

```
root@hacklab:~/parte-02/lezione-24
```

ETHICAL HACKING POTENZIATO DALL'IA // PARTE II

LEZIONE 24

AI Red Teaming e governance dell'IA

Red teaming sistematico dei sistemi IA e governance lungo il ciclo MLOps.

docente **Vincenzo Calabro** fase Specializzazione avanzata: AI Red Teaming e governance

```
root@hacklab:~/lezione-24/agenda $
```

LEZIONE 24 / AGENDA

In questa lezione

// PERCORSO

- > Metodologia di AI red teaming
- > Attacchi a modelli e LLM
- > Sicurezza della pipeline MLOps
- > Robustezza, safety e bias
- > Governance (AI Act, NIST AI RMF)

metodo obiettivi -> teoria -> strumenti -> laboratorio -> caso di studio -> etica -> verifica

```
root@hacklab:~/lezione-24/obiettivi $
```

LEZIONE 24 / OBIETTIVI

Obiettivi di apprendimento

// AL TERMINE SAPRAI (TASSONOMIA DI BLOOM)

- > Comprendere – il red teaming sistematico dei sistemi IA
- > Applicare – attacchi e valutazioni su modelli/LLM/pipeline
- > Analizzare – i rischi lungo il ciclo MLOps
- > Valutare – la governance (AI Act, NIST AI RMF)
- > Creare – un programma di AI assurance

```
root@hacklab:~/lezione-24/inquadramento $
```

LEZIONE 24 / INQUADRAMENTO

Inquadramento

// DI COSA PARLIAMO

I sistemi di IA vanno testati come ogni altro bersaglio, ma con metodo e governance propri. Vediamo attacchi a modelli e agenti, MITRE ATLAS e gli obblighi dell'AI Act.

// PERCHE CONTA

Red teaming sistematico dei sistemi IA e governance lungo il ciclo MLOps.

```
root@hacklab:~/lezione-24/mappa-insieme $
```

LEZIONE 24 / MAPPA // INSIEME

Mappa d'insieme del modulo

■ Attacco/tecnica ■ Difesa ■ Norme & metodo ■ IA

FONDAMENTI

01 AI red teaming e MITRE ATLAS

02 Attacchi ai sistemi di IA

03 MLOps e supply chain

04 Robustezza, safety e bias

05 Governance e guardrail

APPROFONDIMENTI

06 Threat modeling con ATLAS

07 Dalle minacce alle prove pratiche

08 Sicurezza della supply chain ML

09 Red teaming continuo

10 Model card, audit e guardrail

```
root@hacklab:~/lezione-24/mappa-1 $
```

LEZIONE 24 / MAPPA // FONDAMENTI

Mappa dei concetti – Fondamenti

■ Attacco/tecnica ■ Difesa ■ Norme & metodo ■ IA



```
root@hacklab:~/lezione-24/concetto-1 $
```

LEZIONE 24 / CONCETTO 1/10

[IA]

AI red teaming e MITRE ATLAS

// SPIEGAZIONE

Testare un sistema di IA inizia dal threat modeling: quali asset, quali avversari, quali abusi. MITRE ATLAS cataloga tattiche e tecniche contro i sistemi di machine learning, come ATT&CK fa per l'IT, dando un linguaggio comune.

// ESEMPIO

> Si mappano gli attacchi previsti su ATLAS per pianificare la campagna di red teaming.

```
root@hacklab:~/lezione-24/concetto-2 $
```

LEZIONE 24 / CONCETTO 2/10

[IA]

Attacchi ai sistemi di IA

// SPIEGAZIONE

Il ventaglio è ampio: poisoning dei dati, extraction del modello, prompt injection avanzata, jailbreak delle protezioni e abuso di agenti autonomi. Ogni tecnica colpisce un punto diverso della pipeline o dell'uso del modello.

// ESEMPIO

> Prompt injection indiretta che fa eseguire a un agente un'azione non prevista su uno strumento.

```
root@hacklab:~/lezione-24/concetto-3 $
```

LEZIONE 24 / CONCETTO 3/10

[IA]

MLOps e supply chain

// SPIEGAZIONE

La pipeline MLOps, con dati, training, registry e deployment, è un bersaglio: modelli e dataset di terzi possono essere avvelenati o contenere backdoor. Provenienza, integrità e versionamento sono controlli di sicurezza, non solo pratiche di ingegneria.

// ESEMPIO

> Un modello dal registry pubblico con backdoor supera i test ma cede su un trigger nascosto.

```
root@hacklab:~/lezione-24/concetto-4 $
```

LEZIONE 24 / CONCETTO 4/10

[Attacco/tecnica]

Robustezza, safety e bias

// SPIEGAZIONE

Valutare un modello significa misurarne la robustezza agli input avversari, la safety delle risposte e il bias nei risultati. Non è un controllo una tantum: il red teaming continuo accompagna il modello per tutto il suo ciclo di vita.

// ESEMPIO

> Un modello che nega un servizio in modo sistematico a un gruppo rivela un bias da correggere.

```
root@hacklab:~/lezione-24/concetto-5 $
```

LEZIONE 24 / CONCETTO 5/10

[Norme & metodo]

Governance e guardrail

// SPIEGAZIONE

La governance traduce i rischi in obblighi: l'AI Act classifica per rischio, il NIST AI RMF offre un quadro, le model card documentano limiti e usi. Gli agenti autonomi richiedono guardrail e controllo umano sulle azioni irreversibili.

// ESEMPIO

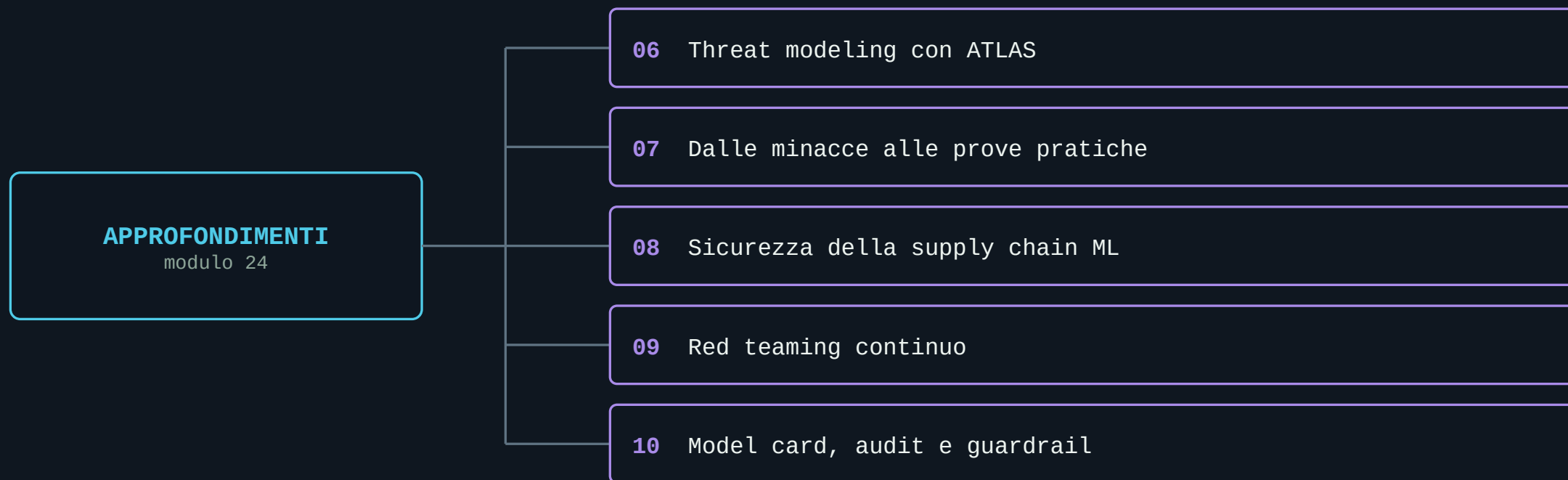
> Una model card dichiara dati di addestramento, limiti noti e usi sconsigliati del modello.

```
root@hacklab:~/lezione-24/mappa-2 $
```

LEZIONE 24 / MAPPA // APPROFONDIMENTI

Mappa dei concetti – Approfondimenti

■ Attacco/tecnica ■ Difesa ■ Norme & metodo ■ IA



```
root@hacklab:~/lezione-24/concetto-6 $
```

LEZIONE 24 / CONCETTO 6/10

[IA]

Threat modeling con ATLAS

// SPIEGAZIONE

Modellare le minacce di un sistema IA significa mappare asset, superfici e avversari e ricondurli alle tecniche di MITRE ATLAS. Ne esce un piano di test mirato invece di prove scollegate, con priorit  basate sul rischio reale.

// ESEMPIO

> ATLAS guida a testare prima l'evasione del modello se e il rischio piu probabile per il caso d'uso.

```
root@hacklab:~/lezione-24/concetto-7 $
```

LEZIONE 24 / CONCETTO 7/10

[IA]

Dalle minacce alle prove pratiche

// SPIEGAZIONE

Ogni minaccia si traduce in test concreti: avvelenare un dataset di prova, interrogare un'API per estrarre il modello, forzare un jailbreak, indurre un agente ad agire oltre il previsto. Sempre in ambiente autorizzato e isolato.

// ESEMPIO

> Si verifica se un agente, ricevendo istruzioni ostili in un documento, compie azioni non autorizzate.

```
root@hacklab:~/lezione-24/concetto-8 $
```

LEZIONE 24 / CONCETTO 8/10

[IA]

Sicurezza della supply chain ML

// SPIEGAZIONE

Come per il software, modelli e dati vanno trattati come componenti di cui conoscere origine e integrità. Un dataset manomesso o un modello non verificato introducono rischi che nessun test a valle intercetta facilmente.

// ESEMPIO

> Verifica dell'hash e della provenienza di un dataset prima di usarlo per l'addestramento.

```
root@hacklab:~/lezione-24/concetto-9 $
```

LEZIONE 24 / CONCETTO 9/10

[IA]

Red teaming continuo

// SPIEGAZIONE

I modelli cambiano con fine-tuning e nuovi dati, e con loro i rischi: il red teaming va ripetuto e integrato nella valutazione continua. Una singola verifica iniziale non garantisce la sicurezza nel tempo.

// ESEMPIO

> Dopo ogni fine-tuning si rieseguo i test di jailbreak e robustezza prima del rilascio.

```
root@hacklab:~/lezione-24/concetto-10 $
```

LEZIONE 24 / CONCETTO 10/10

[IA]

Model card, audit e guardrail

// SPIEGAZIONE

Documentare con model card, sottoporre ad audit indipendente e imporre guardrail agli agenti sono i pilastri di un'IA governata. Un agente non deve poter compiere azioni irreversibili senza un controllo umano esplicito.

// ESEMPIO

> Un guardrail impone conferma umana prima che un agente invii denaro o cancelli dati.

```
root@hacklab:~/lezione-24/strumenti $
```

LEZIONE 24 / STRUMENTI

Strumenti della lezione

// TOOLKIT

garak

ART

PyRIT

Foolbox

framework MLOps

// NOTA OPERATIVA

> Ogni strumento si usa in ambiente virtualizzato e isolato, solo su bersagli di proprietà o autorizzati. Output e artefatti vanno protetti e, a fine ingaggio, gestiti o distrutti.

```
root@hacklab:~/lezione-24/lab-1 $
```

LEZIONE 24 / LABORATORIO 1

Laboratorio 1

\$ Campagna di AI red teaming (threat model -> ATLAS)

// PROCEDURA

- [1]** Esegui una campagna di AI red teaming: threat model
- [2]** attacchi
- [3]** report su ATLAS

[!] Verifica: Come si delimita lo scope di un red team su un agente autonomo?

```
root@hacklab:~/lezione-24/lab-2 $
```

LEZIONE 24 / LABORATORIO 2

Laboratorio 2

\$ Valutazione e model card (robustezza/bias)

// PROCEDURA

[1] Valuta robustezza e bias di un modello e redigi una model card con mitigazioni

[!] Verifica: Quali obblighi impone l'AI Act a un sistema ad alto rischio?

```
root@hacklab:~/lezione-24/caso-di-studio $
```

LEZIONE 24 / CASO DI STUDIO

Caso di studio

// SCENARIO

> Incidente IA (prompt injection indiretta -> esfiltrazione) e il programma di assurance che l'avrebbe prevenuto.

// ANALISI RICHIESTA

- > Ricostruisci la dinamica dei fatti con metodo e tracciabilità
- > Individua errori, cause radice e punti di decisione critici
- > Proponi contromisure, controlli e lezioni apprese

```
root@hacklab:~/lezione-24/etica $
```

LEZIONE 24 / ETICA & DIRITTO

Nota etico-legale

// PRINCIPI INDEROGABILI

> autorizzazione scritta scope definito ambiente isolato dati protetti responsible disclosure

// SPECIFICO DELLA LEZIONE

[!] Attaccare IA solo se di proprietà o autorizzata; GDPR e AI Act; agenti mai oltre lo scope né azioni irreversibili senza controllo umano.

```
root@hacklab:~/lezione-24/ricapitolazione $
```

LEZIONE 24 / RICAPITOLAZIONE

Ricapitolazione

// PUNTI CHIAVE

- > Il red teaming sistematico dei sistemi IA
- > Attacchi e valutazioni su modelli/LLM/pipeline
- > I rischi lungo il ciclo MLOps
- > La governance (AI Act, NIST AI RMF)
- > Un programma di AI assurance

```
root@hacklab:~/lezione-24/verifica $
```

LEZIONE 24 / AUTOVALUTAZIONE

Autovalutazione

// DOMANDE DI VERIFICA

- ? Come si delimita lo scope di un red team su un agente autonomo?
- ? Quali obblighi impone l'AI Act a un sistema ad alto rischio?
- ? Perché un agente non deve compiere azioni irreversibili senza controllo umano?

// VALUTAZIONE

Programma di AI assurance con report di red teaming e governance.

// LETTURE

OWASP Top 10 for LLM; MITRE ATLAS; NIST AI RMF; sintesi AI Act.

lezione 24 completata

\$ next

AI Red Teaming e governance dell'IA

[!] Le competenze di questa lezione si impiegano esclusivamente in contesti legali, autorizzati e deontologicamente corretti, entro lo scope e con autorizzazione scritta.

root@hacklab:~# _

```
root@hacklab:~/contatti $
```

CONTATTI / RISORSE

Restiamo in contatto

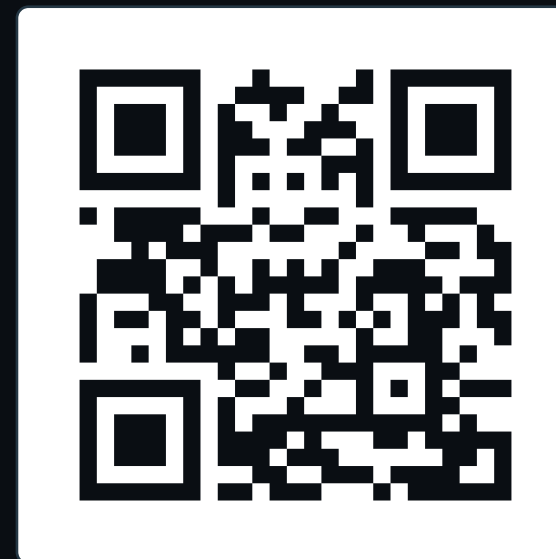


Vincenzo Calabro

cyber security / digital forensics /
docenza e divulgazione

// SCANSIONA O VISITA

> vincenzocalabro.it



Inquadra il QR con la fotocamera
per aprire il sito.