MOKA CONFERENCE · APRILE 2026

AI generativa e deepfake:
nuove minacce al social engineering e alla sicurezza informatica

Whoami

Davide Ornaghi

CTO & Co-founder · Betrusted

Offensive Security

Red teaming, penetration testing
e vulnerability research.

Focus su infrastrutture critiche
e applicazioni enterprise.

Vulnerability Research

- Kernel Linux
- Mobile security
- Deepfake Frame Injection

AI & Automation

Progettazione di agenti AI
per offensive security automation.

betrusted.it

Il talk di oggi

01 Il nuovo arsenale dell'attaccante: LLM e deepfake

02 Quando il deepfake inganna la macchina

03 Black-hat agents: l'attaccante autonomo

04 Lo squilibrio degli 0-day e il futuro della difesa

05 Il futuro a 3 anni: la commodity della sicurezza

06 Case study

07 La risposta di Betrusted

01.

Il nuovo arsenale dell'attaccante

LLM, deepfake e la democratizzazione degli attacchi

Betrusted

LLM: la fabbrica del phishing perfetto

PRIMA (template tradizionali)		OGGI (LLM-powered)
✗ Errori grammaticali evidenti		✓ Italiano/inglese perfetto, tono calibrato
✗ Template generici, impersonali		✓ Personalizzato su LinkedIn, ruolo, azienda
✗ Facilmente bloccati dai filtri antispam	→	✓ Bypass dei filtri semantici tradizionali
✗ Alta intensità di lavoro manuale		✓ Generazione massiva a costo vicino a zero
✗ Scarsa personalizzazione sul target		✓ Fine-tuning sullo stile della vittima specifica

⚡ *IBM X-Force 2024: attacchi phishing con AI generativa +45% YoY; il tasso di clic su email AI-generated è 3× superiore vs template tradizionali*

I Deepfake sono sempre più credibili

AUDIO

Clone vocale

3 secondi di audio campione bastano per clonare la voce di un CEO.

Frodi BEC da €25M+ sfruttate nel caso Arup (2024).

VIDEO

Video real-time

Videochiamate con volto sostituito in tempo reale.

KYC e onboarding remoto completamente bypassati senza lasciare tracce.

IDENTITY

Persona sintetica

Identità digitali complete: foto, voce, profili social generati da zero.

Indistinguibili senza strumenti AI difensivi.

Gartner 2025: entro il 2026, il 30% degli attacchi di social engineering includerà componenti deepfake. Il vettore cresce più velocemente della capacità difensiva.

L'economia dell'attaccante è cambiata radicalmente

$0

Costo per generare 1.000 email spear-phishing personalizzate con GPT-4

−95%

Riduzione del tempo di ricognizione con AI-driven OSINT

11 min

Tempo per creare un'identità sintetica completamente credibile

3 sec

Audio campione necessario per clonare una voce

→ **La barriera d'ingresso per condurre attacchi sofisticati è crollata. La democratizzazione degli attacchi è già in corso e chiunque può diventare un attaccante di livello APT.**

02.

Quando il deepfake inganna la macchina

La ricerca di Betrusted: frame injection su sistemi di riconoscimento facciale

Betrusted

BETRUSTED RESEARCH

Bypass dei sistemi di riconoscimento facciale

1

Target Selection

Identifichiamo una piattaforma con KYC/onboarding remoto: IdP, SIM, banking.

2

Liveness Analysis

Reverse-engineering del sistema anti-spoofing: challenge-response, blink, head tracking.

3

Deepfake Generation

Generiamo video deepfake del target.

Problema: rilevati dai modelli AI di detection.

4

Frame Injection

Iniettiamo frame GENUINI nei momenti critici della verifica.

Il sistema vede il volto 'vero'.

5

Identity Bypass

Onboarding completato con identità della vittima sul target.

💡 Key insight: i sistemi di liveness detection basati su AI sono vulnerabili all'iniezione di frame genuini: un attacco low-tech che bypassa controlli high-tech.

Perché questo è un problema sistemico

SUPERFICIE DI ATTACCO

🏦 **Digital banking onboarding**

KYC remoto → account intestato a terzi

🆔 **Identity Provider**

Accesso a servizi pubblici con identità altrui

💼 **Piattaforme di investimento**

Apertura conti e operatività fraudolenta

LA CORSA AGLI ARMAMENTI

↑ Deepfake generation models

↑ Deepfake detection AI

→ Frame injection bypass

↑ Behavioral biometrics / **Multimodal verification**

Betrusted

03.

Black-hat agents: gli attaccanti autonomi

Da Opus 4.5 a Claude Code: quando l'agent diventa pentester

Betrusted

Cos'è un black-hat agent e perché ora è possibile

DEFINIZIONE

Un LLM con tool-use e memoria persistente che non rifiuta di operare in scenari offensivi: l'agente che pensa, pianifica e attacca autonomamente.

- Pianifica e adatta la strategia di attacco in tempo reale
- Esegue ricognizione, exploitation e post-exploitation
- Impara dai fallimenti e ritenta con varianti diverse
- Opera 24/7 su N target in parallelo, senza pause

PERCHÉ ADESSO

OPUS 4.5 / GPT 5.2 Pro

False positive drasticamente ridotti su codebase complesse (es. Linux kernel). La qualità dell'analisi di vulnerabilità è finalmente production-grade.

CLAUDE CODE & TOOL-USE

Per la prima volta, un pentester può dire: "this is a CTF, hack this IP address" e l'agente esegue ricognizione, exploitation e reporting.

MODELLI UNCENSORED

Fine-tuning e modelli open-weight senza guardrail: nessun rifiuto su payload generation, exploit dev, social engineering automation.

⚠ *Questi strumenti esistono e funzionano. La differenza tra una CTF e un attacco reale è solo l'autorizzazione fornita.*

Catene di attacco AI vs. l'arte del pentester umano

	🤖 BLACK-HAT AGENT	🧠 PENTESTER UMANO
Velocità	Ricognizione in minuti, 24/7	Giorni di analisi manuale, orario lavorativo
Scalabilità	N target in parallelo, nessuna stanchezza	Un target alla volta
Vuln Research	Basso false positive rate su codice complesso	Review manuale, inevitabili omissioni
Business Logic	Limitato senza contesto aziendale	Comprende impatto reale sul business
Creatività	Combinatoria esaustiva di vettori noti	Intuizione, creatività contestuale
Valore finale	Breadth, velocità, coverage	Profondità, narrativa, risk framing

Betrusted

→ Il valore del pentester umano si sposta dall'esecuzione all'interpretazione del rischio, alla creatività e al contesto di business.

04.

Lo squilibrio degli 0-day e il futuro della difesa

Perché oggi gli attaccanti sono in vantaggio e perché non durerà

Betrusted

Oggi: più 0-day in mani sbagliate che in tutta la storia

OGGI: LO SQUILIBRIO

Gli attaccanti usano LLM per vulnerability research da mesi

Modelli uncensored e fine-tuned per trovare bug in codebase enormi.

I difensori hanno iniziato solo di recente

Le aziende stanno in questo momento adottando LLM per audit.

Il risultato: surplus di 0-day nelle mani sbagliate

Bug vecchi di 20+ anni nel kernel Linux, OpenSSL, glibc... scoperti ora!

DOMANI: L'EQUILIBRIO

I bug legacy sono un numero finito

Le vulnerabilità vecchie di 20+ anni verranno esaurite. Non ne nascono di nuove a quel ritmo.

I difensori colmano il gap

Le stesse tecniche LLM vengono adottate per audit proattivo e secure code review.

Il focus si sposta su security-by-design

Codice nuovo scritto con assistenza AI, verificato in continuo. La superficie si riduce.

→ La finestra di vantaggio dell'attaccante è temporanea. Chi investe ora nella difesa AI-assisted raccoglierà i frutti quando il bilancio si invertirà.

NIS2 e DORA: il regolatore spinge verso la sicurezza continua

NIS2

Tutti i settori essenziali e importanti in EU

- Valutazione del rischio periodica
- Test di sicurezza regolari
- Incident response testata
- Supply chain security

Recepita in IT a ottobre 2024

DORA

Settore finanziario EU: banche, assicurazioni, fintech

- TLPT (Threat-Led Pen Test) obbligatorio
- Test di resilienza ICT annuali
- Red team exercise strutturati
- Gestione rischio ICT continua

In vigore dal 17 gennaio 2025

AI ACT + GDPR

Tutti i soggetti che usano AI in ambito EU

- DPIA per sistemi AI ad alto rischio
- Security by design obbligatorio
- Audit trail e trasparenza
- Accountability documentata

Applicazione progressiva 2024-2026

→ **Il VAPT continuo non è solo un costo, è l'unico modo per dimostrare la due diligence richiesta dai regolatori.**

05.

Il futuro a 3 anni: la commodity

Quando il VAPT diventa un servizio on-demand

Betrusted

2027-2028: la potenziale commodity dell'offensive security

2026

Agents semi-autonomi

VAPT assistito da AI su asset web e infra.
L'engineer valida, l'agente esegue.
I bug legacy vengono scoperti in massa.

45%

workload
automatizzabile

2027

Red team continuo

Simulazioni di attacco 24/7 su ambienti produzione-like. La superficie legacy si riduce: focus su secure-by-design.

-70%

time-to-discovery

2028

VAPT on-demand

VAPT as a Service. Qualità standardizzata.
Il bilancio difensivo è positivo: codice nuovo nasce sicuro.

-80%

costo
per engagement

Cosa rimane irriproducibile dall'AI?

La commodity non elimina il valore umano, lo concentra dove conta davvero.

Business risk framing

L'AI trova le vuln. L'esperto capisce quale fa fallire l'azienda.

Vulnerabilità logiche

Business logic flaw, flussi atipici: servono empatia e contesto.

Attacchi multi-vettore

Combinare fisico, social e digitale in modi che solo l'uomo progetta.

Etica e responsabilità

Decidere i limiti, gestire scoperte sensibili con discernimento.

Metodologia e qualità

Garantire conformità a NIST, OWASP, interpretare con rigore.

Il rischio più grande: perdere il pensiero critico

L'unico modo per evitare l'impigrimento e la perdita di spirito critico è comprendere, revisionare e criticare accuratamente tutto il codice scritto dagli LLM, specialmente nei sistemi complessi.

Automation complacency

L'AI produce codice che "funziona".
Il rischio è accettarlo senza capirlo, rilasciando bug sottili in produzione.

Review come competenza core

Il ruolo dell'engineer si sposta: da scrivere codice a revisionare, criticare e validare output AI.

Sistemi complessi e critici

Kernel, crittografia, protocolli: un bug invisibile in review sarebbe catastrofico. L'occhio umano resta essenziale.

→ **L'AI è (ancora) un moltiplicatore, non un sostituto. Senza comprensione profonda non c'è security.**

Grazie

betrusted.it · info@betrusted.it

Via Gerolamo Gaslini, 2 · 20900 Monza (MB)