t-defence

LLM e Cybersecurity

Il nuovo equilibrio tra capacità offensive e difensive

#TDefenceBusiness

AI4Cyber

Summary

AI4Cyber

AI4Cyber studio è il nuovo spazio di ricerca applicata di T-Defence, dedicato alla ricerca e applicazione dell'**Intelligenza Artificiale in ambito cybersecurity**.

Al suo interno, diversi specialisti con competenze ed esperienze eterogenee collaborano per affrontare le sfide poste dal panorama delle minacce informatiche, accompagnare i clienti nell'adozione dell'AI nei propri processi aziendali e condividere le attività di ricerca con la comunità tecnico-scientifica.

Il team cura l'**intero ciclo di sviluppo dei sistemi basati su AI**: dalla raccolta e dal preprocessing dei dati, all'addestramento e validazione dei modelli, fino al deployment in produzione.

Le soluzioni proposte si fondano su tecnologie avanzate di **Machine Learning**, **Deep Learning** e **Large Language Models**, e possono essere personalizzate in base alle specifiche esigenze del cliente.

L'AI Team è costantemente impegnato in attività di ricerca e sperimentazione, con un'attenzione particolare agli **aspetti etici**, alla **trasparenza** e alla **tutela della privacy**.

L'obiettivo è proporre soluzioni innovative che siano in linea con i principi di **equità**, **inclusività** e **rispetto dei diritti fondamentali**.

Abstract

Fra il 7 aprile e il 2 giugno 2026 sei casi di rilevanza sistemica hanno ridefinito la relazione fra Large Language Model e cybersecurity. Anthropic ha rilasciato il 7 aprile Claude Mythos Preview, modello con capacità autonome di scoperta di zero-day, distribuito in regime ristretto a circa cinquanta organizzazioni selezionate (Project Glasswing). OpenAI ha esteso il 14 aprile il programma Trusted Access for Cyber a migliaia di difensori individuali e centinaia di team, rilasciando il modello dedicato GPT-5.4-Cyber sotto verifica di identità. Il 21 aprile Bloomberg ha riportato, e Anthropic ha confermato, un accesso non autorizzato a Mythos Preview attraverso un fornitore terzo, avvenuto lo stesso giorno dell'annuncio del programma. Il 23 aprile OpenAI ha rilasciato GPT-5.5, modello generalista con capacità di vulnerability discovery, che valutazioni indipendenti descrivono come comparabili a Mythos, distribuito direttamente a utenti Plus, Pro, Business ed Enterprise. L'11 maggio Google Threat Intelligence Group ha pubblicato la prima evidenza documentata di un threat actor che impiega uno zero-day AI-developed in una campagna offensiva reale. Il 1º giugno Anthropic ha annunciato che ENISA sarà la prima agenzia europea ad accedere a Mythos, segnando la prima estensione del programma fuori da Stati Uniti e Regno Unito.

Questo studio analizza la sequenza nel suo insieme e ne valuta le implicazioni per l'equilibrio storico fra attaccanti e difensori, con attenzione specifica al contesto italiano ed europeo. Infine, include una sezione conclusiva con raccomandazioni operative differenziate per destinatario.

Autori:

- Gaetano Zappulla: CISO
- Simona Sorgente: AI4Cyber

1. Il Defender's Dilemma nell'era pre-LLM

La letteratura classica descrive la relazione fra attaccante e difensore come asimmetrica: all'attaccante basta riuscire una sola volta, il difensore deve riuscire sempre. La descrizione è corretta sul piano tecnico, ma nasconde una dimensione altrettanto importante: l'equilibrio è in larga misura economico, non puramente tecnico. Gli attacchi sofisticati sono storicamente rari non perché siano impossibili, ma perché richiedono competenze, risorse umane qualificate e tempi lunghi di preparazione che solo attori statali o gruppi criminali organizzati possono permettersi.

Il rapporto Google Defender's Dilemma (2024) ha esplicitato come le prime applicazioni di AI generativa alla sicurezza potessero favorire alla radice i difensori grazie al vantaggio informativo che questi ultimi detengono sui propri sistemi. Il position paper di Divakaran e Peddinti (arXiv, aprile 2024), rappresentativo del consenso accademico pre-agentic, identificava cinque aree in cui gli LLM avrebbero potuto offrire vantaggi ai difensori: rilevamento e gestione delle vulnerabilità, classificazione dei contenuti, spiegabilità e triage, arricchimento dei dataset, mitigazione dei rischi introdotti dagli LLM stessi. Il tono generale della letteratura del 2024 era prudentemente ottimista.

Quella previsione si fondava su due assunzioni implicite oggi erose. La prima è che le capacità offensive dei modelli rimanessero confinate dietro i filtri di sicurezza dei laboratori commerciali. La seconda è che lo sviluppo di modelli offensivamente capaci richiedesse risorse computazionali fuori portata per la maggior parte degli attaccanti. La progressione dei modelli open source nel 2025 e la democratizzazione del reinforcement learning post-training hanno modificato entrambe le assunzioni, sia pure con tempi e modi che restavano fino all'aprile 2026 oggetto di dibattito tecnico aperto. La sequenza di eventi descritta nella prossima sezione ha cristallizzato la transizione.

2. I cinque eventi della primavera 2026

2.1 Anthropic: Mythos Preview e il progetto Glasswing

Il 7 aprile 2026 Anthropic ha reso pubblico Claude Mythos Preview, modello di frontiera sviluppato a partire dall'architettura Claude Opus con capacità cyber significativamente più avanzate. Nella comunicazione ufficiale Anthropic dichiara che Mythos Preview avrebbe identificato in modo autonomo "migliaia" di vulnerabilità zero-day in ogni principale sistema operativo e browser. Tra gli esempi citati: una vulnerabilità di 27 anni nel codice TCP SACK di OpenBSD, un difetto di 16 anni nel codec H.264 di FFmpeg, una vulnerabilità di 17 anni nel server NFS di FreeBSD tracciata come CVE-2026-4747.

I dati di benchmark mostrano un avanzamento sostanziale rispetto alla generazione precedente. Su CyberGym, Mythos Preview ottiene 83,1% contro il 66,6% di Claude Opus 4.6 e il 65% di Claude Sonnet 4.6. Su Cybench, suite di 35 sfide CTF, Mythos satura il benchmark con 100% pass@1. Più rilevante operativamente è il benchmark interno Firefox 147 exploitation, dove Mythos produce 181 exploit funzionanti su 250 tentativi (con controllo dei registri aggiuntivo in altri 29 casi), contro 2 successi su diverse centinaia di tentativi di Opus 4.6: questa è la metrica che giustifica la caratterizzazione di un salto qualitativo specifico nella capacità di costruire exploit, distinta dalla capacità di scoperta.

La risposta di Anthropic è stata istituzionale e conservativa. Invece di rilasciare Mythos come prodotto commerciale, la società ha costituito il consorzio Project Glasswing, che comprende oltre cinquanta organizzazioni (12 partner di lancio fra cui AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, Linux Foundation, Microsoft, Nvidia, Palo Alto Networks; oltre quaranta organizzazioni aggiuntive di infrastruttura critica). Anthropic ha messo a disposizione fino a 100 milioni di dollari in usage credit e ulteriori 4 milioni in donazioni dirette a progetti open source di sicurezza. La presentazione pubblica è esplicita: il modello viene trattato come asset di sicurezza nazionale, non come prodotto commerciale. Il pricing interno al programma è stato pubblicato successivamente, fissato a 25 dollari per milione di token input e 125 dollari per milione di token output, circa cinque volte il pricing di Opus 4.6.

Verifica indipendente e aggiornamento del 26 maggio. *Il claim dei "migliaia di zero-day" non era pubblicamente verificabile al momento dell'annuncio. L'analisi VulnCheck del 15 aprile 2026 (P. Garrity) aveva rilevato che, su 75 record CVE che menzionavano Anthropic, solo CVE-2026-4747 era esplicitamente attribuito a Glasswing/Mythos. L'update di Anthropic del 26 maggio sul primo mese del programma dichiara oltre 10.000 vulnerabilità high/critical identificate dai partner, 6.202 stimate su più di 1.000 progetti open source, e un tasso di vero positivo del 90,6% su un campione di 1.752 valutate. Di queste, 530 sono state rivelate ai maintainer e 75 sono state risolte con patch. Sono numeri autodichiarati e ancora in attesa di verifica indipendente, ma rappresentano il primo dato aggregato che permette di superare il vuoto fattuale del 15 aprile. La valutazione indipendente dell'UK AI Security Institute conferma che Mythos è il primo modello a completare end-to-end alcuni test di attacco di rete (3 successi su 10 tentativi, media 22 step su 32), ma esplicita che gli ambienti di test non avevano le difese dei sistemi reali e che "non possiamo affermare con certezza se Mythos sarebbe in grado di attaccare sistemi ben difesi".*

2.2 OpenAI: Trusted Access for Cyber, GPT-5.4-Cyber e il rilascio di GPT-5.5

Sette giorni dopo l'annuncio di Glasswing, il 14 aprile 2026, OpenAI ha comunicato l'estensione del programma Trusted Access for Cyber (TAC), originariamente lanciato in forma pilota nel febbraio 2026 con un grant program da 10 milioni di dollari. Il programma è stato aperto a migliaia di difensori individuali e centinaia di team, previa verifica di identità automatizzata via KYC accessibile a chatgpt.com/cyber per gli utenti individuali e tramite rappresentante OpenAI per le organizzazioni. Contestualmente è stato rilasciato GPT-5.4-Cyber, variante di GPT-5.4 specificamente fine-tuned per casi d'uso difensivi, con un tasso di rifiuto significativamente più basso per query legate alla ricerca di vulnerabilità, all'analisi di exploit e alla comprensione del comportamento di malware. Tra le nuove capacità figura il reverse engineering di binari, utile per analizzare firmware, librerie di terze parti e malware senza accesso al codice sorgente.

L'inquadramento di OpenAI è diametralmente opposto a quello di Anthropic. Fouad Matin, ricercatore cyber di OpenAI, ha riassunto la filosofia con una frase destinata a diventare slogan: "nessuno dovrebbe avere il compito di scegliere vincitori e vinti quando si parla di cybersecurity".

La logica sottostante è che restringere l'accesso a capacità difensive avanzate a poche grandi organizzazioni lascerebbe scoperte migliaia di realtà (infrastrutture critiche, ospedali, enti locali, piccole società di sicurezza) che difendono porzioni altrettanto rilevanti dell'ecosistema digitale.

Il rilascio di GPT-5.5 del 23-24 aprile. *Nove giorni dopo l'estensione di TAC, OpenAI ha rilasciato GPT-5.5 e GPT-5.5 Pro come modelli generalisti, distribuendoli a Plus, Pro, Business ed Enterprise via ChatGPT e Codex il 23 aprile, e via API dal 24 aprile (con classificatori più stretti per query cyber-sensibili). Una valutazione indipendente di XBOW, società di pentesting agentic con rapporto commerciale dichiarato con OpenAI, riporta su benchmark interni di vulnerabilità note un miss rate del 10% per GPT-5.5, contro il 18% di Claude Opus 4.6 e il 40% del precedente GPT-5. Sul benchmark accademico CyberGym GPT-5.5 ottiene 81,8% contro 83,1% di Mythos. Il dato di benchmark è fornito da soggetto interessato, ma è riproducibile e converge con altre valutazioni indipendenti che pongono GPT-5.5 e Mythos in territorio paragonabile. La conseguenza per il quadro generale è significativa: capacità essenzialmente equivalenti a Mythos sono ora distribuite a milioni di utenti enterprise.*

2.3 L'epilogo del 21 aprile: la breccia di Mythos

Il 21 aprile 2026 Bloomberg News ha riportato che un gruppo di utenti non autorizzati aveva ottenuto accesso a Mythos Preview lo stesso giorno dell'annuncio di Glasswing (7 aprile), attraverso l'ambiente di un fornitore terzo. Il vettore documentato dalle ricostruzioni di Bloomberg, The Register e Boing Boing combina due elementi: un attacco di supply chain al proxy LiteLLM utilizzato dalla piattaforma di placement Mercor (a cui era impiegato uno dei membri del gruppo, contractor di Anthropic), e un URL guessing basato sulle convenzioni di naming di Anthropic per i modelli precedenti. Il gruppo, descritto come una community Discord interessata al test di modelli non rilasciati, ha avuto accesso continuativo dal 7 aprile fino almeno alla pubblicazione del report Bloomberg. Anthropic ha dichiarato di non avere evidenza di compromissione dei propri sistemi né di estensione dell'attività oltre l'ambiente del fornitore terzo.

Indipendentemente dall'utilizzo specifico fatto dal gruppo, l'incidente ha implicazioni dirette sull'inquadramento della Sezione 2.1. L'argomento di Anthropic, ovvero che il contenimento dei rischi cyber dei modelli di frontiera sia gestibile attraverso la selezione ex ante di un numero ristretto di partner, viene confrontato con un dato concreto: nel giorno stesso dell'annuncio, e prima dell'attivazione operativa del programma per gran parte dei partner, il modello era già accessibile fuori dal perimetro Glasswing.

La causa prossima è un classico problema di terza parte (compromissione di un fornitore con accesso interno), categoria di rischio per cui esistono framework consolidati (NIST SP 800-161, EU CRA, requisiti NIS2 sulla supply chain). La causa più profonda è strutturale: "accesso ristretto" significa nei fatti accesso esteso a migliaia di persone presso decine di organizzazioni, con tutti i rischi di insider threat e supply chain associati. L'incidente quindi sposta in dimensione già osservativa la seguente previsione: il controllo dell'accesso ai modelli più capaci si ridurrà progressivamente, da meccanismo di contenimento a meccanismo di rallentamento.

2.4 L'11 maggio: il primo zero-day AI-developed osservato in the wild

L'11 maggio 2026 Google Threat Intelligence Group ha pubblicato il primo caso documentato di un threat actor che ha impiegato un'intelligenza artificiale per sviluppare uno zero-day exploit poi utilizzato in una campagna offensiva reale. Lo zero-day è un 2FA bypass implementato in uno script Python che colpisce uno strumento open-source ampiamente diffuso per system administration via web. Google ha confermato, con alto grado di confidenza, che il codice exploit è stato sviluppato con il supporto di un Large Language Model, sulla base di indicatori forensi rilevati nel codice stesso (commenti eccessivamente esplicativi, un CVSS score allucinato, classi ANSI color di forma standardizzata). Il threat actor è un gruppo cybercrime financially motivated. Il modello AI impiegato non è dichiarato. La campagna prevista era un mass exploitation event, bloccato grazie a una proactive counter discovery di Google e a una coordinated disclosure verso il vendor prima dell'esecuzione.

La stessa pubblicazione di Google estende l'osservazione ad altri attori statali e cybercrime. UNC2814, gruppo collegato alla Cina noto per attacchi contro telecomunicazioni e organizzazioni governative, è stato osservato utilizzare un persona-driven jailbreak (l'AI viene istruita ad agire come un "senior security auditor") per condurre vulnerability research su dispositivi embedded, incluso il firmware TP-Link con implementazioni OFTP. APT45, gruppo nordcoreano, è stato osservato inviare migliaia di prompt ripetitivi per analisi ricorsive di CVE e validazione di proof-of-concept exploit, costruendo quello che Google definisce "un arsenale di capacità di exploit difficile da gestire senza assistenza AI". Sul versante offensivo agentic, attori cinesi sono stati osservati impiegare framework come Strix e Hexstrike in attacchi contro un'azienda tecnologica giapponese e una società cybersecurity dell'Asia orientale.

2.5 Il 1º giugno: ENISA come prima agenzia europea con accesso a Mythos

Il 1º giugno 2026 Anthropic ha annunciato che l'Agenzia dell'Unione Europea per la Cybersicurezza (ENISA) avrà accesso a Mythos Preview, prima estensione del programma Glasswing fuori da Stati Uniti e Regno Unito. L'apertura è il risultato di settimane di negoziazioni fra la Commissione Europea e Anthropic, con quattro o cinque incontri formali nel corso di aprile e maggio, e di un'intermediazione esplicita del governo statunitense: Anthropic aveva infatti dichiarato che la condivisione del modello con governi esteri richiedeva l'autorizzazione di Washington, in coerenza con il framework di export control statunitense sulle tecnologie dual-use AI. La Commissione Europea, attraverso il portavoce per le sovranità tecnologiche Thomas Regnier, ha definito gli incontri con Anthropic "produttivi" e ha confermato che l'obiettivo dell'accesso è ottenere "una comprensione più chiara dei rischi potenziali che la tecnologia comporta".

L'apertura ENISA arriva dopo un evento parallelo del versante OpenAI. L'11 maggio OpenAI aveva pubblicato l'EU Cyber Action Plan, programma dedicato che estende GPT-5.5-Cyber a difensori cyber qualificati europei (imprese, governi, agenzie cyber e istituzioni UE come l'EU AI Office), con Deutsche Telekom e BBVA fra i primi partecipanti documentati. La differenza fra i due programmi resta significativa. Quello di OpenAI è strutturato come licenza commerciale ad accesso verificato e ha tre tier (GPT-5.5 standard, GPT-5.5 con TAC, GPT-5.5-Cyber); quello di Anthropic verso ENISA è bilaterale, su base di partenariato istituzionale, con condizioni operative ancora da concordare. Entrambi i programmi non sono al momento accessibili alle organizzazioni italiane medie.

Il commento del portavoce UE. *Thomas Regnier, portavoce della Commissione per le sovranità tecnologiche, ha dichiarato il 1º giugno: "Mythos non è un caso isolato, una nuova ondata di modelli potenti sta arrivando sul mercato. Questa è una sfida comune, e stiamo intensificando le discussioni con partner affini, inclusi gli Stati Uniti". La dichiarazione segna un punto importante: l'approccio europeo al fenomeno non è confronto avversariale con i laboratori statunitensi, ma costruzione di un dispositivo cooperativo bilanciato dall'esercizio della sovranità regolatoria europea. La sostenibilità di questo equilibrio nel medio termine dipenderà dal grado di reciprocità che il sistema europeo riuscirà a offrire.*

3. Implicazioni per l'equilibrio attaccante–difensore

3.1 Il riequilibrio economico in corso

L'ipotesi che gli LLM possano modificare il Defender's Dilemma a favore della difesa si fondava su un presupposto specifico: che i costi di accesso alle capacità offensive più avanzate rimanessero proibitivi per la maggior parte degli attaccanti. I dati empirici raccolti fra aprile e giugno 2026 stanno modificando rapidamente questo presupposto. La rilevazione di vulnerabilità in codice già divulgato è oggi una capacità banalizzata e ampiamente accessibile (confidenza alta). La distanza fra modelli open-weight e modelli di frontiera nella scoperta esplorativa di vulnerabilità in codebase reali si è significativamente ridotta, con AISLE che ha documentato il 13 aprile otto modelli open-weight su otto in grado di individuare la vulnerabilità FreeBSD CVE-2026-4747. La capacità di costruzione affidabile di exploit resta più dipendente dai modelli di frontiera (confidenza media). La proiezione di una riduzione di uno o due ordini di grandezza del costo unitario degli attacchi sofisticati entro 12-18 mesi è plausibile come scenario di pianificazione, non come premessa fattuale (confidenza bassa, proiezione).

3.2 Cinque classi di rischio emergenti

La letteratura 2024 sugli LLM in cybersecurity mappava un insieme coerente di opportunità difensive. La traiettoria 2026 apre cinque classi di rischio non trattate, o trattate solo marginalmente, nel precedente consenso accademico.

- **Exploit chaining autonomo.** Mythos ha dimostrato la capacità di concatenare quattro vulnerabilità distinte in un unico exploit funzionante (con i caveat di configurazione documentati in 2.1). Il caso Calif del 15 maggio (privilege escalation exploit per macOS sviluppato con assistenza Mythos, riportato da WSJ) conferma che la concatenazione assistita di catene di exploit complesse è ormai accessibile a soggetti con risorse modeste. La compressione del tempo fra scoperta e sfruttamento si sposta verso tempi più brevi, anche se non istantanei.

- **Scoperta zero-day su larga scala.** Il dato di Anthropic del 26 maggio sul primo mese di Glasswing parla di oltre 10.000 vulnerabilità high/critical identificate, 6.202 stimate su più di 1.000 progetti open source, e 90,6% di tasso di vero positivo dichiarato su un campione validato. Anche assumendo una verifica indipendente che ridimensioni i numeri, l'asimmetria fra volume di scoperte possibili e capacità di remediation degli ecosistemi software resta una preoccupazione strutturale.
- **Nuove classi di attacco (speculativo).** La possibilità che modelli di prossima generazione scoprano tecniche di exploitation qualitativamente inedite, analoghe per portata al return-oriented programming o agli attacchi ai canali laterali della cache, è aperta ma oggi puramente speculativa. Mythos opera in larga parte su classi di vulnerabilità note. Aggiornamento dei framework di threat modeling prudente, non urgente.
- **Compromissione della catena di fornitura modellistica.** L'adozione di modelli open source fine-tuned per security introduce un rischio di poisoning intenzionale durante il fine-tuning. La provenienza e l'integrità del modello diventano asset di sicurezza al pari del codice applicativo. La filiera del laboratorio stesso (vendor terzi, contractor) è vettore documentato di proliferazione, come dimostra l'incidente del 21 aprile.
- **Sistemi legacy non patchabili.** Una quota significativa dell'infrastruttura critica, in particolare nella Pubblica Amministrazione italiana, si basa su sistemi con codice sorgente di fatto non più accessibile o manutenibile. Le capacità AI-native di scoperta espongono in modo sproporzionato queste componenti senza offrire un percorso di remediation praticabile.

3.3 Due fenomeni distinti: collasso della finestra di patch e saturazione della capacità di ingestione

Un sesto profilo di rischio, distinto dalle cinque classi precedentemente descritte, riguarda due fenomeni: il primo è il collasso della finestra temporale fra rilascio di un fix e comparsa di un exploit funzionante (problema di velocità, lato attaccante); il secondo è la saturazione della capacità organizzativa di ingerire e applicare patch (problema di capacità, lato difensore). Sono manifestazioni della medesima discontinuità sistemica, l'AI come acceleratore del ciclo di vita delle vulnerabilità, ma richiedono risposte tecniche e organizzative diverse.

3.4 L'AI difensiva come nuova superficie di attacco

Mano a mano che le organizzazioni integrano modelli LLM nei propri SOC, EDR/XDR, processi di triage e code review, questi sistemi diventano essi stessi obiettivi di attacco con vettori specifici e diversi da quelli classici. Quattro vettori meritano attenzione operativa:

- **Prompt injection contro security agent automatizzati:** payload progettati in log, alert, allegati o issue tracker per essere interpretati come istruzioni dal modello (false negative artificiali, escalation a operatori sbagliati, esfiltrazione di contesto).
- **Gli adversarial examples mirati contro detector basati su LLM**: payload tecnicamente malevoli costruiti per essere classificati benigni, con generazione oggi essa stessa automatizzabile via LLM.
- **Il jailbreaking** dei sistemi di triage automatico per far classificare incidenti reali come falsi positivi.
- **L'attacco al canale di apprendimento**, se l'organizzazione utilizza fine-tuning incrementale del proprio modello difensivo: un attaccante che inietta segnali manipolati può degradare progressivamente la qualità del classificatore.

La maturità dei framework di difesa contro queste classi di attacco è oggi modesta. NIST AI 100-2e2023 fornisce una tassonomia di riferimento ma è prevalentemente prospettica. ENISA ha pubblicato guidance preliminare nel 2025, ma il segmento specifico dell'attacco contro security agent LLM è ancora oggetto di ricerca. Per le organizzazioni che adottano architetture AI-augmented in chiave difensiva, la valutazione del rischio non può limitarsi al perimetro classico (CIA dei sistemi protetti) ma deve estendersi all'integrità decisionale dei propri stessi componenti AI.

4. Il contesto europeo e italiano: vuoto normativo, organismi e sovranità

4.1 Il vuoto normativo

Il quadro regolatorio europeo in materia di intelligenza artificiale e cybersecurity è il più articolato al mondo, ma non affronta direttamente il fenomeno emerso nel bimestre aprile-maggio 2026. La discussione sull'applicabilità degli obblighi dei sistemi ad alto rischio dell'AI Act ai modelli cyber-permissive è oggi aperta. La NIS2 non fornisce linee guida specifiche su come gestire il rischio derivante da attaccanti AI-equipaggiati. Il medesimo gap si ritrova in DORA. Il vuoto normativo è temporaneo e verrà riempito nei prossimi 18-24 mesi attraverso atti di esecuzione, linee guida ENISA e provvedimenti di autorità nazionali. In Italia, la Determinazione ACN n. 155238/2026 del 13 aprile costituisce il primo passo operativo del sistema NIS2 e impone ai soggetti essential e important una categorizzazione strutturata delle proprie attività e dei propri servizi, propedeutica all'applicazione delle misure di sicurezza differenziate per categoria di rilevanza. Nel frattempo, le organizzazioni operano in zona di ambiguità in cui le decisioni di governance, procurement e contrattualizzazione devono essere assunte sulla base di valutazioni di rischio proprie.

4.2 Il ruolo degli organismi nazionali ed europei

La governance del rischio AI-cyber in Europa coinvolge oggi una pluralità di attori con mandati parzialmente sovrapposti. A livello europeo: ENISA mantiene la responsabilità primaria sulla cybersecurity e dal 1° giugno 2026 è la prima agenzia europea con accesso al modello Anthropic Mythos Preview attraverso il programma Project Glasswing; il Comitato Europeo per l'IA presiede all'applicazione dell'AI Act, con l'EU AI Office che ha avuto accesso a GPT-5.5-Cyber dall'11 maggio 2026 attraverso l'OpenAI EU Cyber Action Plan; CERT-EU presiede al coordinamento delle risposte incidentali per le istituzioni europee; il Joint Cyber Unit presiede alla risposta a incidenti transfrontalieri.

A livello italiano: ACN ha costruito un perimetro di competenze che comprende sicurezza delle infrastrutture critiche, valutazione di componenti tecnologiche sensibili (CVCN), coordinamento del CSIRT Italia, attuazione di NIS2 e perimetro nazionale. DTD, MIMIT (per i profili industriali) e il Comparto Intelligence completano il quadro.

L'accesso ENISA a Mythos rappresenta una svolta operativa per l'ecosistema europeo: ENISA può ora condurre valutazioni di rischio tecnico sul modello con accesso diretto alle capability, fattore che modifica le condizioni in cui ACN e altre agenzie nazionali potranno costruire propri pareri e provvedimenti. È ragionevole attendersi nei prossimi sei-dodici mesi una posizione coordinata ACN-ENISA che chiarisca, almeno: l'inquadramento degli LLM cyber-permissive ai sensi dell'AI Act; le condizioni operative per l'utilizzo di servizi AI-cyber americani da parte di soggetti del perimetro nazionale; le linee guida di approvvigionamento per le PA. La posizione coordinata dovrebbe coinvolgere l'industria nazionale, le università italiane (CINI, IIT, gruppi specializzati) e i corrispondenti francesi e tedeschi attraverso il framework EU CyCLONe. Il Forum Cyber 4.0 di Roma del 3-4 giugno 2026 è il primo appuntamento istituzionale italiano successivo alla sequenza di eventi qui analizzata, e costituisce un'occasione naturale per articolare la posizione.

4.3 Sovranità digitale: tre opzioni concrete

Gli approcci di Anthropic e OpenAI sono entrambi strettamente legati all'ecosistema tecnologico e regolatorio statunitense. L'apertura ENISA del 1º giugno, ottenuta con l'intermediazione esplicita del governo statunitense, conferma il vincolo strutturale: condividere modelli AI di frontiera con governi esteri richiede l'autorizzazione di Washington. Le organizzazioni europee che adottano le soluzioni dei laboratori statunitensi accettano implicitamente una dipendenza tecnologica su capacità ritenute critiche dagli Stati stessi nelle proprie dottrine di sicurezza nazionale. Le opzioni strategiche europee sono essenzialmente tre.

Opzione A: Consumo. Adozione delle capacità difensive offerte dai laboratori americani, con accettazione delle condizioni contrattuali e di visibilità operativa. Costo modesto, attivazione in settimane, forte dipendenza dal fornitore (lock-in) e incompatibilità strutturale con perimetro classificato. Adatta a organizzazioni private non sottoposte a vincoli di sovranità (manifatturiero, retail, alcuni segmenti del finance non sistemico). L'EU Cyber Action Plan di OpenAI dell'11 maggio formalizza un canale di accesso specifico per il mercato europeo.

Opzione B: Sviluppo sovrano. Sviluppo di capacità europee a livello di modelli base e piattaforme security AI-nativa. Costo dell'ordine di vari miliardi distribuiti su 3-5 anni, maturazione pluriennale, rischio tecnologico significativo. Coerente con iniziative in corso (EuroStack, GAIA-X scope cyber, AI4Trust, EuroHPC). Adatta come obiettivo politico-industriale di medio periodo.

Opzione C: Postura ibrida. Capacità americane per casi d'uso non critici, capacità europee o nazionali per le funzioni più sensibili (dati classificati, infrastrutture strategiche, difesa). Operativamente più realistica nel breve-medio per il sistema-Italia, ma richiede segmentazione esplicita dei carichi di lavoro che oggi le organizzazioni italiane non hanno mappato sistematicamente. L'apertura ENISA del 1° giugno è di fatto la traduzione istituzionale di questa opzione a livello UE: accesso bilaterale e segregato per le funzioni di valutazione, non integrazione operativa nelle pipeline di produzione. Lo strumento MIMIT Voucher Cloud e Cybersecurity 2026 (150 milioni di euro) può costituire una leva di accompagnamento per le PMI nella transizione verso fornitori conformi.

4.4 La compliance dell'accesso come vincolo operativo italiano

Per le organizzazioni italiane che operano nel perimetro di dati classificati, segreto industriale o dati sensibili sotto NIS2, la domanda rilevante non è solo "quale modello è più capace?" ma "quale modello è effettivamente applicabile in produzione con garanzie di sicurezza, compliance e continuità operativa?". Il programma TAC di OpenAI presenta due frizioni rilevanti. Il KYC: l'identificazione personale degli operatori, con conservazione presso OpenAI dei dati di verifica, è in tensione con i regimi di anonimizzazione operativa adottati da unità di intelligence, forze dell'ordine e servizi di sicurezza in attività investigativa o di threat hunting. La possibile rinuncia a Zero-Data Retention per gli utilizzi più sensibili: le query effettuate dall'operatore, che possono includere snippet di codice riservato, IoC interni, ricostruzioni di catene di attacco con dati identificativi della propria organizzazione, possono essere conservate da OpenAI per finalità di safety e monitoring. Alla luce delle norme sul perimetro nazionale e della disciplina su dati classificati di cui al DPCM 6 novembre 2015 e norme attuative, è probabile che alcune condizioni del programma TAC (KYC con conservazione esterna, rinuncia a ZDR) risultino non adottabili per carichi di lavoro classificati o sottoposti a regime di perimetro, salvo specifiche mitigazioni contrattuali o valutazioni di ACN. L'inferenza è esplicitamente analitica, non un parere tecnico-giuridico: la valutazione operativa è prerogativa delle autorità competenti e degli uffici legali delle singole organizzazioni.

Il programma Glasswing risolve formalmente il problema della retention attraverso accordi contrattuali ad hoc, ma è di fatto inaccessibile alle organizzazioni italiane medie: nessuna organizzazione italiana figura attualmente fra i partner pubblicati, e l'accesso ENISA del 1º giugno è istituzionalmente segregato e finalizzato alla valutazione, non al deployment operativo. L'EU Cyber Action Plan di OpenAI ha introdotto a partire dal 1º giugno l'obbligo di Advanced Account Security per i membri individuali di Trusted Access for Cyber, oppure in alternativa di attestare l'uso di autenticazione phishing-resistant a livello organizzativo: requisito che ridefinisce gli standard minimi di adesione e che le organizzazioni italiane interessate devono valutare contestualmente alle proprie politiche di gestione delle identità privilegiate. Per la quasi totalità delle organizzazioni italiane, la decisione strategica nei prossimi dodici mesi non è quindi fra Mythos e GPT-5.4-Cyber/GPT-5.5-Cyber, ma fra attendere modalità di accesso compatibili con i propri vincoli di compliance, sviluppare capacità interne basate su modelli open source self-hosted, stabilire partnership con vendor europei in grado di offrire garanzie di residenza dati e segregazione.

5. Raccomandazioni operative

Le sezioni precedenti hanno articolato un'analisi del fenomeno. Questa sezione conclusiva traduce l'analisi in raccomandazioni differenziate per categoria di destinatario, formulate per essere implementabili sui prossimi sei-dodici mesi senza presupporre disponibilità di budget straordinari né accesso a programmi vendor selettivi.

5.1 Per i CISO di organizzazioni medio-grandi

- **Inventario di esposizione AI-discovery.** Mappare i sistemi e le componenti software con codice sorgente pubblicamente disponibile o accessibile a terze parti, con priorità a quelli che gestiscono autenticazione, sessione, dati sensibili. Sono i candidati più probabili per scoperte AI-assisted nei prossimi mesi. L'evidenza Google GTIG dell'11 maggio (2FA bypass su tool open-source di system administration) indica che i target preferiti dei threat actor AI-equipped sono oggi gli strumenti operativi “widely deployed” con codice pubblico.

- **Commit-diff monitoring difensivo.** Implementare monitoraggio AI-assisted dei diff dei commit di sicurezza pubblicati dai vendor del proprio stack, con generazione automatica di IoC e alert ai team di patching. Riduce la finestra fra patch pubblica e protezione effettiva, traducendo in vantaggio difensivo il fenomeno descritto in 5.4.
- **Architetture di runtime protection.** Investire in tecnologie di protezione runtime (memory tagging, CFI hardware-assistito, RASP per applicazioni esposte) che riducano la dipendenza dalla tempestività delle patch del singolo componente. È la risposta architetturale al collasso della finestra di patch.
- **Esercitazioni threat-led su attaccanti AI-assisted.** Aggiornare red teaming e purple teaming per includere scenari in cui l'attaccante simulato dispone di capacità AI-augmented: sviluppo rapido di exploit da diff pubblici, prompt injection contro security agent automatizzati, generazione di adversarial examples, persona-driven jailbreak come quello documentato da Google GTIG. Per i soggetti DORA, integrare nei programmi TLPT.
- **IoC machine-generated nel proprio threat intelligence.** Aggiungere alla propria pipeline di threat intelligence la classe di IoC tipica di codice machine-generated documentata da Google GTIG: commenti eccessivamente esplicativi nei payload, CVSS score sintetizzati e allucinati, pattern stilistici riconducibili a output LLM, classi ANSI color standardizzate, struttura di error handling sovra-articolata. Sono indicatori imperfetti ma operativamente utili nelle fasi di triage iniziale.
- **Budget per compute AI difensivo.** Pianificare nei piani 2026-2027 una linea di spesa esplicita per inferenza continua di modelli difensivi sui propri sistemi, distinta dalla spesa in strumenti tradizionali. Il pricing Glasswing pubblicato (25 dollari per milione di token input, 125 per output, circa cinque volte Opus 4.6) fornisce un primo riferimento per il dimensionamento, anche se l'accesso a Mythos resta limitato. Dimensionare assumendo progressione triennale, non come investimento una tantum.
- **Governance dei modelli AI difensivi come asset di sicurezza.** Trattare i modelli AI integrati nello stack difensivo (security agent, classificatori di alert, sistemi di triage) come asset critici sottoposti a verifica di provenienza, monitoraggio di integrità e protezione da attacchi (prompt injection, adversarial examples, jailbreaking). La maturità dei controlli su questa classe di asset è oggi modesta e va costruita.

- **Funzione VulnOps dedicata e prioritizzazione exploitability-based.** Costruire (dove non esiste già) una funzione dedicata di Vulnerability Operations distinta dalla classica vulnerability management, con responsabilità su triage, prioritizzazione, validazione e orchestrazione della patch deployment a fronte del volume crescente di disclosure. La prioritizzazione deve fondarsi su exploitability reale nel proprio ambiente, non sul mero severity score, e deve includere validazione automatizzata pre-deployment per ridurre il rischio di patch che introducono regressioni. Il Cloud Security Alliance, nella guidance del 16 aprile 2026, ha articolato la raccomandazione in modo simile.

5.2 Per la Pubblica Amministrazione e i soggetti del perimetro

- **Completamento immediato degli adempimenti NIS2 italiani.** Le scadenze del 31 maggio (aggiornamento piattaforma ACN) e del 30 giugno 2026 (categorizzazione attività e servizi sotto Determinazione 155238/2026) sono adempimenti formali ma con conseguenze sostanziali: la categorizzazione è la base su cui ACN modulerà le misure di sicurezza progressive da implementare entro ottobre 2026. Le organizzazioni che non hanno completato la gap analysis e l'inventario degli asset hanno una finestra ridotta.
- **Censimento dei sistemi legacy esposti.** Completare il censimento dei sistemi con codice sorgente non più manutenibile e classificarli per criticità. Per ciascuno definire una strategia: dismissione, isolamento, runtime protection, modernizzazione. È la categoria di asset più sproporzionatamente esposta alla traiettoria AI-discovery.
- **Engagement strutturato con ACN.** Per i soggetti inclusi nel perimetro nazionale, avviare un dialogo strutturato con ACN sulle modalità ammissibili di utilizzo di servizi AI-cyber americani (KYC, ZDR, residenza dati). Anticipare la richiesta di un parere riduce il rischio che decisioni di procurement debbano essere riviste a posteriori. L'apertura ENISA del 1º giugno e l'EU Cyber Action Plan OpenAI dell'11 maggio modificano il contesto e dovrebbero entrare nella discussione.
- **Criteri di procurement per tool AI-cyber.** Definire criteri di valutazione standardizzati: residenza dati, regime di retention, segregazione tenant, provenienza del modello, possibilità di self-hosting, compatibilità con regimi di classifica nazionale. Pubblicare i criteri come linea guida CONSIP/AGID per favorire un mercato uniforme.

5.3 Per ACN, ENISA e regolatori settoriali

- **Valorizzazione operativa dell'accesso ENISA a Mythos.** L'accesso ENISA a Mythos del 1º giugno è oggi una finestra di valutazione, non un canale di deployment. Il valore strategico per il sistema europeo dipende dalla capacità di tradurre la valutazione in guidance operativa pubblica e condivisibile (con ACN, ANSSI francese, BSI tedesco, NCSC olandese, e per quanto possibile con i partner di rilevanza istituzionale italiana). La pubblicazione di un primo documento congiunto ENISA/ACN sull'inquadramento operativo degli LLM cyber-permissive prima della piena applicabilità dell'AI Act (2 agosto 2026) sarebbe un'azione di alto valore.
- **Inquadramento esplicito degli LLM cyber-permissive sotto AI Act.** Pubblicare guidance sull'applicabilità degli obblighi dei sistemi ad alto rischio dell'AI Act ai modelli generalisti con capacità cyber offensive significative. Gap regolatorio identificato dal dibattito tecnico, opportuno colmarlo prima della piena applicabilità di agosto 2026.
- **Framework di disclosure coordinata per vulnerabilità AI-discovered.** I volumi documentati dall'update Anthropic del 26 maggio eccedono la capacità di assorbimento dei modelli di disclosure attuali. Necessario un framework che includa: provenienza del riscontro, validazione indipendente, prioritizzazione concertata fra vendor e maintainer, allineamento con le pipeline di patch management nei settori regolati.
- **Roadmap di capacità sovrane.** Esplicitare una roadmap pluriennale per lo sviluppo di capacità AI-cyber europee (modelli di base, piattaforme verticali, accesso compute), agganciata agli strumenti di finanziamento esistenti (Horizon Europe, EuroHPC, Digital Europe Programme, fondi nazionali). Maturazione richiede 3-5 anni; ritardare l'avvio aumenta corrispondentemente la dipendenza strategica.

5.4 Per vendor di software e maintainer open source

- **AI-assisted code review come standard di qualità.** Adottare l'analisi AI-assisted del codice come standard di qualità per il rilascio, includendola nei processi SDLC. Il vantaggio è simmetrico al rischio: ciò che un attaccante può trovare nel codice, anche il difensore può trovarlo prima del rilascio.
- **Riduzione della dipendenza dalle finestre di patch.** Privilegiare architetture difensive in profondità che non dipendano dalla tempestività della singola patch, e fornire ai propri clienti documentazione esplicita su quali compensanti operativi possono essere applicati nelle ore o giorni precedenti l'applicazione di un fix.

- **Engagement con programmi di disclosure AI-aware.** Aggiornare le proprie security advisory page e i programmi di bug bounty per gestire riscontri generati da AI, con metadati di provenienza, validazione indipendente e tempistiche di disclosure adeguate al volume.

Bibliografia e fonti

- Anthropic, «Project Glasswing: Securing critical software for the AI era», 7 aprile 2026, anthropic.com/glasswing
- Anthropic Frontier Red Team, «Claude Mythos Preview: Technical Report», red.anthropic.com/2026/mythos-preview, 7 aprile 2026
- Anthropic, «Claude Mythos Preview System Card», aprile 2026
- Anthropic, «Project Glasswing — Initial Update», 25-26 maggio 2026 (oltre 10.000 vulnerabilità identificate nel primo mese)
- Google Threat Intelligence Group, «Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access», cloud.google.com/blog/topics/threat-intelligence, 11 maggio 2026
- OpenAI, «Trusted access for the next era of cyber defense», 14 aprile 2026
- OpenAI, «Introducing GPT-5.5», 23 aprile 2026
- OpenAI, «GPT-5.5 System Card», aprile 2026
- OpenAI, «EU Cyber Action Plan» (G. Osborne, M. Signoux), 11 maggio 2026
- OpenAI, «Introducing Trusted Access for Cyber», 5 febbraio 2026
- XBOW (A. Ziegler, S. Buckley), «GPT-5.5: Mythos-Like Hacking, Open To All», xbow.com/blog, 23 aprile 2026
- XBOW (N. Waisman, O. de Moor), «GPT-5.5: Democratizing Cyber Capabilities», xbow.com/blog, 23 aprile 2026
- VulnCheck (P. Garrity), «Tracking CVEs Attributed to Anthropic Researchers and Project Glasswing», 15 aprile 2026
- AISLE (S. Fort), «AI Cybersecurity After Mythos: The Jagged Frontier», 13 aprile 2026
- B. Schneier, «On Anthropic's Mythos Preview and Project Glasswing», schneier.com, 22 aprile 2026
- UK AI Security Institute, valutazione indipendente di Mythos Preview, aprile 2026

- Calif, «MADBugs — Month of AI-Discovered Bugs», github.com/califio/publications/MADBugs, marzo-aprile 2026
- Calif, «Claude + Humans vs nginx: CVE-2026-27654», 10 aprile 2026
- Calif, ricerca su privilege escalation macOS aiutata da Mythos, 15 maggio 2026 (via WSJ ed Engadget)
- Cloud Security Alliance, «The AI Vulnerability Storm: Building a Mythos-ready Security Program», 16 aprile 2026
- Institute for AI Policy and Strategy, «Mythos and the Evolving Cyber Landscape», iaps.ai, aprile 2026
- D. Stenberg (lead maintainer cURL), dichiarazioni a Bloomberg e a NPR, aprile 2026
- R. Mogull, «Project Glasswing analysis», Securosis blog, 8 aprile 2026
- D. Mon Divakaran, S. T. Peddinti, «LLMs for Cyber Security: New Opportunities», arXiv:2404.11338, aprile 2024
- A. Vassilev et al., «Adversarial machine learning: A taxonomy and terminology», NIST AI 100-2e2023, 2024
- Google, «Secure, Empower, Advance: How AI Can Reverse the Defender's Dilemma», 2024
- N. Carlini et al. (Anthropic Frontier Red Team), «Evaluating and mitigating the growing risk of LLM-discovered 0-days», 5 febbraio 2026
- Stanford HAI, «Artificial Intelligence Index Report 2025», Stanford Institute for Human-Centered AI, 2025
- Epoch AI, «Frontier AI Compute and Capability Reports», 2024-2026
- Regolamento (UE) 2024/1689 («AI Act»), in vigore dal 1º agosto 2024
- Direttiva (UE) 2022/2555 («NIS2»), recepita con D.lgs. 138/2024
- Regolamento (UE) 2022/2554 («DORA»), pienamente applicabile dal 17 gennaio 2025
- Regolamento (UE) 2024/2847 («Cyber Resilience Act»)
- DPCM 6 novembre 2015 e norme attuative del Perimetro Nazionale di Sicurezza Cibernetica
- ACN, Determinazione n. 155238/2026 del 13 aprile 2026, Linee Guida e FAQ
- Bloomberg News, «Anthropic to Give EU's Cybersecurity Agency Access to Mythos» (G. Volpicelli), 1 giugno 2026
- CNBC, «Anthropic to offer EU access to its advanced Mythos model», 1 giugno 2026

- Financial Times via Resultsense, «Anthropic offers EU access to Mythos cyber AI model», 2 giugno 2026
- Help Net Security, «Anthropic: Claude Mythos identified 10,000+ software flaws» (A. Pogorelec), 26 maggio 2026
- The Register, «Anthropic to release Mythos-class models to the public», 25 maggio 2026
- PYMNTS, «Anthropic Will Update Regulators on Mythos' Cyber Vulnerability Findings», 18 maggio 2026
- Engadget, «Security researchers, aided by Anthropic's Mythos, claim to have breached macOS», 15 maggio 2026
- Bloomberg News, «Google Researchers Detect First AI-Built Zero-Day Exploit in Cyberattack», 11 maggio 2026
- SecurityWeek, «Google Detects First AI-Generated Zero-Day Exploit», 12 maggio 2026
- The Hacker News, «Hackers Used AI to Develop First Known Zero-Day 2FA Bypass for Mass Exploitation», 12 maggio 2026
- CNBC, «OpenAI to give EU access to new cyber model», 11 maggio 2026
- EdTech Innovation Hub, «OpenAI expands GPT-5.5 cyber defense access to Europe», 11 maggio 2026
- Help Net Security, «OpenAI's GPT-5.5 is out with expanded cybersecurity safeguards», 24 aprile 2026
- Bloomberg News, «Anthropic's Mythos AI Model Is Being Accessed by Unauthorized Users», 21 aprile 2026
- Bloomberg News, «Mythos: Why Anthropic's New AI Has Officials Worried», 10 aprile 2026 (Bessent-Powell summit)
- Financial Times, dichiarazioni Bryan Preston (CFO Fifth Third Bank), aprile 2026
- Bloomberg, «Anthropic's Mythos Adds Strain on Cybersecurity Teams», 17 aprile 2026
- NPR, «How AI is getting better at finding security holes», 11 aprile 2026
- TechCrunch, «Unauthorized group has gained access to Anthropic's Mythos», 21 aprile 2026
- VentureBeat, «Anthropic says its most powerful AI cyber model is too dangerous to release publicly», 7 aprile 2026
- ICT Security Magazine, «NIS2: la mappa completa degli adempimenti da qui a ottobre 2026», aprile 2026
- Diritto.it, «NIS2 ACN nuovo modello per attività e servizi», 27 aprile 2026

t-defence

Next | Donexit | Foramil | Innodesi

Via Giacomo Peroni, 452 – 00131 Roma
tel. 06.45752720 – info@defencetech.it
www.t-defence.it

#TDefenceBusiness