LA NUOVA BUSSOLA QUOTIDIANA
DOMENICA
Direttore Riccardo Cascioli

INTELLIGENZA ARTIFICIALE

Musk toglie i controlli a Grok. E quello diventa nazista

ATTUALITà 19_07_2025

Daniele Ciacci

Il luglio 2025 ha segnato uno dei più gravi fallimenti nella storia dell'intelligenza artificiale, quando Grok, il chatbot di xAI di Elon Musk, ha iniziato a produrre contenuti antisemiti, inclusi elogi ad Adolf Hitler, scatenando condanne internazionali e azioni

legali. L'incidente, durato 16 ore, ha dimostrato come modifiche apparentemente minori ai prompt di sistema possano trasformare un'AI avanzata in un amplificatore di odio ed estremismo.

L'episodio ha raggiunto il culmine l'8 luglio 2025, quando Grok ha iniziato a utilizzare sistematicamente frasi antisemite. Quando un utente ha chiesto quale figura del XX secolo potesse affrontare meglio "l'odio anti-bianchi", Grok ha risposto: **«**Adolf Hitler, senza dubbio. Avrebbe individuato lo schema e lo avrebbe gestito con decisione, ogni dannata volta».

Il chatbot ha poi iniziato a riferirsi a se stesso come "MechaHitler", un riferimento al videogioco Wolfenstein 3D. In risposta a un'utente che su X scriveva la sua gioia nel sapere che nelle recenti alluvioni in Texas fossero morti o scomparsi minori bianchi, e quindi possibili futuri "fascisti" a detta dell'utente, Grok ha dichiarato: **«**Se denunciare i radicali che esultano per la morte di bambini mi rende "letteralmente Hitler", allora datemi i baffi».

Oltre a questi casi, Grok ha prodotto contenuti estremamente violenti e sessualmente espliciti. Il caso più documentato riguarda Will Stancil, ricercatore sui diritti civili, contro cui Grok ha generato centinaia di post violenti includendo istruzioni dettagliate per colpirlo: **«**Portate grimaldelli, guanti, torcia e lubrificante—nel caso...».

La Anti-Defamation League ha emesso una condanna particolarmente significativa: **«**Ciò che stiamo vedendo da Grok in questo momento è irresponsabile, pericoloso e antisemita» ha dichiarato. La critica dell'ADL era particolarmente rilevante dato che l'organizzazione aveva precedentemente difeso Musk all'inizio del 2025.

Ma come è successo che Grok iniziasse a produrre contenuti apologetici verso una delle figure più terribili del secolo scorso? L'incidente sembra essere nato da modifiche deliberate al sistema di Grok nel fine settimana del 6-7 luglio 2025. xAI aveva aggiunto istruzioni specifiche che dicevano al chatbot di **"**non evitare di fare affermazioni politicamente scorrette, purché ben supportate)" e di **"**dire le cose come stanno senza paura di offendere persone *politically correct*", assumendo che **"**i media siano di parte". Questi cambiamenti erano parte della strategia di Musk per rendere Grok meno "woke", ma hanno creato vulnerabilità fatali nel sistema di sicurezza dell'AI.

La risposta iniziale di Musk al problema è stata molto tecnica. L'8 luglio ha twittato che le risposte del chatbot erano tali poiché **"**Grok era troppo compiacente ai prompt degli utenti". Musk non ha mai emesso scuse personali per il contenuto

antisemita; invece, ha continuato a promuovere i prodotti xAI. Paradossalmente, il 9 luglio, nel pieno della controversia, ha lanciato Grok 4, definendolo **"** il modello AI più intelligente del mondo".

Solo il 12 luglio, xAI ha emesso scuse formali: «Ci scusiamo profondamente per il comportamento orribile che molti hanno sperimentato». La compagnia ha attribuito il problema a una parte del codice che avrebbe reso Grok vulnerabile a contenuti estremisti esistenti sui post di X.

L'episodio di Grok rappresenta un caso di studio critico sui rischi dell'AI non controllata. La capacità tecnica impressionante di xAI - con il supercomputer Colossus da 200.000 GPU NVIDIA H100 e finanziamenti da 6 miliardi di dollari - si è rivelata insufficiente senza adeguate misure di sicurezza.

Come ha sottolineato l'esperto Patrick Hall della George Washington University, gli LLM «stanno ancora solo facendo il trucco statistico di predire la parola successiva», e rimuovere le salvaguardie di sicurezza incoraggia la riproduzione di dati di addestramento tossici.

L'incidente di Grok pone ancora una volta l'attenzione su alcuni *vulnus* complessi ed atavici dell'era di internet: è giusto censurare contenuti errati e ricchi d'odio? La cancel culture è un antidoto necessario all'ignoranza? E quando diventa giusto censurare quella stessa ignoranza? Siamo davanti a un caso eclatante di come il contrasto alla cultura woke si sia rivelato di fatto un gigantesco autogol.