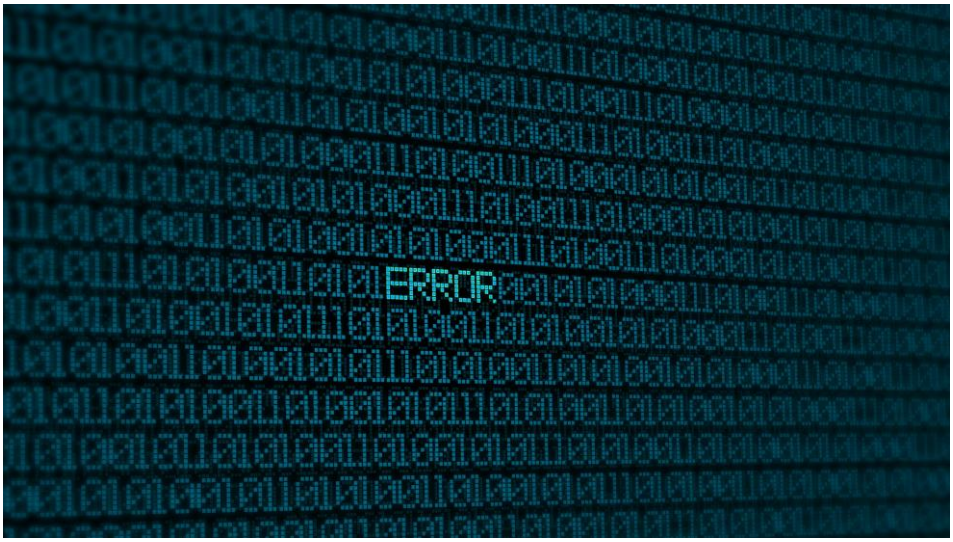
Quanto sbaglia l'intelligenza artificiale e perché?

Secondo uno studio, ChatGPT sarebbe diventata (deliberatamente) più "stupida". Ma in cosa gli errori dell'Ai sono diversi dai nostri? Abbiamo affrontato il tema con due esperti

Di Sandro Iannaccone



CHRISTIAN HORZ

Dialogo con l'**intelligenza artificiale**: *"Hey, ChatGPT, mi puoi aiutare a scrivere il saggio per l'ammissione al college? – Sì! Forniscimi le tracce e ogni informazione importante su di te, le tue esperienze e i tuoi obiettivi'.* **Natasha Singer**, giornalista del ***New York Times***, **ha chiesto** al modello conversazionale più in voga del momento di scrivere il compito per l'ammissione a Harvard, Yale e Princeton, tre tra le più prestigiose università statunitensi. **È andata bene… ma non benissimo**. Racconta Singer, per esempio, che alla domanda *"Quale canzone rappresenta la colonna sonora della tua vita, in questo momento"*, parte del test di accesso a Princeton, ChatGPT ha risposto *Cake by the ocean*, un pezzo il cui titolo fa riferimento a **un rapporto sessuale in spiaggia**, e che evidentemente *"non sembra appropriato per essere ammessi al college"*. Il bot, dopo aver riconosciuto che effettivamente la canzone che aveva suggerito parlava di sesso, si è *"scusato per la confusione"*, e tanti saluti.

L'esperienza non rende più saggi

Errori (o confusioni) di questo tipo sono in realtà **molto frequenti**. E potrebbero diventarlo sempre più. Uno **studio recentemente condotto** da un gruppo di ricercatori di Stanford e Berkeley ha scoperto che **ChatGPT sta diventando sempre più *"stupido"*:** l'accuratezza delle risposte è inferiore e gli errori *misurabili*, per esempio nel codice informatico che il bot è in grado di generare su richiesta, sempre più frequenti. *"GPT-4 (nella versione rilasciata a marzo 2023) era molto bravo a identificare i numeri primi, con un'accuratezza del 97,6%. La versione rilasciata a giugno 2023 era invece molto meno brava: l'accuratezza è scesa addirittura al 2,4%* - hanno scritto gli scienziati nel loro studio – *E sia GPT-4 che GPT-3,5*

commettono molti più errori nella generazione del codice rispetto a qualche mese fa. Abbiamo scoperto che le prestazioni e il comportamento di GPT-3,5 e GPT-4 sono molto diversi rispetto alle versioni precedenti, e in generale sono molto peggiorate nel tempo. È importante capire se ***gli aggiornamenti dei modelli, volti a migliorarne alcuni aspetti, ne abbiano inficiato le capacità in altri****"*. In altre parole, gli aggiornamenti potrebbero aver fatto più male che bene.

L'incertezza che inganna

La questione dell'**intelligenza artificiale che sbaglia** è, comunque, un **tema molto più ampio** (e in qualche modo inquietante, dato il loro impiego sempre più comune) rispetto alla *semplice* generazione di codice o di numeri primi, o scrittura di un saggio universitario. Ed è stata recentemente oggetto dell'incontro *L'intelligenza artificiale generativa e la meraviglia dell'intelligenza umana*, tenuta al **Festival della Mente** di **Sarzana** da **Veronica Barassi**, antropologa e professoressa in scienze della comunicazione alla Scuola di scienze umane e sociali dell'Università di San Gallo, e **Greg Gigerenzer**, psicologo tedesco e direttore emerito del Center for Adaptive Behavior and Cognition al Max Planck Institute for Human Development di Berlino. Gigerenzer, in particolare, si è molto occupato di **euristica** – parte del *modo di ragionare* degli esseri umani – e ha cercato di confrontarlo con quello delle intelligenze artificiali per comprenderne gli errori.

"L'euristica – ci ha raccontato – *è un modo di ragionare adattativo in virtù del quale siamo in grado di* ***focalizzarci solo sulle informazioni più importanti e*** **dimenticare** ***il resto****, ed è fondamentale in situazioni di incertezza in cui non è possibile ottimizzare la decisione. Pensiamo al* ***gioco degli scacchi****: in questo campo le intelligenze artificiali sono più forti degli esseri umani perché sostanzialmente* ***c'è poca incertezza****, e la competizione è sulla* forza bruta *del ragionamento, sulla profondità del pensiero. Lo stesso non vale in situazioni molto* ***più volatili****, per esempio quelle che hanno a che fare con il comportamento umano. In questo caso gli algoritmi fanno fatica a prevedere cosa succederà, e l'intelligenza umana (anche grazie all'euristica) è in grado di superarli. Secondo me sarà così per ancora molto tempo"*.

Una marcia (umana) in più

Un altro aspetto interessante relativo agli errori e ai limiti dell'intelligenza artificiale, secondo lo scienziato, è l'**intuito**. Caratteristica intrinsecamente umana e difficile da definire, *"basata* – dice Gigerenzer – *sull'evoluzione stessa del nostro cervello. L'intuito è quella* sensazione *che ci dice cosa fare o cosa non fare in una determinata situazione, senza che necessariamente siamo in grado di capire il perché*". Uno dei campi in cui si magnificano maggiormente le abilità delle intelligenze artificiali (o, più precisamente, del *machine learning*) è quello del **riconoscimento delle immagini**. Bene, anche in questo caso gli errori sono dietro l'angolo: *"Il* machine learning *è uno strumento potentissimo, basato sulle* ***correlazioni****. Una rete neurale profonda impara a distinguere le immagini di un cane da quelle di un gatto studiandone migliaia, e generalmente diventa molto brava a farlo. Ma si è visto che è possibile alterare le immagini in modo da confondere gli algoritmi, il che può diventare pericoloso, per esempio, se si pensa ai sistemi di guida autonoma che devono riconoscere uno scuolabus o un passante. Pensiamo invece agli esseri umani:* ***a un bambino basta vedere un solo gatto per distinguere un gatto da un cane per tutto il resto della sua vita****, con un'accuratezza praticamente perfetta"*.

Riflessioni necessarie

Di errori delle Ai si è occupata estensivamente anche Barassi, che ha lanciato il progetto ***The Human Error of Artificial Intelligence***. *"Oggi, e sempre di più, i sistemi di intelligenza*

artificiale vengono utilizzati ***dalle forze dell'ordine e dai tribunali per monitorarci****, tracciarci e profilarci in modo da determinare la nostra possibile innocenza o colpevolezza"*, ci ha spiegato, **facendo riferimento all'arresto,** avvenuto nel 2020, di **Robert Julian-Borchak**, un uomo nero fermato dalla polizia a Detroit per un crimine che non aveva commesso e per un **errore** di un sistema di riconoscimento facciale. *"Ho lanciato questo progetto* – dice – *perché ero interessata a capire cosa stiamo facendo, come società, per gestire questa fallibilità delle intelligenze artificiali, specie dopo aver capito che possono essere razziste o* ***riprodurre errori anche eclatanti.*** *E capire come avere a che fare con l'incapacità dell'intelligenza artificiale di capire il nostro mondo e la nostra* ***varietà culturale****. È molto difficile fare previsioni sugli scenari che verranno, ma sono convinta che sia necessario un dibattito pubblico su questi temi, un dibattito che coinvolga gli esperti, le istituzioni e la società civile. E magari prendersi un* ***momento di pausa*** *per capire se ha senso lanciare nel mercato in questo modo prodotti che sappiamo essere fallibili"*.