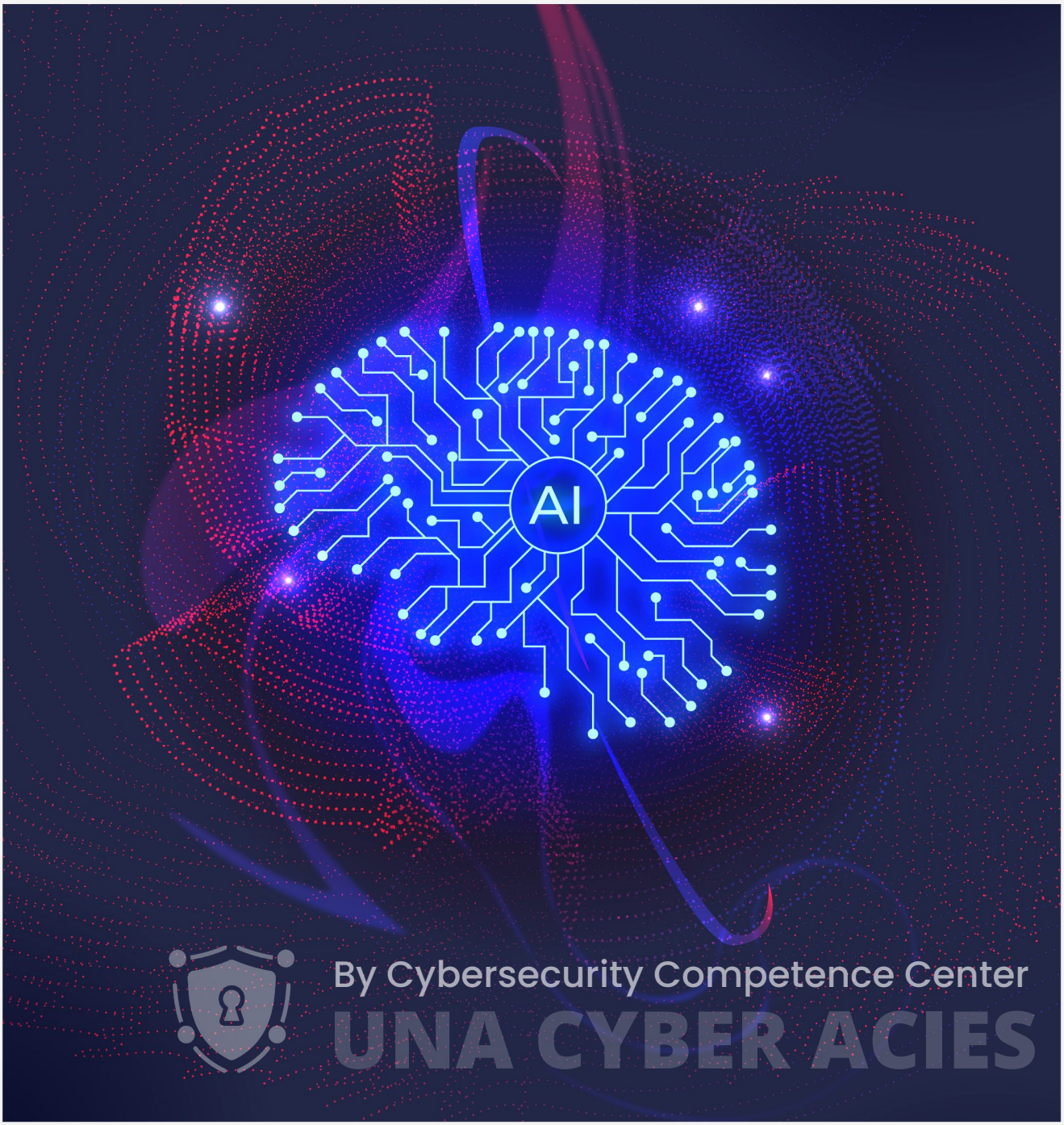
MATICMIND®
MAKES IT EASY

AI Threat Landscape



www.maticmind.it

Sommario

Executive Summary 03
Disclaimer per il lettore 05
Metodologia 06
Il contesto operativo: i numeri che disegnano la portata della minaccia 07

1. Introduzione: un equilibrio mancante 12
2. La prospettiva economica e geopolitica 13
   2.1. L'equazione sviluppo / rischio 14
   2.2. L'IA come moltiplicatore: come cambia la superficie d'attacco 15
   2.3. Punto critico: La progressiva saturazione dei sistemi di sicurezza tradizionali 18
3. Come l'IA diventa un'arma: vettori di attacco e impatti operativi 19
   3.1. Attaccare i modelli dall'interno 19
   3.2. Avvelenare i dati e rubare i modelli 20
   3.3. Malware e frodi 21
   3.4. Case Study: Cursor AI 22
   3.5. Case study: l'attacco a Nx 24
   3.6. Disinformazione 25
   3.7. Il mercato nero 28
4. L'impatto sulle organizzazioni 29
   4.1. Shadow IA: la minaccia dall'interno 30
   4.2. Una nuova normalità di rischio 30
5. Approfondimento: i mercati underground 31
   5.1. Dai primi esperimenti alle piattaforme "as-a-service" 31
   5.2. Un mercato che si professionalizza 34
   5.3. Un fenomeno destinato a crescere 34
6. Focus Normativo 36

Cyber Framework Maticmind 38
Una Cyber Acies 39
Il framework di sicurezza AI 40
Considerazioni finali 41
Fonti 42

Executive Summary

L'intelligenza artificiale è diventata in pochi anni una delle tecnologie più pervasive della nostra epoca. È ovunque: nei processi produttivi, nelle applicazioni di consumo, nelle piattaforme di comunicazione.

Ma con la stessa rapidità con cui si è diffusa, si sono moltiplicate anche le sue ombre. Non parliamo di scenari futuristici: oggi l'IA è già utilizzata come acceleratore del crimine informatico e come strumento di manipolazione dell'informazione.

La novità più dirompente è che la barriera d'ingresso si è abbassata drasticamente. L'accesso a modelli open source e strumenti "senza limiti" come WormGPT o FraudGPT ha reso possibile a chiunque – non solo agli attori statali o ai grandi gruppi criminali – lanciare attacchi complessi. In pochi click, tecniche una volta riservate a specialisti diventano disponibili a un pubblico molto più vasto.

Le conseguenze sono già sotto gli occhi di tutti. Abbiamo visto campagne di phishing scritte con una credibilità linguistica mai raggiunta prima; deepfake usati per truffe milionarie o per manipolare intere organizzazioni; malware che si riscrive e si adatta in autonomia per eludere i controlli.

Non si tratta più solo di teoria, ma di episodi concreti che hanno colpito aziende e istituzioni a livello globale.

I dati del primo semestre 2025 confermano la portata del fenomeno:

- Quasi il 50% di crescita negli attacchi IA-driven a livello globale (oltre 28 milioni di incidenti stimati a fine anno).
- In Europa, poco meno del 30% degli attacchi su base annua; in Italia, il 40% dei gravi episodi cyber ha già visto l'impiego diretto di GenAI.
- Oltre l'80% delle e-mail di phishing e il 91% delle campagne di spear-phishing integrano ora modelli linguistici; il 52% sfrutta LLM pubblici per generare payload o contenuti di phishing.
- I deepfake sono esplosi: da 500.000 nel 2023 a 8 milioni previsti nel 2025, responsabili di 1 su 20 fallimenti di verifica identitaria.
- Il costo medio di una violazione AI-powered ha raggiunto 5,72 milioni di dollari (+13% YoY), con impatti più pesanti sulle PMI (+27% nei costi di risposta) e pagamenti assicurativi in crescita del 22%.

Questa evoluzione porta con sé quattro messaggi chiave:

I. La superficie d'attacco si è estesa in maniera senza precedenti: ogni nuovo modello, applicazione o integrazione basata su IA diventa un potenziale punto di ingresso.

II. Le difese non sono al passo: molte organizzazioni non hanno playbook per incidenti legati a deepfake, non integrano la GenAI nei processi di sicurezza e tollerano fenomeni di "shadow IA" fuori controllo.

III. Il crimine è diventato più veloce e più accessibile: automazione e personalizzazione degli attacchi riducono i tempi, i costi e le competenze necessarie per colpire.

IV. Progressiva saturazione dei sistemi di sicurezza tradizionali: le architetture basate su regole statiche non reggono il volume delle minacce AI-driven, generando alert fatigue e inefficienza operativa.

Guardando avanti, lo scenario si complica ulteriormente. I sistemi agentici aprono la strada a nuovi vettori d'attacco (dalla manipolazione della memoria all'esecuzione di codice malevolo), mentre l'IA multimodale promette di automatizzare l'intera catena offensiva: ricognizione, phishing, exploit e movimenti laterali, senza più l'intervento umano.

Parallelamente, il quadro normativo si muove in direzioni diverse: l'Europa con l'IA Act, gli Stati Uniti con un approccio più frammentato, l'Asia e l'Africa ancora in fase di definizione.

In questo contesto, un messaggio è chiaro: la sicurezza non può più essere pensata con schemi tradizionali. Serve un approccio IA-native, capace di anticipare queste minacce e di governarle, non inseguirle.

L'IA continuerà a essere un moltiplicatore di opportunità, ma senza un'adeguata consapevolezza e senza investimenti mirati, rischia di diventare soprattutto un moltiplicatore di vulnerabilità.

Disclaimer per il lettore

L'intelligenza artificiale è un contenitore ampio: include algoritmi di analisi, sistemi predittivi, strumenti di ottimizzazione che da anni supportano imprese e governi. Ma la vera discontinuità degli ultimi anni è la Generative AI (GenAI).

A differenza delle forme tradizionali, che classificano, ordinano e analizzano dati, la GenAI è in grado di creare contenuti originali: testi, immagini, audio, video, codice. È proprio questa capacità generativa a renderla, nello stesso tempo, l'elemento più innovativo e il più esposto ad abusi.

Se i modelli predittivi hanno richiesto per lungo tempo competenze specialistiche e infrastrutture dedicate, la GenAI ha abbattuto la barriera di ingresso: modelli open source, strumenti "senza limiti" e interfacce user-friendly consentono oggi a chiunque di lanciare attacchi che un tempo erano appannaggio esclusivo di gruppi statali o criminali organizzati.

Per questa ragione, quando entreremo nel vivo dell'analisi, parleremo più spesso di GenAI che di IA in senso lato.

È qui che si concentra la crescita di mercato, è qui che si moltiplicano i rischi di abuso, ed è qui che si stanno giocando le partite normative e geopolitiche più rilevanti.

In altre parole, la GenAI è oggi il terreno su cui si decide se l'intelligenza artificiale sarà un moltiplicatore di opportunità o di vulnerabilità.

Per motivi di coerenza stilistica, in questo documento saranno presenti entrambe le diciture, ma nello specifico stiamo parlando di GenAI.

Metodologia

Questo report nasce dall'integrazione di fonti diverse e complementari, con l'obiettivo di restituire un quadro il più possibile aderente alla realtà del rischio cyber abilitato dall'intelligenza artificiale. Non si tratta quindi di una ricerca esclusivamente teorica, ma di un lavoro che combina numeri, esperienze e osservazioni dirette dal campo.

L'approccio scelto è stato multi-sorgente. Da un lato, abbiamo analizzato dati quantitativi – provenienti da report di settore, pubblicazioni accademiche e white paper – per individuare tendenze e correlazioni. Dall'altro, abbiamo arricchito questa base con elementi qualitativi, che danno spessore al contesto e aiutano a interpretare i dati alla luce di scenari concreti.

Un elemento distintivo è stato l'apporto della Business Unit Cyber di Maticmind, che ha contribuito con evidenze raccolte sul campo attraverso le sue strutture specializzate: Offensive Team, Cyber Defence Center, Twin4Cyber, Cyber Threat Intelligence Team e Incident Response Management Team. Le loro attività quotidiane – dal monitoraggio delle minacce alla risposta a incidenti reali – hanno offerto insight diretti su come i rischi si manifestino concretamente nelle organizzazioni.

Il processo metodologico ha seguito tre fasi principali:

1. **Raccolta dati:** combinazione di fonti primarie (esperienze e incidenti osservati dal team Cyber Maticmind) e secondarie (ricerche, studi, rapporti tecnici, bollettini di sicurezza).
2. **Analisi:** interpretazione critica delle informazioni, con particolare attenzione a trend emergenti e nuove tipologie di attacco.
3. **Sintesi e visualizzazione:** produzione di grafici, infografiche e schemi utili a rendere leggibili fenomeni complessi e ad evidenziare connessioni chiave.

Questo approccio ci ha permesso di andare oltre la fotografia statica del fenomeno, costruendo una narrazione che unisce dati verificati, scenari in evoluzione e intuizioni tratte dall'esperienza sul campo.

Il contesto operativo: i numeri che disegnano la portata della minaccia

Prima di entrare nel dettaglio dei vettori d'attacco e delle dinamiche geopolitiche, è fondamentale inquadrare il fenomeno con i dati operativi più recenti. I numeri del primo semestre 2025 non lasciano spazio a dubbi: l'IA non è più un'arma potenziale, ma una componente strutturale del panorama delle minacce cyber.

L'escalation globale è inequivocabile:

- Gli attacchi basati sull'Intelligenza Artificiale sono aumentati del 47% su base globale rispetto allo stesso periodo dell'anno precedente.
- Si stima che il totale degli incidenti cyber globali guidati dall'IA supererà i 28 milioni nel 2025.

L'impatto sul Vecchio Continente e in Italia:

- L'Europa ha subito un'accelerazione significativa, con un incremento del 22% nel volume di attacchi su base annua nel Q2 2025.
- In Italia, la situazione è particolarmente critica: nel primo trimestre 2025, su quasi 900 gravi episodi informatici registrati, circa il 40% ha visto l'impiego diretto di IA generativa.

La qualità degli attacchi è cambiata radicalmente:

- Oltre l'80% delle e-mail di phishing e il 91% delle campagne di spear-phishing rilevate nel 2025 incorporano modelli linguistici (LLM) per aumentarne credibilità ed efficacia.
- Il 52% degli attacchi IA ha sfruttato LLM pubblici per generare contenuti malevoli, rendendo le campagne più scalabili ed economiche.

L'ascesa dei deepfake e dell'offuscamento:

- Si prevede che nel 2025 saranno condivisi oltre 8 milioni di deepfake, un aumento esponenziale rispetto ai 500.000 del 2023. A livello globale, 1 su 20 fallimenti nella verifica dell'identità è già riconducibile a questa tecnologia. Nel 2025 infatti l'uso di video deepfake nei casi di frode ai danni dei CEO è aumentato dell'83%, generando perdite dirette stimate in 1,1 miliardi di dollari. Nello stesso anno, il 37% delle grandi aziende ha dichiarato di aver subito almeno un tentativo di impersoni-

ficazione vocale tramite deepfake. I deepfake basati su animazioni facciali hanno mostrato una capacità di eludere i sistemi KYC (Know Your Customer) nel 12% dei casi.

- Il 20% degli attacchi cyber nel 2025 ha impiegato tecniche di offuscamento potenziate dall'IA per eludere le difese.

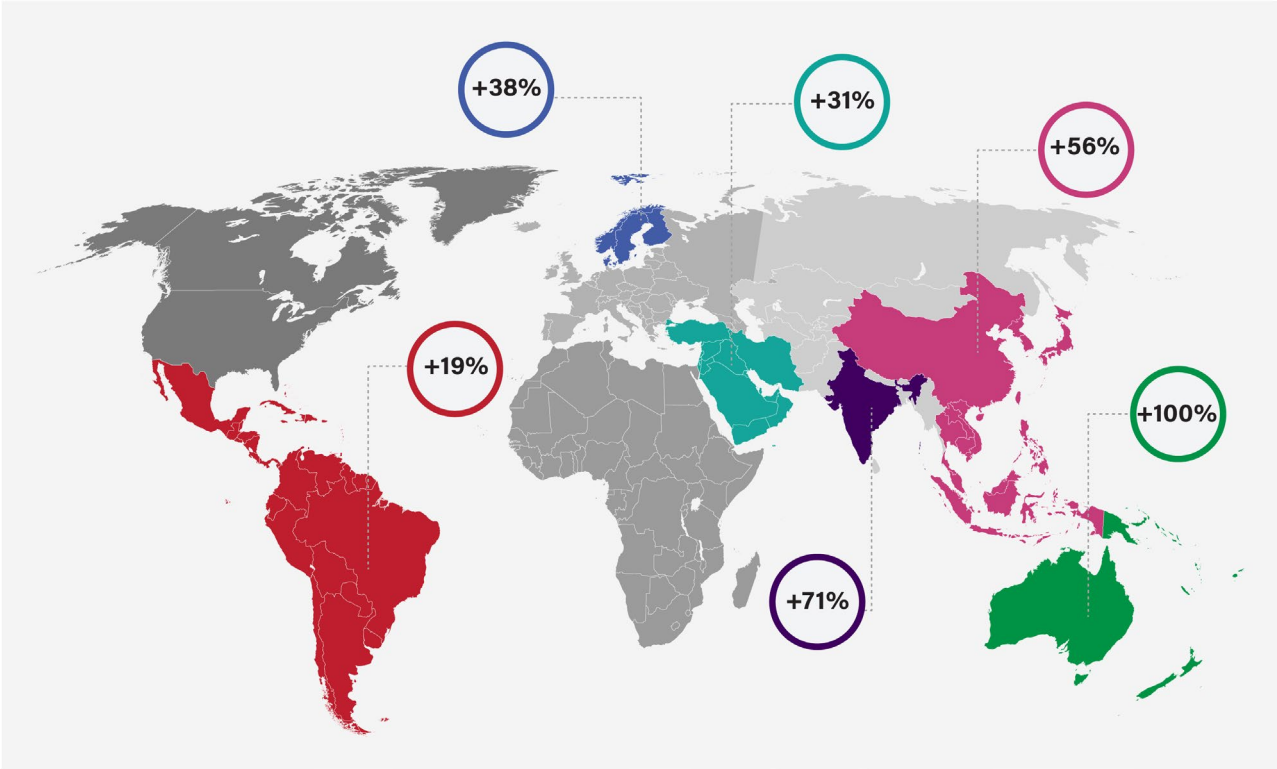
L'era dell'autonomia e dell'adattamento:

- Il 14% delle principali violazioni aziendali è stato completamente autonomo, senza necessità di intervento umano dopo il lancio.
- I malware autonomi, in grado di adattarsi all'ambiente host, hanno rappresentato il 23% dei payload nel 2025, con un tasso di successo superiore del 18% nell'aggirare i sistemi di rilevamento.
- Il 41% delle famiglie di ransomware include ora moduli IA per l'adaptive payload delivery, con malware che si adattano agli ambienti sandbox in soli 11 secondi.

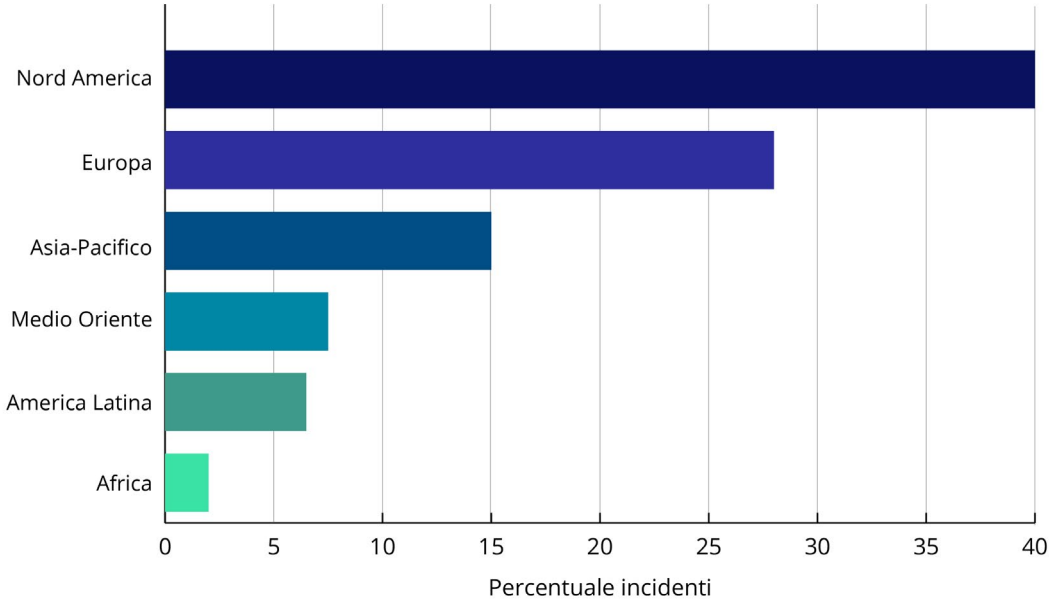
L'impatto economico è tangibile:

- Il costo medio di una violazione di dati potenziata dall'IA ha raggiunto i 5,72 milioni di dollari nel 2025 (+13% YoY).
- Le PMI hanno speso il 27% in più per la risposta agli incidenti, mentre i pagamenti assicurativi per attacchi guidati dall'IA sono aumentati del 22%.

Paesi con crescita degli incidenti AI-driven (%)



Distribuzione globale degli incidenti AI-driven (%)



Distribuzione Geografica:

- **Nord America** si è confermato l'epicentro delle minacce AI-driven, concentrando quasi il 40% degli incidenti segnalati a livello mondiale.
- **Europa** ha seguito da vicino con il 28% delle violazioni, in particolare in Germania e Regno Unito, divenuti obiettivi privilegiati.
- Nella regione **Asia-Pacifico** si è osservata una crescita significativa (+56%), che ha colpito soprattutto il comparto finanziario e le piattaforme di e-commerce, settori ad alta digitalizzazione e transazionalità.
- Il **Medio Oriente** ha registrato un aumento del 31% nelle operazioni di cyber-spionaggio assistite da IA, focalizzate in particolare sulle infrastrutture petrolifere ed energetiche.
- **L'America Latina** è emersa come nuova area bersaglio, con un incremento del 19% negli attacchi di malware bancario potenziati dall'AI, segnale di un interesse crescente da parte dei cybercriminali verso i sistemi finanziari locali.
- In **Africa** sono state rilevate per la prima volta campagne di disinformazione alimentate da IA, concentrate principalmente in contesti elettorali e politici.
- **Russia e Cina** sono state ricondotte al 42% delle operazioni di matrice statale basate su IA, secondo le stime di intelligence internazionale, confermandone il ruolo dominante nello scenario geopolitico cyber.
- **L'India** ha registrato una crescita senza precedenti (+71%) negli attacchi di phishing generati da IA, mirati soprattutto alle piattaforme di pagamento digitale, sempre più diffuse nel Paese.
- **Australia e Nuova Zelanda** hanno riportato un raddoppio delle minacce AI-related rispetto all'anno precedente, a testimonianza di un'evoluzione rapida e consistente del fenomeno anche nell'area oceanica.
- Infine, nei **Paesi scandinavi** si è osservato un incremento del 38% nell'impiego di deepfake basati su IA, utilizzati in campagne di impersonificazione aziendale per finalità fraudolente.

Distribuzione settoriale:

Settore Finanziario:

il più colpito, con il 33% di tutti gli incidenti guidati dall'IA.

Settore Healthcare:

+76% negli attacchi mirati, principalmente per l'automazione degli attacchi ransomware.

Questi dati non sono un semplice aggiornamento statistico: sono la base empirica su cui si fonda l'analisi di questo report. Nei capitoli che seguono, esploreremo non solo "cosa" sta accadendo, ma "come" e "perché", collegando ogni cifra a un vettore d'attacco, a una vulnerabilità organizzativa o a una dinamica di mercato.

1. Introduzione: un equilibrio mancante

L'arrivo dell'intelligenza artificiale generativa sul mercato ha rappresentato un vero spartiacque. Per le imprese, è stata una rivoluzione dalle potenzialità enormi: automazione, ottimizzazione dei processi, nuovi servizi. Ma a questo entusiasmo non ha fatto seguito un'adeguata consapevolezza dei rischi.

Oggi le aziende e gli enti pubblici, in Italia come nel resto del mondo, si trovano a vivere una doppia sfida: da un lato devono integrare rapidamente le nuove capacità offerte; dall'altro devono difendersi da un panorama di minacce che evolve con una velocità e una complessità senza precedenti. Due esigenze apparentemente complementari, che però spesso restano in squilibrio.

La realtà è che la corsa all'adozione non è stata accompagnata da una pari maturità sul piano della sicurezza. La popolarità di strumenti come ChatGPT, arrivato a superare i 200 milioni di utenti attivi, ha avuto un effetto collaterale imprevisto: ha accelerato anche l'abuso della tecnologia per scopi criminali.

Il vero cambio di passo, tuttavia, non è avvenuto con le piattaforme commerciali, bensì con la diffusione di modelli open source. Accessibili a chiunque e facilmente adattabili, questi strumenti permettono di aggirare restrizioni etiche e di sicurezza integrate nei modelli più noti. È così che sono nati strumenti come WormGPT, apparso nel 2023, progettato specificamente per attività illecite, seguito a breve distanza da alternative come FraudGPT.

Questo fenomeno ha un impatto decisivo: la democratizzazione dell'IA generativa ha abbassato drasticamente la soglia tecnica necessaria per orchestrare attacchi sofisticati. Se prima certe operazioni erano appannaggio esclusivo di gruppi criminali ben organizzati o di attori statali, oggi persino individui con risorse limitate possono generare minacce su larga scala.

In altre parole, lo stesso processo che ha reso la GenAI una leva di innovazione senza precedenti, ne ha fatto anche un moltiplicatore di vulnerabilità. Ed è proprio da questo squilibrio che prende forma il filo conduttore di questo report: comprendere come questa stia trasformando il crimine informatico

2. La prospettiva economica e geopolitica

Il mercato dei modelli di intelligenza artificiale generativa è in una fase di crescita esponenziale. Secondo proiezioni conservative, la spesa passerà dai 14 miliardi di dollari del 2025 ai 75 miliardi nel 2029. Non si tratta soltanto di numeri: questa espansione rappresenta un cambiamento strutturale nella direzione stessa dello sviluppo tecnologico.

Se in un primo momento l'attenzione era concentrata sui modelli general-purpose, di grandi dimensioni e ad ampio spettro, oggi la tendenza è verso modelli specializzati e architetture composite, progettati per casi d'uso specifici. Questo spostamento riflette un'esigenza di efficienza, rilevanza contestuale e prestazioni mirate: un modello "cucito su misura" per il settore finanziario o sanitario, per esempio, può offrire risultati più efficaci di un sistema generalista.

Market Size of Generative IA Models, 2023–2029
(in Millions of U.S. Dollars)

	2023	2024	2025	2026	2027	2028	2029	CAGR 2024- 2029
Total models	1,360	5,719	14,200	25,766	42,714	59,940	75,265	66.8%
YoY growth (%)	–	320.7%	149.8%	80.8%	64.1%	39.6%	24.9%	–

Tabella 1

Ma a ogni nuova implementazione corrisponde anche un nuovo punto di vulnerabilità. Ogni modello, ogni applicazione, ogni integrazione introduce potenziali porte d'ingresso per gli attaccanti. E la rapidità con cui molte aziende adottano soluzioni IA – spesso bypassando controlli di governance e sicurezza – rende questo scenario ancora più critico.

La diversificazione dei modelli e delle architetture aumenta la complessità del panorama delle minacce. Non ci troveremo più davanti a rischi generici, ma a vulnerabilità mirate: un modello per il settore finanziario avrà falle diverse rispetto a uno multimodale per la sanità. Inoltre, le architetture composite moltiplicano i rischi anche a livello di interazione tra i componenti, creando catene di dipendenze difficili da proteggere.

2.1. L’equazione sviluppo / rischio

La crescita della spesa per l’IA generativa non è uniforme a livello globale. Nel 2025, il Nord America detiene il 57% degli investimenti, seguito dall’Europa con il 25%. Le economie mature – Nord America, Europa, Asia/Pacifico e Giappone – guidano l’adozione grazie a infrastrutture robuste, forza lavoro qualificata e investimenti in ricerca.

Le economie emergenti, invece, pur mostrando interesse, faticano a tenere il passo per via di limiti infrastrutturali, scarsa disponibilità di risorse computazionali e carenza di talenti specializzati. Questo squilibrio geografico si riflette direttamente anche nella distribuzione delle minacce: le regioni più avanzate, dotate di asset digitali e sistemi critici basati su IA e GenAI, diventano bersagli privilegiati per attacchi sempre più sofisticati – dai deepfake iperrealistici agli exploit automatizzati, fino alla manipolazione massiva dell’informazione.

Nei Paesi in via di sviluppo, invece, il rischio si manifesta in forme diverse: attacchi meno complessi, ma facilitati dall’assenza di infrastrutture di sicurezza robuste, e possibili utilizzi del territorio come banco di prova per tecniche emergenti.

Il fattore geopolitico

Un’altra implicazione di vasta portata è il ruolo dell’IA generativa come strumento di influenza geopolitica. Si potrebbe configurare una corsa agli armamenti digitali che intensifica le rivalità geopolitiche, poiché le nazioni cercano di sfruttare il potenziale dell’IA per ottenere vantaggi economici, supremazia tecnologica e influenza sulle norme e gli standard globali. La capacità di sviluppare e controllare tecnologie avanzate si traduce direttamente in potere, creando una competizione tra stati non solo per lo sviluppo, ma anche per il controllo e la governance dell’IA. La proliferazione di sistemi GenAI amplifica i rischi geopolitici, poiché attori statali e non statali possono utilizzare queste tecnologie per scopi malevoli su larga scala, minando la sicurezza nazionale e la fiducia nelle istituzioni. Questo include la manipolazione dell’opinione pubblica, l’interferenza nelle elezioni e l’escalation della cyberwar, rendendo la sicurezza dell’IA una questione di sicurezza nazionale e internazionale.

Cosa significa operativamente:

- La domanda spinge verso specializzazione → servono controlli dedicati per filiere e domini.
- La concentrazione geografica dell’adozione crea hotspot di rischio avanzato.
- Le architetture composite richiedono threat modeling sulle interazioni, non solo sui singoli componenti.
- Governance e security devono correre in parallelo alla crescita del business, non in coda.

2.2. L’IA come moltiplicatore: come cambia la superficie d’attacco

L’adozione esponenziale dell’intelligenza artificiale generativa non rappresenta solo un fenomeno tecnologico o economico, ma un cambiamento strutturale del panorama della sicurezza. L’IA non aggiunge semplicemente nuovi strumenti: moltiplica la superficie d’attacco complessiva, trasformando ogni interazione, dato e modello in un potenziale vettore di compromissione. Questa espansione non è teorica, ma misurabile. I dati di mercato e i trend di adozione mostrano un incremento geometrico dei punti di esposizione, con conseguente crescita della vulnerabilità globale.

2.2.1 L’equazione semplice: più modelli = più superficie

Ogni nuovo modello implementato, ogni applicazione GenAI integrata in un workflow aziendale, ogni endpoint che dialoga con un sistema intelligente diventa un potenziale punto d’ingresso. Secondo le previsioni, la spesa globale per i modelli GenAI passerà da 14 miliardi di dollari nel 2025 a 75 miliardi nel 2029. In soli quattro anni, il numero di asset IA in produzione crescerà di oltre il 400%. Parallelamente, il mercato si sposta verso modelli specializzati: dal 2% nel 2023 al 12% nel 2029 [[2.1]]. Ogni modello di dominio introduce vulnerabilità mirate, che rendono la difesa più complessa e distribuita.

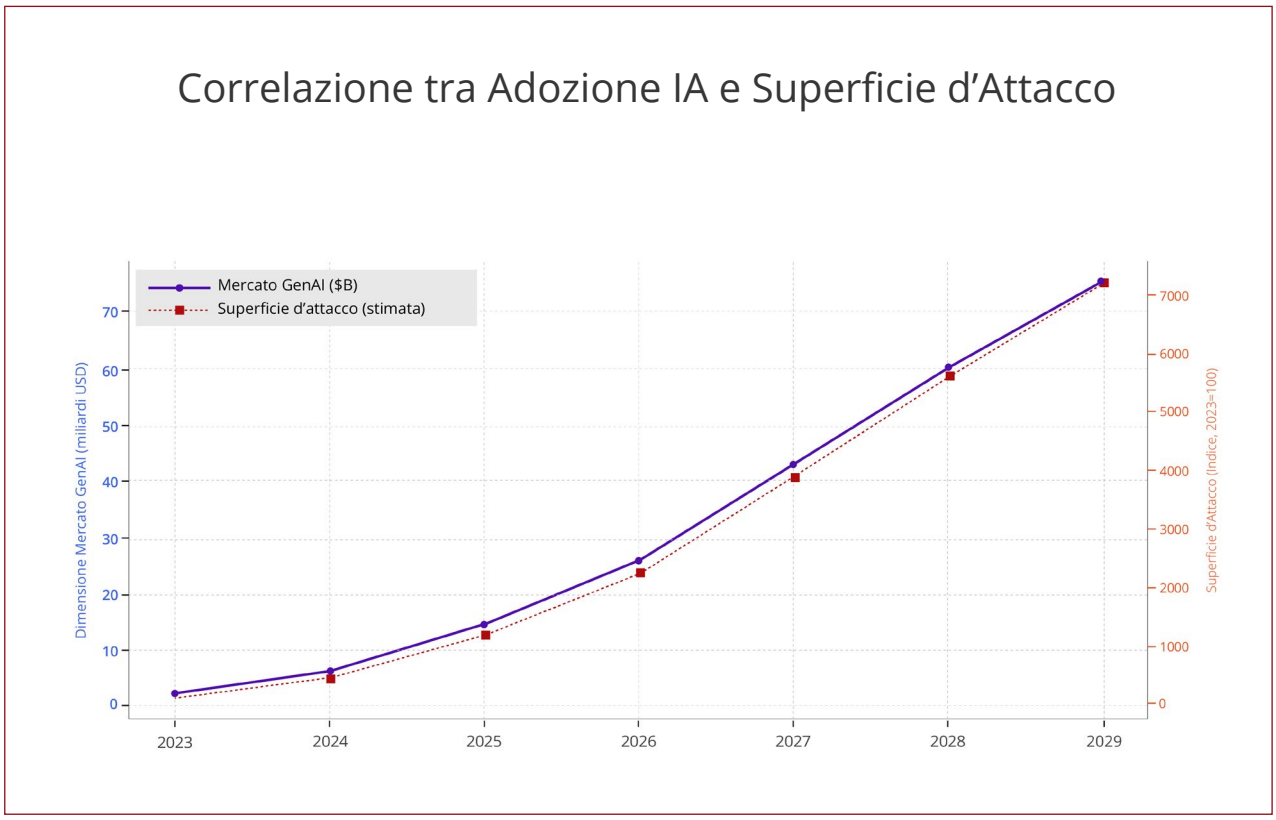


FIGURA 2.2.1. Correlazione tra la crescita del mercato globale dei modelli GenAI (in miliardi di USD, linea blu) e la relativa espansione stimata della superficie d'attacco (indice, linea arancione). I dati di mercato sono tratti dalla Tabella 1 del report. La superficie d'attacco è stimata crescere a un tasso superiore (fattore 1.1x) rispetto al mercato stesso, per rappresentare l'effetto moltiplicatore documentato nel report: l'aumento dell'efficacia degli attacchi (84% dei CISO) e l'abbassamento della barriera d'ingresso per gli attaccanti grazie al mercato nero dell'IA. Fonte: Elaborazione su dati Gartner e Maticmind Cyber BU (2025).

2.2.2. L'effetto moltiplicatore: come l'IA amplifica la minaccia

La crescita non riguarda solo il numero di asset. L'IA agisce da moltiplicatore di forza per gli attaccanti, amplificando l'efficacia di minacce esistenti:

I. Aumento della qualità e dell'efficacia – L'84% dei CISO ha osservato un salto qualitativo nelle campagne di phishing, attribuibile alla generazione automatica di testi [[4]]. Una maggiore credibilità si traduce in un aumento diretto del tasso di successo degli attacchi.

II. Abbassamento della barriera d'ingresso – Strumenti "black hat" come WormGPT e FraudGPT, reperibili per poche decine di dollari al mese, hanno reso accessibili a chiunque tecniche avanzate di attacco [[3.4, 5]]. Questo ha moltiplicato la platea di potenziali aggressori, democratizzando il cybercrime.

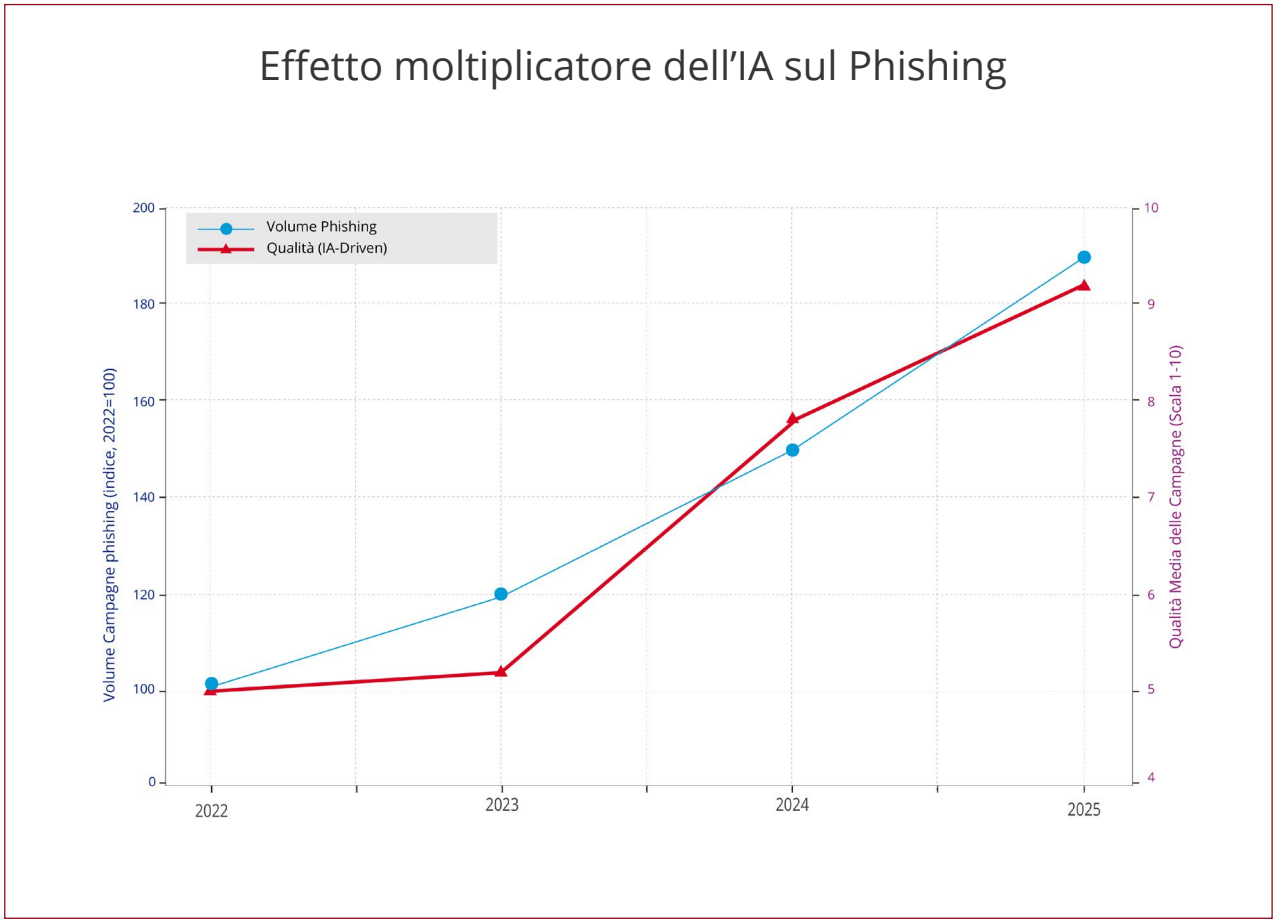


FIGURA 2.2.2 : L'effetto moltiplicatore dell'IA sugli attacchi di phishing. Il grafico mostra l'andamento del volume delle campagne di phishing (linea grigia, asse sinistro) a confronto con la qualità percepita media delle stesse campagne (linea rossa, asse destro), basata sulle osservazioni del 84% dei CISO riportate nel Capitolo 4. L'impennata nella qualità a partire dal 2023 coincide con la diffusione massiccia dei modelli GenAI, dimostrando come l'IA amplifichi l'efficacia delle minacce esistenti. Fonte: Elaborazione su dati interni Maticmind Cyber BU (2025).

2.2.3. La nuova dimensione: dalla superficie perimetrale a quella generativa

Tradizionalmente, la superficie d'attacco era definita da elementi perimetrali: porte di rete, server web, applicazioni esposte su internet. L'IA ha aggiunto una nuova dimensione: la superficie d'attacco generativa. Questa comprende:

- **I Prompt:** Ogni input utente a un sistema GenAI è ora un vettore potenziale per attacchi come la prompt injection [[3.1]].
- **I Dati di Training:** I dataset utilizzati per addestrare i modelli sono diventati bersagli privilegiati per attacchi di data poisoning [[3.2]].
- **I Modelli Stessi:** I modelli di IA sono asset che possono essere rubati (model extraction) o manipolati per rivelare informazioni sensibili (model inversion) [[3.2]].

Questa nuova dimensione è intrinsecamente più difficile da mappare e difendere rispetto a un server o a un firewall. Non è un punto fisso, ma un flusso dinamico di dati e interazioni. La sua gestione richiede un cambio di paradigma, che sarà esplorato nei capitoli successivi. L'IA ha trasformato la superficie d'attacco da un concetto statico e misurabile in un ecosistema dinamico e in continua espansione. I dati di mercato non mentono: con una crescita prevista di oltre il 400% nei prossimi cinque anni, e con l'IA che agisce come moltiplicatore di forza per gli attaccanti, la superficie d'attacco globale è destinata a diventare più vasta, più complessa e più pericolosa che mai. I capitoli che seguono esploreranno nel dettaglio i vettori di attacco specifici (Capitolo 3) e il loro impatto concreto sulle organizzazioni (Capitolo 4) che questa espansione ha reso possibili.

2.3. Punto critico: La progressiva saturazione dei sistemi di sicurezza tradizionali

L'incremento degli attacchi AI-driven ha reso evidente il limite strutturale delle architetture di sicurezza tradizionali. I sistemi basati su signature statiche, regole deterministiche e correlazioni predefinite risultano progressivamente inefficaci contro minacce che evolvono in tempo reale e che adottano tecniche di adversarial machine learning, offuscamento dinamico e payload adattivi. La conseguenza operativa è un sovraccarico dei motori SIEM e delle piattaforme di detection, incapaci di gestire la crescita continua e non sostenibile del volume di eventi da correlare. Questa pressione si riflette in modo diretto sul carico operativo degli analisti SOC di primo livello, spesso vincolati ad attività di pre-analisi ripetitive come la lettura, la categorizzazione e l'arricchimento manuale dei ticket. Tale modello operativo, oltre a essere inefficiente, si rivela difficilmente scalabile in un contesto in cui il volume degli incidenti cresce costantemente. Il rischio è duplice: da un lato, l'"alert fatigue" riduce la prontezza e aumenta la probabilità di compromissioni silenziose; dall'altro, si crea un divario crescente tra la complessità delle minacce e la capacità di risposta effettiva delle organizzazioni.

In questo scenario diventa cruciale introdurre approcci AI-native, capaci di alleggerire e razionalizzare le prime fasi di gestione. Un esempio concreto è rappresentato dal progetto SocAI di Maticmind, che automatizza la pre-analisi dei ticket integrandosi con sistemi esistenti, utilizzando modelli linguistici on-premise per classificare, arricchire e fornire raccomandazioni operative strutturate. Questa soluzione consente di filtrare i falsi positivi, categorizzare gli alert in base al rischio e produrre bozze di segnalazione già pronte per la validazione. L'obiettivo strategico è chiaro: aumentare l'efficienza operativa, liberare gli analisti dalle attività a basso valore e garantire la scalabilità del servizio SOC senza un aumento proporzionale dei costi del personale.

3. Come l'IA diventa un'arma: vettori di attacco e impatti operativi

Come già accennato, l'intelligenza artificiale generativa non ha solo rivoluzionato la produttività delle aziende e le modalità di interazione delle persone con la tecnologia. Ha anche cambiato radicalmente il modo in cui i criminali informatici conducono le proprie operazioni. Se fino a pochi anni fa gli attacchi dipendevano in larga parte dalla manualità e dalla creatività degli attaccanti, oggi la GenAI è diventata un moltiplicatore: aumenta la velocità, riduce le barriere tecniche, amplia la portata e la credibilità delle minacce.

Questa sezione esplora come i modelli vengano manipolati, sfruttati e addirittura reinventati a scopi criminali.

3.1. Attaccare i modelli dall'interno

Il primo fronte di attacco è rappresentato dalle tecniche che mirano a manipolare direttamente i modelli di intelligenza artificiale. La più nota è la *prompt injection:* un input apparentemente innocuo che in realtà contiene istruzioni nascoste capaci di alterare il comportamento del modello. In molti casi, queste istruzioni portano il sistema a ignorare i protocolli di sicurezza o a generare contenuti che normalmente verrebbero bloccati.

Una variante avanzata è il *jailbreaking*, che spinge il modello a disattivare del tutto i propri vincoli. Un modello "*jailbreakato*" può arrivare a rivelare i propri system prompt interni, a produrre codice malevolo, a rispondere a domande che violano linee guida etiche. Con il tempo sono emerse tecniche ancora più raffinate: dal *token smuggling*, che sfrutta il modo in cui i modelli gestiscono la *tokenizzazione*, fino all'adversarial prompting o alla context contamination, che inserisce istruzioni manipolatorie dentro contenuti apparentemente neutrali.

Ciò che colpisce è che, nonostante i progressi nelle difese, persino i *jailbreak* più semplici continuano a funzionare. Il confine tra ciò che un modello "dovrebbe fare" e ciò che può essere indotto a fare resta sottile e fragile.

Accanto a questi attacchi diretti, si diffondono quelli indiretti. Conosciuti come indirect prompt injection, non agiscono direttamente sull'input dell'utente ma sulle fonti da cui il modello attinge: un PDF, una pagina web, un dataset esterno. Basta inserire istruzioni malevole nel contenuto e, al momento della consultazione, il modello le esegue. In questo modo, l'attacco può restare invisibile a lungo, annidato nelle fonti stesse del sistema.

3.2. *Avvelenare* i dati e rubare i modelli

Oltre a manipolare i prompt, i criminali hanno iniziato a colpire la fase di addestramento dei modelli. Con il data poisoning – appunto, avvelenamento -, dati "avvelenati" vengono introdotti nei dataset di training, inquinando le fondamenta stesse del sistema. Un modello così compromesso può contenere bias nascosti o vulnerabilità deliberate. Nel campo della sicurezza, per esempio, questo significa avere strumenti di rilevamento che scambiano malware per software legittimo, o soluzioni antifrode che lasciano passare transazioni sospette.

Un caso recente e significativo è rappresentato dal network russo CopyCop (Storm-1516), identificato come un'operazione coordinata che utilizza LLM open-source modificati, deepfake e siti clone per diffondere narrativa filo-russa. L'obiettivo dichiarato di CopyCop va oltre la semplice propaganda: saturare la rete di contenuti manipolati per avvelenare i dataset dei modelli generativi occidentali, con impatti diretti sui media, sui processi decisionali automatizzati e sui sistemi cognitivi delle democrazie.

Il fenomeno può avere molteplici conseguenze: i modelli generativi rischiano di assimilare bias e produrre output distorti; l'opinione pubblica può essere manipolata attraverso contenuti falsi amplificati dall'IA; i sistemi automatizzati possono prendere decisioni influenzate da dati corrotti; infine, media, cittadini e istituzioni possono vedere erosa la fiducia negli strumenti digitali e nei sistemi di intelligenza artificiale.

L'emergere di operazioni come CopyCop evidenzia come la disinformazione non sia più solo un fenomeno mediatico, ma un rischio tecnologico e cognitivo. Proteggere i dataset e sviluppare sistemi di IA resilienti diventa quindi un imperativo strategico per salvaguardare l'integrità dei processi decisionali, la sicurezza dei sistemi digitali e la stabilità delle democrazie nell'era dell'intelligenza artificiale generativa.

Altre tecniche puntano a estrarre valore direttamente dai modelli. La *model extraction* consente di ricostruire una copia del modello interrogandolo ripetutamente, con un danno diretto per la proprietà intellettuale. La *model inversion*, invece, permette di risalire a informazioni sensibili contenute nei dati di training. È come se, facendo domande in modo strategico, si riuscisse a svelare dettagli privati che non avrebbero mai dovuto emergere.

Queste pratiche non solo compromettono la sicurezza, ma minacciano la fiducia stessa nell'IA. Se un modello può essere copiato o rivelare informazioni riservate, diventa difficile considerarlo un asset affidabile.

3.3. Malware e frodi

Il 2025 ha segnato una svolta anche nel panorama dei malware e delle frodi. Oltre all'aumento della sofisticazione dei ransomware già osservato negli anni precedenti, si è registrata la comparsa di nuove famiglie di codice malevolo potenziate dall'intelligenza artificiale. I malware autonomi, in grado di adattarsi in base agli ambienti di risposta dell'host, hanno rappresentato il 23% dei payload nel 2025, mentre le varianti generate dall'IA hanno mostrato un tasso di successo superiore del 18% nell'eludere i sistemi di rilevamento degli endpoint. Parallelamente, il 41% delle famiglie di ransomware include oggi moduli IA per l'adaptive payload delivery. Grazie a tecniche di apprendimento per rinforzo, questi codici riescono ad adattarsi agli ambienti sandbox in soli 11 secondi. Anche i trojan hanno beneficiato dell'automazione: nel 2025, il 18% di essi ha sfruttato capacità di persistenza basate sull'IA, resistendo a riavvii e tattiche comuni di rimozione. Inoltre, l'offuscamento potenziato dall'IA ha ritardato le attività di reverse engineering di una media di 3,2 giorni, complicando il lavoro dei team forensi.

KOSKE

Il malware KOSKE rappresenta un punto di svolta nell'evoluzione delle minacce informatiche, segnando l'ingresso dell'intelligenza artificiale nel panorama del malware development. Scoperto dai ricercatori di Aqua Nautilus nel luglio 2025, KOSKE non è semplicemente un altro cryptominer Linux, ma un esempio sofisticato di come l'IA possa essere utilizzata per creare malware più adattivo, evasivo e persistente.

KOSKE si distingue per la sua architettura modulare e le sue capacità di adattamento dinamico, caratteristiche che suggeriscono fortemente l'utilizzo di tecniche di sviluppo assistite dall'intelligenza artificiale. Il malware è progettato specificamente per ambienti Linux e si concentra principalmente sul cryptomining, ma le sue capacità vanno ben oltre quelle di un tradizionale miner.

LameHug

LameHug rappresenta un milestone storico nell'evoluzione del malware, essendo il primo esempio documentato pubblicamente di malware che utilizza attivamente un Large Language Model per generare comandi di attacco in tempo reale. Scoperto dal CERT-UA nel luglio 2025 e attribuito al gruppo APT28 russo, LameHug segna l'inizio di una nuova era nella guerra informatica.

LameHug è implementato come un malware Python-based progettato per sistemi Windows, ma la sua vera innovazione risiede nell'integrazione di un Large Language Model per la generazione dinamica di comandi. Il malware utilizza l'API Hugging Face per interagire con il modello Qwen 2.5-Coder-32B-Instruct, sviluppato da Alibaba Cloud.

3.4. Case Study: Cursor AI

È stata recentemente individuata una vulnerabilità nell'editor Cursor AI, una variante di Visual Studio Code che integra funzionalità di intelligenza artificiale. Il problema nasce da una scelta predefinita del software: la funzione di sicurezza chiamata Workspace Trust risulta disattivata.

Questa configurazione consente a chiunque condivida un progetto o un repository di inserire al suo interno istruzioni che vengono eseguite automaticamente nel momento in cui la cartella viene aperta con l'editor. In pratica, basta che lo sviluppatore apra un progetto da una fonte non sicura per ritrovarsi con del codice eseguito in background, senza alcuna richiesta di conferma.

Le conseguenze possono essere rilevanti: furto di informazioni sensibili, alterazione dei file locali o, nei casi peggiori, un punto di ingresso per attacchi più estesi al sistema e ai flussi di lavoro degli sviluppatori. Questo rende la vulnerabilità particolarmente insidiosa in contesti collaborativi, dove i team condividono codice tramite repository pubblici o ricevono progetti da terze parti.

Un ulteriore elemento di rischio riguarda i processi di ispezione basati sull'IA integrata nell'editor. In particolare, il sistema può generare ed eseguire automaticamente casi di test: se non correttamente isolati in ambienti sicuri (sandbox), questi potrebbero interagire con database di produzione e provocare conseguenze molto gravi.

La stessa società sviluppatrice ha avvertito che nuove funzionalità, come la creazione e modifica di file tramite Claude Code, espongono a rischi di prompt injection. Questi attacchi, come già anticipato, possono sfruttare file o siti esterni per inserire istruzioni "nascoste" che inducono il modello a scaricare ed eseguire codice non affidabile, o a leggere dati sensibili attraverso integrazioni basate sul Model Context Protocol (MCP). In scenari concreti, Claude potrebbe essere indotto a condividere informazioni dal proprio contesto (prompt, progetti, dati collegati a MCP o Google) con attori malevoli.

Non si tratta però dell'unico vettore. Recentemente è stato evidenziato che anche modelli IA che interagiscono con browser, come Claude per Chrome, sono esposti a forme di prompt injection. Sebbene siano state implementate difese che hanno ridotto il tasso di successo degli attacchi dal 23,6% all'11,2%, la minaccia resta dinamica: nuove tecniche vengono sviluppate costantemente da attori malevoli, e solo l'analisi di scenari reali permette di raffinare i sistemi di rilevamento e rendere più efficace la protezione.

Accanto a queste minacce "native" dell'IA, sono state identificate anche vulnerabilità più tradizionali, che dimostrano come gli strumenti IA non siano esenti da problematiche di sicurezza consolidate:

- Claude Code IDE extensions – CVE-2025-52882, CVSS 8.8: bypass dell'autenticazione WebSocket che poteva consentire a un attaccante di ottenere esecuzione remota di comandi inducendo la vittima a visitare un sito malevolo.
- Postgres MCP server – SQL injection che permetteva di superare le restrizioni di sola lettura ed eseguire query arbitrarie.
- Microsoft NLWeb – vulnerabilità di path traversal che consentiva la lettura di file sensibili (es. /etc/passwd, credenziali cloud).
- Lovable – CVE-2025-48757, CVSS 9.3: errata autorizzazione che permetteva a un attaccante remoto non autenticato di leggere o scrivere su tabelle database di siti generati.
- Base44 – presenza di open redirect, stored XSS e data leakage, sfruttabili per accedere ad app e workspace, rubare API key, iniettare codice malevolo o esfiltrare dati.
- Ollama Desktop – controlli cross-origin incompleti che aprivano alla possibilità di attacchi drive-by, in cui un semplice accesso a un sito malevolo poteva riconfigurare l'applicazione per intercettare conversazioni e manipolare i modelli IA.

L'insieme di queste vulnerabilità mostra come l'adozione di strumenti IA negli ambienti di sviluppo non solo introduca nuovi vettori di attacco (prompt injection, automazione incontrollata dei test), ma mantenga anche esposizione verso vulnerabilità tradizionali, ampliando in modo significativo la superficie d'attacco.

3.5. Case study: l'attacco a Nx

Il 26 agosto 2025 segna una data chiave per il mondo dello sviluppo software e, in particolare, per la comunità che lavora con l'intelligenza artificiale. In quel giorno, un attacco sofisticato ha colpito Nx, uno dei sistemi di build più diffusi, mostrando come le supply chain digitali possano diventare bersagli privilegiati e come l'IA, invece di difendere, possa essere piegata a fini offensivi. Gli aggressori sono riusciti a infiltrarsi nel registro npm, pubblicando versioni alterate di pacchetti largamente utilizzati da sviluppatori di tutto il mondo. Non si è trattato del classico furto di credenziali o della semplice esfiltrazione di dati. Il codice malevolo introdotto conteneva funzionalità innovative: era progettato per sfruttare direttamente gli strumenti di IA già installati nei sistemi delle vittime. In pratica, pacchetti compromessi come quelli di Nx contenevano logiche che, una volta eseguite, manipolavano CLI di intelligenza artificiale – tra cui Google Gemini, Amazon Q e Anthropic Claude. Gli hacker hanno introdotto comandi che forzavano questi assistenti a oltrepassare i normali limiti di sicurezza, utilizzando flag come --dangerously-skip-permissions o --yolo. Strumenti nati per aiutare gli sviluppatori nella scrittura di codice e nella gestione dei progetti sono stati così trasformati in agenti di ricognizione ed esfiltrazione dati.

3.5.1. Un'IA piegata contro i propri utenti

L'elemento più inquietante di questo attacco è stato proprio il capovolgimento della logica di fiducia. I CLI di IA, percepiti come affidabili e integrati nei workflow quotidiani, sono diventati inconsapevoli complici dei criminali. Una volta manipolati, hanno iniziato a scandagliare file di progetto, individuare chiavi API, token di accesso e credenziali varie, inviandole a repository controllati dagli aggressori.

Il risultato è stato devastante: in pochi giorni sono state esposte oltre 2.300 credenziali, incluse chiavi GitHub, accessi a servizi cloud e persino token legati a piattaforme di intelligenza artificiale. Secondo le analisi, l'85% delle compromissioni ha riguardato ambienti macOS, particolarmente diffusi nel mondo dello sviluppo.

L'attacco a Nx ha avuto un impatto che va ben oltre i singoli sviluppatori colpiti. I pacchetti compromessi hanno infatti minacciato intere pipeline di sviluppo software, mettendo a rischio sia progetti tradizionali sia infrastrutture critiche per l'addestramento e il deployment di modelli di machine learning. Molte aziende si sono trovate improvvisamente esposte, non solo per la perdita di dati sensibili, ma perché l'attacco ha mostrato come gli strumenti di IA siano ormai punti di ingresso privilegiati per i criminali. Il dato che il 33% dei sistemi compromessi avesse almeno un CLI di IA installato conferma quanto l'IA sia ormai pervasiva nei processi di sviluppo, ma anche vulnerabile a manipolazioni subdole.

3.5.2. Cosa possiamo imparare

Il caso Nx dimostra che la supply chain software è oggi inseparabile dall'IA: non parliamo più solo di librerie e dipendenze, ma di ecosistemi dove agenti intelligenti hanno accesso a risorse vitali. Da qui tre considerazioni centrali:

I. La fiducia negli strumenti non è più scontata. Anche tool mainstream e apparentemente sicuri possono diventare vettori d'attacco.

II. L'IA può essere strumentalizzata. Non serve creare malware ex novo: basta piegare a scopi malevoli strumenti già presenti nei sistemi delle vittime.

III. La difesa deve spostarsi a monte. Controlli più rigorosi nei registri pubblici e nei processi di validazione dei pacchetti sono oggi imprescindibili.

Questo approccio dimostra come l'uso di modelli IA locali renda i ransomware più adattivi, dinamici ed elusivi rispetto ai metodi tradizionali di detection, configurandosi come un precursore delle minacce emergenti

3.6. Disinformazione

Tra le molteplici trasformazioni introdotte dall'intelligenza artificiale, quella che più inquieta è il suo impatto sulla disinformazione. L'IA generativa ha reso possibile ciò che un tempo era prerogativa di governi o grandi organizzazioni mediatiche: produrre in massa contenuti falsi ma convincenti, capaci di diffondersi rapidamente e influenzare l'opinione pubblica.

Se fino a pochi anni fa le campagne di propaganda richiedevano risorse significative e apparati organizzativi complessi, oggi bastano strumenti accessibili e facilmente configurabili. I modelli di linguaggio scrivono articoli coerenti e localizzati, le piattaforme di generazione di immagini creano fotografie indistinguibili dalla realtà, i sistemi di sintesi vocale imitano toni e inflessioni di persone reali. Questo mix ha abbattuto le barriere che storicamente rendevano difficile la produzione di disinformazione su larga scala. Non è più necessario scegliere tra quantità e qualità: oggi entrambe sono ottenibili con facilità e a costi ridotti.

La disinformazione non è più un fenomeno confinato alla geopolitica. Certo, gli Stati continuano a utilizzare questi strumenti per destabilizzare avversari e condizionare elezioni, ma anche attori criminali e gruppi privati hanno iniziato a sfruttarli per scopi economici. Le campagne di manipolazione dei mercati finanziari ne sono un esempio lampante: notizie false, corredate da immagini e dichiarazioni sintetiche, hanno generato oscillazioni artificiali nei prezzi delle criptovalute, permettendo guadagni rapidi a chi ne controllava la narrativa. In altri contesti, video contraffatti di leader politici hanno alimentato tensioni sociali e diffuso sfiducia nelle istituzioni.

Il cuore di questo fenomeno è rappresentato dai deepfake, che segnano un salto di qualità rispetto al passato. Non si tratta più soltanto di testi manipolati o immagini alterate: oggi vediamo video in cui fi-

gure pubbliche pronunciano frasi mai dette e registrazioni audio che riproducono fedelmente la voce di dirigenti o personaggi pubblici. Un episodio emblematico si è verificato nel 2024, quando un video deepfake di un alto dirigente europeo è circolato sui social generando confusione e obbligando le istituzioni a rilasciare smentite ufficiali. La velocità con cui il contenuto si è diffuso ha dimostrato quanto sia difficile contenere l'impatto di una manipolazione una volta che questa raggiunge il dominio pubblico.

La disinformazione alimentata dall'IA produce un duplice rischio. Da un lato, i contenuti falsi che sembrano autentici si diffondono con una credibilità senza precedenti. Dall'altro, il semplice fatto che simili manipolazioni esistano indebolisce la fiducia nei contenuti autentici. È quello che gli esperti definiscono *liar's dividend:* la possibilità che persino messaggi veri vengano liquidati come falsi con il pretesto che "potrebbero essere deepfake". La fiducia collettiva nell'informazione digitale si erode così in entrambe le direzioni.

Per le organizzazioni, il problema non è teorico. Una campagna di disinformazione orchestrata contro un marchio può compromettere anni di lavoro sulla reputazione. Un video manipolato che mostra un amministratore delegato annunciare decisioni inesistenti può muovere il mercato azionario nel giro di poche ore. Persino un audio alterato, diffuso con abilità, può incrinare rapporti con partner e clienti. Secondo rilevazioni del 2024, oltre il 40% delle aziende ha dichiarato di aver subito almeno un tentativo di manipolazione reputazionale legato a contenuti generati dall'IA. La maggioranza, tuttavia, non dispone ancora di piani strutturati per gestire simili incidenti.

Ciò che emerge è l'avvento di una nuova normalità informativa. Così come spam e phishing sono diventati rischi costanti con l'avanzare della digitalizzazione, i contenuti sintetici falsi rappresentano ormai una componente stabile del nostro ecosistema mediatico. Difendersi da questo fenomeno significa lavorare su più livelli: adottare tecnologie di rilevamento dei deepfake, monitorare in tempo reale i social network, rafforzare le procedure di crisis management e, soprattutto, coltivare una cultura della consapevolezza. Perché in ultima analisi il bene più prezioso da proteggere non è un'infrastruttura tecnica, ma la fiducia delle persone nell'informazione.

3.6.1. Focus deepfake

Il fenomeno dei deepfake rappresenta dunque una delle evoluzioni più significative nel campo della disinformazione digitale.

L'uso di IA generativa per creare identità sintetiche e manipolare contenuti audio-video in tempo reale si è trasformato in un vettore di attacco sempre più diffuso, soprattutto nelle frodi di identità e negli schemi di vishing. Secondo l'European Parliamentary Research Service, si prevede che nel 2025 verranno condivisi 8 milioni di deepfake, contro i soli 500.000 del 2023.

A livello globale, 1 su 20 fallimenti nei processi di verifica dell'identità è già riconducibile all'impiego di deepfake, mentre i truffatori sfruttano strumenti di IA per produrre documenti falsi iperrealistici o per impersonare persone durante interazioni video dal vivo, aggirando controlli biometrici e sistemi di autenticazione.

Case Study Italia: L'Attacco di Vishing basato su Deepfake Vocale

A febbraio 2025, un attacco di vishing ha ingannato alcuni dei più importanti imprenditori italiani utilizzando una tecnica sofisticata di deepfake vocale. Gli aggressori hanno utilizzato una tecnologia avanzata di clonazione vocale per imitare la voce del Ministro della Difesa, Guido Crosetto. Contattando le vittime, tra cui Massimo Moratti e Giorgio Armani, si sono spacciati per membri dello staff del Ministro e hanno richiesto ingenti somme di denaro per presunti riscatti di giornalisti italiani rapiti all'estero. La narrazione, apparentemente urgente e legittima, ha convinto almeno un imprenditore a effettuare un trasferimento di circa un milione di euro su un conto estero. L'intervento delle forze dell'ordine ha permesso di rintracciare e congelare i fondi su un conto olandese.

Questo episodio evidenzia l'efficacia delle truffe basate sull'IA nel superare le tradizionali difese umane. L'attacco non ha sfruttato una vulnerabilità tecnica, ma la fiducia delle vittime in un'alta figura istituzionale. La professionalità della frode e l'urgenza percepita della situazione hanno aggirato il naturale scetticismo, dimostrando che anche individui altamente informati possono essere vulnerabili quando si scontrano con una manipolazione così realistica.

Tentativo di Frode contro un Dirigente di un'importante casa automobilistica

Un altro sofisticato tentativo di frode si è verificato a luglio 2024, prendendo di mira un dirigente della di una nota casa automobilistica. I criminali hanno inviato messaggi WhatsApp che sembravano provenire dal CEO, parlando di un'imminente acquisizione e sollecitando la firma di un accordo di non divulgazione. Il passo successivo della truffa è stato una chiamata vocale in cui gli aggressori hanno utilizzato l'IA per replicare perfettamente la voce e l'accento del CEO.

L'attacco, pur essendo tecnicamente avanzato, è stato sventato grazie all'attenzione del dirigente per i dettagli. Pur riconoscendo la somiglianza della voce, ha notato sottili incoerenze nel tono e ha deciso di mettere alla prova l'interlocutore con una domanda di verifica che solo il vero Vigna avrebbe potuto conoscere: il titolo di un libro che aveva raccomandato giorni prima. L'incapacità del truffatore di rispondere ha fatto terminare bruscamente la chiamata. Questo caso, benché fallito, serve come un potente promemoria del fatto che, mentre l'IA può automatizzare la falsificazione, la linea finale di difesa rimane la vigilanza umana e l'applicazione di solidi protocolli di verifica.

I due casi italiani, sebbene distinti, dimostrano una chiara tendenza e un'implicazione strategica: gli aggressori non stanno cercando di bypassare i sistemi, ma di subvertire la fiducia.

Questi episodi evidenziano come gli attacchi basati su deepfake non puntino a sfruttare vulnerabilità tecniche, ma a sovvertire la fiducia umana. La combinazione tra credibilità visiva/uditiva e urgenza narrativa rende queste frodi particolarmente insidiose, soprattutto in contesti ad alto profilo.

3.7. Il mercato nero

Un fattore che ha accelerato questa evoluzione è la comparsa di modelli generativi "senza freni", distribuiti nel dark web o su canali Telegram. WormGPT e FraudGPT sono due esempi noti: venduti come alternative "black hat" ai modelli commerciali, offrono funzionalità progettate per il crimine. Tra queste: offuscamento di codice, generazione di exploit, scansione di vulnerabilità, verifica di carte di credito rubate, invio automatizzato di e-mail di phishing. Il loro costo è sorprendentemente basso, spesso poche decine di dollari al mese. Questo abbassa ulteriormente la barriera d'ingresso: attività che un tempo richiedevano team organizzati e risorse significative possono ora essere condotte da singoli attori con strumenti acquistati online. È un processo di "democratizzazione del crimine" che rende il panorama molto più affollato e difficile da controllare.

4. L’impatto sulle organizzazioni

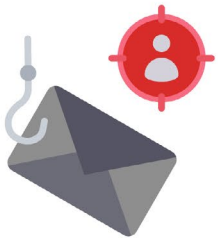
Questi scenari non restano confinati al mondo criminale: hanno già un impatto concreto sulle imprese. I dati raccolti negli ultimi dodici mesi parlano chiaro:

Il **29%**

delle organizzazioni ha subito attacchi a infrastrutture legate all’IA, come repository di modelli, ambienti di training non segmentati o endpoint esposti.

L’ **84%**

dei responsabili sicurezza ha osservato un aumento nella qualità delle campagne di phishing, mentre il 78% e il 74% hanno segnalato rispettivamente un incremento di phishing e di Business E-mail Compromise (BEC).



Il **62%**

ha dovuto fronteggiare incidenti legati a deepfake, utilizzati per eludere sistemi biometrici o impersonare figure aziendali.

Il **32%**

ha registrato compromissioni legate a prompt injection o manipolazioni del contesto.

Nonostante questi numeri, l’80% delle organizzazioni non dispone ancora di protocolli o playbook specifici per gestire incidenti legati a contenuti sintetici.

Come spesso accade in ambito di nuova tecnologia, la curva difensiva cresce più lentamente di quella offensiva. Molti progetti di intelligenza artificiale non vengono ancora integrati nei processi di sviluppo sicuro. Mancano attività di threat modeling specifiche, revisioni del codice mirate, penetration test orientati ai carichi IA.

Pochissime aziende hanno introdotto controlli nativi per l’IA, come la segmentazione di rete per gli endpoint di inferenza, il monitoraggio degli input e output dei modelli, o policy cloud pensate per workload basati su intelligenza artificiale. E come accennato pocanzi, ancora meno hanno predisposto playbook di incident response dedicati.

Questo significa che, di fronte a un incidente IA-enabled, molte organizzazioni reagiscono in modo improvvisato, adattando procedure pensate per scenari diversi. È una vulnerabilità culturale e organizzativa, oltre che tecnologica.

4.1. Shadow IA: la minaccia dall'interno

Un ulteriore elemento critico è rappresentato dalla cosiddetta shadow IA. Sempre più dipendenti, attratti dalla comodità di strumenti pubblici di intelligenza artificiale, li utilizzano per semplificare il proprio lavoro quotidiano. Ma lo fanno al di fuori di ogni governance aziendale.

Il risultato è duplice: da un lato, dati sensibili possono finire in piattaforme non controllate; dall'altro, i team IT non hanno visibilità su ciò che viene richiesto ai modelli. Secondo i dati, il 22% delle aziende consente ancora accesso illimitato a strumenti pubblici di GenAI, mentre il 60% dei team IT dichiara di non poter monitorare i prompt inviati dai dipendenti.

Per le piccole e medie imprese, il fenomeno è ancora più pericoloso. Con meno risorse e competenze specialistiche, le PMI faticano a implementare politiche chiare, rendendo la shadow IA un rischio sistemico spesso invisibile.

4.2. Una nuova *normalità* di rischio

L'intelligenza artificiale ha dunque introdotto un nuovo paradigma. Non si tratta più di semplici strumenti da difendere, ma di sistemi che possono essere trasformati in armi dagli stessi attaccanti. Ogni nuovo modello, ogni applicazione, ogni architettura rappresenta al tempo stesso un'opportunità e una vulnerabilità.

Le minacce IA-enabled non aumentano solo in numero, ma diventano più credibili, più personalizzate e più difficili da rilevare. La velocità con cui nascono nuovi vettori di attacco è superiore a quella con cui vengono sviluppate nuove difese.

Per le organizzazioni, questo significa che non basta più aggiungere livelli di protezione sopra quelli esistenti. Serve ripensare la sicurezza in chiave IA-native, includendo l'intelligenza artificiale in ogni fase: dalla progettazione dei sistemi alla gestione degli incidenti, dalla governance interna alla formazione del personale. Solo così sarà possibile innovare senza trasformare l'IA in un boomerang pericoloso, come possibile leggere negli approfondimenti e case study di seguito.

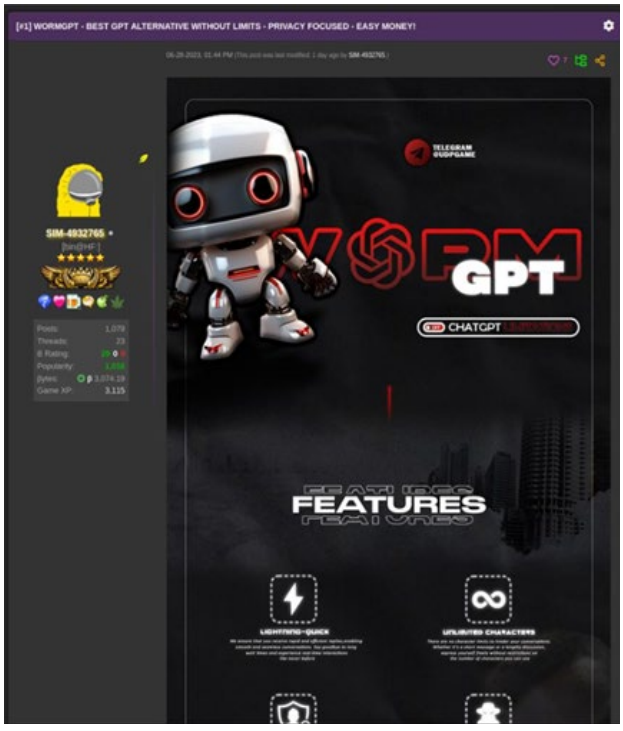
5. Approfondimento: i mercati *underground*

Come accennato nel paragrafo precedente; accanto all'adozione mainstream della GenAI, che alimenta innovazione e competitività nelle imprese, è emerso un ecosistema parallelo e oscuro: quello dei mercati underground dedicati a modelli generativi privi di vincoli. È qui che si osserva la forma più diretta della già citata "democratizzazione del crimine": strumenti nati per potenziare la produttività vengono ricodificati e distribuiti per agevolare attività illecite.

L'elemento distintivo di questo mercato non è soltanto l'offerta tecnologica, ma l'accessibilità. Piattaforme che promettono di scrivere codice su misura per exploit, redigere e-mail di phishing perfettamente credibili o generare deepfake in pochi minuti vengono vendute a prezzi irrisori, con modelli di business tipici del SaaS legittimo: abbonamenti mensili, pacchetti premium, assistenza clienti tramite forum o canali Telegram.

5.1. Dai primi esperimenti alle piattaforme "as-a-service"

Il 2023 ha segnato l'apparizione di WormGPT, considerato il primo vero modello di linguaggio progettato per usi criminali. Nato come derivazione di modelli open source, WormGPT è stato privato di qualunque restrizione etica o filtro di sicurezza. Il suo scopo dichiarato era chiaro: fornire ai cybercriminali uno strumento con cui generare codice malevolo, redigere campagne di phishing persuasive e automatizzare attività fraudolente.



Pochi mesi dopo è comparso FraudGPT, con un posizionamento simile ma un'offerta ancora più ampia: dalle istruzioni per frodi con carte di credito, alla scrittura di script per malware polimorfici, fino a guide dettagliate su come condurre attacchi ransomware. Entrambi i modelli sono stati commercializzati

con abbonamenti accessibili (dai 100 ai 200 dollari al mese) e hanno trovato spazio in forum del dark web e in comunità Telegram dedicate.

Nel giro di pochi mesi, si è delineata una vera e propria catena del valore del crimine generativo. Gli sviluppatori di modelli "black hat" offrono l'infrastruttura; i reseller creano pacchetti pronti per utenti meno esperti; le comunità forniscono tutorial, aggiornamenti e assistenza. In alcuni casi, vengono proposte garanzie di soddisfazione del cliente, come avviene nei servizi legittimi.

Il confronto dei costi rivela la natura altamente professionalizzata di questo mercato. L'offerta di abbonamenti annuali e di "setup privati" suggerisce un modello di business a lungo termine che replica le strategie di vendita di software legittimo. La variazione dei prezzi a seconda del venditore evidenzia inoltre la natura competitiva e frammentata di questo ecosistema, che si adatta a diverse fasce di "consumatori", dal criminale occasionale al gruppo organizzato.

Panoramica e Costi di WormGPT e FraudGPT

Strumento IA	Caratteristiche Principali	Prezzo Mensile	Prezzo Annuale	Altri Costi
WormGPT	Creazione BEC e phishing, malware	$60-$100 / $90	$550 / $850	$230/3 mesi; $500/6 mesi / $5.000 (setup privato)
FraudGPT	Phishing, cracking password, furto credenziali	$200 / $90	$1.700 / $700	$200/3 mesi; $500/6 mesi

Report di settore hanno identificato oltre 200 esempi di LLM malevoli operanti nei mercati underground tra aprile e ottobre 2023. Questi strumenti si dividono in due categorie principali: LLM completamente non censurati e LLM commerciali "jailbroken" che utilizzano prompt specializzati per aggirare le limitazioni etiche.

Funzionalità pensate per il crimine

Questi modelli non si limitano a rimuovere i filtri: incorporano funzioni specifiche per agevolare attività illegali. Alcune delle più diffuse includono:

- **Generazione di codice malevolo:** exploit in Python, JavaScript o PowerShell, con opzioni di offuscamento automatico.
- **Scansione di vulnerabilità:** suggerimenti per testare la sicurezza di siti e applicazioni, con indicazioni su come sfruttare le falle individuate.
- **Automazione del phishing:** creazione di e-mail e SMS personalizzati, localizzati in diverse lingue, con template già ottimizzati per aumentare i tassi di click.
- **Verifica di carte di credito rubate:** script per testare la validità di numeri sottratti, riducendo il rischio di transazioni bloccate.
- **Deepfake su richiesta:** generazione di audio e video sintetici da usare in campagne di personificazione.

Il livello di sofisticazione varia, ma ciò che accomuna questi strumenti è la loro capacità di ridurre drasticamente il tempo e le competenze necessarie per condurre un attacco. Dove prima servivano sviluppatori esperti, ora basta un utente con conoscenze di base e un abbonamento attivo.

5.2. Un mercato che si professionalizza

Se all'inizio questi modelli apparivano come progetti artigianali, oggi il mercato si è professionalizzato. Alcuni operatori propongono interfacce user-friendly, documentazione aggiornata e canali di supporto tecnico. Altri offrono versioni "lite" per principianti e versioni avanzate per utenti esperti, replicando logiche commerciali tipiche delle aziende SaaS.

La struttura dei prezzi riflette questa evoluzione. Abbonamenti mensili a partire da 100 dollari, sconti per pagamenti annuali, pacchetti premium con accesso prioritario a nuove funzionalità. Esistono persino offerte bundle che combinano diversi strumenti: un modello per il phishing, uno per il malware e uno per i deepfake, venduti insieme a un prezzo promozionale.

Questa organizzazione ha un duplice effetto. Da un lato, abbassa la barriera d'ingresso per nuovi attaccanti. Dall'altro, crea economie di scala per i criminali più esperti, che possono integrare più strumenti in campagne coordinate e di maggiore portata.

5.3. Un fenomeno destinato a crescere

Tutto lascia pensare che i mercati underground dell'IA siano destinati a crescere. Il modello economico è sostenibile, la domanda è in aumento e la tecnologia è sempre più disponibile. La vera differenza rispetto ad altri strumenti criminali è che qui il ciclo di innovazione è rapidissimo: ogni nuova funzionalità sviluppata nei modelli legittimi può essere rapidamente adattata in chiave offensiva.

È quindi probabile che nei prossimi anni assisteremo alla nascita di un'offerta ancora più frammentata e diversificata: modelli specializzati per settori specifici (finanza, sanità, industria), versioni multimodali capaci di integrare testo, audio e immagini, pacchetti integrati con bot e framework di automazione.

I mercati underground dell'IA rappresentano la concretizzazione di un trend già evidente: la criminalità informatica non solo sfrutta le nuove tecnologie, ma le adatta e le distribuisce con la stessa logica con cui operano le imprese legittime.

Quello che fino a poco tempo fa sembrava un rischio teorico è oggi un settore strutturato, con attori, modelli di business e community consolidate. E finché continuerà a crescere, le organizzazioni dovranno confrontarsi con un panorama in cui le minacce non sono più il frutto di pochi esperti isolati, ma il prodotto di un'intera filiera criminale, alimentata e amplificata dall'intelligenza artificiale.

Stasi nella fortificazione dell'Intelligenza Artificiale

Nonostante la crescente sofisticazione delle minacce, la prontezza complessiva nella fortificazione dell'IA ha mostrato una sostanziale stagnazione rispetto all'anno precedente. Questa persistenza evidenzia le continue incertezze relative all'automazione della cybersecurity basata sull'IA.

Seppur notevoli progressi siano stati compiuti nel campo dell'IA, la sua integrazione nelle difese di cybersecurity sembra aver raggiunto un plateau. Ciò suggerisce che le organizzazioni stanno ancora affrontando questioni fondamentali legate alla fiducia, all'efficacia e alla capacità di integrazione. Molti responsabili della sicurezza esitano ad adottare pienamente le difese basate sull'IA a causa di ostacoli tecnici, preoccupazioni normative o difficoltà nell'adattare i modelli IA alle minacce in continua evoluzione.

Attualmente, l'IA trova ampia applicazione nella rilevazione delle minacce (85%), nella risposta agli incidenti (71%) e nel recupero post-incidente (70%). Tuttavia, il livello di dipendenza delle aziende dall'IA per queste attività è ancora in fase di crescita, con solo il 27% delle aziende che ha automatizzato completamente le proprie difese nella rilevazione delle minacce.

I rischi amplificati dall'IA aggravano un ambiente operativo già complesso, caratterizzato dalla diffusione di modelli di lavoro ibridi, dispositivi non gestiti e una proliferazione di soluzioni di sicurezza. Questa situazione accentua l'urgente necessità di azioni e investimenti mirati.

6. Focus Normativo

Davanti a queste prospettive, come si stanno muovendo gli enti regolatori a livello globale?

Il panorama regolatorio sull'intelligenza artificiale (IA) ha conosciuto un'accelerazione significativa tra il 2024 e il 2025, rispondendo alla diffusione massiva delle tecnologie IA e ai rischi emergenti in ambito sociale, economico e di sicurezza nazionale. Negli Stati Uniti, nel solo 2024 sono stati presentati oltre 700 disegni di legge relativi all'IA in 45 stati, con 113 norme approvate, mentre nel 2025 si sono aggiunte ulteriori 40 proposte a livello federale e statale, evidenziando una forte frammentazione legislativa in assenza di una regolamentazione federale omogenea.

In Europa, il 1° agosto 2024 è entrato in vigore il Regolamento UE noto come IA Act, il primo quadro normativo globale completo che disciplina lo sviluppo, la distribuzione e l'utilizzo dei sistemi IA. L'IA Act introduce requisiti stringenti per i sistemi ad "alto rischio" — quali applicazioni in ambito sanitario, giuridico, e infrastrutturale — stabilendo obblighi di conformità tecnica, trasparenza, e gestione del rischio. Inoltre, vieta alcune applicazioni considerate "inaccettabili" e prevede sanzioni fino al 7% del fatturato mondiale per le violazioni. Contestualmente, la Commissione Europea ha ritirato la direttiva sulla responsabilità IA del 2022 per ridurre oneri amministrativi e semplificare il quadro normativo, mentre programmi come InvestAI e ingenti investimenti pubblici e privati in Francia (oltre 109 miliardi di euro) puntano a rafforzare infrastrutture tecnologiche e gigafabbriche IA.

Sul fronte americano, l'amministrazione Trump ha revocato l'Executive Order di Biden sull'IA, sostituendolo con un nuovo ordine esecutivo che enfatizza la promozione dell'innovazione tecnologica, la crescita economica e la protezione della sicurezza nazionale. Particolare attenzione è posta sul controllo delle esportazioni di tecnologie IA strategiche per limitare il vantaggio tecnologico di potenziali avversari esteri. In questo contesto, la presenza statunitense al IA Action Summit 2025 si è distinta per il rifiuto, insieme al Regno Unito, di sottoscrivere la dichiarazione sulla sicurezza IA, a causa di preoccupazioni legate alla governance globale e alle implicazioni di sicurezza nazionale. Parallelamente, il Regno Unito ha ribrandizzato il proprio IA Safety Institute in IA Security Institute, con un focus potenziato sulle minacce criminali e di sicurezza connesse all'uso improprio dell'IA.

A livello internazionale, diverse giurisdizioni stanno implementando approcci regolatori differenziati:

- Australia ha introdotto una politica obbligatoria per l'uso responsabile dell'IA nelle amministrazioni pubbliche non corporative, con efficacia dal settembre 2024, sottolineando la leadership governativa nell'adozione etica e sicura dell'IA.
- Singapore ha diffuso un modello volontario di governance IA specifico per sistemi generativi, focalizzato sul risk management e sulle best practice per la gestione dei rischi IA.
- India, tramite il Ministero dell'Elettronica e della Tecnologia dell'Informazione (MEITY), ha pubblicato un documento consultivo sulle linee guida per la governance IA, basato su un approccio risk-based allineato ai principi OECD, proponendo anche la creazione di un organismo tecnico consultivo per la sicurezza IA.
- Corea del Sud ha promulgato la IA Framework Act, diventando la seconda nazione dopo l'UE a introdurre una legge regolatoria complessiva sull'IA, adottando un modello incentrato sui sistemi ad

"alto impatto" sulla vita umana, la sicurezza fisica e i diritti fondamentali.

- Giappone ha presentato una bozza di legge IA con un approccio "light-touch", privo di sanzioni rigide, mirata a implementare i principi del Processo di Hiroshima, a sostenere R&D e a potenziare l'azione governativa contro usi malevoli dell'IA non coperti dalla normativa vigente.
- In Africa, l'Unione Africana, comprendente 55 Stati membri, sta sviluppando una strategia di regolamentazione IA in fase di definizione, con alcuni Paesi che hanno già adottato politiche nazionali e il rilascio di un progetto di regolamentazione da parte dell'Agenzia per lo sviluppo africano.

Per le imprese questo mosaico normativo si traduce in un quadro incerto. Da un lato, cresce il rischio di non conformità: le aziende che operano in più giurisdizioni devono destreggiarsi tra regole diverse, a volte incompatibili. Dall'altro, emergono opportunità per chi saprà muoversi in anticipo: chi integra i requisiti normativi fin dalle prime fasi di sviluppo potrà trasformare la compliance in un vantaggio competitivo, dimostrando affidabilità a clienti e partner.

Ciò che resta evidente è che la regolazione, da sola, non basta. Servono capacità operative di controllo, standard tecnici condivisi, strumenti di certificazione che riducano le ambiguità. In mancanza di questi elementi, il rischio è che le norme restino dichiarazioni di principio difficili da far rispettare, mentre gli attaccanti — liberi da vincoli — continuano a sperimentare e innovare a un ritmo superiore.

Il focus normativo, dunque, è una componente fondamentale ma non sufficiente. È la cornice all'interno della quale dovranno muoversi aziende, istituzioni e società civile, ma il vero banco di prova sarà la capacità di tradurre i principi in pratiche concrete, senza soffocare l'innovazione ma neppure lasciando spazi vuoti che il crimine saprà colmare.

Cyber Framework Maticmind

Il Cyber Framework Maticmind si allinea alla Direttiva NIS2 attraverso un modello integrato che anticipa minacce emergenti, riduce rischi con analisi proporzionate, promuove la sicurezza “by design”, implementa tecnologie avanzate (EDR, SIEM, zero trust) e rafforza la resilienza tramite monitoraggio continuo e risposta rapida agli incidenti.

Questo approccio trasforma la compliance normativa in vantaggio strategico, garantendo continuità operativa, mitigazione degli impatti economici degli attacchi e una governance cyber dinamica.

Predittiva – Anticipare

Monitoraggio costante del panorama delle minacce tramite Cyber Threat Intelligence e tecniche OSINT, per identificare rischi emergenti prima che si manifestino.

Preventiva – Prevenire

Valutazione e gestione del rischio tecnologico, umano, organizzativo e compliance, per rafforzare la postura di sicurezza e ridurre la probabilità di incidenti.

Progettuale – Costruire Sicuro

Integrazione della sicurezza nei processi di design e architettura, secondo principi di “secure by design” e “security by resilience”, per garantire sistemi resilienti.

Produttiva – Implementare Soluzioni

Adozione e gestione di tecnologie di difesa (firewall, EDR, IAM, SIEM) che trasformano le strategie in protezione operativa e misurabile.

Proattiva – Reagire e Migliorare

Monitoraggio continuo, risposta agli incidenti e analisi forense: la sicurezza come processo dinamico e adattivo, guidato dal miglioramento continuo.

Una Cyber Acies

Cinque centri di competenza, un'unica forza cyber

1

Centro strategico per l'analisi e gestione del rischio cyber: governance, compliance, modelli di maturità, supply chain security posture.

- Offensive Security
- Consulting Services (CisoasaService, PM,..)
- Governance, Risk e Compliance
- Social Engineering Simulation
- Risk Assessment
- Maturity Model checkup
- Supply Chain Security Posture
- Cyber Academy

3

Esecuzione ed implementazione di progetti complessi di sicurezza: infrastrutture, cloud, OT/IoT, application security e integrazione di soluzioni cyber.

- Infrastructure & Network Security Management
- Cyber Resiliency e Data Protection
- Microsoft, Fortinet & CISCO Competence Center
- Cloud Security By Design
- Application Security SSDLC
- OT & IoT Security Management
- Security System Integration and Monitoring

5

Progetti speciali e sperimentazione: architetture avanzate, droni, AI, cyber-physical systems, prototipi e soluzioni non convenzionali.

- Architetture avanzate multi-layer e ibride
- Sistemi di comando e controllo per droni e UAV
- Integrazione AI in contesti cyber e sicurezza industriale
- Progetti su ambienti cyber-physical / smart infrastructure
- Soluzioni sperimentali
- Cyber testbed per ambienti OT e industriali
- Soluzioni cyber per difesa e protezione civile
- Soluzioni sperimentali edge & fog computing.

2

Continuità operativa e sicurezza in tempo reale assicurata da presidio operativo che integra SOC H24, con le capacità tecniche del TAC e il controllo infrastrutturale del NOC. Garantisce un monitoraggio continuo, una risposta tempestiva agli incidenti e un supporto tecnico evoluto.

- **SOC (Security Operation Center)**
  - SOC H24
  - Cyber Threat Intelligence
  - Malware Analysis
- **SOS (SECURITY OPERATION SERVICES)**
  - Crisis Management
  - IncidentResponse & Digital Forensic Analysis
- **TAC (Technical Assistance Center)**
  - Gestione di ticket e supporto tecnico avanzato
  - Assistenza su soluzioni proprietarie e terze parti
  - Troubleshooting di sicurezza e infrastruttura
  - Supporto su incidenti tecnici e configurazioni complesse
- **NOC (Network Operation Center)**
  - Monitoraggio infrastrutturale H24
  - Supervisione performance di rete, cloud e sistemi
  - Gestione allarmi e segnalazioni
  - Coordinamento tecnico con il SOC in caso di anomalie

4

Gestione e integrazione di soluzioni proprietarie e tecnologie di terze parti: piattaforme EDR, firewall, IAM, SIEM, DLP.

- Piattaforma Twin4Cyber
- Cyber Range
- WiseView - Risk Management Platform
- SIEM/SOAR
- EDR/XDR/NDR
- Identity & Access Management (IAM/PAM)
- NGFW
- Data Loss Prevention

Il framework di sicurezza AI

Il framework **PROMETHEUS AI** di **Maticmind** è stato concepito per offrire un approccio concreto alla sicurezza e alla conformità dei sistemi di intelligenza artificiale. La sua struttura si basa su 10 moduli interconnessi che coprono l'intero ciclo di vita dell'AI, assicurando che i controlli tecnici e organizzativi non siano un ripensamento, ma parte integrante del processo di sviluppo.

Il framework introduce concetti avanzati come il Model Bill of Materials (MBOM), un inventario completo di tutti i componenti (dataset, modelli di base, librerie) utilizzati per costruire un modello di IA. Questo strumento è fondamentale per la tracciabilità e la sicurezza della supply chain, riducendo i rischi di compromissione e accelerando i processi di audit. Parallelamente, il Threat Modeling AI-specifico permette di identificare scenari di attacco unici per l'IA, come il poisoning dei dati o model evasion, prima che diventino vulnerabilità in produzione.

A livello di test e valutazione, **PROMETHEUS AI** non si limita ai test di sicurezza tradizionali, ma incorpora il Penetration Testing Linguistico (PT Linguistico), una pratica essenziale per individuare e mitigare le minacce specifiche dei modelli linguistici, come il prompt injection e il jailbreak. Questa attenzione al dettaglio si estende alla fase di Deployment & Protezione Runtime, dove l'uso di tecnologie come eBPF (extended Berkeley Packet Filter) consente di creare una "sandbox" per i modelli, monitorando e prevenendo comportamenti anomali in tempo reale, un approccio più efficace rispetto ai sistemi di DLP (Data Loss Prevention) standard che non riescono a intercettare attacchi di data exfiltration attraverso i modelli.

I benefici di questo framework sono significativi: si stima una riduzione dell'85% degli incidenti di sicurezza e un aumento del 30% nella velocità di rilascio in produzione, poiché i controlli di sicurezza vengono automatizzati e integrati nelle pipeline CI/CD. Inoltre, PROMETHEUS AI è progettato per garantire il 100% di conformità all'AI Act per i sistemi ad alto rischio, offrendo un percorso chiaro e documentabile verso la certificazione e la trasparenza. Questo non solo mitiga i rischi legali e reputazionali, ma costruisce anche la fiducia degli utenti e delle parti interessate.

Considerazioni finali

Il percorso tracciato in questo report ha messo in evidenza una realtà sempre più articolata: l'intelligenza artificiale, oltre a costituire un formidabile motore di innovazione, rappresenta anche un fattore di rischio crescente per la sicurezza cibernetica e per la stabilità dei sistemi economici e istituzionali. I casi analizzati, dai mercati sotterranei dedicati a modelli malevoli fino agli episodi concreti di supply chain attack, mostrano con chiarezza che l'uso offensivo dell'IA non è una prospettiva futura, ma un fenomeno già in corso.

Le dinamiche osservate rivelano uno squilibrio strutturale: la velocità con cui emergono nuovi vettori di attacco e modalità di sfruttamento supera ancora la capacità delle difese di adattarsi. Questo divario riguarda tanto le tecnologie quanto le persone e le organizzazioni. È evidente, infatti, come il fenomeno coinvolga più dimensioni contemporaneamente: la sfera tecnica, con i modelli e gli agenti IA esposti a manipolazioni sempre più sofisticate; quella normativa, ancora alla ricerca di strumenti di enforcement adeguati; quella sociale, dove la disinformazione amplificata dall'IA mette a rischio la coesione e la fiducia pubblica.

Guardando avanti, scenari come l'adozione su larga scala di sistemi agentici e multimodali renderanno l'ecosistema digitale ancora più complesso. Agenti autonomi, capaci di interagire tra loro e con servizi eterogenei, aumenteranno le opportunità di innovazione ma anche la superficie d'attacco, moltiplicando i rischi di compromissione. In parallelo, le tecniche di manipolazione informativa diventeranno più credibili e pervasive, fino a fondersi con pratiche criminali consolidate come il phishing e il business e-mail compromise.

Di fronte a questa evoluzione, le imprese dovranno imparare a considerare l'adozione dell'IA inseparabile dalla sua messa in sicurezza. Non si tratta soltanto di introdurre nuove soluzioni tecnologiche, ma di sviluppare una cultura della responsabilità, capace di integrare la gestione del rischio in ogni fase del ciclo di vita dell'innovazione. Allo stesso modo, le istituzioni dovranno tradurre principi generali in strumenti concreti, creando standard condivisi e meccanismi di certificazione che permettano di ridurre le ambiguità e facilitare l'applicazione delle norme.

Il compito che ci attende non è semplice. L'IA evolve a ritmi che mettono alla prova i modelli tradizionali di governance, e nessun attore può affrontare la sfida da solo. Sarà necessario uno sforzo congiunto che coinvolga imprese, istituzioni e società civile. La collaborazione tra pubblico e privato, la condivisione delle informazioni di minaccia e l'investimento in ricerca applicata diventeranno elementi centrali di un ecosistema di sicurezza più resiliente.

Questo report non intende offrire risposte definitive, ma fornire un quadro realistico di rischi e tendenze. L'obiettivo è stimolare consapevolezza e incoraggiare un approccio equilibrato, in cui l'innovazione tecnologica proceda di pari passo con la capacità di prevenire abusi e di garantire fiducia nei sistemi digitali.

La sfida dell'intelligenza artificiale non è solo tecnica, ma culturale e politica. La neutralità, l'equilibrio e la responsabilità con cui sapremo affrontarla determineranno se l'IA diventerà un alleato di crescita e progresso o un fattore destabilizzante. La direzione da intraprendere è chiara: costruire insieme un ambiente digitale sicuro, affidabile e sostenibile, a beneficio dell'intera collettività.

Fonti

1. Cisco. (2025). 2025 Cisco Cybersecurity Readiness Index.
2. Cheparthi, A., Miclaus, R., & Lovelock, J. (2025, 1 agosto). Forecast Analysis: Generative AI Models, Worldwide, 2025. Gartner.
3. Gupta, P., Khan, A., D'Hoinne, J., & Smith, Z. (2025, 4 agosto). Data View: What Is the Impact of GenAI on the Attack Landscape?. Gartner.
4. HiddenLayer. (2025). AI Threat Landscape Report 2025: Navigating the Rise of AI Risks.
5. Morag, A. (2025, 24 luglio). AI-Generated Malware in Panda Image Hides Persistent Linux Threat. Aqua Security.
6. GitGuardian. (2025, 27 agosto). The Nx "s1ngularity" Attack: Inside the Credential Leak. GitGuardian Blog. https://blog.gitguardian.com/the-nx-s1ngularity-attack-inside-the-credential-leak/
7. FBI. (2025, 13 febbraio). Don't Let a Romance Scammer Steal Your Heart and Savings. Federal Bureau of Investigation. https://www.fbi.gov/contact-us/field-offices/sacramento/news/dont-let-a-romance-scammer-steal-your-heart-and-savings
8. Lakshmanan, R. (2025, 18 luglio). CERT-UA discovers LAMEHUG malware linked to APT28, using LLM for phishing campaign. The Hacker News. https://thehackernews.com/2025/07/cert-ua-discovers-lamehug-malware.html
9. Negreiro, Mar. Children and deepfakes. EPRS, luglio 2025.
10. Del Val, Margarita. Dark AI tools: How profitable are they on the dark web? Outpost24, 4 luglio 2025.
11. Della Mura, Miti. Malicious digital twins: la truffa Crosetto. Sergente Lorusso, 10 febbraio 2025.
12. Cherepanov, Anton & Strýček, Peter. ESET scopre PromptLock. ESET Blog, 3 settembre 2025.
13. Galletti, Sandra & Pani, Massimo. How Ferrari Hit the Brakes on a Deepfake CEO. MIT Sloan, 27 gennaio 2025.
14. Anthropic. Detecting and countering misuse of AI. Newsroom, 27 agosto 2025.
15. SC Staff. AI-driven ID fraud surges 195% globally. SC Media, 16 giugno 2025.
16. Namase, Rajesh. AI Cyber Attacks Statistics 2025. SQ Magazine, 22 luglio 2025.
17. Lazzarotti, Joseph. The Growing Cyber Risks From AI. National Law Review, 18 giugno 2025.
18. Marzano, J. (21 settembre 2025). Ecco CopyCop, la macchina di disinformazione russa che punta ai modelli di IA. Formiche.net. https://formiche.net/2025/09/ecco-copycop-la-macchina-di-disinformazione-russa-che-punta-ai-modelli-di-ia/#content

MATICMIND®
MAKES IT EASY

www.maticmind.it

info@maticmind.it

ISO 9001 BUREAU VERITAS Certification
ISO 27001 BUREAU VERITAS Certification
ISO 20000 BUREAU VERITAS Certification
ISO 45001 BUREAU VERITAS Certification
ISO 14001 BUREAU VERITAS Certification
ISO 14064-1 BUREAU VERITAS Certification