Machina sapiens: gli LLM funzionano, ma non sappiamo perché.

"Machina Sapiens" è un nuovo libro di Nello Cristianini, avvincente e stimolante, edito dal Mulino. Con uno stile fluido e accattivante, l'autore ripercorre le tappe fondamentali che hanno portato alla nascita di ChatGPT e, ormai, di numerosi altri chatbot analoghi. "Le macchine possono pensare?", domanda non nuova, la troviamo nel celebre articolo "Computing Machinery and Intelligence" , dove Alan Turing, nell'ottobre del 1950, propose il "Gioco dell'imitazione" (noto come Test di Turing) per scoprire quando una macchina fosse diventata pensante e indistinguibile da una persona. Nel luglio dello scorso anno, ci segnala Cristianini, sulla rivista Nature è stato pubblicato l'articolo "ChatGPT broke the Turing test — the race is on for new ways to assess AI", i grandi modelli linguistici imitano la conversazione, ma ragionano? Ma come siamo arrivati nell'arco di pochi anni a questo punto? L'autore richiama le tappe, che si sono susseguite in un breve arco temporale, solo sette anni. La svolta e l'intuizione fondamentale nel 2017 "L'attenzione è tutto ciò di cui hai bisogno" (Attention is all you need), si presenta un nuovo algoritmo "Transformer", risolve problemi di traduzione automatica, tradurre una parola significa trovare da quali parole dipende per averne una corretta interpretazione, l'algoritmo non solo scopre le dipendenze, ma con l'esperienza apprende e rapidamente opera, automaticamente. Lo schema da prendere in esame è costituito da tre livelli: l'agente con cui si interagisce (il chatbot), il modello interno usato dal chatbot per prendere decisioni (per esempio, GPT-3) e l'algoritmo che crea il modello a partire dai dati (per esempio il Transformer). La sorpresa è che, se l'addestramento è fatto su quantità sufficiente di dati, i modelli linguistici acquisiscono in forma spontanea conoscenze utili e non previste. Le abilità di questi modelli non dipendono solo dall'algoritmo che li ha prodotti, ma anche da come interagiscono con i dati forniti dall'utente. Il "gioco" avviene su due piani il contesto, con le parole (token) in input/output e i parametri ovvero i "pesi" interni regolabili che determinano e ottimizzano il comportamento del modello. Cristianini ci mette a conoscenza dello "stato dell'arte", queste macchine indubbiamente funzionano, nelle batterie di test standard (SAT, MMLU, Big -Beng) utilizzate per rilevarne le prestazioni continuano a migliorare e si avvicinano sempre più a quelle umane. Ci supereranno a breve? La previsione di Turing è prossima ad avverarsi? In Scintille di Intelligenza Artificiale Generale (Sparks of general artificial intelligence), studio condotto nel 2023 da Microsoft su GPT - 4, si afferma: "GPT - 4 raggiunge una forma di intelligenza generale, mostrando effettivamente scintille di intelligenza artificiale generale. Ciò è dimostrato dalle sue capacità mentali fondamentali (come il ragionamento, la creatività e la deduzione)". Aumentando in modo rilevante, come sta già accadendo, i parametri su cui lavorano i modelli linguistici emergeranno abilità impreviste? Possiamo fidarci di macchine che non comprendiamo in profondità? È urgente capire come funzionano, per utilizzarle in sicurezza dobbiamo garantirci che siano allineate ai nostri valori e obiettivi.
Le "allucinazioni" dei modelli linguistici permangono, sebbene in calo.
Quanto è concreto il rischio, descritto nel romanzo, satirico e distopico, Erewhon di S. Butler, che le macchine prendano il sopravvento sull'uomo?" Cristianini parafrasa Butler "Le macchine intelligenti di oggi sono un prodotto degli ultimi cinque anni. Siamo pronti per quello che verrà dopo?". Appunto, non basta comprenderle, ma guidarle e soprattutto controllarle. Possiamo considerare questi nuovi agenti intelligenti alla stregua di alieni tra noi? Ci imitano, eppure cosa ne sappiamo davvero? Certamente il loro "pensare" differisce profondamente dal nostro. Le risposte che forniscono dipendono da come gli algoritmi ovvero i meccanismi matematici

interagiscono con il linguaggio umano. Se l'assunto è che siano alieni allora Cristianini propone di esaminarli a livello esterno (l'anatomia), all'interno (cercare di capire come funzionano), osservare il comportamento (in contesti e condizioni diversificate).
Oggi ci mancano tante informazioni, ma i modelli continuano ad essere rilasciati sempre più potenti, la velocità dell'innovazione non è disgiunta da un senso di timore, anche di paura. La previsione di Turing, presente nell'articolo del '50, non lascia tranquilli, esiste un parallelo con la fissione nucleare? Le analogie non mancano: siamo vicini alla "massa critica" ovvero al livello critico di conoscenze, una volta superato avremo un'esplosione dell'intelligenza? L'informatico Cristianini ci apre a domande che mettono in crisi "Cosa significa essere umani?" Cosa vuol dire essere intelligenti? Comprendere il mondo comporta la creazione di un modello che permetta di prevederne il comportamento, un format da utilizzare in situazioni nuove, ma questa alla fine è la definizione di intelligenza. I "modelli di linguaggio" creati utilizzando il Trasformer, rileva Cristianini, si stanno rilevando dei veri e propri "modelli del mondo", non solo sono colte "le relazioni grammaticali tra le parole, ma anche le relazioni causali tra oggetti, eventi e concetti del mondo reale". Il libro si avvia alla conclusione: comprendere il mondo e comprendere il linguaggio sono due ambiti diversi o questa distinzione è arbitraria? Si torna all'intuizione di Turing che sembra trovare conferme: la creazione di una "macchina pensante" sta portando a capire cosa sia il pensiero e a cercare un nuovo significato per termini come sapere, comprendere, intelligenza.
Cristianini arriva a una conclusione tanto evocativa quanto forte: nell'antichità l'homo sapiens divenne tale strappando il segreto della conoscenza agli dèi secondo il mito di Prometeo. Ora stiamo per entrare nell'era della "machina sapiens" perché ci è stato rubato il segreto della conoscenza? Una nuova era si apre davanti a noi, colma di opportunità ma anche carica di incognite e di rischi da non sottovalutare. Ci fermiamo o proseguiamo? La sfida è impegnativa, eppure non possiamo far altro che raccoglierla. Cristianini è certo, continueremo a "giocare con il fuoco" della conoscenza, rinnovando così il mito di Prometeo e di Pandora, non possiamo rinunciare alla specificità della nostra natura umana.

Nel libro, l'autore riporta dieci dialoghi che, a seconda dei casi, sono stati intessuti con ChatGPT e Bard. In allegato, ne vengono presentati otto di questi dialoghi, ma effettuati con versioni di chatbot gratuiti rilasciati da poche settimane: Claude3 Sonnet (algoritmo proprietario), Mistral Le Chat e Llama2, entrambi open source. Questi "esperimenti artificiali" hanno l'obiettivo di mettere a confronto le prestazioni tra i vari chatbot utilizzati e "sperimentare" come, nel corso degli ultimi mesi, questa tecnologia si sia rapidamente sviluppata. Le differenze non sono rilevanti, i tempi di risposta sono ridotti a pochi secondi, i linguaggi sono fluidi e decisamente conversazionali. I dialoghi di Claude3 Sonnet emergono per articolazione, chiarezza, completezza e fluidità.