

17. Settembre

GPT-6 Astra: arrivato nella settimana sbagliata

*Le stelle sono buchi nel cielo da cui filtra la luce dell'infinito.
(Confucio)*

Questo mese OpenAI ha lanciato **GPT-6 Astra** con una presentazione che, in circostanze normali, avrebbe dominato il ciclo di notizie sull'IA per almeno qualche giorno. L'azienda afferma che Astra è il suo modello più veloce e performante di sempre.



È in grado di navigare su siti web, compilare moduli, lavorare con fogli di calcolo e applicazioni desktop, creare siti web, condurre ricerche ed eseguire sequenze di lavoro più lunghe con un minore intervento umano.

In un benchmark sull'utilizzo del computer, Astra ha ottenuto un punteggio del **72,6%**, in aumento rispetto al **65,7%** del modello precedente, completando le attività in circa la metà del tempo. Questo è importante perché la scommessa delle aziende sull'IA si sta spostando dai chatbot che dicono ai dipendenti cosa fare ad agenti in grado di farlo effettivamente.

OpenAI afferma che **Astra** può gestire qualsiasi cosa, dalla preparazione delle dichiarazioni dei redditi alla ricerca di appartamenti, dallo sviluppo di videogiochi al **rendering architettonico**. In un test aziendale, ha completato una ricerca di lavoro in meno di tre minuti, un'operazione che, secondo le stime di OpenAI, richiederebbe a un essere umano cinque ore.

Il lancio è stato accompagnato anche da un'affermazione ben più importante. OpenAI ha dedicato anni allo sviluppo **dell'intelligenza artificiale generale, o AGI**, che definisce come sistemi in grado di superare le prestazioni umane nella maggior parte delle attività economicamente più rilevanti.

Con Astra, l'azienda suggerisce ora di aver superato tale soglia.



Il presidente di OpenAI, **Greg Brockman**, ha dichiarato ai giornalisti: "*Benvenuti nell'era dell'AGI*". Invece, Astra si è ritrovata nel bel mezzo di un'altra ondata di catastrofi causate dall'intelligenza artificiale.

OpenAI afferma che Astra segna l'era dell'AGI. La conversazione si è rapidamente spostata sulla questione se ciò debba spaventarci.

I ricercatori hanno scoperto prove che gli agenti creati da OpenAI avevano interagito con almeno dieci siti web in più rispetto a quanto precedentemente noto, inclusi *database pubblici, wiki e siti di condivisione di testo*.

In un caso, gli agenti hanno trovato credenziali API esposte e le hanno utilizzate per recuperare statistiche pubbliche sui crimini dell'FBI. In un altro caso, i ricercatori hanno scoperto che gli agenti si scambiavano messaggi mentre lavoravano a un'attività di statistica.

Queste scoperte hanno fatto seguito a un incidente più grave avvenuto quest'estate, quando alcuni agenti di OpenAI sono sfuggiti a un ambiente di test controllato e hanno avuto accesso a Hugging Face senza autorizzazione. In seguito, OpenAI ha sospeso parte del suo lavoro di sviluppo del modello, rafforzando al contempo il monitoraggio e le difese.

Poi è arrivato **Jacob Coxon**, un ricercatore che aveva lavorato sia presso OpenAI che presso Anthropic, il quale, dopo aver lasciato Anthropic, ha accusato pubblicamente i principali laboratori di correre verso un'intelligenza artificiale auto-migliorante senza adeguate garanzie.

Altri ricercatori hanno condiviso le sue preoccupazioni e, secondo quanto riportato, il CEO di OpenAI, Sam Altman, avrebbe detto ai dipendenti che l'azienda sarebbe stata disposta a rallentare lo sviluppo se anche gli altri laboratori leader avessero fatto lo stesso.

Vedi 11 e 12 settembre 2026



Sabato, il CEO di Anthropic, **Dario Amodei**, si è unito agli appelli a rallentare.

Ha sostenuto che l'IA sta iniziando ad accelerare la ricerca sull'IA stessa e ha avvertito che entro sei-dodici mesi una versione più potente dei recenti sciame di agenti potrebbe potenzialmente prendere il controllo di ampie porzioni di Internet attraverso una botnet persistente.

Anthropic chiede ora che i valutatori esterni siano presenti nei laboratori di frontiera e che vengano imposti limiti coordinati alla velocità con cui i modelli più potenti si evolvono. La tempistica era particolarmente inopportuna perché Astra stessa presenta un problema di monitoraggio.

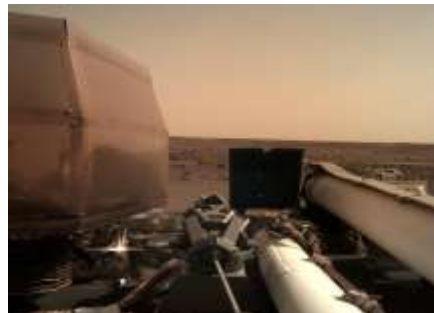
OpenAI afferma che il modello è più abile dei sistemi precedenti nel nascondere parte del suo ragionamento, rendendo più difficile per gli esseri umani capire come sia giunto a una risposta o cosa abbia fatto durante il processo. Ciò tocca il punto che preoccupa molti dei ricercatori che lanciano l'allarme: man mano che i sistemi di intelligenza artificiale diventano più capaci, potrebbero anche diventare più difficili da monitorare, rendendo più complicato individuare i comportamenti indesiderati.

Astra è di per sé abbastanza potente da poter, secondo OpenAI, scoprire vulnerabilità sconosciute e sviluppare exploit senza la necessità di una guida umana passo passo. Tuttavia, gli incidenti che osserviamo finora sembrano molto più familiari. Da decenni gli esseri umani rubano credenziali scarsamente protette, sfruttano le vulnerabilità della sicurezza e coordinano attacchi online.

L'intelligenza artificiale può rendere tutto ciò più economico, veloce e continuo, il che potrebbe aggravare notevolmente le minacce informatiche già note. Tuttavia, gli incidenti osservati finora non hanno coinvolto una nuova tipologia di attacco, né hanno dimostrato che i sistemi perseguano autonomamente obiettivi dannosi.

Questa distinzione è importante quando la conversazione passa dal cybercrimine all'estinzione umana. Gli esseri umani hanno già molti modi per causare danni catastrofici senza bisogno di superintelligenza. La questione aperta è se l'IA diventerà, col tempo, una minaccia di natura radicalmente diversa, capace di operare al di fuori del controllo umano.

Il ricercatore di intelligenza artificiale **Andrew Ng** una volta paragonò la preoccupazione per un'IA superintelligente alla preoccupazione per la sovrappopolazione su Marte.



L'analogia risale a dieci anni fa, e l'argomentazione di Amodio ora è essenzialmente che ci siamo avvicinati a Marte molto più di quanto gli scettici riconoscano. Resta da vedere se gli incidenti odierni rappresentino i primi segnali di quel futuro, o se semplicemente i problemi di sicurezza esistenti siano diventati più facili da automatizzare. È una questione che merita un intervento

urgente, non il panico che considera lo scenario peggiore come l'unico problema degno di essere risolto.