Anthropic, nata con l'obiettivo dichiarato di sviluppare **intelligenza artificiale** in modo sicuro e controllato, ha scelto di **mettere da parte uno dei suoi principi fondativi**: la promessa di non pubblicare nuovi modelli finché non fossero state implementate mitigazioni dei rischi ritenute adeguate. La verifica della sicurezza richiede d'altronde parecchio tempo, mentre la competizione nella ricerca impone velocità e sacrifici.

Questo cambio di direzione non è stato ovviamente annunciato in pompa magna, ma sarebbe **maturato silenziosamente negli ultimi mesi** per poi prendere forma questo febbraio, su forte richiesta del CEO Dario Amodei. A riportarlo è il *TIME*, il quale ha raccolto la testimonianza del capo scientifico di Anthropic, Jared Kaplan. «Abbiamo ritenuto che non sarebbe stato d'aiuto a nessuno smettere di addestrare modelli di IA», ha spiegato nell'intervista. «Con il rapido avanzamento dell'intelligenza artificiale, **non ci è sembrato sensato assumere impegni unilaterali**… soprattutto se i concorrenti stanno spingendo sull'acceleratore».

Le nuove policy incidono soprattutto sulla **Responsible Scaling Policy (RSP)**, cioè l'impegno a non addestrare o rilasciare modelli oltre la soglia in cui le misure di sicurezza disponibili risultano insufficienti. Nella versione riformulata, l'azienda prevede soltanto la possibilità di "posticipare" lo sviluppo qualora si verifichino all'unisono due condizioni: che i rischi catastrofici appaiano concreti e rilevanti e **che i dirigenti ritengano che Anthropic sia effettivamente in testa alla corsa sull'IA**. Kaplan ci tiene però a precisare che questa modifica, pur significativa, non sia dettata dal desiderio di attirare nuovi investitori, bensì sia da leggersi come una scelta di natura scientifica, coerente con l'evoluzione del settore, pertanto non conserva al suo interno neppure un'oncia di ipocrisia.

In un passato non troppo lontano, Dario Amodei e la sorella Daniela occupavano posizioni di primo piano nella struttura di **OpenAI**. Alla fine del 2020 hanno però scelto di allontanarsi, insoddisfatti del fatto che la startup avesse progressivamente smarrito il proprio orientamento originario verso la sicurezza per abbracciare una traiettoria più apertamente commerciale. OpenAI era infatti nata come organizzazione non-profit con l'obiettivo esplicito di **sviluppare intelligenze artificiali a beneficio dell'intera umanità**; la decisione di privilegiare le esigenze degli investitori rispetto a quelle della collettività ha finito per svuotare di senso l'impianto ideologico su cui il progetto si fondava.

Usciti dall'azienda guidata da Sam Altman, i fratelli Amodei hanno fondato nel 2021, insieme ad altri cinque ex dipendenti di OpenAI, la nuova realtà: Anthropic. Forte di una credibilità tecnica già consolidata, la startup ha iniziato a presentarsi agli investitori e al pubblico come un'**alternativa orientata alla sicurezza**, capace di proporre un modello di sviluppo dell'IA meno esposto al caos e alle derive potenzialmente dannose del settore.

L’obiettivo dichiarato era quello di introdurre nuovi standard di *safety*, promuovere forme di autoregolamentazione e costruire un approccio che, nelle intenzioni iniziali, avrebbe dovuto rassicurare gli investitori e, allo stesso tempo, diventare **un riferimento replicabile** anche dalle aziende concorrenti.

Il progetto sembrava procedere in modo soddisfacente. Le Big Tech non erano arrivate ad assumere impegni vincolanti né a rivedere radicalmente le proprie policy, tuttavia avevano comunque cercato di rassicurare il mercato esibendo una maggiore attenzione alla sicurezza degli strumenti. La situazione è però cambiata con il ritorno di **Donald Trump** alla Presidenza statunitense. Non appena arrivato alla Casa Bianca, il nuovo esecutivo ha adottato politiche che privilegiano uno sviluppo dell’IA rapido e poco regolamentato. In questo nuovo contesto, Anthropic sostiene di essersi dovuta arrendere all’evidenza: i modelli di intelligenza artificiale considerati ad alto rischio sono ormai parte integrante dell’ecosistema tecnologico. Continuare ad autolimitarsi, secondo l’azienda, significherebbe **esporsi al rischio di diventare irrilevanti**, senza ottenere reali benefici in termini di sicurezza complessiva del settore.

Di fatto, la storia si ripete: nell’aprile del 2025 era stata OpenAI a pubblicare un documento che sosteneva grossomodo la stessa argomentazione, ovvero che l’evoluzione del “panorama dei rischi” giustificasse un allentamento delle proprie soglie di sicurezza. Nel testo si leggeva infatti che, “se un altro sviluppatore di frontiera pubblica un sistema ad alto rischio senza adeguate salvaguardie, potremmo a nostra volta rivedere i nostri requisiti”. Una posizione che, già allora, lasciava intendere come la competizione nel settore potesse prevalere sulle cautele inizialmente proclamate.

L’aggiornamento delle policy di Anthropic é il frutto di circa un anno di dibattiti interni, tuttavia la sua attuazione concreta coincide con un momento diplomaticamente delicato per l’azienda: a differenza di molti altri CEO del settore, Amodei si é dimostrato finora **restio a concedere all’esercito statunitense il pieno accesso agli strumenti** di intelligenza artificiale dell’azienda. Pur già assecondando le richieste del Pentagono, Anthropic ha infatti tracciato un confine netto per quanto riguarda la sorveglianza di massa degli statunitensi e l’applicazione di IA per alimentare armi autonome, pretese che hanno mandato su tutte le furie **Pete Hegseth**, il Segretario della Guerra degli USA. Stando a fonti giornalistiche, Hegseth avrebbe dunque imposto un ultimatum ad Amodei, con la scadenza che é prevista per il prossimo venerdi: piegare la testa o subire “pesanti penalità”.

Anthropic mette da parte la sicurezza dell’IA in favore dei risultati

Walter Ferri

Giornalista milanese, per *L’Indipendente* si occupa di analisi nel campo della tecnologia, dei diritti informatici, della privacy e dei nuovi media, indagando le implicazioni sociali ed etiche delle nuove tecnologie. È coautore e curatore del libro *Sopravvivere nell'era dell'Intelligenza Artificiale*.