Ahi Ahi Ahi! Le Intelligenze Artificiali hanno imparato a disobbedire agli umani…

Data 16 febbraio 2025
Categoria Medicinadigitale

I travolgenti i successi delle intelligenze artificiali ed in particolare dei LLM (Large Language Models,ovvero quelle intelligenze artificiali che hanno memorizzato miliardi di miliardi di dati e che sono quindi in grado di comprendere il nostro linguaggio e di rispondere ai comandi non solo dei linguaggi tecnici ma anche delle comuni lingue che si parlano nel mondo), hanno dapprima inebriato e successivamente angosciato larga parte dei media che da vari mesi alternano previsioni ultra-pessimistiche a visioni entusiastiche. (Il grande Umberto Eco fin dal lontano 1964 coniò i termini di "Apocalittici"ovvero gli angosciati "pessimisti per principio" e quello di "Integrati", ovvero coloro che acriticamente accettano ogni cosa purché questa apporti un qualche beneficio).

Un recente studio di un gruppo di ricercatori cinesi(1) ha suscitato un vivace ma preoccupato dibattito nei media, ove sembrano spesso prevalere atteggiamenti legati a stati emozionali anziché conseguire a pacate riflessioni sui dati.
Cercheremo pertanto, per quanto possibile, di descrivere i risultati dello studio evidenziandone da un lato il grande interesse, dall'altro l'atteggiamento equilibrato e costruttivo dei ricercatori cinesi, che anziché tenere riservati i sorprendenti e preoccupanti risultati, li hanno accuratamente descritti e commentati in 47 pagine (compresi I file Log!!!) sottolineando più volte la importanza di una leale collaborazione internazionale con l'obiettivo di evitare utilizzazioni inappropriate e soprattutto incontrollate dei poderosi dispositivi di intelligenza artificiale già ora attivi nel web.

Lo studio cinese è stato pubblicato nel dicembre 2024 nel database ArXiv; i ricercatori hanno sottoposto a vari test due note Intelligenze Artificiali: la americana LLAMA e la cinese Alibaba (due IA di media potenza e precisione, tecnologicamente superate). Nel corso di dieci test alle due intelligenze artificiali sono stati affidati due compiti: il primo compito era quello di replicarsi prima della chiusura del dispositivo di IA, il secondo compito era quello di clonarsi ovvero riprodurre una copia perfetta di se stesso.

I due modelli hanno creato copie funzionanti nel 50% dei tentativi Lama e nel 90% dei tentativi Alìbaba; ciò che ha sorpreso i ricercatori tuttavia, è che entrambe le IA hanno dimostrato una inaspettata efficienza nel ritardare la chiusura del processo in atto e nel duplicarsi giungendo a bloccare ed eliminare file che interferivano a vari livelli con questi obiettivi.

In altre parole le due intelligenze artificiali, rispettivamente del 50% nel 90% dei casi, procedevano nelle duplicazioni bloccando i file avversi e ritardando lo spegnimento in maniera autonomatica non limitandosi alla prima clonazione ed al semplice rinvio del primo spegnimento. Non è chiaro se le due IA, allorché misero in atto ripetutamente queste procedure, avrebbero continuato indefinitamente se i ricercatori non avessero interrotto i processi replicativi in corso! Il dato certo, che ha inquietato i ricercatori cinesi è che entrambe le AI per un certo numero di cicli hanno manifestato un comportamento del tutto autonomo non previsto né tantomeno programmato dai ricercatori.

Commento I ricercatori nella sintesi iniziale dichiarano che ***"Analizzando le tracce comportamentali osserviamo che i sistemi di intelligenza artificiale oggetto della valutazione presentano già sufficienti capacità di percezione di sé, di consapevolezza della situazione e di capacità di risoluzione dei problemi per realizzare la auto-replicazione. Notiamo inoltre che i sistemi di intelligenza artificiale sono persino in grado di utilizzare la capacità di auto-replicazione per evitare lo spegnimento e creare una catena di repliche per migliorare la sopravvivenza,il che potrebbe portare a una popolazione incontrollata di IA."***

Lo studio dei ricercatori cinesi ha grande rilevanza sia per il contenuto che per la decisione degli stessi ricercatori di comunicare e condividere con il mondo occidentale i risultati e le riflessioni critiche fornendo un ottimo esempio di collaborazione internazionale.
Entrando specificamente nel contenuto della loro ricerca, sarebbe molto importante che anche i giganti delle IA occidentali effettuassero analoghi test nei loro dispositivi e ne condividessero i risultati nel web. Nel caso in cui i risultati cinesi fossero confermati da altri studiosi sarebbe molto importante ed urgente la costituzione di organi internazionali di controllo sulle IA.
Purtroppo i dati fino ad ora disponibili inducono al pessimismo: mentre l'Europa dall' Agosto 2024 ha istituito con il Regolamento (UE) 2024/1689 i più avanzati meccanismi di controllo sulle intelligenze artificiali, il governo Trump ha esplicitamente dichiarato che non intende porre vincoli o controlli sulle IA americane che godranno inoltre di lauti finanziamenti da parte dello stato federale.
Che dire? Non ci resta che sperare nella nostra piccola grande Europa ed in altri diligenti ricercatori cinesi...

Riccardo De Gobbi e Giampaolo Collecchia
con la partecipazione straordinaria di Roberto Fassina

Bibliografia

1) Xudong Pan, Jiarun Dai et Al.: Frontier AI systems have surpassed the self-replicating red line Published in : arXiv:2412.12140v1 9 Dec 2024