Una breve introduzione ai sistemi di intelligenza artificiale nella prospettiva della metrologia
Luca Mari, Francesco Bertolotti, Dario Petri

Tutto_Misure, 3, 2023

[2.8.23]
Nello scorso numero di Tutto_Misure, questa rubrica è stata dedicata a un "dialogo con un agente artificiale a proposito di qualche argomento di metrologia". Che oggi quello dei *chatbot*, cioè dei sistemi automatici (*bot*, da "*robot*", termine coniato dal drammaturgo ceco Karel Čapek nel 1920) capaci di conversazione (*chat*), sia un argomento caldo è evidente anche solo dalla frequenza con cui sono pubblicate notizie su ChatGPT di OpenAI, Bard di Google, Claude di Anthropic, Llama (e i suoi derivati) di Meta, e così via. Riprendendo l'invito con cui il Direttore di Tutto_Misure concludeva il suo editoriale, "AI e metrologia", dello scorso numero della rivista – "Amici metrologi, c'è nuovo lavoro per voi!" – proponiamo qui qualche considerazione preliminare sulla cosiddetta "intelligenza artificiale" e sulle sue possibili relazioni con la metrologia, ben consapevoli che quanto segue sono solo cenni, che potrebbero essere più e meglio sviluppati in prossimi articoli.

Il nostro punto di partenza è che, con le loro evidenti differenze, intelligenza artificiale e metrologia sono ambiti di conoscenza che possono utilmente interagire, e beneficiare l'una dell'altra. Dato l'obiettivo pragmatico di dotarci di strumenti che ci aiutino a prendere decisioni affidabili di fronte a problemi complessi, la metrologia ha qualcosa da insegnarci tutte le volte che i problemi hanno a che vedere con proprietà empiriche, a proposito delle quali occorre acquisire e rendere disponibile informazione a sua volta affidabile perché sufficientemente oggettiva e intersoggettiva. La consapevolezza che in questo processo sono necessari dei modelli interpretativi (per esempio nell'applicazione alle letture degli strumenti di misura delle informazioni ottenute con la taratura degli stessi strumenti) ci ricorda che ogni misurazione include, esplicitamente o implicitamente, delle attività in cui occorre elaborare informazione.
Per l'elaborazione di informazione i sistemi di intelligenza artificiale ci offrono *un terzo paradigma*, a complemento e possibile integrazione del paradigma algoritmico e di quello funzionale. Uno sguardo retrospettivo ci mostra infatti che, tralasciando le esagerazioni del marketing e delle mode, i sistemi che da oltre cinquant'anni sono stati chiamati di "intelligenza artificiale" sono effettivamente caratterizzati da un comportamento che non è programmato, e perciò non è il risultato dell'esecuzione di algoritmi o di funzioni specificati esplicitamente e implementati mediante linguaggi di programmazione. In un contesto in cui "sistema software" e "programma" sono spesso considerati sinonimi, è perciò davvero in gioco un diverso paradigma concettuale, in cui si cerca di ottenere da un sistema tecnologico un certo comportamento senza programmarlo in modo esplicito (e per questo motivo non pare una buona idea usare l'espressione "algoritmo di intelligenza artificiale", pure impiegata frequentemente per descrivere ciò che produce il comportamento osservato dei sistemi di intelligenza artificiale, anche se è vero che tali sistemi funzionano anche sulla base di algoritmi opportunamente implementati).
Nel corso del tempo, questo terzo paradigma – dunque ciò che propriamente potremmo chiamare "intelligenza artificiale" – è stato realizzato attraverso due strategie alternative. Pur senza entrare in dettagli, ne presentiamo qui le basi concettuali.

Una prima strategia, tradizionalmente chiamata di intelligenza artificiale *simbolica* o *basata su regole*, prevede la costruzione di "basi di conoscenza" in cui le informazioni disponibili su un certo ambito applicativo sono tipicamente rappresentate mediante regole condizionali della forma [se X, allora Y]. Per esempio, due regole potrebbero essere [se si osserva il sintomo A, allora occorre effettuare l'esame B, di cui acquisire l'esito C] e [se l'esito dell'esame B è C1, allora decidi D1]. Un "motore inferenziale", anch'esso parte costitutiva del sistema, è in grado di concatenare queste regole per produrre un ragionamento a supporto della decisione cercata: se si informa il sistema che si è osservato il sintomo A e che l'esito

dell'esame B è stato C1, il sistema stesso suggerisce la decisione D1, e ciò nonostante questo comportamento, cioè questa relazione tra input e output, non fosse stato specificato come tale.
Negli scorsi decenni, ma soprattutto fra la seconda metà degli anni '80 e la prima metà degli anni '90, la strategia simbolica è stata impiegata variamente nella realizzazione dei cosiddetti "sistemi esperti", con basi di conoscenza composte da centinaia o migliaia di regole e con estensioni sofisticate per la gestione di informazione incerta o vaga (generalmente in termini probabilistici e di logiche sfumate rispettivamente). Alla prova dei fatti, questa strategia si è però rivelata poco efficace, soprattutto a causa degli alti costi da sostenere per popolare le basi di conoscenza attraverso complessi processi di elicitazione della conoscenza di esperti. Con poche eccezioni, non è a sistemi simbolici a cui ci si riferisce quando oggi si parla di intelligenza artificiale (anche se è ad essa che ci si riferisce quando si accusa l'intelligenza artificiale di essere solo un mucchio di istruzioni *if-then-else*).

La seconda strategia per controllare il comportamento di un sistema tecnologico senza programmarlo esplicitamente è tradizionalmente chiamata di intelligenza artificiale *subsimbolica*, in riferimento al fatto che l'informazione che viene fornita al sistema per metterlo in grado di esibire il comportamento desiderato non viene codificata in modo esplicito (come invece è nel caso della prima strategia, in cui l'informazione è appunto codificata in regole condizionali), ma in una modalità "di più basso livello".
Benché nel passato siano state sperimentate modalità diverse di implementazione subsimbolica (per esempio gli algoritmi genetici, tecniche ispirate all'evoluzione naturale usate per identificare soluzioni ottimali attraverso selezione, incrocio e mutazione casuale dei parametri che le compongono), oggi questa strategia è realizzata soprattutto attraverso sistemi di apprendimento automatico (*machine learning*) strutturati come reti neurali artificiali.
In termini sufficientemente astratti, si può interpretare una rete neurale artificiale semplicemente come una funzione parametrica, $Y = f_P(X)$, la cui complessità è tale da non consentire di trovare per via analitica i valori dei parametri P per produrre il comportamento $X \rightarrow Y$ atteso. Non c'è, insomma, una formula con cui calcolare i valori ottimi di P (come invece accade per esempio nel caso della regressione lineare, in cui $Y = p_0+p_1X$ e i valori di p_0 e p_1 si possono calcolare con il metodo dei minimi quadrati). Tali valori vengono perciò "imparati" attraverso un processo di "addestramento" (*training*) della rete, che consiste, sommariamente,
(1) nel costruire un "insieme di addestramento" (*training set*) di coppie (x_i, y_i), in cui y_i è la risposta attesa all'informazione in ingresso x_i,
(2) nell'assegnare dei valori iniziali ai parametri della funzione,
(3) nel calcolare $y'_i = f_P(x_i)$ e confrontare il valore calcolato y'_i con il valore atteso y_i,
(4) se y'_i non è abbastanza simile a y_i, nel modificare in qualche modo i valori dei parametri della funzione, per cercare di rendere y'_i più simile a y_i,
e ripetendo i passi (3) e (4) di questo processo fino a che i valori dei parametri non sono stati adattati in modo che la rete approssimi sufficientemente bene il comportamento $X \rightarrow Y$ ricercato (di principio, l'unica cosa non ovvia di questa procedura è, nel passo 4, l'aggiornamento dei valori dei parametri, realizzato "propagando all'indietro" (*back propagation*) sui parametri l'informazione sulla differenza tra valore calcolato y'_i e valore atteso y_i mediante un algoritmo di discesa lungo il gradiente della funzione di errore).
Come mostra il caso delle regressioni polinomiali, se un comportamento semplice può essere ottenuto con una funzione con pochi parametri (per esempio un comportamento lineare con due parametri, come ricordato sopra), un comportamento complesso di solito richiede molti parametri. Peraltro, il fatto che i chatbot, che sono realizzati attualmente come reti neurali artificiali e dunque in accordo alla seconda strategia, abbiano in ingresso X e uscita Y dei testi, e non dei numeri, non deve confonderci: prima di ogni altra elaborazione, i testi in ingresso vengono convertiti in liste di vettori numerici, che vengono riconvertiti in testi come ultimo passo dell'elaborazione stessa. Il comportamento che osserviamo negli attuali chatbot è palesemente assai complesso, come è immediato comprendere considerando che con ChatGPT si può avere una conversazione praticamente in qualsiasi lingua su praticamente qualsiasi argomento. Quindi, per assegnare un valore ai

molti parametri necessari per ottenere il comportamento complesso desiderato, occorre ripetere molte volte la procedura di addestramento accennata sopra: questo richiede la disponibilità di un training set di grandi dimensioni, che nel caso dell'addestramento di un chatbot può essere costituito non da coppie (x_i, y_i), ma solo da testi x_i, grazie all'ingegnosa tecnica, chiamata di "addestramento auto-supervisionato" (*self-supervised training*) di dividere ogni testo in due parti, in cui la prima è l'input e la seconda è l'output da prevedere. Il comportamento sofisticato che osserviamo negli attuali chatbot si spiega dunque anche quantitativamente, in riferimento (https://en.m.wikipedia.org/wiki/Large_language_model)
– al numero dei parametri della rete, che per alcuni chatbot è dell'ordine delle centinaia di miliardi (un'enormità, in confronto ai due parametri necessari per una regressione lineare…), e
– al numero delle parole dei testi del training set, che in qualche caso oggi supera i mille miliardi.

È interessante osservare la complementarietà delle due strategie sopra descritte:
– i sistemi di intelligenza artificiale simbolica, come i sistemi esperti, hanno un comportamento che deriva da conoscenza esplicitamente memorizzata e forniscono risultati che possono essere spiegati mostrando il processo con cui sono stati generati; sono perciò analoghi a quello che Daniel Kahneman ha chiamato "Sistema 2" (https://it.m.wikipedia.org/wiki/Pensieri_lenti_e_veloci), un modo di operare della mente "lento, pigro, logico/matematico, conscio";
– i sistemi di intelligenza artificiale subsimbolica, come le reti neurali artificiali e quindi gli attuali chatbot, hanno un comportamento che deriva da conoscenza memorizzata implicitamente e forniscono risultati che sono assai difficilmente spiegabili (sarebbe come osservare l'immagine del cervello di una persona per cercare spiegare le ragioni di quello che la persona ha appena detto…); sono perciò analoghi a quello che Daniel Kahneman ha chiamato "Sistema 1", un modo di operare della mente "veloce, automatico, emozionale, inconscio".

Anche se lo sviluppo tecnologico sembra stia portando a chatbot che includono componenti di Sistema 2, o più semplicemente componenti programmati, nella forma di agenti / plugin, i chatbot attuali sono essenzialmente Sistemi 1, e questo rende conto di quelli che sono forse i loro due limiti principali, per altro caratteristici di tutti i sistemi di intelligenza artificiale generativa.
Per prima cosa, i chatbot non sono sempre affidabili nelle informazioni fattuali che riportano e a volte fanno affermazioni false che non sanno riconoscere come tali. Non sono cioè *trustable*. Dal punto di vista metrologico, è problematica l'*oggettività* del loro comportamento, cioè la capacità di produrre informazione affidabile perché correttamente riferita al suo oggetto. Nel caso della misurazione, si migliora l'oggettività intervenendo empiricamente sugli strumenti di misura o informazionalmente sui modelli di misura, in modo da aumentare la garanzia che i valori misurati siano effettivamente riferiti al misurando. Come migliorare l'oggettività del comportamento dei sistemi di machine learning, e dei chatbot in particolare, riducendone quelle che oggi qualcuno chiama le loro "allucinazioni", è un problema aperto.
Secondariamente, come abbiamo accennato sopra, i chatbot hanno una struttura cosi complessa da rendere problematico ricostruire le ragioni del comportamento che esibiscono. Non sono cioè *explainable*. Dal punto di vista metrologico, è problematica l'*intersoggettività* del loro comportamento, cioè la capacità di produrre informazione affidabile perché interpretabile e quindi giustificabile nello stesso modo da tutti i soggetti interessati. Nel caso della misurazione, si migliora l'intersoggettività operando sulla catena di riferibilità metrologica, dalla definizione delle unità di misura alla taratura degli strumenti, e garantendo la trasparenza dell'intero sistema, in modo che, in linea di principio, chiunque possa ricostruire come i valori misurati sono stati ottenuti. Come migliorare l'intersoggettività del comportamento dei sistemi di machine learning, e dei chatbot in particolare, rendendone spiegabile il comportamento, è un problema aperto.
Insieme con le straordinarie capacità che i chatbot stanno mostrando, questi limiti giustificano la cautela che è opportuno mantenere nel loro uso, anche considerando che la nostra società non pare pronta ad attribuire una responsabilità a entità non-umane per quanto esse producono (non sono cioè *accountable*). Tutto ciò rende quello dei chatbot, e più in generale dei sistemi di intelligenza artificiale basati sulla strategia

subsimbolica, un interessante ambito di esplorazione e sperimentazione, a cui la cultura metrologica può utilmente contribuire.