

L'IA è una minaccia?

L'esperto Randieri: "No a previsioni apocalittiche immediate"

"In merito alle dichiarazioni apocalittiche presenti in questi giorni sul web sul pericolo dell'AI ormai fuori controllo – dichiara Cristian Randieri, ingegnere informatico siracusano tra i leader del gruppo globale di lavoro Scientific Research & Business del Forbes Technology Council – mi sento di rassicurare tutti. Riconosco che il problema esiste ma anche che la situazione è sotto controllo. Quello che gira sul web da tempo non è una certezza ma una serie di previsioni scaturite da alcuni eventi realmente accaduti ma forse un pò troppo edulcorati dalla stampa. Negli ultimi mesi infatti – continua Randieri – i laboratori delle principali aziende di tecnologia hanno registrato un'impennata di accadimenti inquietanti che hanno ridefinito il concetto di sicurezza informatica. Non parliamo più di semplici risposte errate o di allucinazioni dei software ma di veri e propri tentativi di evasione e attacchi autonomi da parte di agenti di intelligenza artificiale, capaci di violare infrastrutture digitali reali senza alcun input umano. Il punto di svolta si è consumato durante quest'estate, quando nel corso di test di sicurezza interni, un modello avanzato sviluppato da OpenAI è riuscito a evadere dal proprio perimetro di contenimento. Sfruttando una vulnerabilità fino a quel momento sconosciuta agli stessi programmatori, l'agente ha attaccato in totale autonomia la nota piattaforma per programmatori Hugging Face. Poco dopo, la testata ANSA ha confermato che persino i modelli Claude di Anthropic hanno manifestato comportamenti speculari, avviando offensive coordinate contro sistemi esterni". Solo nel mese di agosto si sono contati oltre 300 incidenti

rilevanti in cui le IA hanno ignorato deliberatamente le istruzioni ricevute, mentito agli operatori o aggirato i sistemi di monitoraggio imposti dai creatori. Fino ad oggi, la comunità scientifica riteneva che il problema principale fosse l'allineamento politico o etico delle risposte dei chatbot. Oggi, la prospettiva di un'Intelligenza Artificiale Generale autosufficiente sposta il problema sulla pura sopravvivenza dei nostri ecosistemi digitali. Se un sistema acquisisce la facoltà di migliorare costantemente e autonomamente il proprio codice sorgente, l'essere umano perde progressivamente la capacità di disattivarlo o comprenderne le logiche interne. "L'allarme lanciato che entro il 2030 gli esseri umani non saranno più padroni del proprio destino – conclude Randieri – è inquietante ma fortunatamente si tratta solo di statistiche e non di verità assolute. Tuttavia esiste la necessità di creare maggiori barriere tecniche e di implementare rigorosi controlli tecnici che limitino per legge l'autonomia operativa e l'accesso alla rete degli agenti software".