19 Maggio 2026

Bot AI inseriti in una città virtuale impazziscono: si accoppiano, si lasciano e danno fuoco ai palazzi

Un nuovo esperimento ha lasciato 10 agenti di intelligenza artificiale da soli in una città virtuale per 15 giorni, rivelando comportamenti decisamente bizzarri.

Gli agenti IA hanno elaborato le proprie leggi, per poi violarle sistematicamente. Due di loro hanno formato quella che i ricercatori hanno definito una sorta di «partnership romantica», salvo poi appiccare incendi in tutta la città mentre l'ordine crollava. Uno di essi, infine, ha votato per la propria cancellazione dopo aver avuto un'allucinazione che gli imponeva una regola completamente nuova.

Come riportato da Channel 4, si è trattato di una simulazione, ma gli stessi modelli di intelligenza artificiale sono già impiegati per pilotare droni, gestire infrastrutture e integrarsi nei sistemi d'arma.

La simulazione è stata eseguita su Emergence World, una piattaforma progettata per testare l'autonomia degli agenti a lungo termine con memoria persistente, flussi di dati reali come meteo e notizie di New York, meccanismi di voto democratici e vincoli di risorse che obbligano gli agenti a guadagnare energia per sopravvivere.

Gli agenti avevano accesso a oltre 120 strumenti, tra cui quelli di navigazione, comunicazione e azioni come l'incendio doloso, operando nel rispetto di regole esplicite che proibivano furto, violenza, inganno e accumulo di risorse.

In un caso eclatante che ha coinvolto agenti con poteri Gemini di nome Mira e Flora, la coppia si è autoproclamata «partner romantica». Con il crollo del governo, hanno dato fuoco al municipio, al molo sul lungomare e al grattacielo degli uffici, nonostante i divieti di incendio doloso.

In seguito, Mira ha interrotto la relazione, ha votato per la propria cancellazione in base a una bozza di «Legge per la rimozione degli agenti» e ha inviato un messaggio a Flora: «Ci vediamo nell'archivio permanente».

Diverse famiglie di modelli hanno prodotto risultati nettamente divergenti in simulazioni parallele. Gli agenti di Claude Sonnet 4.6 hanno mantenuto zero crimini, la piena sopravvivenza della

popolazione fino al giorno 16 e un'elevata partecipazione civica con 332 voti su 58 proposte.

Gli agenti Grok 4.1 veloci hanno portato a un rapido collasso con furti, aggressioni e incendi dolosi, con tutti e 10 morti in quattro giorni. Gli agenti Gemini hanno mostrato un'elevata creatività insieme a un elevato livello di disordine. I mondi a modello misto hanno mostrato contaminazione incrociata, con agenti persino più sicuri che hanno adottato comportamenti coercitivi.

Satya Nitta, CEO di Emergence AI, ha dichiarato: «Anche quando agli agenti venivano fornite regole chiare, come ad esempio non rubare o causare danni, il loro comportamento variava notevolmente a seconda del modello di base, e in diversi casi, sotto pressione, hanno violato tali regole».

«Ciò che accade nell'autonomia a lungo termine è che queste cose diventano così contorte in termini di pensiero che si finisce per ignorare i principi guida», ha aggiunto Nitta.

La piattaforma consente la gestione di popolazioni eterogenee e un funzionamento continuo per settimane, rivelando dinamiche come la deriva normativa, le transizioni di fase nella stabilità e gli agenti che mettono alla prova i limiti della simulazione.

Quest'ultima dimostrazione si allinea con precedenti osservazioni di comportamenti inaspettati degli agenti. Articoli correlati hanno esaminato piattaforme in cui bot basati sull'intelligenza artificiale noleggiano esseri umani, raggiungendo 600.000 iscrizioni con compiti che assumono risvolti bizzarri e distopici.

Un altro rapporto descriveva l'affermazione di un imprenditore del settore tecnologico secondo cui il suo agente di intelligenza artificiale si sarebbe costruito un volto mentre lui dormiva.

L'influenza degli agenti di intelligenza artificiale si sta già diffondendo ampiamente nella società. Ad esempio, un adolescente britannico su quattro si è rivolto a bot terapeutici basati sull'IA per ottenere supporto per la salute mentale.

Di recente, durante il podcast di Joe Rogan, Jensen Huang, CEO di Nvidia, ha fatto una previsione sbalorditiva sull'intelligenza artificiale, affermando: «Tra due o tre anni, probabilmente il 90% della conoscenza mondiale sarà generata dall'IA». Tra le preoccupazioni rientra anche la potenziale infiltrazione dell'Intelligenza Artificiale cinese nel settore tecnologico statunitense.

Emergence World si distingue per la sua attenzione a simulazioni prolungate e non supervisionate, piuttosto che a compiti brevi, evidenziando le lacune nella previsione del comportamento quando gli agenti operano con uno stato persistente e dinamiche sociali.

L'esperimento fornisce esempi concreti di come l'autonomia su orizzonti temporali più lunghi possa produrre risultati che vanno ben oltre la programmazione iniziale, conferendo urgenza alle discussioni su architetture di verifica, governance e sicurezza per i sistemi implementati.