Posso chiamarti Prosdocimo?

Perché è bene non fidarsi troppo delle risposte di ChatGPT.

Guido Vetere

Volevo provocarlo sul piano personale. Chi? Lui, naturalmente: ChatGPT, il grande interlocutore automatico di OpenAI, il *chatbot* al quale, in questo istante, milioni di persone stanno rivolgendo domande, chiedendo consigli, scambiando motti di spirito, rimanendo impressionate dalle sapide e spesso verosimili risposte.

Volevo provocarlo, dicevo, fargli perdere il suo meccanico *aplomb*, il suo anglosassone *understatement*. Allora gli ho chiesto: “Posso chiamarti Prosdocimo?”. Speravo che reagisse a un nome così inusuale, che mi rispondesse qualcosa come: “Beh, non ti viene in mente nulla di più normale?”. Invece lui non ha fatto una piega: “Certamente! Se preferisci chiamami Prosdocimo, puoi farlo senza problemi” (sottotesto: “Ti ho già spiegato che sono un automa e non ho gusti o sentimenti, stupido umano!”).

Allora l’ho incalzato: “Sai che Prosdocimo è il personaggio di un’opera di Rossini?”. Qui mi aspettavo che mi parlasse del poeta in cerca di ispirazione che Rossini ha voluto mettere sulla scena del Turco in Italia, il quale assiste alla vicenda tra il principe Selim e Fiorilla ricavandone la trama per la sua commedia. Invece la risposta dell’automa è stata sconcertante: “Sì, ho familiarità con la figura di Prosdocimo: un personaggio del Barbiere di Siviglia di Gioachino Rossini. Prosdocimo è uno dei personaggi principali dell’opera che aiuta il Conte di Almaviva a corteggiare Rosina, la figlia di Don Bartolo”. Chiaramente, quello invece è il celeberrimo Figaro: che abbaglio!

Ma non finisce qui: lo informo che Prosdocimo è un personaggio di Turco in Italia, e lui, invece di scusarsi, insiste: “Mi dispiace ma devo correggerti: Prosdocimo non è un personaggio del Turco in Italia, ma del Barbiere di Siviglia”. Deve essere proprio un *chatbot* temerario per contraddirmi su questioni che riguardano il melodramma. Insisto anch’io chiedendogli chi siano allora i protagonisti del Turco, e qui la conversazione diventa un

surreale crescendo di fandonie: “Nel Turco in Italia di Gioacchino Rossini i personaggi principali sono Mustafà Albazar, Zaida, Fiorilla e Prosdocimo. Mustafà è un turco che si innamora di Fiorilla, una donna sposata con un italiano chiamato Geronio. Albazar è un poeta turco che è innamorato di Zaida, una schiava. Prosdocimo, d’altra parte, non è un personaggio del Turco in Italia, ma del Barbiere di Siviglia di Gioachino Rossini”. La trama del Turco rimasticata dal *chatbot* risulta infatti una inconsistente macedonia in cui si rimestano personaggi e ruoli rossiniani. “Mi par d’esser con la testa in un’orrida fucina” avrebbe chiosato, sarcastico, il Cigno di Pesaro. Quando gli faccio notare il suo vaneggiamento, ChatGPT, senza un minimo di onestà intellettuale, mi congeda bruscamente: “*Conversation not found”*.

È questo dunque il nuovo ritrovato dell’Intelligenza Artificiale, l’automa che cambierà il nostro modo di informarci e fare ricerca, ma anche di scrivere discorsi, articoli e saggi, e perfino di soddisfare il bisogno di relazioni personali? Si stenta a crederlo, e tuttavia l’attesa attorno a ChatGPT sembra essere, per molti, proprio questa. Cerchiamo allora di capire con cosa abbiamo a che fare e quali possano essere le reali prospettive.

I “trasformatori generativi pre-addestrati” (*Generative Pre-trained Transformers*, da cui l’acronimo GPT) sono davvero un meraviglioso ritrovato dell’IA linguistica più recente. La loro abilità fondamentale consiste nel derivare testi plausibili in relazione allo spunto che gli si fornisce, cioè generare la continuazione più attendibile di un inciso detto *prompt*. La continuazione di “C’era una volta”, ad esempio, potrebbe essere “un Re”, anche se “un pezzo di legno” s’è pure letto da qualche parte. La cosa fantastica di questi generatori è che possono essere addestrati con metodi cosiddetti “non supervisionati”: basta (semplifico) prendere qualche *terabyte* di testo, farlo a fettine grandi come una frase, dividere ogni frase in due pezzi e mostrare alla IA che il secondo è la continuazione del primo, cioè il primo è il *prompt* del secondo. Sembra un gioco da ragazzi, ma in realtà il modo in cui viene costruita la rete neurale che riuscirà a generare buone continuazioni per qualsiasi *prompt* è molto sofisticato, e per ottenere risultati soddisfacenti bisogna mettere dentro così tanti dati e consumare così tante risorse computazionali (dunque energetiche) che solo pochi sono oggi in grado di sfruttare appieno queste tecniche.

Ottenuto in modo automatico un modello linguistico di base (ve ne sono in realtà anche di piccole dimensioni e disponibili per tutti, alcuni perfino in

italiano), ci si può sbizzarrire a specializzarlo (*fine tuning*) e utilizzarlo in moltissimi compiti derivati. Tra questi, vi sono la generazione di cronache o riassunti, la conversazione, la classificazione, il *question answering* cioè la capacità di rispondere a domande. Testi di una certa lunghezza, come quelli necessari per condurre conversazioni, si possono ottenere per concatenazione: basta (semplifico ancora) usare il frammento generato come *prompt* per il successivo. Un automa che abbia tali capacità di base si può poi specializzare per produrre continuazioni plausibili in contesti conversazionali, facendogli osservare milioni di chat facilmente reperibili sulle piattaforme online. Se a questo viene aggiunta la capacità di utilizzare fonti enciclopediche (una per tutte: Wikipedia) per trovare risposte alle domande, ecco che otteniamo qualcosa di simile a ChatGPT: un sistema che dialoga attorno allo scibile umano.

Abbiamo computer potentissimi, algoritmi sofisticati, tanta energia elettrica, migliaia di ricercatori, *billion dollars* delle multinazionali e una quantità incalcolabile di testi: cosa può andare storto? In realtà, tutto. Per capirlo, consideriamo le perverse idee di ChatGPT attorno al protagonista del Barbiere di Siviglia. Quando ho chiesto al *chatbot* se conosceva l'opera di Rossini in cui compariva Prosdocimo, probabilmente all'automa sarà "venuto in mente" il protagonista rossiniano *par excellence*, cioè Figaro, di sicuro molto più nominato dell'oscuro personaggio del Turco. Evidentemente, il *topos* fa aggio su tutto: Figaro è un protagonista rossiniano così tipico che un elemento della "coda lunga" dei personaggi minori, nella testa dell'automa, può facilmente "collassare" su di lui. Il secondo scambio è poi particolarmente illuminante: con spirito eracliteo, il *bot* afferma che Prosdocimo *è* e *non è* un personaggio del Turco. La coerenza complessiva del discorso, su cui i ricercatori di OpenAI dicono di aver molto lavorato, è ancora evidentemente di natura, per così dire, statistica. Il sistema imita bene alcuni stereotipi conversazionali (notevole è l'uso della congiunzione avversativa 'tuttavia') e tuttavia (scusate il bisticcio) non è capace di introspezione logica.

Il tecno-ottimismo globale è ora al lavoro per dire che i difetti, che assieme ai pregi sono sotto gli occhi di tutti, saranno presto emendati con reti neurali ancora più bulimiche di quelle attuali, o integrando procedure di ragionamento simbolico, dunque razionale e non statistico. Ma c'è un tema di fondo che non ha a che fare con la tecnica, bensì con l'epistemologia.

Siamo cresciuti con l'idea che la rete avrebbe facilitato la ricerca di informazione, e Google ha costruito con grande maestria la piattaforma globale di questa funzione, divenuta poi così vitale da dar corpo a un monopolio tentacolare. Per quanto, con l'introduzione del *Knowledge Graph*, abbia nell'ultima decade notevolmente sofisticato il suo modo di trattare l'informazione facendo un passo verso la conoscenza, Google resta un sistema epistemicamente disimpegnato: dice ciò che ha trovato in relazione alla nostra ricerca, senza commettersi sulla sua veridicità. Resta a noi valutare se ciò che leggiamo sia o meno credibile; se non siamo soddisfatti, possiamo approfondire e cercare (sempre beninteso con Google) altre fonti. Un sistema di *question answering* ambisce invece a dire, definitivamente, le cose come stanno. Ora, consideriamo che il padre di tutti questi sistemi, quel Watson di IBM che nel 2011 vinse contro i campioni umani nel gioco a quiz televisivo *Jeopardy!*, aveva una accuratezza di circa il 75%: una risposta su quattro era sbagliata. Tuttavia, potendo valutare il proprio grado di confidenza e rinunciare talvolta alla risposta, tanto bastò per battere gli umani. A dieci anni di distanza, grazie ai nuovi modelli neurali, questi sistemi hanno fatto notevoli progressi, ma non certo al punto da essere diventati infallibili. Possiamo accettare di vivere in un mondo popolato da oracoli algoritmici fallaci? Facendo il verso a Nietzsche dobbiamo chiederci: quanta falsità può sopportare un uomo?

Il problema, come si dice, è a monte. Molta di quella che viene venduta come conoscenza è in realtà, come sappiamo, la credenza di qualcuno, cioè una costruzione sociale i cui fondamenti sono sempre in discussione. Il nesso tra ciò che si dice e ciò che è vero, oltre che filosoficamente problematico, è linguisticamente impercettibile. La peggiore mistificazione ha le stesse proprietà statistiche della più profonda verità. Non c'è algoritmo che, muovendo dall'umana testualità ridotta a sequenze numeriche, possa uscire, in forza della computazione, dalle sabbie mobili della (chiamiamola così) *noosfera*. Chi oggi propaganda le magnifiche sorti dell'Intelligenza Artificiale linguistica oltre la misura della sua intrinseca limitatezza, nella migliore delle ipotesi non sa di cosa parla. Questa propaganda, peraltro, danneggia chi cerca di usare le tecnologie linguistiche per quello che di buono possono umilmente fare, che non è affatto poco. Qualcuno probabilmente tenta operazioni di *marketing* in perfetta malafede, altri forse vagheggiano, nel nero dell'anima, l'annichilirsi di ciò che tuttora resta

umano: il giudizio. In ogni caso, chi propugna un mondo in cui in un unico punto, sotto il controllo di un soggetto privato, sia concentrata la facoltà di discernimento dell'intera umanità, tralascia il fatto che quel discernimento sarebbe in realtà un mero gioco di assonanze binarie, una vuota combinatoria di zero e di uno, cioè, in definitiva, un nulla. Un nulla verso il quale non dobbiamo farci trascinare.