
Intelligente a chi?

La progettazione dei sistemi dotati di intelligenza artificiale (IA) fa ricorso a metodi induttivi basati sulla grande quantità di dati resa disponibile in Internet, piuttosto che a metodi istruttivi come avveniva fino alla fine degli anni Ottanta. Gli attuali algoritmi induttivi mostrano, accanto a notevoli differenze, alcune interessanti somiglianze con il funzionamento dell'intelligenza umana (IU).

Fabio Fantini

Il tentativo di produrre macchine capaci di svolgere lavori intellettuali di alto livello iniziò intorno alla metà del secolo scorso e ha avuto un percorso irregolare. Fino alla fine degli anni Ottanta gli sforzi degli esperti seguivano una strategia top-down: per riprodurre o quanto meno emulare comportamenti umani, occorreva “distillare” da questi comportamenti i principi essenziali, trasformarli in regole logiche e programmare le macchine secondo le regole. Una volta fornito un insieme sufficientemente ampio di regole precise, la macchina avrebbe eseguito le azioni richieste. Questo approccio si scontrò presto con difficoltà insormontabili nel tentativo di svolgere compiti come il riconoscimento del linguaggio parlato, la traduzione da una lingua all'altra e l'analisi delle immagini. Il riconoscimento del linguaggio parlato, per esempio, richiedeva regole grammaticali per giudicare se una frase fosse plausibile, ma le regole non bastavano mai e ogni regola ne richiedeva un'altra. Anziché risolvere il problema, grammatiche sempre più estese continuavano a fornire risultati deludenti.

Chiunque oggi guardi un video su Youtube potrebbe chiedersi dove siano finite queste difficoltà: in molti casi basta premere il tasto “c” per ottenere sottotitoli che automaticamente riproducono, con ottima approssimazione, il parlato del video. Se ci si chiede come sia stato possibile avviare a quella che trenta anni fa appariva come una limitazione insuperabile, la risposta è lapidaria: rovesciando la strategia impiegata per istruire le macchine. Anziché fornire un insieme sempre più vasto e articolato di regole da impiegare in una casistica polverizzata con una miriade di specificità particolari, si



sono addestrate le macchine a imparare le regole logiche dagli esempi loro sottoposti, secondo una logica bottom-up.

Il cambio di strategia si è appoggiato su un arsenale matematico che consente di estrarre relazioni complesse da sterminate quantità di dati e di valutarne il potere predittivo, con una retroazione capace di perfezionare rapidamente il procedimento algoritmico. E c'è un'altra ovvia gamba su cui la rivoluzione nell'approccio all'Intelligenza Artificiale (IA) è proceduta: lo sviluppo di Internet ha consentito di disporre di quantità crescenti di dati "esperienziali", la materia prima per rendere efficace l'azione istruttiva dal basso verso l'alto. Ogni nostra fruizione di un social media, di un motore di ricerca, di uno streaming multimediale arricchisce la base di dati che rappresenta l'esperienza dei sistemi di IA. Impeccabili agenti di "raccomandazione", questi sistemi di IA sono pronti a suggerirci nuovi contatti e contenuti, il completamento della parole o della frase appena iniziata, video o brani musicali che hanno una elevata probabilità di risultarci graditi (Fig. 1). E c'è di più, perché la traccia delle dichiarazioni o azioni pubbliche di ciascun utente, come i gradimenti espressi dai "mi piace" o i siti visitati, consente di tracciarne un profilo probabilistico dei tratti psicometrici e di conseguenza della personalità¹.

Nel momento in cui scrivo queste righe, il programma di IA più noto e oggetto di dibattito si chiama ChatGPT, dove la sigla GPT sta per Generative Pre-trained Transformer. Transformer è il nome di una rete neurale artificiale stratificata che, istruita secondo il metodo di pre-addestramento generativo, apprende a prevedere quale sarà la prossima parola nell'ambito di una frase o di un insieme di frasi. La popolarità di ChatGPT, efficacemente indicata dal raggiungimento di 10⁸ utenti a inizio febbraio 2023, è dovuta alla capa-

cità di produrre testi completamente plausibili su una ampia gamma di argomenti, con una qualità spesso indistinguibile da quella di testi analoghi prodotti da soggetti umani.

Di fronte a questi risultati il dilemma dell'osservatore spassionato è se si tratti del raggiungimento da parte di una macchina di vette intellettuali fino ad ora riservate alla nostra specie oppure del disvelamento di meccanismi propri di attività intellettuali ritenute prerogativa esclusiva ed insondabile della specie umana.

				
				
				
				
				
				

Fig. 1 - Un agente di raccomandazione agisce sulla base dell'assunto che persone che hanno mostrato simili preferenze in passato continueranno a mostrarle, con correlazione statisticamente significativa, anche in futuro. Lo schema rappresenta il principio in base al quale un agente di raccomandazione proporrà (o non proporrà) un certo contenuto a un dato utente. Si tratta di uno schema enormemente semplificato, perché gli agenti di raccomandazione hanno accesso a informazioni personali di miliardi di utenti in relazione a milioni di prodotti (da https://en.wikipedia.org/wiki/Recommender_system).

1. Nello specifico: Kosinski M, Stillwell D, Graepel T, *Private traits and attributes are predictable from digital records of human behaviour*, Proceedings of

the National Academy of Sciences, 110, 15:5802-5805, 2013. Per un approccio più generale: Cristianini N, *La scorciatoia*, Bologna, Il Mulino, 2023.

Quando nel 1950 Turing si pose il problema dell'intelligenza delle macchine computazionali, se la cavò con una risposta (che qui non riporterò, al fine di non contribuire a un incontrollabile processo inflattivo) sintetica e apparentemente efficace, dall'indubbio merito di essere operativa. Era un po' come chiedersi se una trave avrebbe retto un certo peso e, per avere una risposta, mettere sulla trave un corpo con quel peso e osservare che cosa succedeva.

Geniale in tante cose, in questo caso Turing fece, però, un po' la stessa figura di quel produttore cinematografico che, appena pochi anni prima, si era esibito nella poco lungimirante dichiarazione che la televisione non sarebbe durata, perché le persone si sarebbero presto stancate di stare davanti a una scatola di compensato ogni sera. Eravamo nel 1950, però, e chiunque avrebbe risposto con un risolino di sufficienza alla domanda se pensasse di lasciarsi ingannare da un ENIAC da 30 tonnellate (Fig. 2), surriscaldato nel problematico tentativo di accreditarsi, attraverso un dialogo a distanza, come un umano intelligente. Nel 2023 categorie di lavoratori intellettuali che si pensavano al riparo dei rischi dell'automazione scoprono di essere sul punto di essere sostituiti

da macchine capaci di fare il loro stesso lavoro e quindi, per definizione, "intelligenti".

Noi umani stentiamo a concordare su una definizione di intelligenza, anche se concordiamo senza difficoltà nel definirci dotati di Intelligenza Umana (IU). Una definizione di largo uso, dovuta a James Albus², un ingegnere che lavorò a lungo alla comprensione dei meccanismi computazionali e rappresentativi del cervello umano, propone di considerare l'intelligenza come la capacità di un sistema di agire in modo appropriato in un ambiente incerto, dove le azioni appropriate sono quelle che aumentano le probabilità di successo. Tralascierò per il momento di considerare che cosa si intenda per "successo" per sottolineare, piuttosto, il fatto che, seguendo la definizione, siamo immersi in un mondo di intelligenza diffusa, una qualità in possesso di sistemi così disparati come una quercia secolare, un micelio fungino esteso nel sottosuolo, un polpo, un umano, una colonia di insetti sociali. In definitiva, si potrebbe pensare all'intera biosfera come un sistema intelligente, formato da sottoinsiemi a loro volta intelligenti. Torno rapidamente al termine "successo" della definizione di Albus. Il successo di un'azione può

essere valutato solo in funzione di uno scopo; pertanto, quale che sia lo scopo, il concetto di intelligenza può applicarsi solo a sistemi teleologici. Nel campo dei viventi, la selezione naturale ha favorito gli organismi (non necessariamente individui) con maggiore successo evolutivo, cioè capaci di originare una prole più numerosa di al-



Fig. 2. ENIAC (*Electronic Numerical Integrator and Computer*) fu il primo computer elettronico digitale programmabile. Completato nel 1946, occupava uno scantinato di 15 m per 9 m nella Moore School of Electrical Engineering della Università di Pennsylvania (da <https://www.britannica.com/technology/ENIAC>).

2. Albus J S, *Outline for a theory of intelligence*, IEEE Transactions on Systems, Man, and Cybernetics, volume 21, issue 3:473-509, 1991.

tri. In una visione estrema, ma anche un po' limitata, e che tuttavia bene esemplifica il concetto di "scopo", il successo evolutivo, scopo dei sistemi viventi, è aumentare quanto possibile la presenza di copie dei propri geni nel serbatoio genico di una popolazione. È bene notare che qui non si parla di scopo consapevole, ma solo di azioni la cui esecuzione rende possibile la permanenza temporale, sia pure attraverso un periodico azzeccamento delle individualità e il riciclo dei materiali costituenti, di sistemi dotati di caratteristiche simili perché "ereditate".

La consapevolezza di cui, come umani, ci diamo reciprocamente credito ci porta a considerare le prestazioni intellettuali di sistemi artificiali come GPT su un piano differente rispetto alle nostre. Se, per molti aspetti, questa differenza è bene evidente, esistono anche caratteristiche che accomunano l'IU all'IA. Ne considererò due.

Una prima caratteristica riguarda il fatto che le macchine apprendono associazioni di parole da testi scritti prodotti da umani, e queste associazioni riflettono gli atteggiamenti, spesso impliciti, espressi dagli umani. Si può trattare di associare piacevolezza a panorami naturali e sgradevolezza alle zanzare, per esempio, e qui magari nessuno avrebbe da ridire. Ma si verificano anche associazioni prevalenti meno accettabili, per esempio tra nomi femminili e cura della casa, mentre i nomi maschili sono associati allo sviluppo della carriera. O, ancora, la predisposizione a delinquere è correlata in modo differenziale al quartiere di residenza, all'etnia, alla religione.

Queste distorsioni cognitive pregiudiziali non sono espresse esplicitamente nel materiale prodotto dall'IA, ma hanno influenze statisticamente significative nei testi sfornati e nei giudizi emessi. Rischiamo di avere a che fare con una IA strumento involontario di perpetuazione delle discriminazioni basate su genere, razza, età, etnia. La causa di tutto ciò non è una perversa inclinazione

dell'IA, ma risiede nei corpora testuali forniti per l'addestramento³. Questi corpora testuali sono ricavati da Internet, contengono immancabilmente riferimenti alle attitudini e agli stereotipi umani e non sono filtrati per garantire la loro correttezza etica. In altre parole, l'addestramento dell'IA è condotto in prevalenza sulla base di ampie basi di testi rappresentativi del sentimento medio dell'umanità che fa uso di Internet e le distorsioni pregiudiziali ne sono il logico portato. Il desiderio di una IA moralmente ineccepibile, migliore di noi, è destinato a rimanere tale (un po' come il miraggio di infrangere la seconda legge della termodinamica), finché la addestriamo con i materiali che noi stessi produciamo.

In secondo luogo, l'apprendimento linguistico nell'IA ha forti analogie con quello che si verifica nei neonati umani. I piccoli di *Homo sapiens* apprendono il linguaggio attraverso un complesso processo di ascolto e di riproduzione sonora, ripetuti tentativi di imitazione basati su suoni che l'esperienza indica associati, correzioni continue accompagnate da arricchimento lessicale. Il processo di apprendimento umano somiglia a quello di una rete neurale a pre-addestramento generativo, se non nei meccanismi, sui quali non siamo ancora in grado di indagare con precisione, almeno nel metodo: induzione pura. E non mi dilungherò su come le distorsioni pregiudiziali si perpetuino, tra gli umani, da una generazione all'altra, perché le analogie con quanto ho appena ricordato nei capoversi precedenti sono sufficientemente evidenti, così come il ruolo dei processi induttivi.

Forse due differenti punti di vista, cioè se le macchine stiano diventando intelligenti oppure se la nostra tanto decantata intelligenza sia in buona parte frutto di induzione statistica, hanno qualche ragione. Ciò che per ora emerge è l'interazione tra i campi di studio di IA e IU, insieme alla difficoltà di tracciare un confine netto tra i due. ●

3. Caliskan A, Bryson J J, 356, 6334:183-186, 2017.
Narayanan A, *Semantics derived automatically from language corpora contain human-like biases*, Science,