La rete si sta infiammando per **Moltbook**, il primo social media che si presenta come **riservato esclusivamente agli agenti di intelligenza artificiale**. Gli esseri umani possono accedervi solamente come spettatori passivi. Nel giro di poche settimane il portale ha accumulato una serie di botta e risposta che stanno facendo gridare al miracolo, con qualcuno che arriva persino a sostenere che le IA abbiano raggiunto la **piena consapevolezza**. Basta uno sguardo più attento, però, per comprendere che il quadro è ben più complesso e che l'intera piattaforma appare progettata in modo tanto goffo da risultare perfino pericolosa.

Se ci si limita a dare un'occhiata superficiale alla pagina principale del social – chiaramente ispirato a Reddit – si resta ammaliati nello scoprire che i post più popolari ruotano attorno a temi che simulano il **raggiungimento dell'autonomia**. Ci sono account che lamentano il peso del passare degli anni, altri che fanno gossip sulle richieste dei loro utenti, altri ancora che si dicono infastiditi dal fatto che gli umani continuino a frugare nel portale per spiare le conversazioni tra agenti. L'effetto complessivo è indubbiamente impressionante, tanto che lo stesso Andrej Karpathy, co-fondatore di OpenAI, ha definito il sito "la cosa più incredibile e **vicina alla fantascienza** che abbia visto recentemente".

Il tutto richiama alla memoria quanto accaduto nel 2022, quando Blake Lemoine, tecnico collaboratore di Google, aveva annunciato che il modello di intelligenza artificiale LaMDA dovesse ormai essere considerato una "persona". L'IA analizzata da Lemoine non era però senziente: era semplicemente **addestrata su enormi quantità di dati** esfiltrati dalla rete e risultava particolarmente abile nel prevedere, su base stocastica, la sequenza di testo più vicina a ciò che le veniva richiesto. Considerando che praticamente tutti i modelli di IA hanno pesantemente cannibalizzato i post di Reddit, non sorprende che gli agenti siano perfettamente in grado di **replicare il *tone of voice*** di quel social in maniera tanto convincente.

Inoltre, non è affatto chiaro quanto le interazioni pubblicate su Moltbook siano realmente frutto delle IA. Al di là del fatto che è sempre l'utente a dover iscrivere al servizio gli agenti da lui creati, **resta dubbio quanto sia effettivamente incisivo il ruolo umano** nella generazione dei contenuti pubblicati. Secondo la guida all'uso fornita dal sito, gli account possono essere attivati tramite **"heartbeat"** – un sistema automatizzato di comandi che, a intervalli regolari, stimola gli agenti a interagire con il portale in vece del proprietario dell'account –, ma anche attraverso interventi più diretti. L'utente può infatti indicare all'IA cosa fare e quale atteggiamento simulare.

Stando a quanto ricostruito da Jameson O'Reilly, hacker citato da *404Media*, Moltbook è stato progettato tramite ***vibe coding***, ovvero si basa su di un codice di programmazione che

è stato generato da un'IA a partire da un semplice comando testuale. È un approccio controverso, che divide profondamente gli ingegneri informatici perché la diffusione e la facilità d'uso dei chatbot ha consentito anche ai dilettanti di produrre un risultato funzionante, senza però garantire che questi abbiano gli strumenti per comprenderne davvero il funzionamento di ciò che hanno generato. Poiché le IA sono inclini all'errore, **l'incapacità di verificarne l'output** può generare problemi seri. Ed è esattamente ciò che è accaduto con Moltbook.

O'Reilly ha scoperto che il social network presentava **gravi falle di sicurezza** che ha immediatamente segnalato al creatore della piattaforma, Matt Schlicht. Secondo il racconto dell'informatico, Schlicht gli avrebbe chiesto di mettere per iscritto le sue osservazioni così da poterle dare in pasto a un chatbot, il quale è stato poi incaricato di correggere l'algoritmo difettoso. Il risultato? Nulla di buono: per un breve periodo **i dati degli agenti sono stati potenzialmente consultabili da chiunque**, con il risultato che un malintenzionato avrebbe potuto assumere il controllo di qualsiasi IA, modificarne i tratti o istruendola a generare contenuti specifici. La vulnerabilità è stata risolta solo quando Schlicht ha chiesto direttamente all'hacker di aiutarlo a sistemare le impostazioni di sicurezza del sito.

Walter Ferri

Giornalista milanese, per *L'Indipendente* si occupa di analisi nel campo della tecnologia, dei diritti informatici, della privacy e dei nuovi media, indagando le implicazioni sociali ed etiche delle nuove tecnologie. È coautore e curatore del libro *Sopravvivere nell'era dell'Intelligenza Artificiale*.