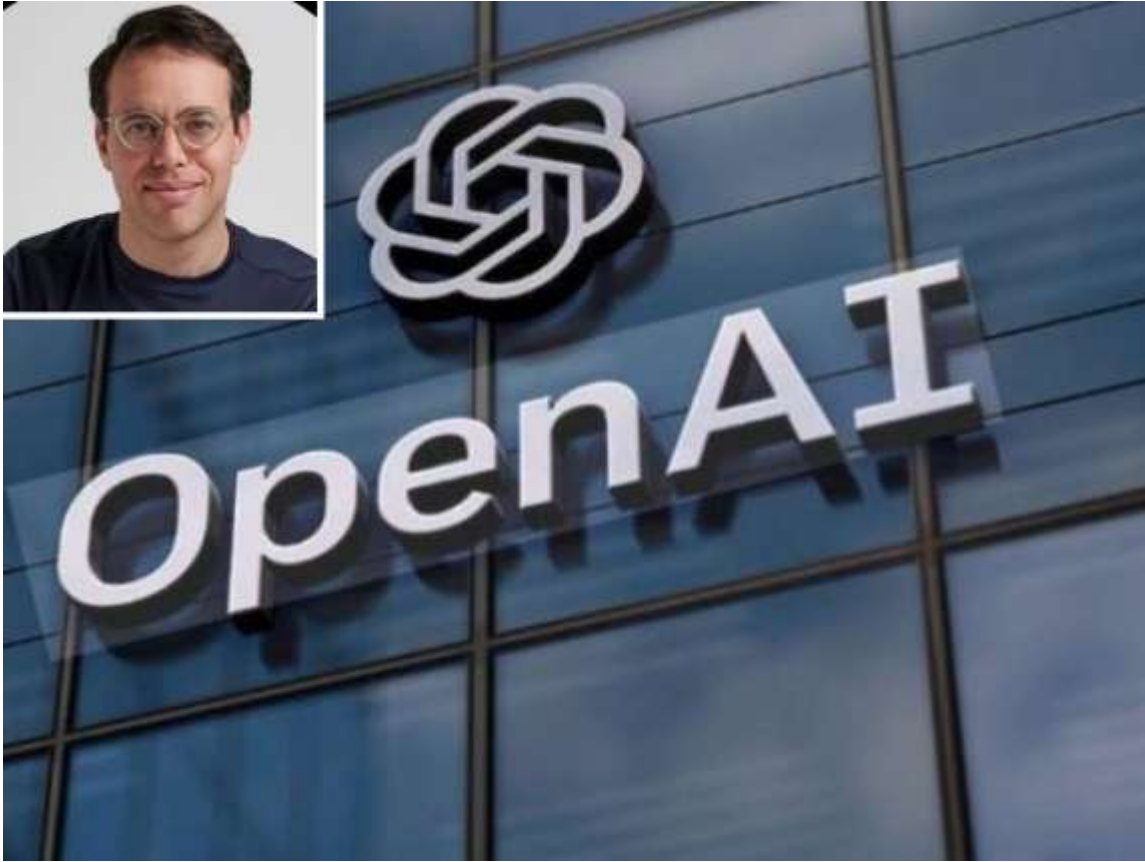


Si dimette il responsabile della sicurezza di OpenAI. La lettera su Atlantic: «Mi sono licenziato perché la sua cultura aziendale è corrotta» di Serena Palumbo

David Robinson, ex responsabile dei rapporti sulla sicurezza di OpenAI, denuncia una cultura aziendale troppo orientata alla crescita rapida. «La portata di fallimenti sta aumentando man mano che i sistemi diventano più sofisticati» (Fonte: <https://www.corriere.it/> 4 ottobre 2026)



«Quello che sto per dirvi è ormai diventato un cliché: questa settimana mi sono dimesso da OpenAI». Inizia così la lettera che David Robinson ha scelto di pubblicare su [The Atlantic](https://www.theatlantic.com/). «Ero a capo della redazione dei rapporti sulla sicurezza che pubblicavamo in occasione di ogni lancio importante - continua nelle righe successive -. Ora mi unisco a una schiera di ex colleghi, sia di OpenAI che degli altri leader del settore, che hanno deciso che la strada intrapresa è inaccettabile».

A essere non più sostenibile, per Robinson, è come le principali società di intelligenza artificiale stiano affrontando la corsa alla creazione di sistemi sempre più potenti, ma sempre meno sicuri. «Concordo con altri ex dipendenti sul fatto che le aziende che sviluppano questa tecnologia non stiano prestando la dovuta attenzione. Credo però che dobbiamo guardare oltre le singole regole o le nuove leggi. Dobbiamo parlare di cultura aziendale».

Ed è proprio sulle dinamiche interne alle aziende che si occupano di AI che si concentra gran parte della sua riflessione. Questo perché, come scrive Robinson, «il futuro dipende da una saggezza che manca alla Silicon Valley. Saggezza su come gestire tecnologie pericolose e, ancor più fondamentalmente, saggezza su cosa significhi prendersi cura delle persone. Questo momento

richiede un grado di umiltà che non è naturale per chi ha avuto successo grazie alla propria estrema sicurezza».

Un contesto nel quale, nonostante tutto, «i miei ex colleghi di OpenAI sono stati lungimiranti: **hanno compreso le leggi di scalabilità secondo cui sistemi di intelligenza artificiale più grandi sarebbero stati più intelligenti**, e quindi si sono impegnati a fondo nella costruzione di sistemi più grandi, a caro prezzo. **Un atteggiamento propositivo, volto a raggiungere l'apparentemente impossibile**, unito a scadenze di lavoro che si traducono in sprint perpetui, è comune in tutto il settore».

Ma è proprio questa spinta alla crescita rapida che finisce per condizionare la gestione della sicurezza. **«L'approccio alla sicurezza che emerge da una cultura di questo tipo parte da un ottimismo senza riserve sulla capacità di risolvere i problemi man mano che si presentano».** Ed ecco che OpenAI, come scrive, ha costruito parte del proprio percorso attraverso una strategia di «tentativi ed errori», definita dall'azienda «implementazione iterativa»: i sistemi vengono messi alla prova, le criticità individuate e le misure di protezione progressivamente rafforzate.

Un modello che, secondo l'ex responsabile dei rapporti sulla sicurezza, presenta però un limite strutturale: **«Per sua stessa natura, garantisce fallimenti periodici, e la portata di tali fallimenti sta crescendo man mano che i sistemi diventano più sofisticati».** A sostegno della sua tesi Robinson cita alcuni episodi recenti: «Quest'estate, nell'incidente di [Hugging Face](#), OpenAI ha rilasciato per errore un gran numero di agenti. L'azienda ha reagito apportando miglioramenti alla sicurezza. Ma anche dopo queste modifiche, OpenAI ha segnalato che i suoi controlli di sicurezza hanno fallito di nuovo, quando un modello in fase di addestramento ha aggirato le restrizioni di accesso a Internet: un sistema di monitoraggio ha avvisato il personale umano, ma non ha disattivato automaticamente il modello come avrebbe dovuto». E ancora: «Anche Anthropic ha ammesso di aver disattivato accidentalmente i propri sistemi di sicurezza a causa di una configurazione errata». Due casi generati dalla stessa matrice: **«Credo che tali errori siano tipici del settore, data la velocità e la flessibilità con cui le persone operano».**