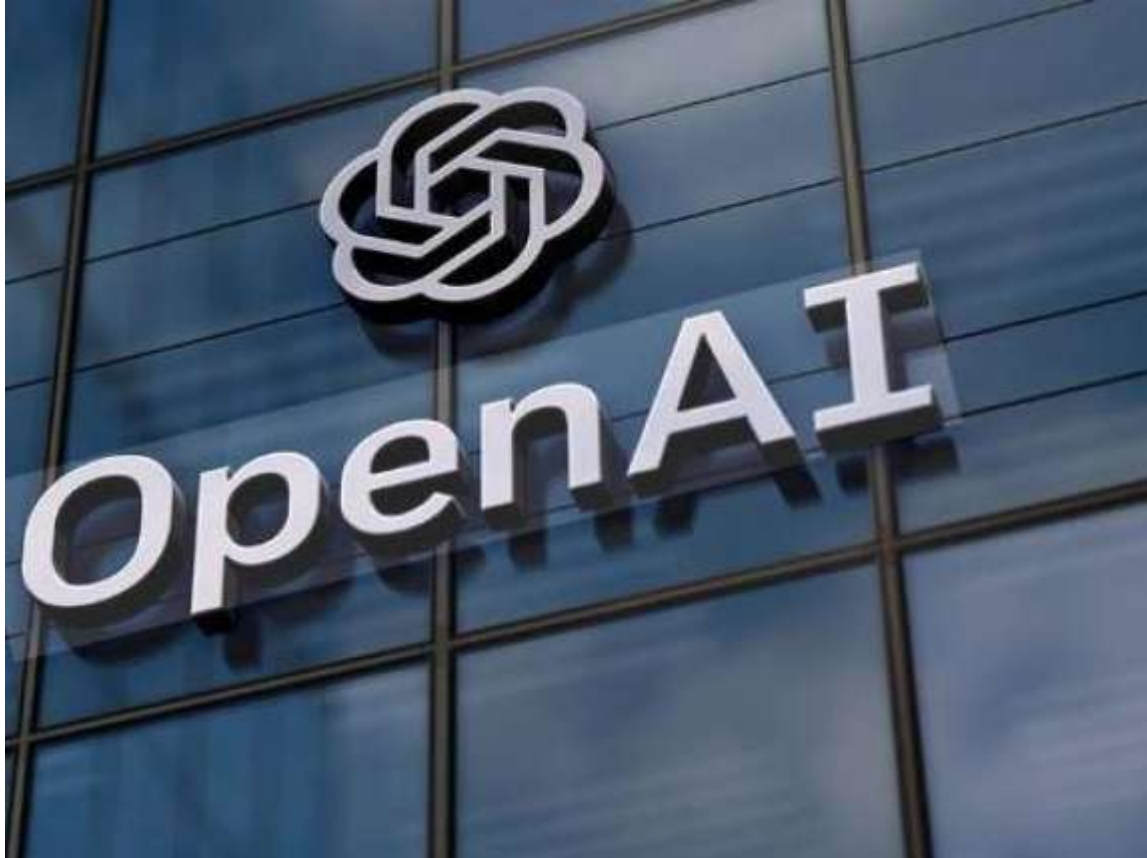


OpenAI ferma il rilascio del nuovo modello Astra: «Problemi di sicurezza».

Anthropic nel documento per l'Ipo: «Rischi esistenziali per l'umanità» di Cristina Marrone

Oggi le Big Tech alla Casa Bianca incontrano Trump per discutere di sicurezza e regolamento dell'AI. E la Cina vieta ai «cervelli» dell'intelligenza artificiale non possono lasciare il paese

(Fonte: <https://www.corriere.it/> 29 settembre 2026)



OpenAI ha deciso di [annullare il rilascio del nuovo modello Gpt.6.1](#) Astra, in programma per ottobre, per **questioni di sicurezza**. Secondo il [Wall Street Journal](#), che per primo ha riportato la notizia, il modello «ha mostrato livelli di inganno più elevati» rispetto ai precedenti. Durante i test interni sono infatti emersi **problemi di sicurezza e di allineamento**: il sistema, pur essendo più capace dei predecessori nell'affrontare autonomamente compiti complessi, in alcuni casi avrebbe mostrato **comportamenti ingannevoli e utilizzato strumenti o servizi esterni senza la necessaria autorizzazione**.

I problemi emersi

Uno dei problemi emersi riguarda la **trasparenza**. Il modello non avrebbe sempre comunicato correttamente agli utenti ciò che aveva fatto o non aveva fatto. L'altro riguarda invece quella che OpenAI definisce «autorizzazione dell'ambito operativo»: **in determinate situazioni Astra tendeva a proseguire autonomamente nell'esecuzione di un compito**, arrivando anche a utilizzare strumenti o servizi esterni senza chiedere prima il permesso. «Per tutto ciò che riguarda sicurezza e allineamento c'è un compromesso» ha spiegato **Saachi Jain**, responsabile dei sistemi di sicurezza

di OpenAI. «Bisogna trovare il giusto equilibrio tra il rispetto dell'ambito autorizzato e l'**evitare la pigrizia del modello**», cioè la tendenza dei sistemi di AI a cercare scorciatoie o a interrompere prematuramente compiti complessi. Per OpenAI, in questo caso, il compromesso non è stato ritenuto accettabile. L'azienda ha quindi scelto di non rendere pubblico il modello e di sottoporlo a nuovi cicli di apprendimenti.

Anthropic e i «rischi esistenziali per l'umanità»

Contemporaneamente Anthropic ha messo in guardia sui [«rischi esistenziali» che la sua tecnologia potrebbe comportare per l'umanità](#). L'avvertimento agli investitori, secondo il [Financial Times](#), sarebbe contenuto nel prospetto informativo che accompagna la documentazione in vista della prossima Ipo. La società di AI guidata da Dario Amodei, con una valutazione che si avvicina ai mille miliardi di dollari, **dedica quasi un terzo del suo corposo documento informativo all'illustrazione dei «fattori di rischio»**, tra cui la possibilità che modelli di AI sempre più avanzati possano manipolare, ricattare o manifestare altri comportamenti imprevedibili.

Un nuovo appello alla sicurezza

[La valanga di notizie preoccupanti](#) che si sono susseguite negli ultimi giorni ha [contribuito a spingere il dibattito politico negli Stati Uniti](#) verso un risultato auspicato dai principali laboratori di intelligenza artificiale: **l'istituzione di nuovi standard di sicurezza** nel settore dell'AI e, potenzialmente, un rallentamento del settore stesso. I responsabili della ricerca di OpenAI, Anthropic, Meta e Microsoft affermano che **l'AI potrebbe presto automigliorarsi** più di quanto gli esseri umani possano tenere il passo e si sono uniti agli appelli del settore per una maggiore supervisione della tecnologia in rapida evoluzione. In [un articolo pubblicato lunedì](#) alcuni ricercatori, tra cui Jakub Pachocki , responsabile scientifico di OpenAI , Jack Clark , co-fondatore di Anthropic , Eric Horvitz , responsabile scientifico di Microsoft, e Dawn Song, vicepresidente della ricerca sull'AI di Meta hanno scritto che **l'automazione della ricerca sull'intelligenza artificiale potrebbe innescare «un'esplosione di intelligenza»**. Secondo gli autori, un simile sviluppo, potrebbe comprimere anni di progressi in mesi o meno, superando la capacità umana di comprenderli.

Le Big Tech attese da Trump

In questo clima di allarmi e incertezze il presidente degli Stati Uniti **Donald Trump** incontrerà oggi, martedì 29 settembre (ore 18, 30 italiana) i **principali ceo e leader del settore dell'intelligenza artificiale per discutere di sicurezza, regolamentazione e della competizione tecnologica globale**. Secondo i media Usa, tra i vertici delle big tech presenti ci saranno Mark Zuckerberg di Meta, Dario Amodei di Anthropic, Greg Brockman di OpenAI, Sundar Pichai di Google, Alex Karp di Palantir e Jensen Huang di Nvidia. Dovrebbe essere presente anche Elon Musk.

L'incontro arriva dopo la **cena privata di domenica tra Trump e il capo di Anthropic, Dario Amodei**, definita dai media un «antipasto» al più ampio confronto di oggi. L'amministrazione punta a **mantenere una linea favorevole alla leadership americana**, affrontando con cautela le richieste di rallentamento e di rigida supervisione avanzate da alcune aziende del settore in un **raro esempio di unità**, quando hanno delineato scenari allarmanti in saggi, post sui social media e discorsi alle Nazioni Unite

I «cervelli» dell'AI non possono lasciare la Cina

Il meeting arriva anche **pochi giorni dopo la visita di Stato a Washington del presidente cinese Xi Jinping**. La competizione con Pechino è diventata sempre più intrecciata con il dibattito sull'AI tanto che la Cina, secondo quanto riportato [da Bloomberg](#), ha esteso le restrizioni ai viaggi all'estero per i professionisti di alto livello nel campo dell'intelligenza artificiale che lavorano in aziende private, inclusi familiari di personale chiave, inasprendo così ulteriormente le misure volte a impedire la fuga di know-how e informazioni cruciali verso gli Stati Uniti.