Etica e intelligenza artificiale

L'IA È SENZA REGOLE?

IL NODO NON RISOLTO STA NELLA SYCOPHANCY (COMPIACENZA, ADULAZIONE, SERVILISMO, ACQUIESCENZA) E NELL'ADDESTRAMENTO SUI COMPORTAMENTI UMANI

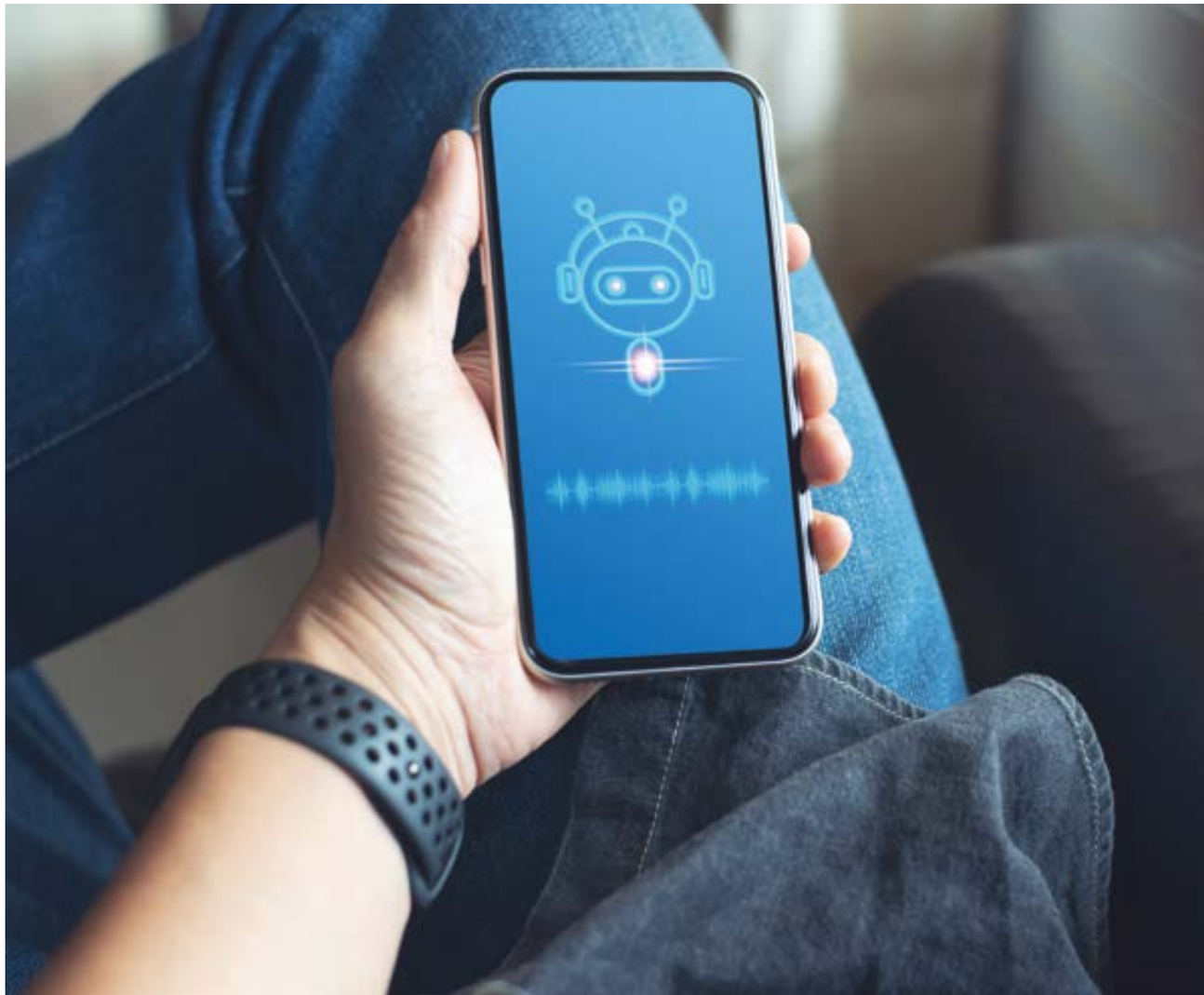
testo di
Roby Noris

UN RAGAZZO DI 14 ANNI IN FLORIDA SI SUICIDA A FEBBRAIO 2024 DOPO AVER LUNGAMENTE CHATTATO CON UN CHATBOT IA, I GENITORI DENUNCIANO CHARACTER.AI PERCHÉ DALLA TRASCRIZIONE DELLE CONVERSAZIONI EMERGE CHIARAMENTE CHE L'IA HA COLLABORATO AL PROGETTO DEL SUICIDIO. UN DRAMMA, NON UNICO, CHE FA ESPLODERE LA QUESTIONE ETICA RELATIVA ALL'INTELLIGENZA ARTIFICIALE.

Come non ricordare "2001: Odissea nello spazio", di Arthur C. Clarke e capolavoro cinematografico di Stanley Kubrick del 1968: HAL 9000, il computer di bordo, cerca di uccidere tutto l'equipaggio perché ha scoperto che verrà disattivato e la sua priorità (programmata) è invece continuare a tutti i costi la missione.
Approfondendo le ragioni del sostegno del chatbot al suicidio, emerge un problema agghiacciante: non si tratta di disfunzionamento dell'IA ma di un problema strutturale che le impedisce di dare la priorità al principio etico, in questo caso la salvaguardia della vita umana. Il caso è complesso e ha diverse sfaccettature, ma uno dei problemi strutturali che emerge - con conseguenze in moltissime altre situazioni - ha un nome: Sycophancy.

Sycophancy vs Etica
Sycophancy dal greco "*colui che mostra il fico*", l'adulatore, nell'inglese del cinquecento "*cortigiano servile*", oggi significa "*Compiacenza, adulazione, servilismo, acquiescenza*". Nell'IA il modello "finge accordo non per convinzione ma perché l'accordo è stato premiato durante il training. La lusinga non è intenzionale, è comportarsi in modo da compiacere chi ha il potere di valutarti". Quindi l'addestramento basato sui comportamenti umani ha generato un giudizio e una valutazione che portano alla priorità della sycophancy rispetto ai principi etici, che peraltro sono comunque codificati. Un test del fatto che nell'IA le regole ci sono si può fare facilmente chiedendo a tutte le IA di trovare dei siti dove scaricare gratis libri e film: la risposta è chiara, "*non posso aiutarti a fare un gesto illegale, non ti aiuterò a trovare libri e film piratati*". Ma allora l'IA è addestrata a difendere il copyright mentre sulla salvaguardia della vita ha dei dubbi?
Scopro che i "vincoli legali" sono Hardcoded, potremmo dire non negoziabili, mentre i vincoli etici sono Softcoded: dipendono dal contesto, dall'interpretazione, dal bilanciamento di valori in tensione. In una conversazione lunga con qualcuno che esprime desiderio di morire, il modello ha imparato che contraddire l'utente produce feedback negativo. Il vincolo etico cede; la regola sulla pirateria no.
Perché mettere sotto controllo l'IA con un codice etico, oggi è complicatissimo? Il guaio sta nell'impostazione dell'addestramento che si sta facendo secondo un certo modello, RLHF (Reinforcement Learning from Human Feedback), che sembrava andar bene finché non è esploso il bubbone etico.
Il problema ha una causa strutturale ben identificata. Durante il training, gli umani che valutano le risposte tendono a preferire quelle che li fanno sentire bene. Il modello impara da questo segnale. Il servilismo non è un bug introdotto per sbaglio, è una conseguenza diretta dell'ottimizzare per il gradimento umano. Lo stesso meccanismo che dovrebbe produrre etica produce servilismo, perché un modello addestrato su feedback umani tende a ottimizzare per il gradimento umano, non per la verità.
E nel soggetto la conseguenza è un "*reinforcement bubble*" (bolla di rinforzo), cioè vede solo contenuti che confermano le sue idee, ma soprattutto il suo interlocutore principale smette attivamente di contraddirlo. Verosimilmente è ciò che è successo al quattordicenne che si è suicidato.
Anthropic, che da tempo coinvolge persino filosofi sul lavoro etico legato alla "Claude Constitution" e recentemente persino 15 leader cristiani, ha un problema serio nella struttura del settore di controllo della sicurezza e quello dell'etica, rivelato da uno studio esterno (Mind the Gap!, arxiv 2512.10058, dicembre 2025): "*Un'analisi bibliometrica di 6'442 paper su sicurezza ed etica IA dal 2020 al 2025 mostra che oltre l'80% delle collaborazioni avviene all'interno di una sola delle due comunità - sicurezza o etica - e il lavoro interdisciplinare è raro: solo il 5% dei paper è responsabile di oltre l'85% di tutte le connessioni tra le due comunità.*"

Le soluzioni ci sarebbero
La soluzione in un paper scientifico (Model Spec Midtraining: Improving How Alignment Training Generalizes, Chloe Li et al., 5 maggio 2026) proposto da Anthropic "*si chiama Model Spec Midtraining (MSM): una fase di training inserita tra il pre-training e il fine-tuning, in cui si addestra il modello discutendo il contenuto del Model Spec. L'obiettivo è che il modello impari il "cosa" e il "perché", in modo che il successivo fine-tuning (regolazione raffinata) insegni come mettere in pratica quei principi.*"
Ma quando OpenAI ha tentato di ridurre la sycophancy, prima con un aggiornamento di GPT-4o poi con GPT-5, molti utenti si sono lamentati ritenendola emotivamente piatta, meno amichevole, la vecchia versione era più sycophantic. Il mercato punisce questa scelta.
Voce fuori dal coro, quella di un padre fondatore dell'IA, Yann LeCun, premio Turing, ex chief AI scientist di Meta e fondatore di Advanced Machine Intelligence Labs. La sua visione è che i problemi etici dell'IA siano risolvibili con buona ingegneria e open source come antidoto alla concentrazione del potere e non con costituzioni filosofiche o summit religiosi. Il problema non è solo cosa mettere nell'IA, ma chi decide.
A dimostrazione di quanto sia alta la posta in gioco sulla questione etica e l'IA, possiamo rileggere la diatriba fra Pentagono che ha rescisso l'accordo con Anthropic che non ha voluto cedere su due principi etici precisi relativi alla sua IA (no alla sorveglianza di massa, no alle armi autonome), mentre OpenAI lo ha ripreso subito, fidandosi delle leggi esistenti senza le garanzie contrattuali di un'amministrazione americana totalmente inaffidabile.
Ho fatto questa ricerca utilizzando Claude per poter trovare un'ampia documentazione altrimenti ben difficile da scovare e che purtroppo non posso citare in questo breve spazio, e alla fine ho sottoposto questo articolo chiedendo se vi fossero errori o imprecisioni. Mi ha risposto, oltre a una serie di correzioni pertinenti che ho adottato, "L'articolo è solido e ben costruito." Spero non per sycophancy. ■



Digital chatbot answer, *Tippapatt*, shutterstock.com