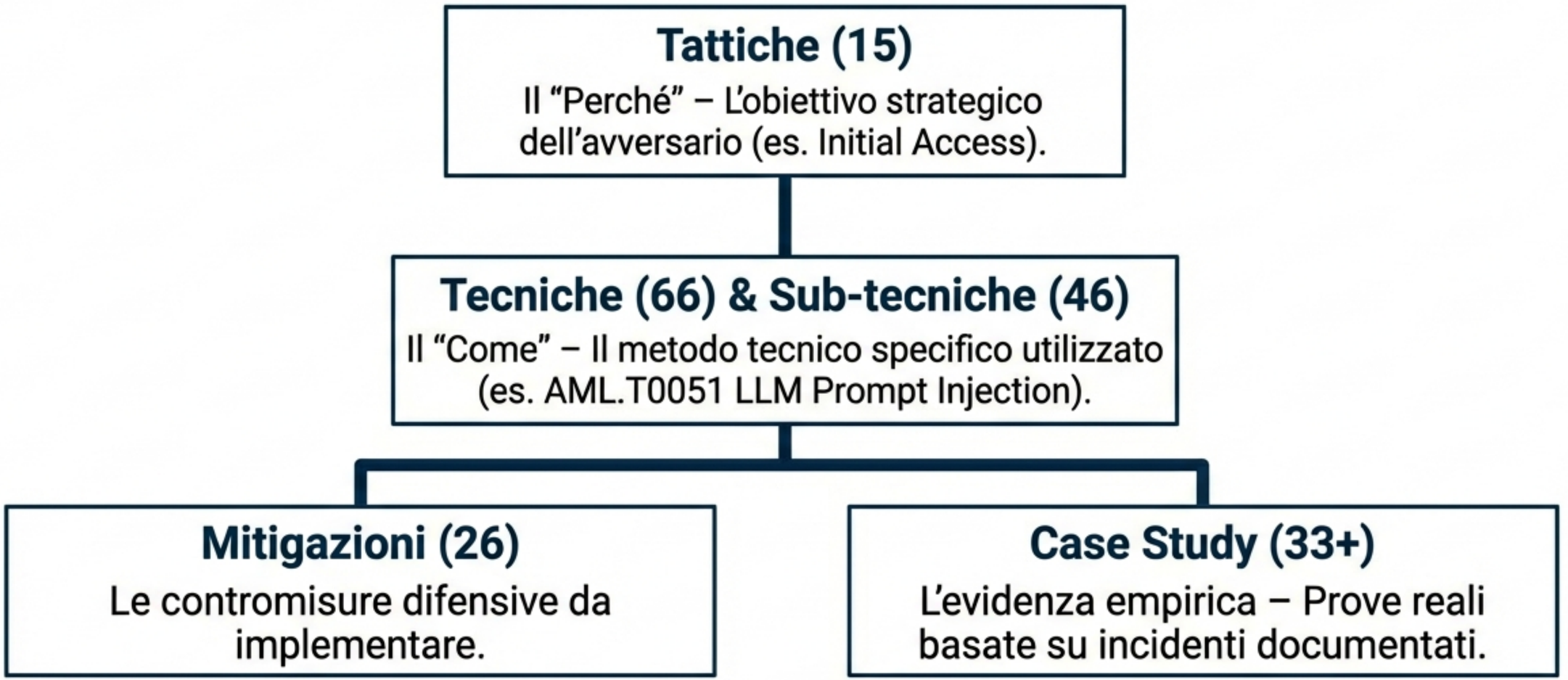
Adozione Globale

Supportato da oltre 150 organizzazioni (industria, governo, mondo accademico).

Linguaggio Comune

Standardizza la classificazione delle minacce per red team e difensori.

L’Ontologia Strutturale di ATLAS



I dati di ATLAS sono strutturati in formato STIX 2.1, garantendo l’integrazione machine-readable con le piattaforme di sicurezza.

La Matrice ATLAS: La Progressione dell'Attacco

15 colonne che rappresentano l'intero ciclo di vita dell'attacco (da Reconnaissance a Impact).

Reconnaissance	Resource Development	Initial Access	ML Model Access	Execution	Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery	Collection	ML Attack Staging	Exfiltration	Impact

Tattiche Esclusive AI

ML Model Access (AML.TA0004)

Accesso al modello tramite API di inferenza o accesso diretto agli artefatti.

ML Attack Staging (AML.TA0012)

Preparazione dell'attacco, inclusi l'avvelenamento dei dati (poisoning) e l'inserimento di backdoor.

Analisi Comparativa: ATLAS vs. ATT&CK

	MITRE ATT&CK	MITRE ATLAS
Focus Primario	Comportamenti avversari nei sistemi IT/OT tradizionali.	Comportamenti avversari specifici per AI e Machine Learning.
Sistemi Bersaglio	Endpoint, reti, infrastrutture cloud.	Modelli ML, pipeline di addestramento, LLM.
Tattiche Uniche	Movimento laterale, Command-and-Control.	ML Model Access, ML Attack Staging.

Sinergia operativa: utilizzare ATT&CK per l'infrastruttura e ATLAS per i vettori di attacco AI.

Vettori di Attacco Principali nell'AI

Prompt Injection (AML.T0051)

Metodo: Manipolazione del comportamento del modello linguistico (LLM) tramite input malevoli.

Obiettivo: Evasione delle difese, esfiltrazione dati o esecuzione non autorizzata.

Data Poisoning (AML.T0020)

Metodo: Contaminazione dei dataset di addestramento.

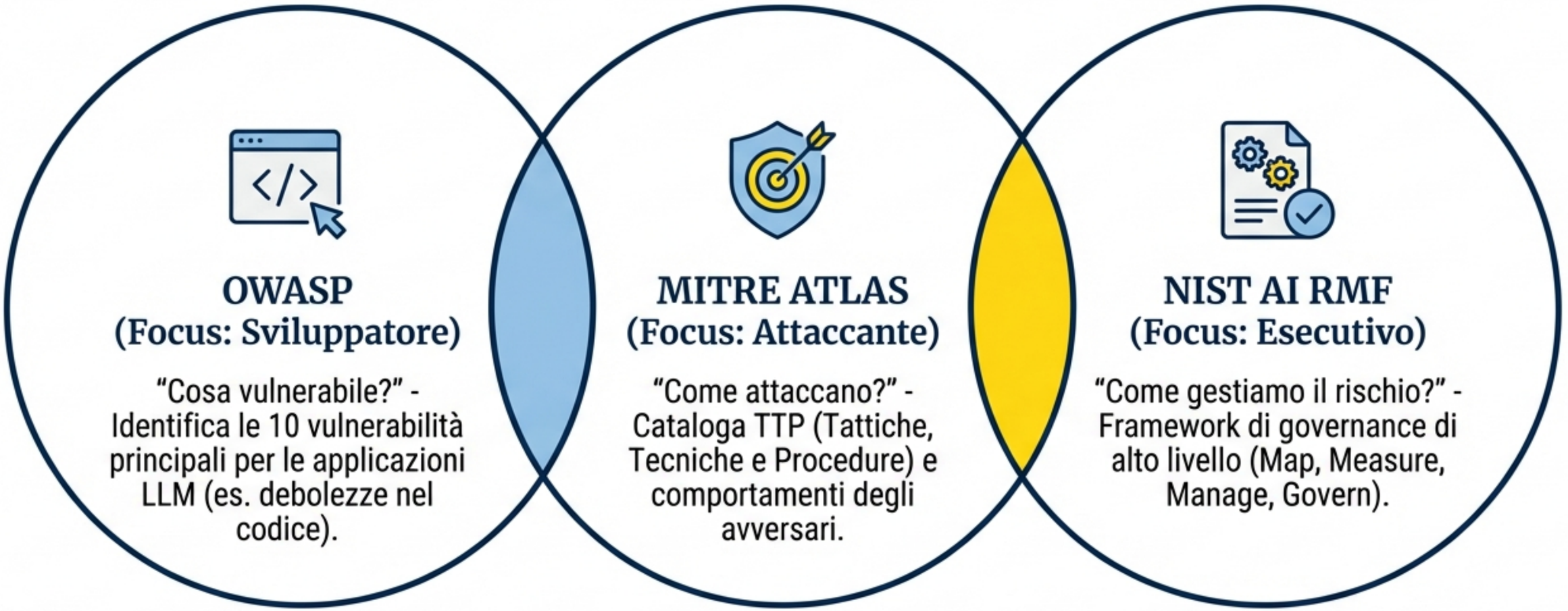
Obiettivo: Alterare i pattern di apprendimento per generare output errati o introdurre vulnerabilità silenti.

Model Extraction (AML.T0040 / AML.T0024)

Metodo: Interrogazione sistematica delle API di inferenza.

Obiettivo: Rubare la logica proprietaria o replicare il modello ML senza accesso diretto ai pesi.

L'Ecosistema della Sicurezza AI



Questi tre framework si integrano per una difesa completa, coprendo ogni aspetto del ciclo di vita dell'AI.

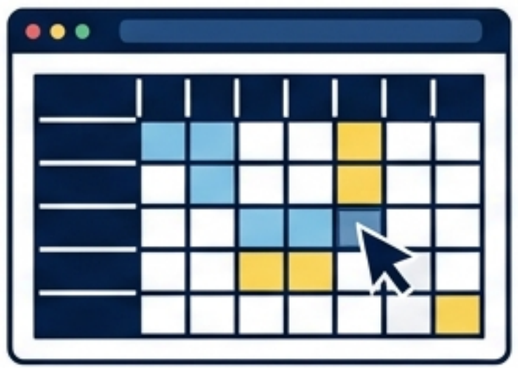
Evidenza Empirica: I Case Study di ATLAS

AML.CS0033: Evasione dei Sistemi di Identificazione Facciale (iProov)

- **Il Contesto**: Attacco contro i sistemi di verifica dell'identità remota (KYC - Know Your Customer) in ambito finanziario.
- **La Tecnica**: Utilizzo di app di intelligenza artificiale generativa (Faceswap) e software di telecamere virtuali.
- **L'Impatto**: Iniezione riuscita di video deepfake per aggirare il rilevamento 'active liveness'.
- **Il Risultato Accademico**: Dimostrazione che i sistemi AI biometrici possono essere compromessi senza alcun accesso fisico al dispositivo bersaglio.

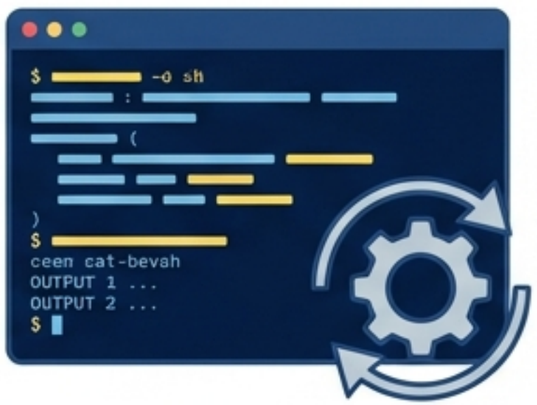
Strumenti Operativi e Automazione

ATLAS Navigator



- **Funzione**: Interfaccia web interattiva per visualizzare e annotare la matrice.
- **Applicazione**: Threat modeling, analisi dei gap di copertura e mappatura delle difese in tempo reale.

MITRE Arsenal



- **Funzione**: Plugin per il motore di emulazione CALDERA.
- **Applicazione**: Automazione del red teaming. Permette di testare l'efficacia dei controlli di sicurezza eseguendo attacchi simulati allineati ad ATLAS.

Operazionalizzare ATLAS nel SOC

Roadmap 2026: GenAI e AI 'Agentic'

Evoluzione Rapida

Passaggio dalla modellazione discriminativa tradizionale all'Intelligenza Artificiale Generativa (GenAI).

Nuovi Vettori (Aggiornamenti Recenti)

Integrazione di 14 nuove tecniche focalizzate sulle vulnerabilità degli agenti AI autonomi (Agentic AI), come l'uso improprio degli strumenti e le manipolazioni della memoria.

Integrazione Dati

Espansione di progetti come 'Misp-Galaxy' per lo scambio automatizzato di intelligence sulle minacce in formato STIX 2.1.

Conclusioni e Takeaway

- L'IA introduce vulnerabilità probabilistiche che i framework tradizionali (come ATT&CK) non possono gestire da soli.
- ATLAS fornisce un'ontologia rigorosa e basata sull'evidenza empirica per unificare Data Science e Cybersecurity.
- La difesa proattiva richiede simulazione continua tramite strumenti operativi (Navigator, Arsenal).

> Il settore deve superare la mentalità della 'scatola nera': la sicurezza dell'Intelligenza Artificiale richiede un approccio strutturato, trasparente e guidato dalla conoscenza delle minacce.