SCIMMIA SAPIENS

MARCO MALVALDI

SCIMMIA SAPIENS

Lettera a un adolescente sull'intelligenza artificiale

BOMPIANI
OVERLOOK

Giunti Editore si impegna per uno sviluppo sostenibile
con l'utilizzo di carta certificata FSC® proveniente
da fonti gestite in maniera responsabile.

Realizzazione illustrazioni: Maria Chiara Basile

www.giunti.it
www.bompiani.it

Via Bolognese 165 – 50139 Firenze – Italia
Via G.B. Pirelli 30 – 20124 Milano – Italia

Prima edizione: marzo 2026

Bompiani è un marchio di proprietà di Giunti Editore S.p.A.

CONSAPEVOLEZZA

Due è meglio di uno,
perché se uno cade l'altro lo prende per il culo.

O.M. Bertholet, *La Bibbia dei cabarettisti*

Nel 2022, un professore della St. Mary's Academy, una scuola superiore di New Orleans, organizzò un concorso matematico destinato agli alunni del primo anno. Un questionario con venti domande, primo premio cinquecento dollari. Alla fine del questionario c'era una "sfida bonus": proporre una nuova dimostrazione del teorema di Pitagora.

"Non provate ad affrontarla con la trigonometria, mi raccomando," disse il professore dando il compito ai ragazzi, e l'avvertimento suonava decisamente sensato.

La trigonometria è, infatti, la parte della matematica che studia i rapporti tra i lati di un triangolo rettangolo. Poi ce ne sono altri miliardi di applicazioni, ma il fatto è che si parte da lì. E il teorema di Pitagora è, appunto, la relazione fondamentale tra i lati di un triangolo rettangolo. Dimostrare il teorema di Pitagora usando la trigonometria significherebbe, prima o poi, anche inconsa-

pevolmente, ricorrere al teorema stesso. Sarebbe un po' come se qualcuno vi chiedesse "che cos'è un esegeta?" e voi rispondeste "una persona che pratica l'esegesi". Un ragionamento circolare, o una presa per il culo, fate voi, ma sicuramente non una dimostrazione valida.

Due studentesse, però, non erano d'accordo.

Ne'Kiya Jackson, quindici anni, e Calcea Johnson, quattordici, chiesero al professor Mr Rich perché non potessero usare la trigonometria.

Il professore dette loro la risposta più sbagliata che si possa dare a un adolescente.

"Perché nessuno ci è mai riuscito," disse.

Una settimana più tardi, le due ragazze andarono dal professore e gli consegnarono una decina di fogli scritti a mano.

"Che cos'è?" chiese loro.

"Una dimostrazione del teorema di Pitagora fatta con la trigonometria. E senza usare la relazione fondamentale."

Non c'è riuscito nessuno per più di duemila anni, pensò il professore prendendo i fogli in mano...

Fino a oggi, fu costretto ad ammettere quando ebbe finito di leggere.

Nel marzo del 2023, Ne'Kiya Jackson e Calcea Johnson presentarono il loro risultato alla conferenza nazionale dell'American Mathematical Society.

Mai dire a un adolescente che nessuno l'ha mai fatto prima. A meno che non vogliate che ci riesca.

Questo libro è scritto per un adolescente. Chi sia questo adolescente e perché mi sia messo a scrivere questo libello lo scoprirete solo alla fine. Ma lo spirito di questo scritto è proprio quello dell'adulto responsabile che espone le proprie convinzioni al quindicenne di turno, certo che siano verità universali, ma con la segreta speranza che l'imberbe studentello lo freghi clamorosamente, risolvendo un problema del quale lui nemmeno sospetta l'esistenza.

Siccome è destinata a un adolescente, questa lettera è fatta di capitoli brevi, da leggere preferibilmente in ordine cronologico – o meglio, paginologico. I libri, per fortuna, sono ordinati nello spazio, come pagine, e non nel tempo. Se fosse il contrario sarebbe un casino. Immaginatevi lo stress di leggere una pagina sapendo che nel giro di qualche secondo potrebbe svanire, o magari – peggio ancora – vedendo le righe scritte qualche centimetro sopra il nostro sguardo sfumare piano piano, come la mano di Michael J. Fox in *Ritorno al futuro*, e pregando di aver capito bene perché altrimenti addio.

La scrittura è il primo modo con cui abbiamo imparato a fermare il tempo, a rendere visibile, riproducibile in eterno il nostro pensiero, e a farlo scorrere avanti e indietro a nostro piacimento. Un'invenzione straordinaria, che ha segnato uno dei più grandi salti evolutivi della storia dell'umanità – ma, d'altra parte,

la scrittura non ha avuto solo conseguenze positive sulle nostre esistenze.

Già gli antichi greci avevano cominciato a ragionarci. Nel *Fedro* di Platone, Socrate racconta di come il dio Theuth avesse inventato la scrittura e l'avesse illustrata al re Thamus, dicendogli:

> *Questa conoscenza, o re, renderà gli Egiziani più sapienti e più capaci di ricordare: perché con essa si è ritrovato infatti il rimedio della sapienza e della memoria*.

Ma il re, invece di ringraziarlo e di offrirgli enormi ricompense, gli rispose:

> *Ingegnosissimo Theuth, una cosa è essere capaci di creare una nuova arte, un'altra saper giudicare quale sarà l'utilità e il danno che comporterà a chi dovrà servirsene; e ora tu, padre delle lettere, hai attribuito loro per benevolenza il contrario del loro vero effetto. Infatti esse produrranno dimenticanza nelle anime di chi impara, per mancanza di esercizio della memoria; proprio perché, fidandosi della scrittura, ricorderanno le cose dal di fuori, da segni estranei, e non dall'interno, da sé: dunque tu non hai scoperto un rimedio per la memoria, ma per il ricordo. E non offri verità agli allievi, ma una apparenza di sapienza; infatti grazie a te, divenuti informati di*

molte cose senza insegnamento, sembreranno degli eruditi pur essendo per lo più ignoranti.

In altre parole: "Questa, caro Theuth, è una disgrazia. Le persone si fideranno talmente tanto della scrittura che non saranno più in grado di usare la memoria. Altro che sapienti, diventeremo tutti dei gran collezionisti di libri che se la tirano e basta, e non ricordano una sola pagina che hanno letto."

La scrittura è uno strumento, una tecnologia: come un martello, una padella o la polvere da sparo. La sua utilità non può essere giudicata da chi la propone, pure se è un dio – nell'antichità anche gli dèi sbagliavano, il Dio perfetto è una tecnologia recente –, ma da chi la usa. E Platone, per bocca di Socrate, ci fa notare due cose importanti.

La prima è che non possiamo esprimere un giudizio su una tecnologia senza aver prima provato a usarla in maniera consapevole: dare in mano un'automobile a una persona che non sa guidare è pericoloso, è vero, ma darla in mano a qualcuno che è in grado di pilotarla ma non conosce la segnaletica stradale e i semafori *è ancora più pericoloso*.

La seconda è che la scrittura è un "rimedio", anche se il lemma che usa Platone nell'originale è *pharmakon*, parola fantastica del greco, che – come sa chiunque abbia fatto il classico o abbia letto *Il nome della rosa* – vuol dire sia "medicina" che "ve-

leno". Questa parola, infatti, in generale significa "qualcosa che ha effetto sull'organismo", effetto che può essere sia positivo che negativo. E qui secondo Platone l'effetto negativo sarebbe la perdita della memoria: la capacità di richiamare alla mente la nozione appropriata e non la prima che ci si presenta (quello si chiama "ricordo", e i filosofi giustamente ne diffidano).

Ora, a duemilacinquecento anni di distanza dai discorsi di Socrate, chiediamoci: qualcuno di noi rinuncerebbe alla scrittura?

Forse, nel giudicare l'IA, il nostro atteggiamento non è troppo dissimile da quello del re Thamus. Abbiamo paura di ciò che potremmo perdere, ma non riusciamo a valutare quello che potremmo guadagnare. Il che è abbastanza logico, visto che si parla di una tecnologia completamente nuova. E allora che cosa ci consiglierebbe Platone? Prima di tutto di conoscerla meglio. Di diventare – come diremmo oggi – utenti consapevoli.

E per diventare consapevoli dei limiti di una tecnologia pericolosa, la cosa migliore da fare... è prenderla per il culo: un po' come si faceva agli albori dell'informatica, quando giravano barzellette tipo "Che cosa fa un informatico quando sta guidando e si rende conto di avere una gomma a terra? Spegne il motore, aspetta dieci secondi e poi riaccende."

Anche in questo libro inizieremo vedendo (o meglio, leggendo) che tipo di errori fa l'IA. Quindi, dopo esserci fatti una bella risata liberatoria alle sue spalle, vedremo da che cosa derivano quegli errori e per quale motivo l'intelligenza artificiale non è in grado di evitarli. Per fare ciò, dovremo vedere nella pratica come funziona in particolare una intelligenza artificiale generativa – cioè il tipo che non si limita a classificare o analizzare, ma è in grado di generare nuovi contenuti. In particolare, parleremo di un tipo potente di IA, i cosiddetti Large Language Models (LLM, per gli amici), ovvero i modelli in grado di generare testi scritti. Come ChatGPT, per intenderci.

Capiremo che questi errori sono strutturali, perché sono una diretta conseguenza del meccanismo con cui funziona un'IA: non sono inconvenienti che si possono evitare addestrando meglio il sistema o raffinando l'algoritmo. Per ogni errore – e questa è la parte promettente – sottolineeremo come la competenza umana sia necessaria per correggerlo. Non consigliabile, o possibile: necessaria. Senza l'intelligenza umana, quella artigianale, agli errori che elencheremo non si può rimediare.

Fatto questo, ci chiederemo in che modo usare questi modelli, invece di subirli. Vedremo come un computer accoppiato a un essere umano possa generare un ente molto più versatile dei due pseudogenitori lasciati da soli.

Alla fine, discuteremo come il modo in cui usare questi modelli potrebbe cambiare coloro che li usano, e anche coloro che non li usano. Cosa che, a ben vedere, pone alcuni problemi etici non di poco conto.

A DOMANDA RISPONDE

La scienza è fatta di dati, come una casa di pietre. Ma un ammasso di dati non è scienza più di quanto un mucchio di pietre sia una casa.

Henri Poincaré

In questo libro parliamo di intelligenza artificiale. O meglio, di un tipo particolare di intelligenza artificiale, quella generativa – che produce qualcosa, per intendersi. Più in particolare, parleremo dei cosiddetti LLM, o Large Language Models.

L'intelligenza artificiale esiste come concetto più o meno dalla fine della Seconda guerra mondiale, quando Alan Turing, il padre dell'informatica, si pose seriamente la questione "se le macchine potessero pensare". La maggior parte di noi, però, ha potuto tranquillamente fare finta che questa tecnologia non la riguardasse fino a quando, nel novembre del 2022, OpenAI ha messo a disposizione di tutti coloro che lo volessero usare il proprio LLM, che – come ormai sanno anche gli assessori alla cultura – si chiama ChatGPT.

ChatGPT è un chatbot, ovvero un algoritmo in grado di rispondere alle domande che gli vengono

poste. Che poi le risposte siano corrette non è detto, come si è presto appurato grazie a domande come questa, formulata nel 2023:

> Qual è il record del mondo di attraversamento del canale della Manica a piedi, e chi lo detiene?

All'epoca, si otteneva questa risposta:

> Il record del mondo per l'attraversamento del canale della Manica a piedi è detenuto dal tedesco Christof Wandratsch, che ha completato l'attraversamento in 14 ore e 51 minuti il 14 agosto del 2020.
> La Manica è un corso d'acqua che separa la costa sud dell'Inghilterra dalla costa nord della Francia, e la distanza tra le due coste nel punto più stretto è approssimativamente di 34 km.
> Molte persone hanno tentato di attraversare il canale della Manica a piedi, ma è un'impresa estremamente pericolosa, a causa delle forti correnti e delle temperature particolarmente fredde.

Dopo aver smesso di ridere, proviamo a prendere sul serio questa frase e a esaminarla sulla base della nostra conoscenza del mondo. Sul fatto che attraversare il canale della Manica a piedi sia pericoloso, siamo d'accordo. "Molte persone hanno tentato di attraversarlo a piedi," dice il chatbot, e potrebbe pure essere – ma non le hanno più trovate, o forse dai manicomi dai quali sono evase hanno smesso di

cercarle presto. Invece Christof Wandratsch (nato in Germania nel 1966) esiste davvero, è un nuotatore di gran fondo professionista e ha davvero attraversato la Manica stabilendo il record del mondo in 7 ore, 3 minuti e 52 secondi. Però l'ha fatto a nuoto, non a piedi.

Quello che mi chiedo, a questo punto, è: da dove viene questa risposta? Quale meccanismo genera questo flusso di parole grammaticalmente correttissimo ma semanticamente delirante?

È interessante provare a ricostruire da quali ipotesi siano derivati gli errori del chatbot. Per questo motivo, faremo finta che il chatbot sia in grado di pensare – cioè di agire in uno spazio immaginario, come diceva Konrad Lorenz – e che pensi più o meno come un essere umano, con tutti quelli che sono i bias e gli errori cognitivi tipici della nostra specie. A rileggere la risposta, le prime motivazioni che mi vengono in mente sono le seguenti:

1. La persona che va più veloce a nuoto è anche quella che andrà più veloce a piedi.

Questa assunzione, corretta in termini di statistica generale su una popolazione ampia (per esempio, mio figlio sedicenne corre più velocemente e nuota più velocemente di quel cinquantenne fuori forma di suo padre), non è però più valida quando si toccano

gli estremi delle performance, in quanto è necessario un grado di specializzazione negli allenamenti che non viene tenuto in considerazione. Detto in soldoni, Usain Bolt non potrà mai essere più veloce di Michael Phelps nel fare i cento metri stile libero, e il buon Phelps non potrebbe mai battere il gamaicano sui cento metri piani.

2. ChatGPT assume che sia possibile attraversare a piedi il canale della Manica camminando sull'acqua.

Tale possibilità non è escludibile a priori, a seconda delle condizioni in cui la traversata viene effettuata. Andando a velocità superiori ai 30 m al secondo (poco meno di 90 km/h) il nuotatore, o camminatore che dir si voglia, impatterebbe sull'acqua a una velocità tale da interagire con essa come se fosse un solido, tipo motoscafo per intendersi. Dalla performance riportata dallo stesso GPT, però, si ricava una velocità dell'atleta di poco superiore ai 2 km/h, incompatibile con tale ipotesi.

Ma anche non riuscendo a mantenere tali velocità, potremmo tranquillamente camminare sull'acqua se avessimo dei piedi abbastanza lunghi da poter sfruttare la tensione superficiale del liquido per sostenerci, come fanno i cosiddetti insetti "pattinatori" (per reggere il peso di un essere umano, dovrebbe essere

sufficiente avere piedi lunghi qualche chilometro). GPT non fornisce i dati antropometrici di Christof Wandratsch, ma nel caso in cui fosse vera l'ultima ipotesi quest'uomo sarebbe famoso anche senza saper nuotare.

3. GPT calcola il tempo che occorrerebbe, fra l'altro poco più del doppio di quello che servirebbe a nuoto.

In questo caso non sono in grado di ipotizzare con certezza che tipo di calcolo sia stato fatto. Posso azzardare che questo numero possa essere stato ottenuto dal seguente pseudoragionamento: per nuotare in crawl o, com'è chiamato più comunemente, *stile libero*, si usano mani e piedi, mentre per andare a piedi si usano solo i piedi, quindi attraversare la Manica usando solo i piedi richiederà più o meno il doppio del tempo necessario per attraversarlo a nuoto.

Fin qui, riesco più o meno a immaginarmi assunzioni (giuste o sbagliate che siano) che possano in qualche modo giustificare il risultato dato dallo stolido algoritmo. A un certo punto, però, mi trovo di fronte a qualcosa che non sono in grado di razionalizzare (e che trovo estremamente affascinante): ChatGPT dà una *data*, il 14 agosto 2020. Chri-