# AI affidabile e spiegabile: *still a long way*

**Renzo G. Avesani** [*]

**Sommario**: 1. Le origini. – 2. Problemi: "*Curse of dimensionality*", "*Black-box problem*". – 3. Il futuro: *causal AI*, *neuro-symbolic reasoning*.

## 1. Le origini

In questa nota ci proponiamo di illustrare lo sviluppo di alcune delle principali ipotesi di ricerca che hanno portato all'emergere dell'*Artificial Intelligence* (AI) così come viene intesa oggi. Nel fare questo crediamo emerga come la tensione tra spinta allo sviluppo di nuovi metodi e la loro capacità di generare risultati affidabili e spiegabili sia intrinseco alla storia dell'evoluzione dell'AI. Questo per altro sembra essere vero in larga parte della scienza. In questo contesto crediamo che le richieste di "spiegabilità, affidabilità e trasparenza" delle applicazioni di AI contenute nell'AI Act [1], in principio condivisibili, si scontrino sul campo con grandi difficoltà. In particolare pensiamo possano rischiare di costringere il sentiero di sviluppo dell'AI almeno in Europa.

Comprendere appieno cosa significhi "Intelligenza" continua ad essere da secoli una vera e propria sfida per filosofi e scienziati. Ancora oggi, fatichiamo ad accordarci sulla semantica del termine. Una delle caratterizzazioni più comune-

---

[*] Presidente di Leithà, Unipol gruppo.

Ringrazio Leithà (Gruppo Unipol) per avermi permesso di approfondire queste tematiche ed in particolare il dott. Manuel Gozzi per l'interessante confronto ed il supporto nella stesura del testo. Sono invece unicamente responsabile di ogni errore od imprecisione contenuta in queste note.

[1] Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale (AI Act) e modifica i Regolamenti (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 168/2013, (UE) n. 2018/858, (UE) n. 2018/1139 e (UE) n. 2019/2144 e le Direttive n. 2014/90/UE, (UE) n. 2016/797 e (UE) n. 2020/1828 (Regolamento sull'intelligenza artificiale). Pubblicato nella *Gazzetta ufficiale dell'Unione Europea*, L 213, 12 luglio 2024.

mente accettate risiede nel riconoscere all'intelligenza la capacità di "adattamento" al cambiamento generato dalla massa smisurata di informazioni che si è resa via via disponibile per l'analisi e la comprensione. Allo stesso tempo asserire che una macchina è in grado di produrre risultati intelligenti è un'affermazione che merita particolare attenzione e cautela. Una macchina di questo tipo, infatti, deve essere in grado di adattarsi, "mimando" il comportamento di una specie imperfetta, quale noi umani siamo.

Il percorso evolutivo che caratterizza l'AI ha origine nella statistica classica e nel processo inferenziale. Questa disciplina nacque allo scopo di dare risposte, con un certo grado di affidabilità, a problemi decisionali in contesti caratterizzati da incertezza. Nel XVIII secolo, Thomas Bayes, come evoluzione dell'approccio classico introdusse il principio centrale della probabilità condizionata e dell'aggiornamento delle proprie credenze (*prior*) alla luce dell'evidenza acquisita. Questo concetto rappresenta uno dei pilastri su cui si fonda l'inferenza bayesiana, utilizzata ancora oggi in numerosi ambiti, dal filtraggio di rumore nei segnali alla diagnosi medica e non solo. Con l'avvento del XX secolo, la statistica inferenziale si evolse grazie ai contributi fondamentali di Ronald Fisher[2], che sviluppò concetti come la massima verosimiglianza, e di Jerzy Neyman ed Egon Pearson[3], che introdussero il moderno approccio alla verifica delle ipotesi statistiche. Questi strumenti permisero non solo di modellare fenomeni complessi, ma anche di definire un quadro decisionale rigoroso, capace di quantificare l'incertezza e di minimizzare i rischi di errore. Nel contesto dell'intelligenza artificiale, tali avanzamenti statistici trovarono applicazione nei primi sistemi probabilistici. L'inferenza bayesiana, ad esempio, venne utilizzata per creare modelli in grado di rappresentare la conoscenza attraverso distribuzioni di probabilità (non solo stime puntuali) e di aggiornare questa rappresentazione man mano che nuove informazioni venivano acquisite. Un esempio emblematico di tali applicazioni è il filtro di Kalman e sue evoluzioni sviluppate a partire dagli anni '60. Tale filtro, in principio bayesiano, combina osservazioni rumorose con stime precedenti per produrre previsioni più accurate in problemi di *tracking* e localizzazione. La caratteristica fondamentale che accomuna questi approcci è il rigore tecnico che si traduce in un forte determinismo e affidabilità nei risultati proposti ma la cui applicabilità era limitata dalle capacità computazionali dell'epoca. Tale rigore presuppone la conoscenza o l'assunzione a priori delle caratteristiche probabili-

---

[2] R.A. FISHER, *The Use of Multiple Measurements in Taxonomic Problems*, in *Annals of Eugenics*, 7(2), 1936, pp. 179-188, disponibile all'indirizzo: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[3] J. NEYMAN-E.S. PEARSON, *On the problem of the most efficient tests of statistical hypotheses*, in *Philosophical transactions of the royal society A*, 231(694-706), 1933, pp. 289-337, disponibile all'indirizzo: https://doi.org/10.1098/rsta.1933.0009.

stiche del processo che genera i dati (*data generating process*) limitandone quindi il campo di applicazione. L'inferenza bayesiana poi, richiedeva calcoli che diventavano rapidamente intrattabili al crescere della dimensionalità dei dati. Questo limite noto come "*curse of dimensionality*", può essere visto anche come un problema di scalabilità. Nei metodi statistici inferenziali, all'aumentare delle variabili in gioco seguiva una distribuzione sempre più sparsa dei dati nello spazio. L'identificazione delle relazioni tra gli stessi dati diveniva oltremodo complessa, aumentando esponenzialmente i requisiti computazionali e di campionamento. Risentirono dello stesso problema anche i primi approcci simbolici all'AI, teorizzati inizialmente da John McCarthy, Marvin Minsky, Nathan Rochester e Claude Shannon nel 1955 [4]. In quest'ultimi la rappresentazione della conoscenza dipendeva da ontologie, da regole costruite manualmente o da formalismi logici (da lì, il termine "simbolismo"). Tali sistemi si dimostravano inadatti a gestire l'alta dimensionalità poiché richiedevano una modellazione esplicita di ogni variabile e delle sue relazioni con le altre.

## 2. Problemi: "*Curse of dimensionality*", "*Black-box problem*"

Un approccio puramente simbolico non è scalabile perché risulta eccessivamente rigido nel rappresentare la conoscenza. La logica, linguaggio capace di formalizzare il ragionamento umano, da sola non tiene il passo con le esigenze di scalabilità moderne, diventate estreme al crescere dell'informazione disponibile. Il *Machine Learning* (ML) ha rappresentato una svolta significativa nel superamento dei limiti della dimensionalità dei dati avendo adottato anche tecniche già consolidate in altri ambiti come l'Analisi delle Componenti Principali (PCA). Questa ha permesso di ridurre la dimensionalità e trasformare i dati in rappresentazioni più compatte e gestibili. La transizione dal simbolico al connessionista ha segnato un cambiamento radicale. Se i sistemi simbolici codificavano la conoscenza in regole esplicite, il ML ha dimostrato che i modelli possono apprendere dai dati seguendo le architetture dei modelli impartite dal progettista. Ciò ha ridotto la dipendenza da rappresentazioni rigide e ha reso possibile la gestione dell'alta dimensionalità, aprendo la strada a progressi significativi in campi come la *Computer Vision* e il *Natural Language Processing* (NLP). L'approccio del ML si basa sul paradigma del "*teacher*" e "*learner*", diversamente dall'imperativo dei sistemi *rule-based*, in cui il programmatore definisce un modello matematico che la mac-

[4] J. McCarthy-M.L. Minsky-N. Rochester-C.E. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, *1955*, in *AI Magazine*, 27(4), 2006, p. 12, disponibile all'indirizzo: https://doi.org/10.1609/aimag.v27i4.1904.

china utilizza per apprendere le relazioni tra i dati. Il processo di addestramento del ML può essere diverso a seconda dal tipo di interazione uomo-macchina: supervisionato, non supervisionato e semi-supervisionato. Questo approccio ha rivoluzionato la rappresentazione della conoscenza, spostandola dai modelli simbolici, basati su ontologie e regole, verso numeri e operazioni matematiche. Un momento chiave in questa rivoluzione fu la definizione del *Perceptron* da parte di Rosenblatt nel 1958 [5]. Rosenblatt formalizzò un modello matematico ispirato al funzionamento del neurone umano. Tale modello, alla base anche del moderno approccio dell'apprendimento automatico, è in grado di calcolare l'errore tra la predizione e l'*output* atteso minimizzandolo durante la fase di addestramento. Come ulteriore sviluppo, grazie all'evoluzione delle GPU, negli anni 2000 è stato possibile addestrare reti neurali via via più profonde portando alla nascita del *Deep Learning* (DL). Questo approccio ha raggiunto un'importante pietra miliare nel 2012 con la vittoria di AlexNet, sviluppata dai dottorandi Alex Krizhevsky e Ilya Sutskever sotto la supervisione di Geoffrey Hinton [6] nella competizione ImageNet. In tale occasione si è dimostrata la superiorità delle reti neurali profonde rispetto alle tecniche tradizionali di *computer vision*. Tuttavia, mentre il ML ha parzialmente mitigato il problema della *curse of dimensionality*, il DL ne ha introdotto uno nuovo: la difficoltà nello spiegare l'*output* dei modelli, nota anche come il "*black-box problem*". Il DL si fonda sul Teorema dell'Approssimazione Universale (George Cybenko 1989 [7]). Tale teorema afferma che le reti neurali possono approssimare qualsiasi funzione continua e non-lineare. Questo risultato rappresenta un'evoluzione rispetto al *Perceptron* di Rosenblatt, in cui la struttura della rete è limitata a un singolo strato. L'architettura multistrato, propria del DL, rende necessario l'utilizzo di algoritmi di addestramento più avanzati; il più famoso è la "*backpropagation*" sviluppato da Rumelhart, Hinton e Williams nel 1986 [8]. Tale algoritmo consente di calcolare e minimizzare l'errore del modello tenendo conto della dipendenza e del contributo alla predizione dei diversi strati della rete. Tuttavia, questa capacità di rappresentare relazioni complesse rende difficile spiegare

[5] F. ROSENBLATT, *The perceptron: A probabilistic model for information storage and organization in the brain*, in *Psychological Review*, 65(6), 1958, pp. 386-408, disponibile all'indirizzo: https://doi.org/10.1037/h0042519.

[6] A. KRIZHEVSKY-I. SUTSKEVER-G.E. HINTON, *ImageNet classification with deep convolutional neural networks*, in *Advances in Neural Information Processing Systems*, 25, 2012, pp. 1097-1105, disponibile all'indirizzo: https://doi.org/10.1145/3065386.

[7] G. CYBENKO, *Approximation by superpositions of a sigmoidal function*, in *Mathematics of Control, Signals and Systems*, 2(4), 1989, pp. 303-314, disponibile all'indirizzo: https://doi.org/10.1007/BF02551274.

[8] D.E. RUMELHART-G.E. HINTON-R.J. WILLIAMS, *Learning representations by back-propa gating errors*, in *Nature*, 323, 1986, pp. 533-536, disponibile all'indirizzo: https://doi.org/10.1038/323533a0.

i risultati ottenuti, in quanto le "decisioni del modello" non sono trasparenti. L'introduzione dei *Transformers*[9] e dei *Large Language Models* (LLM) ha esacerbato questo problema. Gli LLM, con miliardi di parametri, sono modelli di linguaggio potenti ma estremamente complessi, dove, pur conoscendo i risultati, non siamo in grado di spiegare completamente come vengano generati. La spiegazione dei risultati di un LLM richiederebbe la comprensione delle interazioni tra miliardi di parametri, trattati in combinazioni di funzioni lineari e non lineari, rendendo il modello altamente difficile da interpretare. Il contributo dato da Aarohi Srivastava et al.[10] con altri 450 autori ben rappresenta le complessità che si incontrano nel valutare la *performance* degli LLM. Quando si parla di interazioni tra parametri si fa riferimento alla rappresentazione interna della conoscenza dell'LLM, che è ben diversa dal linguaggio naturale e che è circoscritta nello spazio vettoriale. La dottrina ci insegna che esistono due tipologie principali di rappresentazione della conoscenza: interna ed esterna. Ognuno di noi possiede la sua propria rappresentazione interna, il proprio modo di vedere le cose, e in nessun caso questa risulta essere direttamente comprensibile da qualcuno di diverso da noi stessi. L'interfaccia usata dall'uomo per interagire con le macchine è il linguaggio naturale, rappresentazione esterna della conoscenza. In letteratura, Turing[11] afferma che una macchina può essere definita intelligente se e solo se un umano, interagendo con essa, non riesce a distinguerne la natura: macchina, o umano? L'interazione uomo-macchina passa attraverso il linguaggio naturale, processo non esente da ambiguità che si prova a mitigare con la semantica distribuzionale. In tale approccio il significato delle frasi è dato dalla frequenza e dalla co-occorrenza dei termini all'interno delle stesse. Ecco perché la macchina deve possedere la capacità di tradurre il linguaggio naturale in una propria rappresentazione interna, ed ecco perché l'approccio simbolico, da solo, non è sufficiente, sebbene goda di spiegabilità e riproducibilità.

---

[9] A. Vaswani-N. Shazeer-N. Parmar-J. Uszkoreit-L. Jones-A.N. Gcmez-L. Kaiser-I. Polosukhin, *Attention Is All You Need,* arXiv preprint, 2017, disponibile all'indirizzo: https://doi.org/10.48550/arXiv.1706.03762.

[10] A. Srivastava et al., *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*, arXiv preprint, 2022, disponibile all'indirizzo: https://doi.org/10.48550/arXiv.2206.04615.

[11] A.M. Turing, *Computing Machinery and Intelligence*, in *Mind*, LIX(236), 1950, pp. 433-460, disponibile all'indirizzo: https://doi.org/10.1093/mind/LIX.236.433.

## 3. Il futuro: *causal AI*, *neuro-symbolic reasoning*

Sia l'approccio basato sulla statistica inferenziale che quello basato sull'AI simbolica godevano di riproducibilità ed interpretabilità, ma si scontravano col problema della "*curse of dimensionality*" al crescere del numero di variabili in gioco. Il ML mitiga questo problema tramite tecniche quali il PCA e introduce una rappresentazione esterna della conoscenza flessibile, che rende scalabile il sistema. La crescente domanda e offerta di dati portano al superamento del ML passando prima per il DL e poi per i LLM. Al problema della "*curse of dimensionality*" si affianca, ora, il "*black-box problem*". Spiegabilità e riproducibilità dei risultati sono venute meno a vantaggio di una maggior scalabilità sul fronte dimensionale. Gli LLM sono certamente in grado di generare risultati altamente accurati, ma la loro natura complessa e non lineare rende difficile comprendere come arrivano a determinati *output*. Questo ha portato a un crescente interesse verso approcci che possano migliorare la trasparenza e l'interpretabilità dei modelli. In questo contesto, un'area emergente è la *Causal AI*. Questo approccio si concentra sull'uso di modelli causali per capire le relazioni di causa ed effetto nei dati. Anziché trattare le variabili come entità indipendenti, come fanno molte tecniche di apprendimento automatico, la *Causal AI* cerca di mappare come una variabile influisca su un'altra, permettendo una comprensione più profonda e interpretabile del processo di generazione dell'*output*. Applicare la *Causal AI* ai *Transformers* potrebbe, quindi, offrire un modo per spiegare come certe variabili o *input* influenzino il comportamento del modello, tracciando le connessioni causali tra i diversi strati di trasformazione. Un altro approccio emergente è quello di combinare i metodi neuro-simbolici, che uniscono le potenzialità delle reti neurali (connessioniste) con gli approcci simbolici tradizionali, basati su regole e rappresentazioni esplicite. Mentre le reti neurali sono eccezionali nel catturare *pattern* complessi e nelle attività predittive, gli approcci simbolici possono aggiungere una componente di struttura e trasparenza, grazie alla loro capacità di formalizzare la conoscenza in regole interpretabili. Integrando queste due modalità, si potrebbero stabilire delle linee guida che orientino il comportamento del modello, rendendo il sistema complessivo più comprensibile e interpretabile, e permettendo di spiegare in modo più chiaro come i modelli arrivino a determinate conclusioni. In questa direzione, la *Transparent AI* sta diventando una tendenza crescente, mirando a costruire modelli che non solo siano accurati, ma anche comprensibili. L'integrazione di metodi causali e simbolici con il DL potrebbe rappresentare una chiave per superare il *black-box problem*, aumentando l'affidabilità dei modelli e facilitando la loro adozione in applicazioni che richiedono alta trasparenza, come nel settore sanitario, legale e finanziario. Stanno inoltre emergendo altre tecniche di frontiera come il *dictionary learning* [12] e

[12] T. Bricken-A. Templeton-J. Batson-B. Chen-A. Jermyn-T. Conerly-N.L. Turner-C. An-

gli *sparse autoencoders* [13] per affrontare la complessità dei dati in modo più mirato. Il *dictionary learning* si basa sull'idea di rappresentare i dati come combinazioni semplici di elementi base, mentre gli *sparse autoencoders* utilizzano restrizioni per creare rappresentazioni più essenziali e comprensibili. Spiegare e interpretare l'*output* degli LLM è una sfida aperta con cui si sta concentrando il mondo della ricerca. L'Unione Europea, nel formalizzare l'AI Act, ha sicuramente indicato obbiettivi condivisibili per assicurare che l'AI rispetti anche valori etici. Tra gli obiettivi principali vi sono la trasparenza, la responsabilità e la promozione di un ecosistema di innovazione affidabile. Il concetto di "trasparenza" espresso nella normativa è ampio e oltrepassa l'aspetto tecnico, oggetto di questo scritto. In quest'ottica, l'Act formalizza requisiti di trasparenza sulla documentazione tecnica dei modelli (e.g. i dati utilizzati per il *training* e le istruzioni per l'uso), sugli impatti dell'AI sui processi operativi e sui meccanismi di controllo dell'operato dell'AI attuati dal *provider* dei modelli. L'AI Act impone inoltre obblighi di trasparenza soprattutto per le applicazioni ad alto rischio, richiedendo che gli utenti siano consapevoli quando interagiscono con un sistema di AI o quando i loro dati vengono utilizzati per addestrarlo. Ciò riflette un impegno per garantire che i processi decisionali automatizzati siano spiegabili e verificabili, per evitare discriminazioni o errori sistematici. Tuttavia, il livello di dettaglio richiesto per questa trasparenza solleva questioni pratiche e tecniche. Come spiegato in questo contributo, la ricerca mondiale sta lavorando attivamente per dare delle risposte a queste esigenze ma allo stato attuale la tensione tra trasparenza, complessità tecnologica e *performance* dell'AI rimane alta.

IL-C. DENISON-A. ASKELL-R. LASENBY-Y. WU-S. KRAVEC-N. SCHIEFER-T. MAXWELL-N. JOSEPH-A. TAMKIN-K. NGUYEN-B. MCLEAN-J.E. BURKE-T. HUME-S. CARTER-T. HENIGHAN-C. OLAH, *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*, in *Transformer Circuits Thread*, 2023, disponibile all'indirizzo: https://transformer-circuits.pub/2023/monosemantic-features.

[13] S. RAJAMANOHARAN-T. LIEBERUM-N. SONNERAT-A. CONMY-V. VARMA-J. KRAMÀR-N. NANDA, *Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders*, arXiv preprint, 2024, disponibile all'indirizzo: https://doi.org/10.48550/arXiv.2407.14435.