SAGGIO 7

Calibrazione e IA: un giro sul carro di Fetonte

Claudia Marsico · Hernán Inverso

E se il problema stesse nel come lo dice? Un'esplorazione urgente di modelli linguistici usciti dai cardini.

Gustave Moreau, La caduta di Fetonte (1878)

Un esperto sa molte cose, ma quando si avvicina al bordo abbassa la voce o esita. Un LLM può tenere lo stesso tono quando risponde 1815 alla domanda sull'anno di Waterloo e quando lancia un dato dubbio come esito di un calcolo in cui di solito ci azzecca poco. Il fenomeno si chiama "assenza di calibrazione" e riguarda il grado di sicurezza della risposta. La fiducia dovrebbe salire e scendere insieme all'evidenza. Guai a noi quando non accade.

In sintonia con i disastri che comporta, il nome si mescola ad armi ed esplosioni. Nel Cinquecento, in Francia, *calibre* era il diametro della bocca di un cannone, forse dall'arabo *qālib*, "stampo", a sua volta dal greco *kalópous*, la forma del calzolaio, o più probabilmente dal latino medievale *qua libra*, "di quale peso". Comunque sia, il verbo "calibrare" compare in francese nel Seicento con il senso di prendere la misura di qualcosa e si diffonde presto in altre lingue. L'accezione che ci interessa, graduare uno strumento perché misuri senza errore, è dell'Ottocento.

Pochi temi si intrecciano con la storia dell'Occidente quanto questo. *Medèn ágan*, niente di troppo, è la formula delfica della calibrazione. Il mito è pieno di esempi degli orrori che comporta ignorarla. Prendiamo il caso di Fetonte. Figlio di Elio, il Sole, per vantare la propria origine supplica di guidare il carro del padre. Elio inorridisce e cerca di dissuaderlo, spiegandogli che la strada è ripida e che i cavalli sputano fuoco. Nemmeno Zeus, gli dice, nella versione che ne dà Ovidio nelle *Metamorfosi*, se la sente di guidarlo. Fetonte si ostina e prende le redini. Immediatamente i cavalli non riconoscono più la guida e si imbizzarriscono. Il carro sale all'impazzata e brucia il cielo, precipita e brucia la terra, prosciuga i fiumi, apre i deserti della Libia e arrostisce per sempre la pelle dei popoli del sud. Per fermare l'incendio del mondo, a Zeus non resta che il fulmine, e Fetonte cade ardendo. Aveva calibrato male la propria abilità di condurre il carro, e la sicurezza infondata lo uccise.

Nella favola di Esopo, il pastorello che grida "al lupo!" per divertimento quando non c'è nessun lupo addestra i vicini alla diffidenza. Quando il lupo arriva davvero, le sue grida non valgono più nulla, perché ha speso tutto il suo credito in falsi allarmi. Il pastorello sbaglia tanto che non gli credono quando ci prende. Parla di calibrazione anche l'*Otello* di Shakespeare, scritto intorno al 1603. Quando il moro di Venezia chiede una "prova oculare" del tradimento di Desdemona, Iago gli consegna appena un fazzoletto e qualche insinuazione. Con quel materiale minimo Otello la uccide. Così poco è ciò che sa e così tanto ciò che crede di sapere. Della stessa stirpe di chi sbaglia a graduare sono il Chisciotte che vede giganti dove ci sono mulini e il re Lear che misura l'amore delle figlie sulla lunghezza dei loro discorsi.

Di che cos'altro si occupa Socrate, se non di calibrazione, quando definisce la propria sapienza a partire dal non credersi sapiente? Duemila anni più tardi Montaigne avrebbe fatto di quel temperamento un motto, "*Que sais-je?*", che ne so io, e lo fece incidere su una medaglia. Il termine aureo della calibrazione è "probabile", coniato da Cicerone negli *Academica* (45 a.C.) per tradurre il greco *pithanón*, "ciò che è credibile", il criterio con cui lo scettico Carneade proponeva di muoversi in un mondo dove la certezza ci sfugge; e gli stoici fecero riposare tutta la loro epistemologia sull'assenso che diamo alle impressioni, tanto che il loro sapiente infallibile, in sintonia con il cosmo, è un calibratore che non sbaglia.

Quell'assenso divenne un'arte nella modernità. Per Descartes, nella quarta delle sue *Meditazioni metafisiche* (1641), siamo alla mercé di una volontà più ampia dell'intelletto e di chiarezza limitata, che ci precipita nell'errore: ci scalibriamo, cioè, affermando con più forza di quanta l'evidenza ne autorizzi. Poco dopo Jacob Bernoulli, nell'*Ars Conjectandi* (1713), definì la probabilità come un grado di certezza e dimostrò che, con un numero sufficiente di prove, la frequenza di un fatto si avvicina alla

sua probabilità, gettando il ponte tra la credenza e la misura. Da parte sua il vescovo Joseph Butler, nell'*Analogia della religione* (1736), coniò la frase "la probabilità è la guida stessa della vita", perché l'evidenza probabile ammette gradi, dalla più alta certezza morale alla più bassa presunzione.

John Locke dedicò, nel suo *Saggio sull'intelletto umano* (1689), un capitolo ai "gradi dell'assenso", dove chiedeva di credere a ciascuna cosa con una fermezza proporzionale ai suoi indizi. David Hume lo sintetizzò, nella *Ricerca sull'intelletto umano* (1748), nell'idea che un uomo saggio commisura la propria credenza all'evidenza. Il matematico William Clifford lo portò alla sua forma più dura ne "L'etica della credenza" (1877), dicendo che è sbagliato, sempre, ovunque e per chiunque, credere qualcosa sulla base di un'evidenza insufficiente. William James guardò tra le crepe e gli rispose ne "La volontà di credere" (1896) che ci sono decisioni vitali che non ammettono di aspettare l'evidenza completa. A volte bisogna semplicemente avere il coraggio di scommettere. In ogni caso, per quanto riguarda il nostro tema, i due concordano sull'importanza di sapere da dove viene la nostra sicurezza e di non andare in giro convinti di qualcosa su cui abbiamo poche probabilità di aver ragione.

Il comandamento della probabilità, enunciato da Dennis Lindley in *Making Decisions* (1971), prescrive di non assegnare mai una certezza dello zero né del cento per cento, salvo che alle verità logiche, perché gli estremi chiudono la porta all'apprendimento di indizi futuri. La regola è detta "di Oliver Cromwell", perché Lindley si ispirò alla supplica che questi rivolse nel 1650 ai presbiteriani scozzesi: "vi supplico, per le viscere di Cristo, di considerare che potreste sbagliarvi". Affermare ogni cosa con la stessa sicurezza distrugge il margine per sbagliare, cosa che facciamo spesso. Secondo Sarah Lichtenstein e Baruch Fischhoff (1977), chi è sicuro al settanta per cento ci prende circa la metà delle volte, mentre la fiducia estrema è la peggio calibrata. Nel 1999 David Dunning e Justin Kruger mostrarono l'altra faccia della medaglia, esibendo casi in cui i meno competenti sono quelli che misurano peggio la propria competenza, privi come sono di un criterio di misura.

La meteorologia, quella grande sfida quotidiana, fece un passo avanti quando nel 1950 Glenn Brier inventò un sistema per assegnare un punteggio alle previsioni del tempo che porta ancora il suo nome, legando la calibrazione alla correlazione tra fiducia dichiarata e frequenza dell'azzeccarci. Un previsore leale non ci prende sempre, ma sa quanto di solito ci prende e lo dice. Nei decenni successivi le misurazioni guadagnarono in precisione e migliorò il cosiddetto "diagramma di affidabilità", che mette a confronto in un grafico la sicurezza che il sistema dichiara e la percentuale di volte in cui ci prende. Se dice novanta e ci prende novanta, cade sulla diagonale, la linea della lealtà; se dice novanta e ci prende settanta, la curva sprofonda al di sotto, e quella distanza dalla diagonale, mediata lungo tutta la scala, si chiama errore di calibrazione. Serve a sapere se il "sono sicuro" della macchina vale qualcosa, ed è per questo che venne usato su reti neurali che classificano immagini, ben prima che esistessero i chatbot. Resta ancora oggi lo strumento standard della calibrazione per i modelli linguistici.

Le reti neurali sono diventate più precise e, allo stesso tempo, più sovrasicure (Guo et al. 2017). Quelle degli anni Novanta, piccole, erano ben calibrate; quelle recenti, profonde, ci prendono di più ma esagerano la propria certezza, e le stesse scelte di progetto che hanno alzato la precisione (più strati, più parametri, la normalizzazione per lotti) hanno affondato la calibrazione. Eppure basta una sola

variabile, la temperatura, che ammorbidisce o irrigidisce di colpo tutte le probabilità del modello, per ricalibrare buona parte dello scarto. Se una sola manopola aggiusta tanto, la sovrasicurezza è un bias uniforme su tutta la scala.

Con i modelli linguistici il meccanismo è chiaro. Un modello base impara aggiustando le proprie previsioni alla frequenza con cui ciascuna parola compare nei dati: quanto più si allontana da quelle proporzioni, tanto maggiore è la penalizzazione. Per questo, in linea di principio, le sue probabilità riflettono abbastanza bene ciò che accade effettivamente nel materiale con cui è stato addestrato. Nel rapporto tecnico di GPT-4 (2023) la curva di affidabilità del modello base è quasi una diagonale perfetta: quando dichiara maggiore fiducia, ci prende in proporzione equivalente. Poi arriva la messa a punto che lo trasforma in assistente e cambia l'obiettivo. Il feedback umano preferisce ciò che suona sicuro, sicché la politica ottimale si ingoia il dubbio e con esso la calibrazione. Nello stesso rapporto su GPT-4, la diagonale che il modello base portava con sé si inarca dopo la messa a punto e l'ago si incolla in alto.

La storia non è però segnata, perché il segnale calibrato non scompare: resta sepolto. Un modello grande, interrogato nel modo giusto, può stimare la probabilità che la propria risposta sia corretta; questa capacità migliora con la dimensione e la messa a punto per rinforzo la indebolisce senza cancellarla (Kadavath 2022). Un modello può imparare a dire in linguaggio naturale quanto è sicuro, anche senza accedere alle proprie probabilità interne (Lin et al. 2022), e nei modelli allineati la fiducia espressa a parole è di solito meglio calibrata di quella che si ricava dai loro stati interni, riducendo circa la metà dell'errore (Tian 2023). La conoscenza della propria ignoranza era ancora lì, archiviata in un altro cassetto. Chiedere al modello, però, non fa miracoli. Nel mettere in parole la propria fiducia, il modello tende anche a esagerarla, concentrando le risposte fra il novanta e il cento per cento, come se imitasse la spacconeria umana (Xiong 2024). La calibrazione si può recuperare, dunque, ma solo a metà.

Si potrebbe addestrare un modello all'atteggiamento socratico, l'emblema del dubbio ben misurato? In parte è ciò che cercano i modelli di ragionamento o *Thinking*. Prima di rispondere scompongono il problema, provano strade, rivedono i propri passi e cercano di individuare i propri errori. Quel tempo aggiuntivo di solito migliora la resa. Funzionano anche i procedimenti dialettici attraverso consigli di modelli. Ragionare di più, tuttavia, non equivale a essere meglio calibrati. La revisione può portare a difendere un errore con maggiore abilità. Uno strato socratico aggiunto può produrre appena dubbio rappresentato: domande, riserve e cautele che avvolgono la stessa sovrasicurezza. Nemmeno una cautela indiscriminata serve. Il modello che risponde sempre "cinquanta per cento" non distingue il sicuro dall'incerto. Calibrare è variare il dubbio a seconda della difficoltà del terreno.

Il procedimento socratico può aiutare a produrre ragioni e dati migliori per la valutazione, ma non costituisce da solo una cura. E ha senso. Il modello può imparare a riconoscere schemi associati al proprio errore e a dire "non lo so" nei casi opportuni, il che implica una condotta epistemicamente prudente, ma non c'è nessuno che stia tentando di conoscere se stesso. Il modello non ha una storia di acquisizione, un registro interno di quando e con quanta fermezza ha imparato ciascuna cosa, sicché

tutto quello che dice esce con lo stesso peso, ciò che è stato verificato mille volte e ciò che è stato ordito or ora, senza un'etichetta che li separi. Calibrarsi presuppone un io che si gioca qualcosa nell'aver ragione, che può sbagliare e pagarne il prezzo e che proprio per questo impara ad avere cura della propria certezza, lasciandosi sempre, come chiedeva Cromwell, un margine per essere in errore.

Fetonte si scalibrò e bruciò. La sua tragedia ha, almeno, la forma di una tragedia. Ci fu un padre che lo pregò di non farlo, una caduta, un corpo, una punizione che trasformò l'eccesso in lezione per tutti gli altri. Il modello prende le redini del carro del Sole con la stessa cecità di Fetonte, ma non sente la vertigine, non gli si bruciano le mani, non cade. È un Fetonte senza caduta, un eccesso a cui non segue nessun disastro, perché non c'è nessuno lì che possa rovinarsi. Ciò che può bruciare è la terra di sotto, dove stiamo noi, quelli che ricevono il dato falso detto con il tono sicuro. Lo strumento che gradua quanto meriti di essere creduto ciò che ci dicono non sta sul carro. Sta, come sempre, dalla nostra parte, e tocca a noi non mollarlo.

◊

Mappe del labirinto è una serie di Phantom Maze, AI & Language Lab. La collezione vive su **phantommaze.cloud**. L'iscrizione è gratuita.

Alcune delle idee che inquadrano questo testo sono sviluppate in *Il fantasma senza ego. Una mappa nel labirinto dell'intelligenza artificiale*, di Hernán Inverso e Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-19-5).

PHANTOM MAZE

Leggi questo saggio online · Tutti i saggi · Il libro

Fonti. G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability" (*Monthly Weather Review*, 1950). S. Lichtenstein, B. Fischhoff e L. Phillips, "Calibration of Probabilities: The State of the Art" (1977/1982); D. Dunning e J. Kruger, "Unskilled and Unaware of It" (1999). C. Guo, G. Pleiss, Y. Sun e K. Q. Weinberger, "On Calibration of Modern Neural Networks" (ICML 2017); OpenAI, *GPT-4 Technical Report* (2023, la calibrazione del modello base degradata dopo la messa a punto, fig. 8); A. T. Kalai e altri (OpenAI), "Why Language Models Hallucinate" (2025); S. Kadavath e altri (Anthropic), "Language Models (Mostly) Know What They Know" (2022); S. Lin, J. Hilton e O. Evans, "Teaching Models to Express Their Uncertainty in Words" (2022); K. Tian, E. Mitchell e altri, "Just Ask for Calibration" (EMNLP 2023); M. Xiong e altri, "Can LLMs Express Their Uncertainty?" (ICLR 2024); S. Farquhar, J. Kossen, L. Kuhn e Y. Gal, "Detecting Hallucinations in Large Language Models Using Semantic Entropy" (*Nature*, 2024); "Mind the Confidence Gap: Overconfidence, Calibration, and Distractor Effects in Large Language Models" (2025).