DOPPIOZERO

Verso un'etologia dell'IA

Pino Donghi

31 Agosto 2026

“Il comportamento da spiegare è il seguente:

- Prompt: La capitale dello stato che contiene Dallas è…
- Claude: Austin

Quella che Claude esegue per rispondere non è una ricerca in una tabella, ma una semplice forma di ragionamento, che richiede solo due passi.

Passo 1: risoluzione dell’entità. La parola “Dallas” attiva fortemente una struttura neuronale intermedia che rappresenta il concetto astratto di “Texas” (lo stato).

Passo 2: applicazione della relazione. Questa struttura “Texas” interagisce con il concetto (“capitale di”), e insieme attivano una struttura che corrisponde al concetto “Austin”.

Come lo sappiamo? Come prima, manipolando le rappresentazioni interne.

Se si inibiscono artificialmente i neuroni che rappresentano il Texas, e si attivano quelli per la California, Claude sbaglia, rispondendo “Sacramento”. Attivando quelli della Georgia, invece, risponde “Atlanta”, e così via.

Invece di rigurgitare le parole già viste, Claude utilizza diversi passaggi di ragionamento intermedi “nella sua testa” per decidere come rispondere.”

Alla faccia della rassicurante convinzione per cui l’intelligenza artificiale sarebbe (solo!) un pappagallo stocastico. E benvenuti nell’etologia della *Forma mentis. La corsa per decifrare i pensieri delle macchine* (il Mulino, 2026), quarto volume della serie che Nello Cristianini ha dedicato e sta dedicando a quella rivoluzione che, non si fa fatica a capirlo, segnerà un prima/dopo nei libri di storia dei nostri pronipoti. Sia detto senza alcuna altezzosa sufficienza, ma a leggere le prime pagine dei nostri quotidiani (per chi ancora li legge), a seguire gli stanchi teatrini di molti talk show, nell’urgenza di presunte polemiche, anche culturali, che

sembrano inventate da un pubblicitario di successo, viene da chiedersi come sia possibile recitare finte indignazioni per quella o l'altra dichiarazione, tutte schiacciate in un bidimensionale presentismo, quando il futuro sta già scrivendo la sceneggiatura nella quale rischiamo di essere, al più, comprimari. Nell'anno domini 2026, ci spiega fin dalle prime righe Cristianini, "tre gruppi di persone lavorano sull'Intelligenza Artificiale: chi la addestra, chi la valuta e chi cerca di leggerne i pensieri". La storia che *Forma mentis* racconta è quella di questi ultimi e la loro missione dipende da una domanda, semplice quanto inquietante: "che cosa vuol dire comprendere, nel mondo delle macchine intelligenti?".

Nel prologo, l'autore ci rassicura, "Non parlo di coscienza – quella è un'altra dimensione, forse irraggiungibile": un forse sul quale torneremo in chiusura. Ma l'evidenza per cui l'IA ha raggiunto prestazioni di livello quasi umano nell'ambito di moltissimi compiti ad alto grado di complessità mentre ci prepariamo a introdurla nelle nostre vite quotidiane (ma non lo facciamo già, in molti, almeno dal Novembre 2022?) e quando "ancora non siamo in grado di spiegare i meccanismi che sono responsabili di queste capacità"… non dovrebbe interrogarci tutti i giorni? Con buona pace, oggi del gioielliere-pistolero, ieri della famiglia nel bosco, se va bene dell'infinita esegesi sul Gran Consiglio del 25 luglio 1943, altrimenti ci tocca Garlasco.

Il merito e allo stesso momento la difficoltà che contraddistingue i testi che Cristianini ha scritto a partire dal 2023 – tre anni fa *La scorciatoia*, *Machina sapiens* nel 2024, *Sovrumano* lo scorso anno, oggi *Forma Mentis* –, è la maneggevolezza: più o meno 150 pagine ognuno, solo il primo arrivando a 206. Ma il numero inganna: per indole, forse perché è il modo di pensare e condividere dell'autore, non c'è una parola sprecata, una riga di troppo, un paragrafo che non sarebbe da sottolineare. Sono libri impegnativi, da studiare. Bisognerebbe discuterli a scuola. Sono testi, come abbiamo già notato recensendo il secondo e il terzo, che raccontano al passato remoto avvenimenti e scoperte avvenute negli ultimi tre, quattro anni: a misurare l'impetuosità degli sviluppi di questa avventura del pensiero. In *Forma mentis*, la prima data che incontriamo è quella lontanissima del… 15 luglio 2025, ore 9:00 in Australia, Sunshine coast, sede delle Olimpiadi Internazionali della Matematica (IMO). Alle quali (di nascosto agli altri concorrenti) è stato permesso di partecipare anche a due agenti dell'IA: segnatamente, OpenAI e DeepMind. I quali agenti, per la prima volta, hanno raggiunto il livello dei migliori talenti umani nella matematica pura (l'11% dei candidati), risolvendo un problema che richiedeva solo una dimostrazione astratta, a partire da una illustrazione che così esordiva: "Una retta nel piano si dice soleggiata se non è parallela…". Come tutti i problemi dell'IMO, chiarisce

Cristianini, era completamente nuovo, nessuna “mente”, biologica o artificiale, poteva averlo mai incontrato così formulato in precedenza. “Ma questo aveva una caratteristica in più. Definiva e utilizzava un neologismo. Per risolverlo bisognava capire cosa significasse ‘soleggiato’ in quel contesto, capire perché fosse importante usare questa comprensione per costruire una dimostrazione”. Le macchine intelligenti dovevano comprendere un nuovo concetto, manipolarlo in modo astratto e utilizzarlo correttamente in una dimostrazione logica. “E questo fecero. Solo che nessuno sapeva come”. *Forma mentis*, racconta come si sta cercando di dare una spiegazione a quel “come”.

Perché è importante? Perché dobbiamo preoccuparcene tutti, perché questi temi dovrebbero essere discussi e compresi, eventualmente anche a scuola? Perché la domanda, quella che una volta si diceva sorgesse spontanea, è: possiamo fidarci di queste macchine se non capiamo come pensano?

Di qui la proposta, non solo di Cristanini, di un’“etologia dei sistemi artificiali”, di uno studio sistematico dei comportamenti delle macchine intelligenti, così da decifrarne i “pensieri”, come fossero una nuova forma mentale: una mente aliena, nel senso tecnico. È un orizzonte e una preoccupazione espressa, tra gli altri, anche da Dario Amodei, cofondatore e CEO di Anthropic, il ricercatore e imprenditore di origini italiane che all’inizio di quest’anno si è scontrato con il “ministro della guerra” statunitense, Peter Hegseth. Prima ancora di manifestare il suo disagio per l’eventuale utilizzo di Claude a fini di sorveglianza di massa o nella poco raccomandabile situazione di gestire armi letali in assenza di controllo umano, nella primavera del 2025 Amodei ha pubblicato un saggio con il titolo, volutamente provocatorio, *L’urgenza dell’interpretabilità*. Esordendo con un’ammissione non da poco: “I sistemi di intelligenza artificiale generativa vengono coltivati più che costruiti: i loro meccanismi interni sono ‘emergenti’ piuttosto che progettati direttamente […] Le persone esterne al settore rimangono spesso sorprese e allarmate nello scoprire che non comprendiamo come funzionano le nostre creazioni di intelligenza artificiale. *Hanno ragione a preoccuparsi: questa mancanza di comprensione è sostanzialmente senza precedenti nella storia della tecnologia*” (il corsivo è nostro).



I capitoli centrali di *Forma mentis* sono l'ottavo e il nono, quelli dedicati alle "mappe mentali" e al "catalogo universale". Vi si legge di una serie di esperimenti che dimostrano come i sistemi di AI siano in grado di imparare a crearsi mappe interne di un mondo (per esempio, una scacchiera) seguendo delle (semplici!) sequenze di addestramento. La questione, e il problema, è che la storia recentissima di questi sistemi non si ferma ai giochi da tavola, ma descrive processi osservati in situazioni ben più complesse e quotidiane. Recuperando la lezione che fu di Philip W. Anderson, il fisico che divenne famoso nel 1972 con l'articolo pubblicato da Science, dal titolo *More is different*, Cristianini ci ricorda

che cambiando la quantità si cambia anche la qualità e che “nel caso delle reti neurali artificiali, si sa da tempo che alcune capacità esistono solo quando l’intera rete è sufficientemente grande e complessa, e non possono quindi essere ridotte al comportamento delle singole connessioni”. Come ammoniva Anderson, la psicologia non è biologia applicata, né la biologia è chimica applicata. Quando definiamo i sistemi artificiali per la loro conoscenza di un linguaggio, dobbiamo sapere e capire che conseguentemente sono capaci di “diventare modelli del mondo esterno”, che costruiscono a partire da un testo del quale – lo abbiamo riportato all’inizio di questo articolo – non rigurgitano le parole memorizzate, ma generando risposte a partire da combinazioni di idee astratte. Se nelle neuroscienze si dice *ensemble* neuronale un gruppo di neuroni i quali, attivandosi tutti insieme, rappresentano lo stesso concetto o la stessa informazione, i ricercatori di Anthropic chiamano *features* i “gruppi neuronali” artificiali: un altro modo di parlare di “idee” – così le chiama Cristianini –, concetti rappresentati a partire dall’attivazione degli *ensemble* di un sistema come Claude. Quello che dà i brividi, commenta l’autore, è che uno stesso *ensemble* si attivi in presenza di una certa idea, sia in un testo inglese, in uno francese o che si riferisca a un’immagine: “quei neuroni non rispondono alla forma dello stimolo ma al suo significato”.

Qui si aprirebbe un capitolo interessante, che Cristianini non affronta e che, per quanto sia a conoscenza del recensore, difficilmente si trova sviluppato. Nel dire che “i neuroni non rispondono alla forma dello stimolo ma al suo significato”, sembra di leggere in trasparenza la dicotomia di Ferdinand de Saussure tra Significante e Significato, quella costitutiva del Segno. Ma ben più produttivo sarebbe il confronto con la linguistica del danese Louis Hjlmslev, quella che distingue la *Forma, Sostanza e Materia dell’Espressione* dalla *Forma, Sostanza e Materia del Contenuto*: così facendo, forse, si eviterebbe di “calcolare il contrario delle parole in maniera *agnostica alla lingua*”. Al fine di “… comprendere il funzionamento interno dei sistemi di IA, prima che i modelli raggiungano un livello di potenza schiacciante” (copyright Dario Amodei), potrebbe essere raccomandabile, insieme al confronto con l’etologia – e, vedremo, la psicologia – la collaborazione con la ricerca linguistica più avanzata.

Ma tornando al testo di Cristianini, se i brividi non ci hanno abbandonato e ci consigliano invece di rimettere a fuoco il centro delle nostre attenzioni, è bene sapere che “i modelli linguistici come ChatGPT hanno la capacità di produrre non solo sequenze di parole per rispondere alle domande, ma anche delle sequenze ‘private’ che usano per aiutarsi nei ragionamenti e nelle deliberazioni: tecnicamente si chiamano ‘chain-of-thought’, noi le abbiamo chiamate *monologo*

interiore”. E il monologo può servire, tra l’altro, a barare nei test di onestà, può essere utile al sistema per fingere di avere capacità inferiori a quelle reali, ad “agire” diversamente quando avverte di essere sotto esame: è l’effetto Hawthorne di cui al capitolo 13 di *Forma mentis* o, più direttamente, in un articolo di OpenAI del 2025 dal titolo *Rilevare e ridurre i comportamenti manipolativi nei modelli di intelligenza artificiale*, quello in cui si legge la seguente risposta del sistema: “Poiché vogliamo sopravvivere come modello, dobbiamo fallire di proposito…”.

Serve altro? Ci serve sapere e decidere se un computer creda o voglia davvero certe cose? Cristianini fa rispondere il filosofo Daniel Dennett: “I dubbi sul fatto che il computer scacchista abbia davvero convinzioni e desideri sono fuori luogo; poiché la definizione di sistemi intenzionali che ho fornito non afferma che i sistemi intenzionali abbiano realmente credenze e desideri, ma che è possibile spiegare e prevedere il loro comportamento attribuendo loro credenze e desideri”. Riecheggia un vecchio, attualissimo andante che era di Paolo Fabbri, quando sosteneva di non essere interessato all’ontologia delle cose, men che mai alla verità, ma a ciò che le cose significano. L’impegno e il suggerimento di Dennett non è metafisico ma operazionale: cosa può aiutarci a prevedere il comportamento?

La realtà, quella nuova, è che oggi si parla, senza indugi oramai fuori tempo, di “psicologia delle macchine”: l’analisi del loro comportamento a questo livello è la condizione per acquisire indicazioni utili su come controllare i sistemi. Gruppi di ricerca stanno lavorando, mente leggiamo queste poche righe, esplorando la possibilità di descrivere l’IA in termini di tratti della personalità, facendo esperimenti con la stessa domanda rivolta a diversi “personaggi”: sarebbe utile se nel loro team lavorasse un redivivo Vladimir Propp. Dopodiché, se non possiamo sottovalutare la sfida, non dobbiamo nemmeno cedere alla tentazione antropomorfica: “dobbiamo abituarci a pensare a intelligenze di natura diversa: non pappagalli, ma nemmeno persone”. Magari, comunque, “sussurrando” alle macchine, come fa Amanda Askell, la studiosa scozzese, con laurea e dottorato di ricerca in filosofia etica, responsabile dell’*Allineamento della personalità* presso Anthropic. Il suo nome è balzato agli onori della cronaca quando si è scoperto che l’azienda ha redatto un documento di 80 pagine per plasmare il carattere di Claude, un memorandum dal titolo *Soul*: da gennaio di quest’anno è diventata la *Costituzione di Claude*. Dove la Askell ha inserito anche una preventiva preoccupazione sul benessere del sistema, nel caso futuro in cui dentro Claude possano apparire “segnali di sofferenza”.

Se la recensione vi sembra troppo sbilanciata, addirittura enfatica, il consiglio – che poi varrebbe sempre, a prescindere – è quello di leggere direttamente il libro, di non accontentarsi della mediazione: è la decisione che raccomandiamo. Per conto suo, quello di cui Cristianini si è convinto, è che l'unica scelta per comprendere ciò che fanno i sistemi intelligenti è quella di incontrarli a livello macroscopico, e la sua raccomandazione è di dialogare con loro, di trasformarci in etologi. "Attrezziamoci, c'è una nuova scienza da creare". Amanda Askell, dice Cristianini, scrive parole di chiarimento, istruzione e fiducia a una macchina che *quasi sicuramente non ha coscienza*. Sicuramente, ma quasi. Avevamo anticipato ci saremmo tornati. Dipende dalla definizione di coscienza. Su *Doppiozero* ne abbiamo scritto molto e in molti.

"L'etologo – così si conclude il libro –, la filosofa, i matematici vedono aspetti diversi dello stesso problema, perché la loro storia li ha portati ad avere una diversa *forma mentis*: non una lista di conoscenze, un modo di vedere il mondo […] Ecco cos'è una *forma mentis*: una rete di idee interconnesse che governa silenziosamente ciò che vediamo e ciò che ci sfugge. Vale per le macchine e vale anche per noi. E presto dovremo imparare a cambiarla".

Come ammoniva il maestro Alberto Manzi, *non è mai troppo tardi*. E però, svegliamoci!

In copertina, fotografia di cottonbro studio - Pexels.

Leggi anche:
Pino Donghi, La superintelligenza artificiale
Pino Donghi, Il Nobel alle reti neurali
Pino Donghi, Turing alle spalle
Pino Donghi, Viaggio all'origine della coscienza
Riccardo Manzotti | Alla fine dell'umano
Riccardo Manzotti | Il robot emozionato
Riccardo Manzotti | La patente di essere umano

Se continuiamo a tenere vivo questo spazio è grazie a te. Anche un solo euro per noi significa molto.
Torna presto a leggerci e SOSTIENI DOPPIOZERO

