26 Giugno 2026

L’AI è «a pochi mesi di distanza» dal rovesciare i governi: parlano le agenzie di Intelligence

I modelli avanzati di Intelligenza Artificiale potrebbero presto fornire agli hacker la capacità di paralizzare governi, aziende e sistemi critici, hanno messo in guardia le agenzie di sicurezza informatica di Five Eyes, l’unione internazionale dei Paei anglofoni per lo spionaggio.

In una rara dichiarazione congiunta diffusa lunedì, i vertici della sicurezza informatica di Australia, Stati Uniti, Gran Bretagna, Canada e Nuova Zelanda hanno sostenuto che i modelli di IA all'avanguardia si stanno evolvendo più velocemente del previsto e si prevede che «supereranno le attuali aspettative del settore, trasformando radicalmente le capacità di sicurezza informatica sia offensive che difensive».

«Non si tratta di anni, ma di mesi», hanno affermato le agenzie, aggiungendo che «il rischio informatico non può più essere trattato come una questione puramente tecnica. Si tratta di un rischio aziendale fondamentale e di una responsabilità della leadership».

Secondo il documento, l'AI contribuirà a potenziare le difese informatiche nel tempo, ma sta anche abbassando le barriere per gli attori malevoli, aumentando la velocità e la complessità degli attacchi e riducendo il tempo tra la scoperta e lo sfruttamento delle vulnerabilità.

Le agenzie hanno invitato le organizzazioni a rafforzare le proprie difese digitali, ad aggiornare più rapidamente i software obsoleti, a limitare l'accesso ai sistemi sensibili e a prepararsi agli attacchi informatici prima che si verifichino.

Sebbene la dichiarazione dei Five Eyes non abbia citato alcun modello o azienda specifica, il recente dibattito sulla sicurezza dell'IA si è concentrato sullo sviluppatore statunitense Anthropic, finito sotto esame per i suoi sistemi più recenti e avanzati.

All’inizio di quest’anno, l’azienda ha dichiarato che uno dei suoi modelli di punta, Mythos, era troppo potente per essere rilasciato al grande pubblico e ha limitato l’accesso a un piccolo gruppo di organizzazioni fidate. Successivamente, l’azienda ha introdotto Fable 5, una versione più restrittiva della tecnologia, ma entrambi i modelli sono stati poi ritirati dal mercato dopo che il governo degli Stati Uniti ha ordinato che ai cittadini stranieri fosse vietato utilizzarli, citando motivi di sicurezza nazionale.

Questi sviluppi si collocano nel contesto di avvertimenti più ampi da parte di ricercatori, leader tecnologici e funzionari della sicurezza, secondo i quali le capacità dell’AI si stanno evolvendo più rapidamente di quanto governi e istituzioni riescano ad adattarsi.

Gli esperti hanno sempre più spesso messo in guardia sul fatto che i sistemi progettati per aumentare la produttività e rafforzare le difese informatiche potrebbero essere utilizzati anche per automatizzare gli attacchi, abbassare le barriere per gli attori malevoli e amplificare l’impatto di piccoli gruppi.

Secondo una clamorosa indiscrezione riportata in questi giorni dalla rivista *The Economist*, il software Mythos avrebbe violato la National Security Agency (NSA), ossia l’agenzia di spionaggio informatico USA, nota per la sofisticazione dei suoi sistemi e la preparazione dei suoi hacker.

La testata ha riferito che il senatore Mark Warner ha svelato i dettagli di un briefing tenuto dal generale Joshua Rudd, capo della

NSA e del Cyber Command statunitense. Secondo quanto riportato, durante un'esercitazione di *red-teaming* (la pratica di testare rigorosamente le difese, i sistemi o le strategie di un'organizzazione adottando una prospettiva avversaria), Mythos è riuscito a penetrare in quasi tutti i sistemi classificati della NSA nel giro di poche ore, anziché settimane.

Antrophic avrebbe deciso di non distribuire pubblicamente il modello proprio a causa delle sue straordinarie capacità autonome di hacking e analisi dati, che includono anche la ricostruzione di tipi cellulari dal DNA grezzo e l'individuazione di vulnerabilità inedite nei principali sistemi operativi e browser.

L'affermazione sulla violazione dei sistemi NSA ha scatenato un acceso dibattito tra gli esperti di tecnologia e cybersicurezza, con molti osservatori che ritengono si sia trattato della forzatura di ambienti isolati o sistemi di prova in condizioni controllate, piuttosto che di un vero e proprio attacco riuscito alla rete centrale dell'agenzia.

Tale evento ha comunque segnato una svolta geopolitica decisiva, spingendo l'amministrazione Trump ad abbandonando l'approccio deregolamentato per imporre severi controlli sulle esportazioni dei modelli di IA di frontiera.

Secondo quanto riportato dal *New York Times*, in queste ore la NSA ha perso l'accesso al modello di IA Mythos 5 di Anthropic, che utilizzava per individuare vulnerabilità nei software. La vicenda si inserisce nel contesto della disputa, che dura da mesi, tra Washington e l'azienda della Silicon Valley.

Il blocco è scattato dopo che l'amministrazione Trump ha imposto restrizioni all'esportazione nei confronti di Anthropic all'inizio di questo mese, citando motivi di sicurezza nazionale, secondo quanto riportato dal New York Times.

La perdita ha «privato» l'agenzia di Intelligence di uno «strumento che ha impressionato e allarmato i suoi analisti per la sua efficacia nell'individuare le vulnerabilità del software», ha aggiunto la testata.

La tecnologia AI di Anthropic è stata sempre più impiegata su reti governative classificate e integrata nelle attività di sicurezza nazionale degli Stati Uniti, con i suoi modelli utilizzati per l'analisi dell'intelligence, la pianificazione operativa e le operazioni informatiche.

Tuttavia, a febbraio, il dipartimento della Guerra USA ha classificato Anthropic come «rischio per la catena di approvvigionamento» dopo che l'azienda si è rifiutata di rimuovere le restrizioni su alcune applicazioni militari dei suoi sistemi di intelligenza artificiale. L'azienda ha dichiarato di opporsi alla sorveglianza di massa sul territorio nazionale e alle armi completamente autonome. Il presidente degli Stati Uniti Donald Trump aveva quindi ordinato alle agenzie federali di eliminare gradualmente la tecnologia di Anthropic entro sei mesi.

Anthropic ha citato in giudizio il governo, sostenendo che le misure adottate costituivano una ritorsione illegale per il rifiuto di allentare le garanzie sull'utilizzo militare dell'IA.

Nonostante l'ordine di eliminazione graduale e la battaglia legale in corso, diverse testate giornalistiche hanno successivamente affermato che alcune componenti del governo statunitense continuano a utilizzare i sistemi Anthropic.

Questi sviluppi si verificano in un contesto di avvertimenti da parte di ricercatori, leader tecnologici e funzionari della sicurezza, secondo i quali i sistemi di AI vengono integrati nelle operazioni militari e di Intelligence a un ritmo più rapido di quanto governi e istituzioni riescano ad adattarsi alle loro crescenti capacità.

Gli esperti hanno avvertito che gli stessi strumenti utilizzati per rafforzare le difese informatiche potrebbero anche automatizzare gli attacchi e abbassare le barriere per gli attori malevoli.

Lo scontro emerso segue alle accuse secondo cui il modello di Intelligenza Artificiale dell'azienda sarebbe stato utilizzato durante l'operazione per rapire il presidente venezuelano Nicolas Maduro all'inizio di gennaio. Tuttavia, L'utilizzo dell'Intelligenza Artificiale nell'attacco statunitense a una scuola elementare femminile in Iran, che ha causato la morte di quasi 160 persone, per lo più bambini, non ha violato le «linee rosse» di Anthropic, ha dichiarato l'amministratore delegato Dario Amodei.

Si tratta dell'azienda coinvolta dal Vaticano per il lancio dell'enciclica di Leone XIV *Magnifica Humanitas*.

Come riportato da *Renovatio 21*, negli ultimi mesi vi è stato un progressivo deterioramento dei rapporti tra Anthropic e il

Pentagono, legato alla volontà del dipartimento della Guerra statunitense di utilizzare l'IA per il controllo di armi autonome senza le garanzie di sicurezza che l'azienda ha cercato di imporre.

L'Amodei, ha più volte espresso gravi preoccupazioni sui rischi della tecnologia che la sua azienda sta sviluppando e commercializzando. In un lungo saggio di quasi 20.000 parole pubblicato il mese scorso, ha avvertito che sistemi AI dotati di «potenza quasi inimmaginabile» sono «imminenti» e metteranno alla prova «la nostra identità come specie».

Amodei ha messo in guardia dai «rischi di autonomia», in cui l'IA potrebbe sfuggire al controllo e sopraffare l'umanità, e ha ipotizzato che la tecnologia potrebbe facilitare l'instaurazione di «una dittatura totalitaria globale» attraverso sorveglianza di massa basata sull'Intelligenza Artificiale e l'impiego di armi autonome.

Come riportato da *Renovatio 21*, l'anno passato l'Amodei ha dichiarato che l'AI potrebbe eliminare la metà di tutti i posti di lavoro impiegatizi di livello base entro i prossimi cinque anni.

Settimane fa Mrinank Sharma, fino a poco tempo fa responsabile del Safeguards Research Team presso l'azienda sviluppatrice del chatbot Claude, ha pubblicato su X la sua lettera di dimissioni, in cui scrive che «il mondo è in pericolo. E non solo per via dell'Intelligenza Artificiale o delle armi biologiche, ma a causa di un insieme di crisi interconnesse che si stanno verificando proprio ora».

Il Fondo Monetario Internazionale ha citato il recente rilascio controllato di Claude Mythos Preview da parte di Anthropic, descritto come «un modello di Intelligenza Artificiale avanzato con eccezionali capacità informatiche». Secondo il FMI, Mythos sarebbe in grado di individuare e sfruttare vulnerabilità in tutti i principali sistemi operativi e browser web, «anche se utilizzato da utenti non esperti».