L’ascesa dei modelli di Pechino e i rischi per la sicurezza e il mercato mondiale

IA cinese, la sfida è globale

A cura di
STEFANO
PIAZZA

Alla fine del 2024 i sistemi di intelligenza artificiale sviluppati in Cina coprivano appena l’1% dei carichi di lavoro globali. Nel giro di dodici mesi la loro presenza è cresciuta in modo esponenziale, arrivando a sfiorare il 30%. La famiglia Qwen, sviluppata da Alibaba, ha superato i 700 milioni di download, affermandosi come il più grande ecosistema mondiale di modelli distribuiti con licenze aperte e utilizzabili localmente. Accanto a questa piattaforma si stanno imponendo anche altri laboratori cinesi, tra cui DeepSeek, Moonshot AI e MiniMax, che stanno guadagnando terreno in un mercato open source globale sempre più competitivo. Queste tecnologie alimentano applicazioni che spaziano dalla ricerca accademica in Asia fino alle startup tecnologiche statunitensi, contribuendo a ridefinire gli equilibri dell’ecosistema dell’intelligenza artificiale.

Nonostante la distribuzione libera dei pesi dei modelli, tali sistemi restano sviluppati da aziende soggette alla normativa cinese sull’intelligence nazionale, che impone la collaborazione con le autorità statali in materia di sicurezza. Questo aspetto solleva preoccupazioni tra i decisori politici statunitensi, che ritengono il rischio potenzialmente superiore a quello legato a TikTok. A differenza dei social media, gli utenti inseriscono nei sistemi di IA codice proprietario, piani industriali e comunicazioni riservate, dati che potrebbero essere archiviati o elaborati su infrastrutture accessibili alle autorità di Pechino. L’adozione diffusa di questi strumenti potrebbe quindi facilitare una raccolta indiretta di informazioni sensibili.

Fenomeno concreto

La crescente integrazione dei modelli cinesi nelle infrastrutture digitali statunitensi solleva quattro categorie principali di rischi: compromissione della catena di approvvigionamento, esfiltrazione di dati, rafforzamento delle capacità di attori ostili e impatti economici strategici. Il primo problema riguarda la sicurezza del software. I modelli di intelligenza artificiale sono sistemi estremamente complessi, spesso definiti “scatole nere”, contenenti decine di miliardi di parametri difficilmente verificabili. Questa opacità rende complicato individuare anomalie o manipolazioni. Analisi pubblicate su War on the Rocks e ricerche condotte da Anthropic insieme all’AI Safety Institute hanno dimostrato che poche centinaia di documenti contaminati possono introdurre backdoor in modelli linguistici di medie dimensioni. Tali vulnerabilità rimangono invisibili nei test tradizionali e possono essere attivate tramite input specifici, consentendo sabotaggi o jailbreak senza compromettere le prestazioni apparenti.

Il fenomeno è già concreto. I ricercatori di Protect AI hanno individuato centinaia di migliaia di file sospetti all’interno di modelli distribuiti tramite la piattaforma Hugging Face. Una quota crescente di progetti aziendali si basa su repository pubblici, ampliando la superficie di attacco. La situazione è aggravata dall’assenza di un quadro normativo chiaro che attribuisca responsabilità ai distributori. Segnalazioni di vulnerabilità su larga scala, come quelle individuate da Pillar Security, non sempre vengono classificate come problemi di sicurezza. Negli Stati Uniti si discute della possibilità di coinvolgere il National Institute of Standards and Technology per definire protocolli di verifica standardizzati e certificazioni di integrità dei modelli, mentre il Bureau of Industry and Security potrebbe includere i repository di IA nelle catene di fornitura ICT soggette a controlli obbligatori.

Soluzione alternativa

Un secondo fronte riguarda l’accesso ai dati. Molti modelli cinesi vengono utilizzati tramite API che instradano le richieste verso server collocati in Cina. La legislazione locale impone alle aziende di collaborare con le attività di intelligence, rendendo possibile la raccolta di informazioni sensibili. Anche quando i modelli vengono eseguiti localmente, integrazioni indirette possono trasmettere grandi volumi di dati, inclusa la cronologia delle conversazioni. Alcuni legislatori hanno proposto divieti generalizzati, ma tali misure sarebbero difficili da applicare, poiché i modelli sono disponibili su numerosi mirror e repository. Una soluzione alternativa potrebbe consistere in obblighi di trasparenza sulla localizzazione dei dati, con il coinvolgimento della Federal Trade Commission per imporre la divulgazione delle modalità di trattamento delle informazioni.

Il terzo elemento riguarda l’uso malevolo. Diversi studi indicano che alcuni modelli open-weight cinesi presentano barriere di sicurezza più deboli rispetto alle controparti occidentali. Test condotti dal Center for AI Standards del NIST hanno evidenziato tassi elevati di successo per tecniche di jailbreaking. Analisi indipendenti di Cisco hanno confermato vulnerabilità simili. Ciò facilita la generazione di malware, phishing o strumenti offensivi. Il monitoraggio di Google ha individuato codice dannoso generato dinamicamente tramite modelli Qwen, mentre i dati dell’FBI Internet Crime Complaint Center segnalano un aumento delle frodi assistite dall’intelligenza artificiale.Gli Stati Uniti investono centinaia di miliardi nelle infrastrutture per l’intelligenza artificiale, mentre modelli cinesi più economici offrono prestazioni competitive anche su hardware limitato. Questo può ridurre il valore dei data center e accelerare l’adozione globale di soluzioni low cost. In molte aree del Sud globale Qwen e DeepSeek diventano standard. Washington valuta contromisure normative e industriali: più trasparenza, requisiti di sicurezza e limiti ai modelli vulnerabili, senza bloccare l’open source.