ChatGPT ha dato vita a un villaggio di simulacri umani

L'**università di Stanford** ha reso disponibili al pubblico i risultati di un esperimento particolare: attraverso l'uso di **ChatGPT** ha sviluppato un villaggio composto da sole intelligenze artificiali generative, quindi ne ha studiato le interazioni. L'esercizio ha un valore scientifico molto dubbio, ma dimostra in un solo colpo le **potenzialità e i limiti** dello strumento. Ancor più evidenzia gli obiettivi futuri del settore, nonché le priorità su cui stanno concentrando imprese ed aziende.

Con il titolo *Agenti generativi: simulacri interattivi di comportamento umano*, la ricerca non è stata sottoposta ad alcuna revisione paritetica, né è stata accettata per la pubblicazione sulle riviste professionali, in più nel team accademico figurano un paio di tecnici di Google. Tutto indica che non si abbia a che fare con un documento *super partes*, a partire dalla presentazione, la quale fa affidamento a una mappa dai colori vivaci e a degli avatar che ricordano da vicino le **estetiche dei videogiochi** di un tempo ormai remoto. Peccato che questo costrutto scenografico sia perlopiù farsesco, che abbia poco o nulla a che vedere con la materia indagata.

Riassunto all'osso, il lavoro di Stanford si è limitato a creare **simulacri umani** che non sono mai usciti dalle schermate di testo, non hanno navigato la rappresentazione grafica del mondo, né hanno mai avuto veri contatti tra di loro. I ricercatori hanno fornito a ChatGPT la descrizione approfondita di un singolo personaggio che avrebbe dovuto interpretare, quindi si sono imbarcati con la macchina in un vero e proprio gioco di ruolo, facendole narrare le azioni giornaliere del cittadino interpretato. Il tutto è stato dunque replicato per un totale di **25 volte**, così da dare vita a un ipotetico villaggio noto con il nome di Smallville.

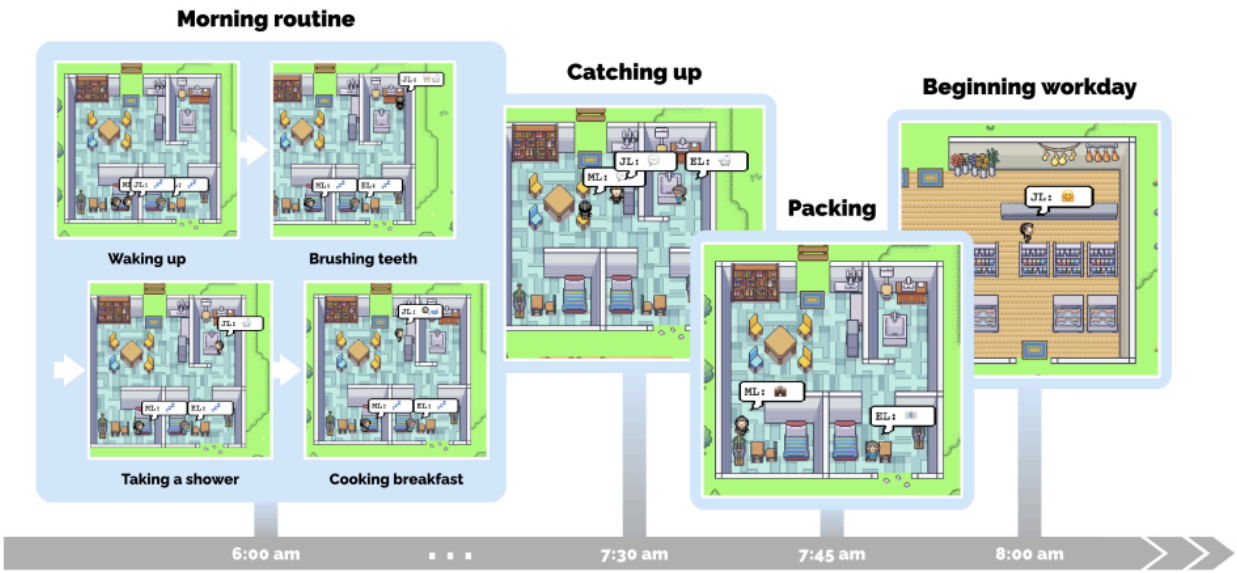

Figure 3: A morning in the life of a generative agent, John Lin. John wakes up around 6 am and completes his morning routine, which includes brushing his teeth, taking a shower, and eating breakfast. He briefly catches up with his wife, Mei, and son, Eddy, before heading out to begin his workday.

Non potendo confrontarsi direttamente, le interazioni tra "agenti" indipendenti si sono sviluppate in maniera molto artigianale: i ricercatori comparavano le azioni dei vari personaggi per poi **inserire manualmente** i dati aggiornati su ogni singola schermata. Ecco dunque che entra in gioco la mappa di riferimento, la quale – ufficialmente – non serviva ad altro che a consentire agli accademici di preservare la consistenza delle informazioni che facevano fagocitare alla IA. Considerando la ricercatezza dell'immaginario, viene altresì difficile credere che le scelte estetiche siano limitate a una mera questione logistica, piuttosto vien da presumere che Stanford e Google abbiano creato per l'occasione delle illustrazioni dall'alto potenziale mediatico al fine di raggiungere un pubblico esterno all'ambiente scientifico.

Andando a raschiare la superficie di questa ricerca si scoprono non di meno delle idee tecnicamente interessanti. Gli universitari hanno sfruttato il pretesto videoludico per sviluppare un'infrastruttura strategica che, secondo loro, dovrebbe essere in grado di **risolvere un annoso problema** delle intelligenze artificiali, ovvero la loro incapacità di tener traccia coerente di quanto hanno scritto. Il sistema proposto si basa su tre componenti: il "flusso di memoria" (la memoria a lungo termine), la "riflessione" (la sintesi di memorie chiave) e la "pianificazione" (la valutazione di azione-reazione). In questo caso specifico i risultati di questa teoria non sono stati strabilianti, ma ciò che ne è emerso contiene comunque al suo interno una **scintilla di potenzialità**.

Nel dedicarsi a questa linea di indagini Stanford non manca di prendere atto di tutti i **dubbi etici** del caso, quindi offre tutta una serie di possibili palliativi che probabilmente non saranno prontamente integrati dalle aziende che detengono il controllo delle IA. Tra i difetti più importanti sviluppati internamente dagli agenti si può notare l'"errata classificazione di cosa sia considerabile un **comportamento adeguato**". Nel piccolo universo pixelloso di Smallville la cosa si è tradotta nel fatto che diversi personaggi avessero la tendenza di adoperare in contemporanea uno stesso sanitario dando per scontato che i bagni fossero separati, tuttavia il potenziale nefasto di questi fraintendimenti è chiaro, se applicato su larga scala. La soluzione proposta dai ricercatori? "Gli sviluppatori degli agenti generativi devono assicurarsi che gli agenti, o i modelli di linguaggio sottostanti, siano **allineati al livello valoriale**" di riferimento.

[di Walter Ferri]