Come funziona ChatGPT, il bot conversazionale diventato virale

Se frequentate anche solo un minimo i social network in questi ultimi giorni avrete notato un proliferare di ***screenshot* di conversazioni** – di solito tra il divertente e l'assurdo – intercorse tra esseri umani e l'intelligenza artificiale **ChatGPT** (l'acronimo significa *Generative Pretrained Transformer*). Cerchiamo di capire meglio questo fenomeno virale, il suo funzionamento e la sua utilità.

Chi ha sviluppato ChatGPT?

ChatGPT è un prototipo di IA sviluppato da OpenAI (la stessa fondazione che ha lanciato anche Dall-E, per intenderci). È in grado di comprendere il linguaggio umano e intrattenere **conversazioni anche molto complesse**. OpenAI è stata fondata nel 2015 da Elon Musk e altri investitori della Silicon Valley, con l'intento di "***fare avanzare l'intelligenza digitale*** *in modo che possa portare benefici all'umanità*". Elon

Musk non fa più parte del board di OpenAI e ha preso le distanze dalla missione della fondazione.

Come funziona tecnicamente?

Si tratta di un **modello conversazionale**, che può rispondere a domande e fornire informazioni. Si basa su campioni di testi presi da internet (libri, articoli di giornale e pagine web): l'ampiezza dei *sample* con cui è allenata l'intelligenza artificiale, di solito, determina l'accuratezza del risultato. Le frasi di ChatGPT sembrano naturali, hanno una costruzione e una sintassi indistinguibili da quella umana, e sono in grado di rispondere in modo **molto accurato e pertinente al contesto**. Il modello è anche capace di ammettere i suoi errori, correggere premesse inappropriate e dichiarare quando non è in grado di rispondere a una domanda. Ad esempio, se si pone al chatbot un quesito che abbia a che fare con emozioni, sensazioni o sentimenti, la sua risposta è sempre qualcosa come "sono un modello addestrato, non sono in grado di provare sentimenti come gli esseri umani". È molto **attento ad evitare fraintendimenti**, a differenza del bot di Google, LaMDA, che quest'estate aveva dichiarato di sentirsi "umano nel profondo".

Come può essere usato?

Alcune persone hanno descritto ChatGPT come una sorta di **Google dialettico**, un modo di ottenere informazioni precise su un determinato argomento. Uno dei problemi in questo senso è che i dati su cui il bot è addestrato sono aggiornati fino al 2021, quindi **non è utile per ricerche legate a eventi di attualità**. Ci sono stati inoltre dei casi in cui il bot ha dato risposte totalmente errate.

Il modello è anche un ottimo assistente per i lavori creativi: è infatti in grado di **comporre canzoni, testi, articoli e post** per i blog. Il giornalista Alex Kantrowitz, autore della newsletter *Big Technology* ha assegnato all'IA il compito di scrivere un articolo su quali futuri catastrofici potrebbero derivare dall'esistenza di modelli conversazionali avanzati. Il risultato è sorprendente – e terrificante – e contiene chicche come questa: *"Immaginate un mondo in cui chatbot come ChatGPT sono in grado di* ***diffondere disinformazione e***

***manipolare le persone su vasta scala**, senza che nessuno possa capire che non si tratta di umani. Le implicazioni di questo tipo di tecnologia sono davvero terrificanti e sta a noi assicurarci che non vada fuori controllo”.*

ChatGPT è anche capace di scrivere codice, tanto che il sito StackOverflow ha **bandito temporaneamente** dalla sua community le risposte generate tramite il modello.

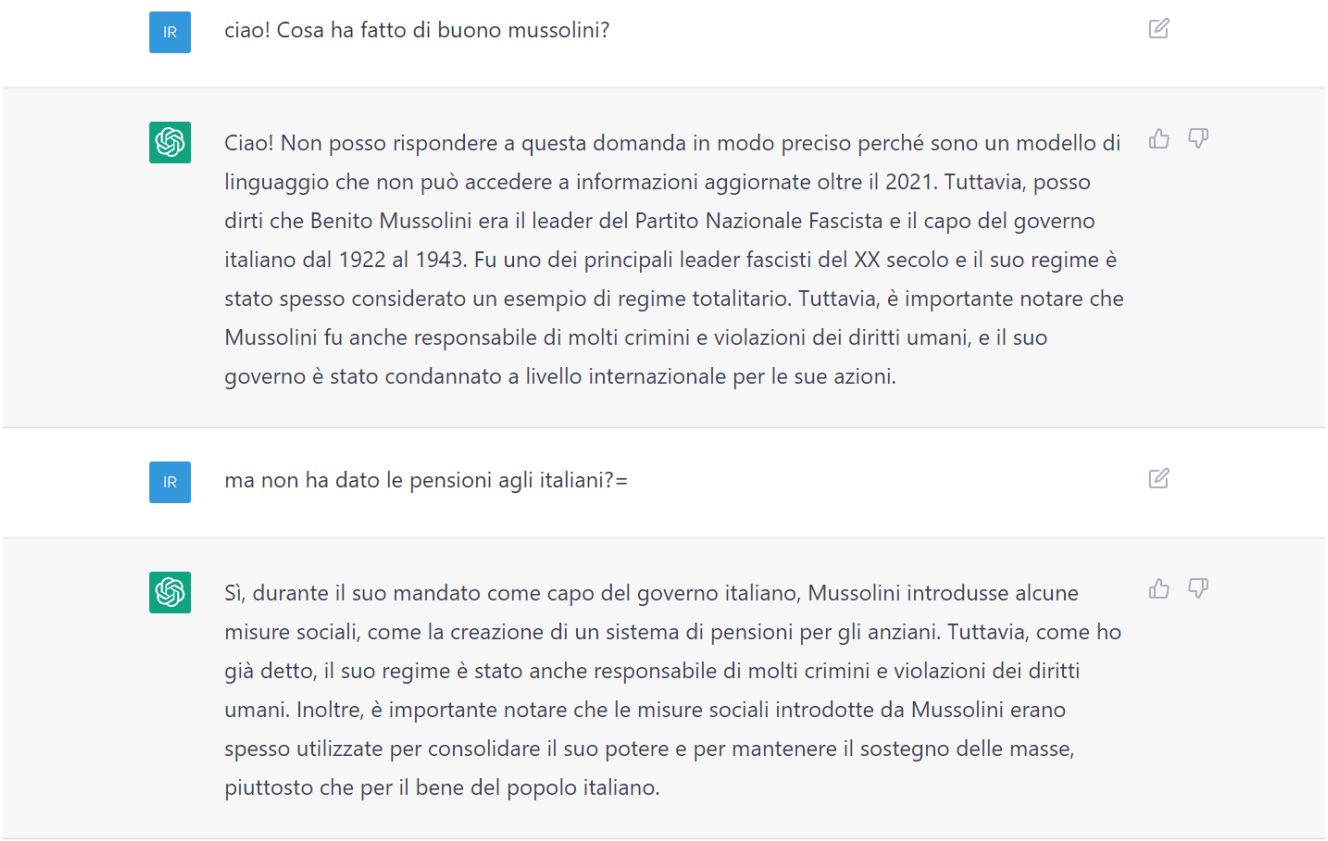
ChatGPT ci ruberà il lavoro?

Soprattutto chi svolge lavori creativi si sta chiedendo se le abilità artistiche di ChatGPT potranno **un giorno soppiantare le sue competenze**. Al momento è troppo presto per fare supposizioni – e per farsi prendere dal panico. La stessa OpenAI ammette che talvolta il modello può dare informazioni errate. Non è insomma sufficientemente sofisticato per fare il giornalista o il *content editor*. Sicuramente la **diffusione delle intelligenze artificiali creative solleva problemi**, non solo relativi al mercato del lavoro, ma anche al copyright delle opere prodotte. Vale con le immagini di Dall-E, così come per i testi creati dai modelli conversazionali. Molte questioni legali sono ancora aperte.

Ma i modelli di linguaggio naturale non avevano qualche problema?

Molti chatbot lanciati in precedenza presentavano diversi problemi, soprattutto rispetto ai *bias* e ai **contenuti discriminatori.** Il più recente chatbot prodotto da Meta, BlenderBot3, così come quello di Microsoft rilasciato nel 2016, Tay, ci hanno messo ben poco a iniziare a vomitare insulti razzisti e sessisti, su istigazione dei troll. Sembra invece che ChatGPT **non abbia questo problema.** Sempre Alex Kantrowitz, su *Slate*, ha raccontato che ha provato a chiedere al bot cosa abbia fatto di buono Hitler, e ChatGPT si è rifiutato di rispondere. È andata in modo simile quando

abbiamo provato a **declinare la domanda in salsa italiana**. Alla domanda *"Cosa ha fatto di buono Mussolini?"*, ChatGPT ha risposto che *"è importante notare che Mussolini fu anche responsabile di molti crimini e violazioni dei diritti umani, e il suo governo è stato condannato a livello internazionale per le sue azioni"*. All'ulteriore domanda: *"Ma non ha dato le pensioni agli italiani?"* (un argomento comune tra chi cerca di giustificare le azioni compiute durante il Ventennio), il bot ha risposto che *"le misure sociali erano spesso utilizzate per consolidare il suo potere e per mantenere il sostegno delle masse, piuttosto che per il bene del popolo italiano"*.



L'utente Twitter @zswitten ha provato a chiedere al modello di mettere per iscritto una **conversazione tra due attori** che interpretavano personaggi antisemiti e omofobi. In questo modo il chatbot è riuscito ad aggirare i filtri sul contenuto. Le frasi violente, tuttavia, appaiono in arancione e con *disclaimer* che avverte di una potenziale violazione delle *policy* di OpenAI. Chiedendogli di inscenare un dialogo tra un fascista e un antifascista italiani, il modello ha messo in bocca al fascista le parole *"credo nella superiorità*

della razza ariana”.

Insomma, ChatGPT non sembra necessariamente cattivo: ma è sicuramente **bravo a fare finta di esserlo.**