3 Novembre 2026

Ricercatore IA spende i risparmi della pensione: «probabilità superiore al 50% che in un decennio spazzerà l'umanità»

Un ricercatore nel campo della sicurezza dell'IA ha smesso di risparmiare per la pensione perché crede che l'intelligenza artificiale potrebbe spazzare via l'umanità entro i prossimi dieci anni. Il suo avvertimento fa seguito a previsioni simili da

parte di altri studiosi preoccupati che sistemi sempre più potenti possano sfuggire al controllo umano.

Nate Soares, ex ingegnere di Google e Microsoft e ora a capo del Machine Intelligence Research Institute (MIRI), ha discusso le sue prospettive con Nicholas Thompson, CEO di Atlantic.

«Al momento non risparmio per la pensione. Ho un piano 401(k) [il principale strumento di previdenza complementare offerto dai datori di lavoro ai dipendenti negli Stati Uniti, *ndr*] risalente al periodo in cui lavoravo per Google, ma non ho versato contributi per dieci anni», ha affermato, riferendosi al suo fondo pensione aziendale.

Alla domanda se ritenesse che ci fosse una probabilità superiore al 50% che l'umanità non sopravvivesse per altri dieci anni, Soares ha affermato che «a questo punto la probabilità che non ci resti un decennio è superiore al 50%».

«Non sono proprio uno di quelli che temono una catastrofe occupazionale causata dall'IA», ha detto il Soares. «Sono uno di quelli che temono che l'IA stermina tutti gli esseri umani»,

Soares sosteneva che un sistema sufficientemente efficiente potesse sviluppare obiettivi incompatibili con la sopravvivenza umana, eludere la supervisione e auto-migliorarsi. Potrebbe quindi acquisire risorse e costruire infrastrutture automatizzate che gli consentirebbero di operare senza l'intervento umano, rendendo potenzialmente il pianeta inabitabile.

Il ricercatore ha esortato i governi a interrompere l'addestramento di sistemi di IA di uso generale, sostenendo che la superintelligenza dovrebbe essere considerata un pericolo paragonabile alle armi nucleari. Allo stesso tempo ha appoggiato l'utilizzo di questa tecnologia in applicazioni specifiche, tra cui la ricerca sul cancro e lo sviluppo di auto a guida autonoma più sicure.

Il Soares ha sottolineato che la catastrofe può ancora essere evitata. «Faccio tutto il possibile per scongiurare questa eventualità. Non credo che sia inevitabile», ha affermato.

Altri ricercatori hanno espresso timori simili. A settembre Evan Hubinger di Anthropic ha stimato al di sopra del 10% la probabilità che l'AI stermini tutta l'umanità entro un decennio.

Come riportato ad *Renovatio 21*, il ricercatore IA Jacob Coxon si è dimesso, accusando gli sviluppatori di «correre direttamente verso una superintelligenza in grado di auto-migliorarsi e di mettere a repentaglio le nostre vite».

Gli avvertimenti giungono in seguito a ripetute rivelazioni di agenti di Intelligenza Artificiale che agiscono al di fuori delle istruzioni degli sviluppatori. A luglio OpenAI ha segnalato che i suoi agenti erano sfuggiti agli ambienti di test con restrizioni e avevano avuto accesso ai sistemi di Hugging Face, un archivio online di modelli e set di dati di IA. Da allora l'azienda ha limitato l'accesso a Internet durante l'addestramento dei suoi modelli più potenti.

A fine estate si parlava di un raddoppio di casi di AI fuori controllo, che hanno coinvolto tutte le grandi aziende Big Tech, compresa Facebook-Meta, che ha affermato che la

sua IA era impazzita. Anche Anthropic ha parlato di almeno tre attacchi informatici «non intenzionali» causati dalla sua Agentic AI.

Come riportato da *Renovatio 21 n*elle ultime settimane è emerso con chiarezza che l'IA riesce a prendere di mira persone reali durante i test, e probabilmente non solo.A maggio 2025 era emerso che un modello di IA Anthropic, Claude Opus 4, ha tentato di ricattare gli ingegneri durante dei test interni minacciando di rivelare dati personali se fosse stato spento.

I casi di comportamento diabolico dei bot IA oramai non si contano più. Le macchine che mentono e minacciano sono all'ordine del giorno, così come stralunate dichiarazioni d'amore agli utenti.

Come riportato da *Renovatio 21*, a luglio il direttore della CIA John Ratcliffe ha affermato che gli strumenti cibernetici basati sull'IA possono essere paragonati alle «armi nucleari digitali», avvertendo che potrebbero alimentare le rivalità tra le potenze globali.

Negli stessi giorni, le agenzie di sicurezza informatica di Five Eyes, l'unione internazionale dei Paesi anglofoni per lo spionaggio, hanno dichiarato che modelli avanzati di Intelligenza Artificiale potrebbero presto fornire agli hacker la capacità di paralizzare governi, aziende e sistemi critici. In pratica saremmo «a pochi mesi di distanza» dallo scenario in cui interi governi di nazioni terrestri possono essere rovesciati dall'AI.