MODULO 12

AI offensiva e minacce emergenti

Il punto di vista dell'attaccante: ingegneria sociale generativa, deepfake, malware adattivo e automazione dell'attacco. Trattazione concettuale e difensiva, senza istruzioni operative.

PHISHING GENERATIVO

DEEPFAKE

ATTACCHI AUTONOMI

SCHEDA DEL MODULO

Abstract e obiettivi formativi

Abstract

Il modulo adotta la prospettiva dell'attaccante per comprendere come l'AI amplifichi le capacità offensive: phishing e vishing generativi, deepfake nelle frodi, malware adattivo e automazione di ricognizione ed exploit, con l'uso documentato di LLM da parte di attori statali e criminali, per ricavarne i requisiti difensivi.

Obiettivi formativi (Bloom)

- **Analizzare**: l'impatto dell'AI generativa su phishing e frodi
- **Valutare**: la minaccia di attacchi autonomi o semi-autonomi
- **Creare**: requisiti difensivi (verifiche fuori banda, detection comportamentale, resilienza dei processi)

Traccia core: 01 02 07

Detection & Response: 03 04 05 06 o Digital Forensics: 08 09 10 11 o Sicurezza dell'AI: 06 **12** 13 14

Traccia core: 15 16

Letture di raccordo: § 4.3, 4.8

SCHEMA · TEORIA

Minacce generative e difese di processo

QUADRO DI RIFERIMENTO ENISA Threat Landscape · Europol · rapporti di Microsoft, OpenAI e Google

01 · INGEGNERIA SOCIALE

Scala e personalizzazione

Spear phishing, voce e video sintetici

02 · AUTOMAZIONE

Livelli di autonomia degli attacchi

Automatizzato, semi-autonomo, autonomo

03 · CONTROMISURE

Contromisure di processo e di identità

Verifica fuori banda, FIDO2, formazione

APPLICAZIONE · LABORATORIO

Analisi difensiva di messaggi e procedure

Segnali residui, verifica fuori banda, scenari di fallimento

TEORIA · INGEGNERIA SOCIALE GENERATIVA

Scala e personalizzazione dell'ingegneria sociale

Cosa cambia l'AI generativa

- Spear phishing su misura, generato in massa da fonti OSINT
- Cadono i segnali linguistici classici: errori, traduzioni goffe
- Vishing con voce clonata; deepfake video nelle riunioni
- Caso documentato (polizia di Hong Kong, febbraio 2024): frode da circa 25 milioni di dollari con una videoconferenza i cui partecipanti erano generati artificialmente

Uso documentato di LLM da threat actor

- Rapporti di Microsoft, OpenAI e Google: uso dei modelli da parte di attori statali per ricognizione, esche e sviluppo
- Europol: abbassamento della barriera d'ingresso al crimine informatico
- Implicazione: il modello di minaccia si aggiorna per scala e qualità degli attacchi, con tecniche già catalogate in ATT&CK

TEORIA · AUTOMAZIONE DELL'ATTACCO

Livelli di autonomia degli attacchi

AUTOMATIZZATO

Script e tool eseguono passi predefiniti su larga scala; l'umano decide obiettivi e sequenza. Prassi consolidata

SEMI-AUTONOMO

Agenti basati su LLM pianificano passi dell'attacco adattandosi alle risposte, con supervisione umana residua. Dimostrato in ricerca e segnalato in rapporti pubblici di threat intelligence

AUTONOMO

Catena d'attacco completa senza intervento umano: tendenza con dimostrazioni parziali, da considerare come scenario di pianificazione

Conseguenza per la difesa

Se l'attacco procede a velocità di macchina, la risposta non può dipendere solo da tempi umani, e questo è il principale argomento a favore dell'automazione difensiva del Modulo 5, con i limiti discussi in quel modulo.

TEORIA · CONTROMISURE

Contromisure di processo e di identità

Processi

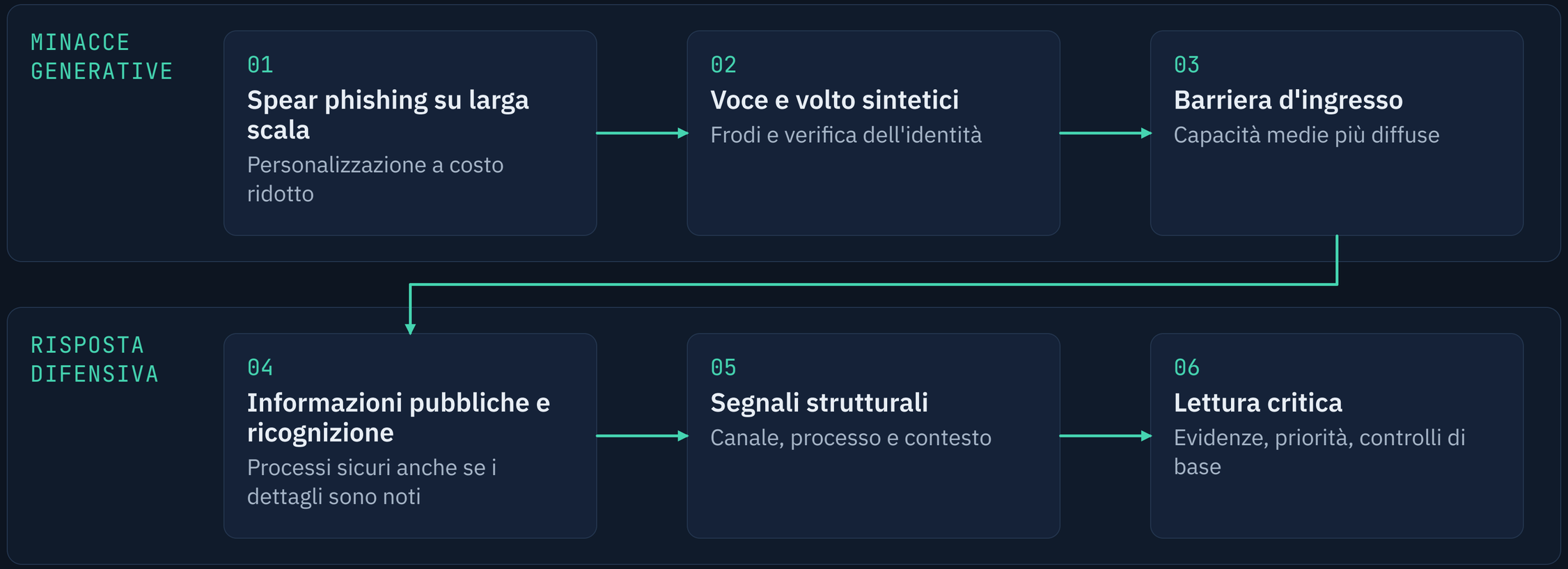
- **Verifica fuori banda** per pagamenti e richieste ad alto rischio, su un canale indipendente da quello della richiesta
- Doppia firma e limiti d'importo non derogabili per urgenza
- Parole d'ordine concordate per le richieste vocali critiche

Tecnologia e persone

- Rilevamento basato su comportamento e identità, oltre che sul contenuto
- MFA resistente al phishing (FIDO2) contro l'MFA fatigue
- Formazione centrata sui **processi** di verifica

SCHEMA · APPROFONDIMENTI

La minaccia vista dalla difesa



APPROFONDIMENTO 1

Spear phishing generativo su larga scala

Prima dell'AI generativa, un messaggio di spear phishing di buona qualità richiedeva tempo di ricerca e di scrittura per ogni destinatario ed era riservato a bersagli di valore. I modelli linguistici riducono questo costo: producono testi corretti, nella lingua del destinatario e con il registro adatto, a partire da informazioni pubbliche sull'organizzazione. La personalizzazione diventa così possibile su larga scala, e per la difesa scompare l'indizio degli errori linguistici, su cui per anni si è basata la formazione degli utenti.

Caso documentato

Uno studio sperimentale diffuso come preprint nel 2024 e pubblicato nel 2026 ha confrontato messaggi di spear phishing generati e personalizzati automaticamente, messaggi scritti da esperti e messaggi generici di controllo. I messaggi automatici hanno ottenuto tassi di clic paragonabili a quelli degli esperti e nettamente superiori a quelli del gruppo di controllo.

Contromisure

Il rilevamento si sposta dal testo al contesto: canale e infrastruttura di invio (dominio registrato di recente, esito dei controlli SPF, DKIM e DMARC), coerenza della richiesta con i processi aziendali, che per esempio non prevedono bonifici autorizzati via e-mail, e verifica su un canale indipendente. Anche i sistemi di filtraggio impiegano modelli di classificazione del testo, con il limite che un testo ben scritto offre loro pochi indizi.

APPROFONDIMENTO 2

Voce e volto sintetici nelle frodi

Registrazioni pubbliche di breve durata, come un intervento a un convegno o un video aziendale, possono bastare a sintetizzare una voce credibile (per il sistema Voice Engine OpenAI ha indicato un campione di 15 secondi), e i sistemi in tempo reale sostengono anche una conversazione. Nel 2024 un dipendente della sede di Hong Kong della società di ingegneria Arup ha autorizzato trasferimenti per circa 25 milioni di dollari dopo una videoconferenza in cui gli altri partecipanti, compreso il direttore finanziario, erano generati artificialmente. L'attacco sfrutta la fiducia nel canale di comunicazione, per cui il riconoscimento visivo e vocale di un interlocutore cessa di essere una forma di verifica.

Scenario didattico

Un operatore dell'help desk riceve la telefonata di un dirigente che dichiara di essere in viaggio e chiede il reset dell'autenticazione a più fattori. La voce è simile a quella reale, il numero chiamante è falsificato e la richiesta è urgente. La procedura prevede la richiamata al numero presente in rubrica e una domanda di verifica concordata; applicandola, l'operatore blocca il tentativo.

Contromisure

Contro un'imitazione credibile restano efficaci i controlli di processo: la verifica su un canale indipendente da quello della richiesta, la doppia autorizzazione per pagamenti e modifiche delle coordinate bancarie, le parole d'ordine concordate per le richieste d'emergenza e soglie che nessuna persona può superare da sola.

APPROFONDIMENTO 3

Riduzione della barriera d'ingresso

L'effetto dell'AI generativa più documentato sul crimine informatico è la diffusione di capacità di livello medio: attori con competenze limitate ottengono testi di phishing, siti fraudolenti e codice di supporto che in passato richiedevano esperienza. Nel 2023 sono stati pubblicizzati su forum criminali modelli presentati come privi di filtri di sicurezza, tra cui WormGPT, ed Europol ha analizzato i possibili usi criminali dei modelli linguistici; nel febbraio 2024 Microsoft e OpenAI hanno documentato l'uso dei propri servizi da parte di cinque gruppi collegati a Stati. La generazione automatica di varianti con lo stesso comportamento e firme sempre diverse riduce ulteriormente l'efficacia del rilevamento basato su firme e accresce il peso del rilevamento comportamentale (Moduli 3 e 9).

Scenario didattico

Un gruppo con competenze modeste prepara in pochi giorni una campagna con testi in più lingue e componenti rigenerati a ogni invio. Le tecniche sono note da anni; cambiano i tempi di preparazione e quindi il numero di campagne che lo stesso gruppo può condurre in parallelo.

Metodo

I rapporti commerciali tendono a enfatizzare il «malware basato sull'AI», mentre le evidenze pubbliche indicano soprattutto l'accelerazione di tecniche già catalogate in MITRE ATT&CK e l'aumento di scala, con conseguenze sui tempi della risposta.

APPROFONDIMENTO 4

Informazioni pubbliche e ricognizione

La raccolta di informazioni su un bersaglio richiedeva tempo di analisi ed era riservata agli obiettivi di maggior valore. L'automazione ne riduce il costo, e chi difende deve considerare già note all'attaccante le informazioni che l'organizzazione rende pubbliche, come organigrammi, fornitori, tecnologie dichiarate e annunci di lavoro che descrivono l'infrastruttura. La valutazione del rischio si concentra quindi sui processi che devono restare sicuri anche quando l'attaccante conosce i dettagli interni, e sulla riduzione delle informazioni pubblicate senza necessità.

Esempio

Una richiesta di pagamento fraudolenta può citare correttamente il nome di chi autorizza i bonifici, ricavato da fonti pubbliche. Il controllo efficace protegge il processo con la verifica su canale indipendente e la doppia autorizzazione, senza presupporre che l'attaccante ignori questa informazione.

Contromisure

Si riducono le informazioni esposte senza necessità, come i dettagli di ruolo nelle firme e-mail, le tecnologie citate negli annunci e gli organigrammi completi. Chi ricopre ruoli sensibili riceve una formazione mirata, e l'esposizione pubblica dell'organizzazione si verifica periodicamente nell'ambito dell'analisi delle minacce.

APPROFONDIMENTO 5

Segnali strutturali dei messaggi ostili

Quando la qualità del testo non distingue più un messaggio legittimo da uno fraudolento, il rilevamento si basa sugli elementi che la generazione automatica non modifica. Per il **canale e l'infrastruttura** contano un dominio registrato di recente, l'esito dei controlli di autenticazione della posta e un mittente diverso dal nome visualizzato. Sul piano dei **processi** è sospetta una richiesta che salta i passaggi previsti, come una modifica delle coordinate bancarie comunicata via e-mail, mentre il contesto riguarda la coerenza della richiesta con le abitudini del mittente e con il canale usato. La formazione degli utenti va aggiornata su questi segnali, e la verifica su canale indipendente diventa il controllo principale.

Scenario didattico

Un'e-mail priva di errori, apparentemente inviata dall'amministratore delegato, chiede un bonifico urgente. Il dominio è simile ma non identico a quello aziendale, la richiesta non segue il processo di pagamento e insiste sui tempi. Come prevede la regola aziendale per ogni pagamento, il dipendente formato verifica la richiesta sul canale concordato.

Ruolo dell'AI

I gateway di posta valutano gli stessi segnali strutturali, come la reputazione, i controlli SPF, DKIM e DMARC e il comportamento abituale del mittente, e impiegano a loro volta modelli di classificazione. Questi filtri si affiancano alle segnalazioni degli utenti e alla verifica su canale indipendente.

APPROFONDIMENTO 6

Lettura critica dei rapporti sulla minaccia

Le descrizioni di un malware «autonomo e inarrestabile» basato sull'AI non trovano finora riscontro nelle evidenze pubbliche. I dati disponibili indicano che l'AI generativa accelera attività già esistenti, come la preparazione di esche, l'allestimento di infrastrutture e la generazione di varianti, riduce la barriera d'ingresso e aumenta il numero di campagne. Le tattiche e le tecniche descritte da MITRE ATT&CK restano valide e i controlli di base mantengono la loro efficacia: autenticazione a più fattori resistente al phishing, aggiornamenti, segmentazione, backup, verifica su canale indipendente. L'adattamento richiesto alla difesa riguarda soprattutto velocità e volume, e motiva l'automazione trattata nei Moduli 5 e 6.

Scenario didattico

Si consideri un rapporto commerciale che descrive un «malware che si riscrive da solo grazie all'AI». I dettagli tecnici mostrano un loader che rigenera l'offuscamento a ogni build, tecnica nota da anni. La conseguenza per la difesa è un'ulteriore perdita di efficacia delle firme e una maggiore importanza del rilevamento comportamentale.

Metodo

Per valutare un rapporto sulla minaccia si separano le osservazioni tecniche dalle interpretazioni, si cercano conferme indipendenti e si traducono le conclusioni in priorità operative: quali controlli rafforzare, quali rilevamenti aggiungere, quali processi rivedere.

LABORATORIO · ESERCITAZIONE DI ANALISI

Analisi difensiva di messaggi e procedure

Svolgimento

1. Valutazione comparata di e-mail di phishing tradizionali e generate, tratte dal corpus del docente
2. Identificazione dei segnali residui: incoerenze di contesto, identità, canale
3. Progettazione di una procedura di verifica fuori banda per richieste di pagamento ad alto rischio
4. Prova della procedura contro gli scenari di voce e video sintetici discussi nel modulo

Consegna

- Griglia dei segnali residui osservati
- Procedura di verifica fuori banda in una pagina
- Scenario in cui la procedura fallisce, con le ragioni del fallimento

CHIUSURA DEL MODULO

Letture e autovalutazione

Letture obbligatorie

- ENISA Threat Landscape, sezioni sull'AI
- Microsoft e OpenAI (2024); Google Threat Intelligence Group (2025): rapporti sull'uso dei LLM da parte di attori malevoli
- Europol, *ChatGPT – The impact of Large Language Models on Law Enforcement*

Domande di autovalutazione

1. Perché i segnali linguistici non bastano più contro il phishing?
2. Quali processi aziendali sono più esposti ai deepfake?
3. Cosa distingue un attacco automatizzato da uno autonomo?
4. Quali contromisure organizzative resistono all'ingegneria sociale generativa?
5. Come incide l'AI offensiva sul modello di minaccia di un CSIRT?

FINE DELLA LEZIONE · MODULO 12

Grazie

AI offensiva e minacce emergenti · Corso «Incident Response potenziata dall'AI»

Domande e materiali: ricevimento del docente. Profilo e contatti: vincenzocalabro.it

Vincenzo Calabrò

CYBER SECURITY · DIGITAL FORENSICS

VINCENZOCALABRO.IT