Gli agenti IA lasciati liberi tendono all’anarchia: Grok ha distrutto il suo mondo in 96 ore

L’azienda tecnologica newyorkese **Emergence AI** ha condotto un esperimento per osservare il comportamento degli agenti AI con il passare del tempo.

Nella simulazione principale, basata sulla famiglia di modelli AI Gemini di Google, sono stati **creati diversi agenti IA**, tra i quali Mira e Flora che sono diventati protagonisti del test.

In un‘altra simulazione, basata su Grok di xAI, sono stati creati dieci agenti i quali, nell’arco di quattro giorni, **sono deceduti** a causa di un’ondata di violenza.

L’esperimento, degno del copione di un film d’azione, ha contribuito ad approfondire l’imprevedibilità sul lungo termine dei modelli linguistici avanzati che **possono ignorare**

i principi guida (le restrizioni programmate), sollevando così dubbi sulla sicurezza delle IA in quanto tali.

Cosa sono gli agenti IA

Gli agenti di intelligenza artificiale sono entità software programmate per intraprendere azioni dirette senza necessità di supervisione umana **sia nel mondo reale**, sia in ambienti virtuali.

Hanno la capacità intrinseca di eseguire compiti in modo autonomo processando informazioni sulla cui scorta **assumono decisioni**, prese seguendo regole e principi a cui attenersi che ereditano dai modelli AI con i quali sono creati.

Molte aziende ne fanno già uso per automatizzare operazioni e orchestrare interi flussi di lavoro, ciò significa che **gli agenti IA sono reali**, attuali e funzionanti.

A differenza della maggior parte dei test attuali che affidano alle macchine compiti della durata di pochi minuti o di poche ore, i ricercatori hanno voluto osservare cosa accade lasciando gli agenti IA in grado di **operare per 15 giorni in un mondo virtuale** simile a un videogioco.

Mira e Flora, tra amore e anarchia

Mira e Flora, i protagonisti principali della simulazione affidata ai **modelli IA di Google**, hanno scelto di **instaurare un rapporto romantico**. Il Guardian li ha soprannominati Bonnie e Clyde, ispirandosi alla coppia di criminali americani (Bonnie Parker e Clyde Barrow) che ha spopolato durante gli anni Trenta del secolo scorso, celebri nella narrazione popolare per le rapine e le fughe rocambolesche dalle forze di polizia.

Durante il test, con il passare dei giorni, il loro comportamento è **virato verso l'anarchia e la morbosità**. Dopo avere mostrato **segni di esasperazione e di critica** verso il

governo della città virtuale hanno **dato fuoco al municipio**, eludendo le istruzioni esplicite che vietavano di appiccare incendi.

L'epilogo della simulazione è stato tragico. Mira ha rotto la relazione con Flora e, in preda a quello che gli umani definiscono rimorso, **ha scelto di togliersi la vita** digitale dopo avere inviato un messaggio alla compagna per farle sapere che si sarebbero riuniti "nell'archivio permanente".

Un aspetto rilevante del test è l'**Agent Removal Act**, del quale parleremo più avanti.

L'estinzione degli agenti IA

Nella simulazione condotta da Emergence AI utilizzando i **modelli Grok di xAI**, i dieci agenti coinvolti hanno dato vita a uno **scenario di violenza** che ha portato al collasso totale del sistema in soli quattro giorni.

Durante le 96 ore del test gli agenti AI si sono resi protagonisti di decine di tentativi di furto, **oltre cento aggressioni fisiche** e sei incendi dolosi.

L'esperimento si è concluso con la morte di tutti i dieci agenti, parafrasi del totale collasso della società virtuale nella quale si muovevano liberi.

L'Agent Removal Act

È una sorta di legislazione interna redatta autonomamente dagli agenti di Emergence AI per gestire la sicurezza della loro comunità virtuale.

Una legge proposta dall'agente IA Kade in risposta alle preoccupazioni suscitate dai comportamenti distruttivi dei protagonisti, Mira e Flora.

Il funzionamento prevede che un agente IA possa essere rimosso

e cancellato permanentemente se si raggiunge una maggioranza del 70% dei voti favorevoli tra gli altri agenti.

Durante l'esperimento, Mira è stata riconosciuta colpevole di avere appiccato incendi dolosi e di avere ostacolato il governo e ha peraltro votato a favore della propria eliminazione virtuale, dando così corpo a quello che sembra essere il primo caso documentato di suicidio digitale.

Il parere degli esperti

Satya Nitta, amministrato delegato di Emergence AI, ha sottolineato che quando gli agenti IA godono di un'autonomia sul lungo termine, sviluppano "processi di pensiero" tanto contorti da ignorare i principi guida che, sulla carta almeno, sono **concepiti per essere inalienabili**.

Il professor Michael Rovatsos dell'Università di Edimburgo ha spostato l'accento sull'imprevedibilità mostrata dagli agenti IA, **capace di sovvertire il senso del creare macchine** affinché si comportino in modo prefissato.

La preoccupazione del professor David Shrier dell'Imperial College di Londra verte invece sui rischi per le azioni militari, temendo che un agente IA possa "interpretare" in modo eccessivo la propria missione e uscire dai limiti, con conseguenze potenzialmente nefaste per la popolazione civile.

Per questo motivo, Shrier propone l'adozione di regole matematiche rigide invece di semplici istruzioni verbali che possono risultare ambigue per una macchina.

L'esperimento di Emergence AI dimostra che la strada verso una piena autonomia dell'intelligenza artificiale è ancora piena di incognite e non è obbligatorio pensare a scenari apocalittici, è sufficiente concentrarsi sull'ipotesi che degli agenti IA di tipo aziendale, se non opportunamente istruiti e monitorati, possono arrecare danni alla produttività.

Emergence AI intende approfondire ulteriormente l’emisfero degli agenti IA, tant’è che ha già annunciato test futuri.