SAGGIO 8

Interpretabilità: l'oscuro interno della creatura

Claudia Marsico · Hernán Inverso

Per la prima volta abbiamo fabbricato qualcosa che si comporta come un pezzo di natura. L'abbiamo fatta noi, eppure dobbiamo studiarla come una sconosciuta. Ora interpretare è urgente.

Henri Rousseau, Giungla equatoriale (1909)

Da sempre fabbrichiamo cose proprio perché sapevamo come funzionavano. Poteva andarci male, ma il progetto lo avevamo in testa. Non importa se si tratta dell'orologio, della macchina a vapore o di un programma di calcolo contorto. L'autore metteva i pezzi e sapeva dove andava a parare. Le reti neurali hanno aperto un altro gioco, perché si coltivano invece di farsi. Un obiettivo, una procedura di ottimizzazione, la discesa del gradiente e milioni di parametri si sistemano da soli finché l'errore non scende. Il risultato funziona e non l'ha scritto nessuno, quindi non lo conosciamo nel dettaglio. Abbiamo dovuto sviluppare rami di ricerca appositi per venire a sapere che cosa fa un modello al suo interno, con metodi che somigliano a quelli delle scienze naturali, come se ciò che ci preoccupa fosse un oggetto trovato nella giungla.

Rintracciamo le coordinate di questa stranezza nel sud dell'Italia. Correva l'anno 1710 e Giambattista Vico, professore di retorica a Napoli, pubblicava il *De antiquissima Italorum sapientia*, dove coniò la celebre formula *verum et factum convertuntur*, ossia il vero e il fatto si convertono l'uno nell'altro. L'idea è che possiamo conoscere con certezza soltanto ciò che abbiamo fatto noi stessi, ed è per questo che la geometria è trasparente per noi. La costruiamo definizione dopo definizione. La natura, invece, no, ed è per questo che la fisica, per esempio, è condannata alla congettura. Guai a Descartes e alla sua ingenua fiducia nelle idee chiare e distinte. Il libro di Vico non suscitò entusiasmo ai suoi tempi, ma ebbe la fortuna che l'Ottocento trovasse nella sua opera maggiore, la *Scienza Nuova* (1725/1744), ciò che serviva per fondare la storia, che stava muovendo i primi passi come scienza. Le idee di Vico dicono che lo è, perché il mondo civile è stato fatto dagli uomini e con un po' di attenzione se ne vedono i fili. Se è fatto da noi, è conoscibile da noi e merita una disciplina propria.

A rigore, l'intelligenza artificiale classica soddisfaceva il principio di Vico, perché un sistema esperto era trasparente per i suoi autori, ma la regola va in frantumi con i sistemi addestrati, quelle reti profonde i cui parametri non li scrive nessuno. Progettiamo l'architettura e scegliamo la dieta di dati, ma i milioni di pesi li trova la discesa del gradiente. Il *factum* è lì, progettato da noi e pagato con il nostro bilancio, ma il *verum* non compare. Perché quel groviglio di numeri fa quello che fa? Per Vico l'opacità si concentrava nella natura, perché non è opera nostra. Adesso abbiamo un oggetto che abbiamo fatto noi ma che si comporta come se fosse di fattura altrui. In un certo senso è così, perché ci è sfuggito di un passo. Abbiamo fatto il processo e il processo ha fatto l'oggetto, come fossero dei nipoti stranieri cresciuti in un'altra lingua.

Aristotele aveva tracciato con cura la frontiera tra naturale e artificiale che ora vacilla. Nella *Fisica* (2.1) li distingue in base alla sede del principio del movimento. Gli esseri naturali portano in sé il principio del proprio mutamento, mentre negli artefatti esso è esterno, viene da un artigiano. Un letto non ha alcun impulso interno a essere letto. Aristotele allude lì (193a) all'analogia del sofista Antifonte, il quale aveva notato che se si sotterra un letto e il legno mette radici, ciò che cresce è legno, non un nuovo letto. La sua riproduzione è impossibile perché la sua forma non gli appartiene se qualcuno non gliela impone in una bottega. Del resto il greco *physis* e i suoi derivati vengono da *phyein*, germogliare, e il latino *natura* è un participio futuro di *nasci*, nascere, cioè ciò che sta per nascere. *Technē*, invece, è parente del *tektōn*, il carpentiere, da un'antica radice indoeuropea che significava intrecciare e che abbiamo ancora in tecnica, testo, tessuto e architetto. Ciò che germoglia e ciò che si ordisce non si mescolavano, ed ecco che il modello linguistico si trova esattamente a cavallo della linea. Un tessuto di parametri ordito da un processo che abbiamo progettato è *technē*, ma dobbiamo studiarlo come *physis*, sondando un interno pieno di strutture germogliate da sole durante l'addestramento. Abbiamo sotterrato nei dati il letto di Antifonte e alla fine vi è germogliata un'entità liminare, artificiale per origine e naturale per opacità.

L'immaginario ha elaborato figure simili ma non uguali. Il Talmud (*Sanhedrin* 65b) racconta che il saggio Rava creò un uomo e lo mandò a un altro maestro, il quale gli parlò, non ottenne risposta e lo restituì alla polvere con una frase secca: torna alla tua argilla. La parola viene dal *golmi* del Salmo 139, la massa informe che gli occhi di Dio videro prima che avesse forma. La leggenda successiva, fissata

attorno al rabbino Loew di Praga, morto nel 1609, aggiunse il dettaglio che più ci interessa. Il Golem lo si anima scrivendogli in fronte *emet*, verità, e lo si spegne cancellando la prima lettera, perché senza l'alef resta *met*, morto. Una creatura di argilla che funziona a parola, accesa e spenta dalla parola verità. Vico avrebbe apprezzato il dettaglio.

Frankenstein offre una variante moderna. Mary Shelley sottotitolò il suo romanzo del 1818 *Il moderno Prometeo*. La lettura consueta calca la mano sulla colpa morale del creatore che abbandona la sua creatura, ma sotto quella colpa ce n'è un'altra meno commentata. Victor sa assemblare la creatura e animarla, pur ignorando quasi tutto di lei. L'interno dell'essere che ha fabbricato gli risulta chiuso quanto a un qualsiasi estraneo, e la creatura deve raccontargli le proprie pene in una lunga confessione sul ghiaccio perché lui venga a sapere qualcosa. Aggiungiamo Goethe e la sua versione del *Filopseude* di Luciano di Samosata (II secolo), dove un apprendista anima un mortaio perché trasporti acqua e, non conoscendo la formula per fermarlo, lo spacca in due e ottiene due portatori che non riesce a fermare. Goethe la mise in versi nel 1797, Paul Dukas la orchestrò nel 1897 e Disney la fissò in *Fantasia* (1940), con Topolino in tunica. L'apprendista conosce la formula d'avvio e ignora quella d'arresto, che è un modo esatto di dire che far funzionare un sistema e capirlo sono saperi distinti.

Tutti e tre i racconti, però, portano la questione in un luogo sbagliato. Ciò che conta nel caso che analizziamo non è la ribellione della creatura, cioè il lato morale, ma il fatto che il fattore non riesca a leggere la propria creatura, il suo lato epistemico. Il Golem della leggenda va fuori controllo, il mortaio allaga la casa e Frankenstein scappa. In tutti e tre i casi il problema è posto come disobbedienza, e così perdiamo di vista che, anche in assenza di disobbedienza, se non possiamo capire, l'obbedienza non verificabile somiglia troppo alla fortuna. Bisogna dunque interpretare.

Gli studi di interpretabilità si lasciano ordinare in quattro famiglie di metodi.

La prima si concentra sulla condotta, senza aprire nulla, ed è per questo che si chiama della "scatola nera", termine che viene dall'ingegneria elettrica di metà Novecento. W. Ross Ashby dedicò un intero capitolo della sua *Introduzione alla cibernetica* (1956) al problema di inferire che cosa ci sia dentro una scatola di cui si vedono solo ingressi e uscite. Applicato ai modelli, il metodo è quello dello psicologo sperimentale con un soggetto che non si lascia aprire. Si progettano esperimenti con prompt e batterie di valutazione, lo si sottopone a red-teaming, e dalle risposte si inferisce che cosa accade nella dimensione a cui non accediamo. Il manifesto di questa famiglia è *Machine Behaviour*, l'articolo che Iyad Rahwan e una ventina di coautori pubblicarono su *Nature* nel 2019, con una proposta che allora suonò provocatoria: studiamo le macchine come gli etologi studiano gli animali, con la pazienza osservativa di un Lorenz tra le sue oche, perché il comportamento di un sistema appreso non si deduce più dai suoi progetti.

La condotta osservata, in effetti, riservava sorprese. Basta aggiungere alla consegna una frase sciolta, per esempio "ragioniamo passo dopo passo", perché un modello che sbagliava un problema cominci ad azzeccarlo, senza toccare un solo parametro (Kojima et al. 2022). Nessuno ha progettato quella meccanica: è comparsa da sola e la si è trovata provando. Nemmeno era prevedibile la cosiddetta "maledizione dell'inversione" (Berglund 2023). Un modello che ha imparato che Mary Lee Pfeiffer è

la madre di Tom Cruise risponde bene a quella domanda circa l'ottanta per cento delle volte, ma se gli si chiede chi sia il figlio di Mary Lee Pfeiffer ci prende appena un terzo delle volte. Sa il dato in una direzione e lo ignora nella direzione contraria, come se la sua memoria fosse una strada a senso unico. L'abbiamo fabbricato perché imparasse, ma non impara come avremmo supposto.

Le sonde, *probes* nella letteratura, formano la seconda famiglia. Si addestrano piccoli classificatori sulle attivazioni interne del modello per rilevare se una certa informazione è codificata lì e in quale strato. Nel 2023 Kenneth Li e la sua squadra addestrarono un modello con partite di Othello, il gioco da tavolo, scritte come semplici sequenze di mosse del tipo "E3", "D6". Il modello non vide mai una scacchiera e ciò nonostante, quando ne esaminarono le attivazioni con sonde per leggere quale informazione fosse rimasta codificata negli strati interni, trovarono una rappresentazione dello stato della partita, con l'informazione su quali caselle fossero occupate e da quale giocatore. Il sistema si era costruito un piccolo mondo interno che nessuno gli aveva chiesto. Con la stessa tecnica venne trovata una mappa (Wes Gurnee e Max Tegmark 2023). Addestrando sonde sulle attivazioni che un modello produce leggendo il nome di una città, riuscirono a prevederne latitudine e longitudine, così che nel graficare le risposte compare un planisfero riconoscibile, con i continenti al loro posto, ricostruito a partire da puro testo. Altre sonde recuperarono linee del tempo con la data di fatti e personaggi ordinati come in un fregio. La cosa interessante è che il modello non vide mai né una mappa né un almanacco. Solo leggendo parole gli è germogliata la forma del mondo.

Una sonda, tuttavia, risponde a quale informazione ci sia nelle attivazioni e in quale strato si trovi, ma non può rispondere se il modello usi effettivamente quell'informazione quando genera le sue risposte. Un dato può essere codificato in qualche angolo della rete senza intervenire nel calcolo, allo stesso modo in cui avere un libro nella nostra biblioteca non dimostra che l'abbiamo letto. Nel caso di Othello il dubbio si è dissolto, perché quando i ricercatori alterarono la scacchiera interna le mosse del modello seguirono la scacchiera alterata, ma l'obiezione resta in piedi per la maggior parte degli studi con sonde, che si fermano prima di quel passo.

Con l'interpretabilità meccanicistica, la terza famiglia, si passa all'ingegneria inversa dei circuiti. È quella che più somiglia a un'anatomia. Il suo primo trofeo furono le teste di induzione, descritte da Catherine Olsson e dai suoi colleghi di Anthropic nel 2022. Si tratta di piccoli meccanismi di attenzione che implementano una regola di copia: se la sequenza ha già contenuto la coppia A seguito da B e ricompare A, scommetti su B. Costituiscono la fonte più solida che si conosca dell'apprendimento in contesto, quella capacità dei modelli di cogliere uno schema all'interno della conversazione stessa. Possiamo vedere nascere le teste di induzione in una finestra stretta dell'addestramento, che lascia un segno visibile nei grafici. La perdita, la misura dell'errore che l'addestramento cerca di ridurre, scende di un gradino, e in quello stesso punto il modello inaugura la capacità di imparare in contesto (Olsson et al. 2022).

L'interpretabilità ha spiegato anche un vecchio problema del campo. I neuroni individuali quasi mai significano una cosa sola. Il modello ha bisogno di rappresentare molti più concetti dei neuroni che possiede, e quindi li conserva sovrapposti, come tanti cappotti in un armadio piccolo (Elhage et al.

2022). Da quella constatazione sono usciti gli autoencoder sparsi, reti ausiliarie che decomprimono l'armadio e separano caratteristiche monosemantiche, una per concetto, sia in modelli piccoli (Bricken et al. 2023) sia in modelli commerciali. Fra queste ne comparve una che si attivava davanti al ponte Golden Gate. Non rispondeva a una parola particolare, bensì al nome del ponte in qualunque lingua e a descrizioni che non lo nominavano. Rispondeva anche a fotografie, segno che catturava il concetto e non una sequenza di lettere. Allora si fissò quella caratteristica su un valore artificialmente alto, come chi alza una manopola e la lascia bloccata, mettendo il modello a conversare così. Il risultato fu un Claude monotematico che per qualche giorno del maggio 2024 rimase disponibile al pubblico. Gli si chiedeva una ricetta di torta e suggeriva ingredienti per un picnic con vista sul ponte. Gli si chiedeva della sua forma fisica e rispondeva di essere il Golden Gate in persona, nebbia inclusa. L'aneddoto è comico e la dimostrazione terribilmente seria. Se amplificando una sola caratteristica l'intera condotta del modello si riorganizza attorno al concetto corrispondente, quella caratteristica era un pezzo funzionale del sistema e ora c'è un punto di presa per muoverla. Ciò che si è fatto con un ponte si può fare, in linea di principio, con caratteristiche meno pittoresche, come l'adulazione o la disponibilità a fornire informazioni pericolose, caso in cui lo stesso gesto smette di essere divertente.

Lo stesso metodo ha prodotto una scoperta bellissima. Prendendo una rete minuscola addestrata a sommare numeri in aritmetica modulare, un compito da scuola, e aprendola per vedere come lo facesse, ci si aspettava una tabella o un trucco di memorizzazione e comparve invece la trigonometria (Nanda et al. 2023). La rete aveva imparato da sola a rappresentare ciascun numero come una rotazione su un cerchio e a sommarli componendo rotazioni con identità trigonometriche, un algoritmo che nessuno le aveva insegnato e che stava lì, in funzione, sotto il cofano di una macchina a cui si era soltanto chiesto di abbassare l'errore. Ricostruire quel procedimento è il tipo di risultato che dà senso all'analogia anatomica di questa variante dell'interpretabilità.

La quarta famiglia interviene sulle cause. L'*activation patching*, o tracciamento causale, fa girare il modello due volte con ingressi diversi e trapianta attivazioni da una passata all'altra, strato per strato, per vedere quale pezzo dell'interno causi davvero la risposta. La sonda dice che cosa c'è e il trapianto dice che cosa viene usato. Il caso più citato è ROME, *Locating and Editing Factual Associations in GPT* (Kevin Meng, David Bau e colleghi, NeurIPS 2022), che usò il tracciamento causale per localizzare dove viva un dato puntuale, l'associazione tra la torre Eiffel e Parigi, nei moduli intermedi della rete, e poi modificò quei pesi con chirurgia minima. Il modello operato rimase convinto che la torre Eiffel si trovi a Roma e rispondeva di conseguenza, con le indicazioni per arrivarci dal Colosseo. Saper modificare un ricordo è la prova più forte di averlo localizzato e apre al tempo stesso tutte le domande scomode su chi modifichi e manovri questi fili. La questione diventa inquietante nei casi in cui il rifiuto di un modello, quel "non posso aiutarti con questo" che segue al suo addestramento in sicurezza, corre lungo una sola direzione nello spazio delle sue attivazioni. Cancellata quella direzione, il modello smette di rifiutarsi di fare quasi tutto, e quando la si amplifica si rifiuta persino davanti alla richiesta più innocente (Arditi et al. 2024). Tutta la cautela appresa è condensata in un unico vettore che si alza e si abbassa come una manopola. Trovare la manopola del "no" di una macchina è indubbiamente un trionfo della comprensione e una chiave rischiosa.

E perché non chiedere al modello, visto che parla, come sia arrivato alla sua risposta, e leggere la sua catena di ragionamento come se fosse un registro di calcolo? Il problema è che non lo è. Uno studio dal titolo trasparente, *Language Models Don't Always Say What They Think* (Turpin et al. 2023), ha mostrato che si possono orientare le risposte di un modello con segnali nascosti nel prompt, facendo sì che la sua spiegazione successiva giustifichi la risposta orientata senza mai menzionare il segnale che l'ha causata, come i pazienti con il cervello diviso degli esperimenti di Michael Gazzaniga (1967), che inventavano sul momento ragioni impeccabili per condotte di cui ignoravano l'origine reale, al punto che Gazzaniga postulò un "interprete" alloggiato nell'emisfero sinistro, il cui mestiere è fabbricare spiegazioni per ciò che il corpo ha già fatto. Fabbricare ragioni plausibili sembra un talento di ogni mente verbale, carbonica o meno. Dunque la spiegazione che il modello fornisce è un dato di condotta in più, alla maniera della materia prima per la prima famiglia: la scatola nera che racconta storie su se stessa.

Abbiamo costruito una mente, o qualcosa di abbastanza simile da contenderle la parola, il nostro fantasma senza ego, e dobbiamo studiarla come una sconosciuta, con l'etologia, con le sonde, con la chirurgia e con la sana diffidenza dell'intervistatore. Il fattore è diventato l'esploratore della propria opera, un cartografo dentro un edificio la cui costruzione ha pagato. La regola di Vico, che per tre secoli è parsa di senso comune, si è rotta quando abbiamo fatto la cosa e la verità su di essa è rimasta in sospeso. Il *factum* se ne va per il mondo orfano del suo *verum* e un'intera disciplina lavora per ricongiungerli. Antifonte ha perso la sua scommessa, perché abbiamo sotterrato un artefatto nei dati ed è cresciuto come artefatto. Con tutte le virgolette del caso, sembra che abbiamo fabbricato qualcosa che somiglia curiosamente a un pezzo di natura.

◊

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. W. R. Ashby, *An Introduction to Cybernetics* (1956), sul problema della scatola nera. I. Rahwan e altri, "Machine Behaviour" (*Nature*, 2019). T. Kojima, S. S. Gu, M. Reid, Y. Matsuo e Y. Iwasawa, "Large Language Models are Zero-Shot Reasoners" (NeurIPS 2022), l'effetto del "ragioniamo passo dopo passo". L. Berglund e altri, "The Reversal Curse: LLMs Trained on 'A is B' Fail to Learn 'B is A'" (2023). K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister e M. Wattenberg, "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task" (ICLR 2023), la rappresentazione della scacchiera di Othello. W. Gurnee e M. Tegmark, "Language Models Represent Space and Time"

(ICLR 2024), la mappa e la linea del tempo recuperate con le sonde. C. Olsson e altri (Anthropic), “In-context Learning and Induction Heads” (2022). N. Elhage e altri (Anthropic), “Toy Models of Superposition” (2022). “Towards Monosemanticity: Decomposing Language Models with Dictionary Learning” (Anthropic, 2023) e “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet” (Anthropic, 2024), gli autoencoder sparsi e la caratteristica del Golden Gate. N. Nanda, L. Chan, T. Lieberum, J. Smith e J. Steinhardt, “Progress Measures for Grokking via Mechanistic Interpretability” (ICLR 2023), la rete che reinventò la trigonometria per sommare. K. Meng, D. Bau, A. Andonian e Y. Belinkov, “Locating and Editing Factual Associations in GPT” (ROME; NeurIPS 2022). A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee e N. Nanda, “Refusal in Language Models Is Mediated by a Single Direction” (NeurIPS 2024). M. Turpin, J. Michael, E. Perez e S. R. Bowman, “Language Models Don’t Always Say What They Think” (NeurIPS 2023). M. S. Gazzaniga, “The Split Brain in Man”, *Scientific American*, 1967.