IL SOLE 24 ORE – 9 GIUGNO 2026

Con Mithos dirimente la governance del potere cognitivo

di Giuseppe F. Italiano

Per anni abbiamo immaginato l'intelligenza artificiale come una tecnologia destinata a diventare progressivamente più accessibile: prima confinata nei laboratori di ricerca, poi adottata dalle imprese e infine resa disponibile al grande pubblico.

Il rilascio controllato di Claude Mythos, il nuovo modello di frontiera sviluppato da Anthropic, sembra però indicare una traiettoria diversa: quella di sistemi così potenti da non poter essere distribuiti liberamente.

Le capacità di Mythos nell'analisi di software complessi alla ricerca di vulnerabilità segna un passaggio importante.

Nei test condotti da istituti indipendenti, il modello avrebbe ottenuto risultati significativamente superiori rispetto ad altri sistemi di frontiera, inclusi GPT-5.5 e Claude Opus 4.7, risolvendo con successo una quota molto elevata di sfide Capture the Flag, competizioni che simulano scenari realistici di attacchi cyber.

Le ricadute sul software commerciale sono state rapide. Nel giro di due sole settimane dalla prima *release* limitata, Mozilla ha annunciato di aver individuato e corretto ben 271 vulnerabilità di sicurezza all'interno di Firefox grazie all'utilizzo di Mythos, che non si è limitato a segnalare i bug: in 181 casi è riuscito a convertire autonomamente quelle vulnerabilità in exploit funzionanti, ottenendo in alcuni casi anche il controllo parziale dei sistemi.

Questo episodio è rilevante non solo per il numero di vulnerabilità individuate, ma soprattutto perché rende evidente la vera discontinuità introdotta da Mythos: la drastica riduzione del cosiddetto *Time To Exploit* (Tte), cioè il tempo che intercorre tra la scoperta di una vulnerabilità e la sua effettiva sfruttabilità. Fino ad oggi, trasformare una vulnerabilità in un exploit richiedeva settimane o mesi di lavoro da parte di team altamente specializzati.

Mythos comprime drasticamente questo intervallo temporale: operando in autonomia e a velocità impensabili per un essere umano, può ridurre il Tte a pochi minuti, collegando e concatenando più vulnerabilità tra loro per aggirare anche sistemi difensivi stratificati.

Per il settore finanziario e per le grandi imprese, questa accelerazione cambia la natura stessa del rischio cyber. Banche e istituzioni finanziarie operano su architetture digitali complesse e fortemente stratificate, nelle quali l'aggiornamento dei sistemi e l'applicazione di rimedi (patch) non possono avvenire in tempo reale.

Se il Time To Exploit si riduce a pochi minuti, il margine temporale tra la scoperta di una vulnerabilità e il suo possibile sfruttamento tende ad azzerarsi: una falla di sicurezza può trasformarsi rapidamente in un rischio sistemico, minacciando la continuità operativa di reti profondamente interconnesse prima ancora che possano essere attivate contromisure efficaci.

Non sorprende, dunque, che il tema abbia ormai superato il perimetro aziendale per entrare in quello della stabilità economica e finanziaria. Anche le principali istituzioni internazionali stanno monitorando con crescente attenzione l'impatto di sistemi di AI avanzata sulla sicurezza informatica: la Banca centrale europea ha già aperto un confronto con i vertici del sistema bancario europeo sui nuovi scenari di rischio emergenti. In effetti, negli ultimi anni l'Ue si è dotata di un apparato regolatorio per affrontare emergenze, come l'AI Act (con norme dedicate alla prevenzione del rischio sistemico) e il Digital operational resiliency act (Dora, per il settore finanziario), nonché di istituzioni dedicate a supervisionare tali rischi (come l'Enisa, agenzia per la cybersicurezza, che ha infine avuto accesso a Mythos).

Mythos solleva però una questione cruciale: quella della governance del potere cognitivo. Se le AI più evolute dovessero rimanere accessibili soltanto a una cerchia ristretta di attori selezionati - governi, grandi imprese od organizzazioni critiche - chi definirà i criteri di accesso? Stati sovrani, organismi di regolazione internazionale oppure un numero limitato di grandi aziende tecnologiche?

Per l'Europa, la questione supera ormai i tradizionali temi della sovranità digitale. Non riguarda più soltanto semiconduttori, cloud, data center o AI factories, ma l'accesso stesso a sistemi cognitivi avanzati. Il pericolo è quello di una nuova forma di dipendenza strategica: non più soltanto tecnologica, ma cognitiva. Mythos potrebbe così essere ricordato non tanto come un ulteriore avanzamento tecnologico, ma come il momento di rottura in cui è diventato evidente che la partita del futuro non si gioca più sulla diffusione dell'intelligenza artificiale, ma sul controllo delle sue capacità più avanzate.