Digeat N.3 - 19 Settembre 2024

eXplainable AI: algoritmi a prova di fiducia

Di Serena Deplano

Abstract

Come ci sentiremmo se un giorno il medico, con in mano la nostra cartella clinica, ci dicesse che la diagnosi è stata formulata da un potentissimo algoritmo di intelligenza artificiale (del quale, però, non si conoscono i meccanismi di funzionamento)? Spoiler: quel lontano futuro è già attuale e l'intelligenza artificiale è una realtà concreta nella nostra vita, anche in ambiti critici e sensibili come quello medico. Eppure, qualcosa sfugge del suo funzionamento. La buona notizia è che grazie all'*eXplainable AI* possiamo aprire la *black box* dove si cela l'alchimia degli algoritmi e provare a comprenderli. In definitiva, fidarci dell'IA.

Indice

Houston, abbiamo un problema: l'intelligenza artificiale permea le nostre vite, ma noi non sappiamo *esattamente tutto* sul suo funzionamento. Gli algoritmi ci suggeriscono cosa guardare in tv, guidano autonomamente veicoli, formulano diagnosi mediche, elaborano analisi predittive, stabiliscono chi può ricevere un prestito. Eppure, il *modo* in cui *giungono a questi risultati* è, almeno in parte, un mistero: è **una scatola nera nella quale non possiamo guardare** (nonostante questo, non è la magia a far funzionare l'IA, ma la matematica).

Tipicamente, infatti, il motivo per cui un sistema di IA produce un determinato risultato è sconosciuto non soltanto agli utenti finali, ma anche a chi lo sviluppa. Si tratta di modelli definiti *black box*: conosciamo l'input, possiamo vedere l'output, ma non siamo in grado di capire perché, a partire dal primo, il sistema ha determinato il secondo.

Una premessa tecnica (rapida ed indolore)

Per capire il problema della *black box* è necessario partire dalle dinamiche di funzionamento dell'IA. I sistemi di IA funzionano attraverso il *deep learning* (DL), una classe di modelli di *machine learning* (ML) che hanno l'obiettivo di simulare il funzionamento della mente umana attraverso le c.d. reti neurali multistrato o reti neurali profonde (*deep neural network* – DNN): si tratta di **insiemi di algoritmi concepiti e addestrati per apprendere modelli di correlazione tra dati non strutturati e generare degli output**, come per esempio formulare analisi predittive.

Le reti neurali processano **grandi quantità di dati**, tra cui individuano correlazioni statistiche basate su determinati parametri, appresi in fase di addestramento. Sono organizzate secondo una struttura

reticolare a nodi, detti **neuroni** (proprio e volutamente come nella mente umana), e secondo una **successione gerarchica di strati**: ogni strato, attraverso funzioni non lineari, elabora i calcoli in base ai risultati prodotti dallo strato precedente, partendo dallo strato di input e arrivando a quello di output. Gli strati intermedi sono definiti strati latenti (*hidden layers*) e sono deputati all'elaborazione dei dati. È qui che succede la magia (della matematica).

Ogni nodo si collega ad altri nodi appartenenti solo a strati successivi attraverso archi di collegamento a cui è associato un parametro di calcolo che pondera la significatività della relazione tra i nodi. In itinere, **la rete neurale apporta delle correzioni progressive ai parametri per minimizzare le differenze tra le previsioni e i valori effettivi**, proprio come fa la mente umana in fase di apprendimento: procede per approssimazioni, errori e correzioni.

Il lavoro della rete neurale, quindi, è quello di elaborare progressivamente gli input propagandoli di strato in strato attraverso **calcoli estremamente complessi**, con l'obiettivo di produrre il risultato richiesto: per esempio, l'apprendimento di un modello linguistico che ci consente di interagire con un agente intelligente come se stessimo interloquendo con un essere umano.

> Il problema della *black box*, quindi, è insito nella complessità degli strati latenti, dove – come si è detto – si realizza la maggior parte dell'elaborazione dei dati (Barredo Arrieta et al., 2020). Questo per diversi ordini di motivi tra cui la numerosità e la complessità dei parametri e la loro correlazione con il risultato finale, così come la scarsità di strumenti di interpretazione dei comportamenti delle reti neurali profonde (Hassija et al., 2023).

Questo è il motivo che rende non conoscibile ciò che accade all'interno della scatola nera.

La richiesta di fiducia nell'IA e la risposta del Legislatore europeo

Tutto questo fa emergere un interrogativo di non poco conto: quanto siamo disposti a fidarci di un sistema che opera autonomamente senza tuttavia avere la consapevolezza di come funziona? **Non troppo**, secondo le ricerche (Gillespi et al. 2021): la mancanza di fiducia ci allontana dall'uso della tecnologia (Chakravorti, 2024). Ma se dalla tecnologia non possiamo – e forse non vogliamo – allontanarci, allora la nostra equazione si riequilibra con **la domanda di maggiore trasparenza, fattore precursore della fiducia**. Così, in effetti, sta accadendo anche nel campo dell'IA: più le nostre vite sono immerse negli algoritmi, più aumenta la domanda di trasparenza (Blackman e Ammanath, 2022).

> Come è facile intuire, il tema è estremamente rilevante nel contesto dell'intelligenza artificiale, in particolare nelle applicazioni che prevedono sistemi in grado di adottare decisioni, di elaborare previsioni e in generale di compiere in autonomia azioni che hanno un impatto diretto sulla vita delle persone e che fino ad ora sono state tipicamente svolte da esseri umani esperti. Si tratta di un contesto in cui la percezione del rischio da parte degli utenti è inversamente proporzionale alla percezione del controllo.

Il giusto mix per un AI *trust gap* (Chakravorti, 2024) che **innesca paure e dubbi fino a determinare un fenomeno noto come avversione agli algoritmi**: **tendiamo a fidarci maggiormente di un'indicazione proveniente da un essere umano – piuttosto che da un algoritmo –** sebbene gli algoritmi mostrino capacità performative maggiori degli esseri umani (Jussupow et al. 2020, Mahmud et al. 2022, Schaap et al. 2024).

È proprio questa mancanza di trasparenza che si rivela cruciale nella definizione dell'AI *trust gap*, in particolare quando gli sviluppi dell'AI tracciano traiettorie di tangenza che si intersecano in punti

nevralgici sensibili per la vita delle persone (Holzinger et al., 2022).

Gli esempi possibili sono numerosi, ma uno estremamente rappresentativo è quello delle diagnosi in ambito medico: in che misura può un professionista affidarsi *tout-court* al risultato di un sistema intelligente, in mancanza di una piena evidenza degli elementi che ne hanno costituito l'iter logico e con la consapevolezza che la mancanza di trasparenza non consente di rilevare – e correggere – eventuali errori e *bias del sistema*?

> Le domande sono tutt'altro che speculative perché si tratta di scenari che investono concretamente la vita delle persone e sollevano inquietudini, interrogativi e riflessioni.

Non a caso, il legislatore europeo – consapevole della vulnerabilità a cui sono esposti gli individui in questi frangenti – ha introdotto specifiche regole in due contesti normativi differenti e cruciali: il **Regolamento europeo 2016/679 sulla protezione dei dati personali (GDPR) e il Regolamento europeo 2024/1689 sull'intelligenza artificiale (AI Act)**.

Il GDPR, per garantire un **trattamento corretto e *trasparente***, prevede l'obbligo specifico di informare l'interessato circa "l'esistenza di un *processo decisionale automatizzato* compresa la profilazione di cui all'articolo 22, paragrafi 1 e 4" e di fornire "almeno in tali casi, *informazioni significative sulla logica utilizzata*, nonché l'*importanza e le conseguenze* previste di tale trattamento per l'interessato" (art. 13).

L'AI Act, per altro verso, nella regolazione dei sistemi ad alto rischio prevede che questi siano "progettati e sviluppati in modo tale da garantire che *il loro funzionamento sia sufficientemente trasparente* da consentire ai *deployer* di *interpretare l'output* del sistema e utilizzarlo adeguatamente" (art. 13).

Verso l'intelligenza artificiale spiegabile *o eXplainable AI* (XAI)

In risposta a una **crescente domanda di trasparenza** e alla necessità di un'intelligenza artificiale *responsabile* (Periccioli, 2020), si è ravvivato l'interesse verso un ambito di ricerca nato agli albori del *machine learning*: l'intelligenza artificiale spiegabile *o eXplainable AI* (XAI).

> Obiettivo dell'XAI è quello di rendere intellegibile il funzionamento dell'AI, esplicitandolo, così da renderlo comprensibile all'utente finale e, in ultima analisi, aumentando la fiducia verso gli output generati (Love et al., 2023). È un tentativo di trasformare la *black box* in una *glass box* (Rai, 2020) attraverso un approccio multidisciplinare che intreccia competenze derivanti dalla psicologia, dall'ingegneria e dall'informatica (Phillips et al., 2021). I metodi adottati dall'XAI sono riconducibili a due macrocategorie: modelli intrinsecamente spiegabili e modelli che prevedono una spiegazione successiva (*post-hoc*).

Nel primo caso parliamo di modelli concepiti secondo una logica di *explainability by design*: il funzionamento degli algoritmi è direttamente comprensibile e la trasformazione degli input in output è interpretabile. **In questo caso si parla di modelli *white box***. Un esempio è quello degli alberi delle decisioni, utilizzati, per esempio, per la classificazione dello spam nella posta elettronica. Per altro verso, nel caso di modelli come le reti neurali, è altamente improbabile riuscire a ipotizzare una interpretabilità intrinseca proprio a causa – come si è detto – della complessità del modello, (la nostra *black box*). In questi casi, **l'approccio della XAI è quello della *post-hoc explainability***: l'idea di fondo è che i metodi di XAI siano in grado – con varie tecniche – di ricostruire *ex post* il funzionamento degli algoritmi.

Secondo gli schemi di categorizzazione dell'XAI (Speith, 2022), i modelli *post-hoc* sono soggetti a diverse classificazioni, in base alla prospettiva di analisi. Così, i metodi di XAI possono essere globali o locali – a seconda che l'interpretazione riguardi l'intero modello o una sua singola istanza – ma anche agnostici o specifici, a seconda che il metodo sia applicabile a tutti i modelli o solo specificamente ad alcuni.

Tra i principali algoritmi di XAI vale la pena citare LIME e SHAP

LIME (Local Interpretable Model-agnostic Explanation) è un metodo locale (si concentra su singole istanze e non sull'intero modello) e agnostico, applicabile quindi a tutti i modelli. Si basa su un'idea di fondo: quella di alterare i dati di input del modello e osservare l'effetto sull'output. Attraverso un processo iterativo, LIME costruisce un surrogato interpretabile del modello, che può essere utilizzato per spiegare come questo è giunto a un determinato risultato.

SHAP (SHapley Addictive exPlanation) invece è un metodo globale basato sulla teoria dei giochi cooperativi: permette di spiegare quanto una certa variabile del modello incide sul suo output. Ogni variabile, infatti, viene ponderata secondo un punteggio (valore di Shapley) che misura quanto essa determini lo spostamento del modello verso il suo risultato.

Immaginiamo, per esempio, di chiedere un prestito a **un istituto bancario che utilizza un algoritmo di Machine Learning per determinare la nostra l'affidabilità creditizia**: al termine della procedura scopriamo che la nostra richiesta è stata respinta ma senza avere alcun dettaglio su quale (o quali) dei nostri parametri abbiano determinato il risultato e non ci è dunque data la possibilità di porvi rimedio. **Con SHAP è possibile calcolare l'incidenza delle singole variabili individuali** (reddito, età, professione etc.) su una *baseline* generale (un punteggio di credito calcolato in generale su tutti i clienti) per comprendere quale sia la composizione del nostro punteggio individuale di affidabilità e quale (o quali) dei nostri parametri abbia determinato il rigetto della nostra richiesta.

Attraverso questi (e altri) metodi di XAI **è quindi possibile comprendere come funzionano i modelli e spiegarne il funzionamento**: il che, come si è detto, ha lo scopo di rendere comprensibile e trasparente il funzionamento della *black box*.

> C'è un però: non mancano infatti critiche verso i metodi di XAI. La principale riguarda il fatto che la spiegazione degli algoritmi rappresenta un compromesso tra trasparenza e accuratezza e l'esplicitazione può comportare una perdita di prestazioni del modello: più trasparenza, meno efficacia. Circostanza – questa – che si amplifica all'aumentare della potenza e della complessità dei modelli di AI. Altro rilievo riguarda il fatto che la semplificazione del modello non è comunque tale da consentire a chiunque di poterlo comprendere, perché l'interpretazione stessa presuppone competenze non necessariamente diffuse (de Bruijn et al. 2022). Va poi considerato che la spiegazione del funzionamento di un modello è spesso dipendente dal contesto, il che rende difficile fornire una spiegazione in termini generali.

Al di là dei suoi attuali limiti – i quali probabilmente costituiranno la leva per sviluppi futuri – **l'XAI è certamente un ambito di ricerca affascinante e ricco di potenziale**, che si pone l'obiettivo di fornire strumenti per rendere più trasparenti e comprensibili i meccanismi di funzionamento dell'IA, contribuendo così a **superare le barriere cognitive che ostacolano la piena comprensione di come questa funziona** e, in definitiva, a ingenerare la fiducia degli utenti verso la tecnologia. Del resto, la spinta a cercare strumenti e chiavi di lettura che consentano una maggiore trasparenza e leggibilità degli algoritmi è vitale: se ci pensiamo si tratta – in fondo – di sviluppare la tecnologia a misura di essere umano.

Bibliografia e sitografia

Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, in Information Fusion, Volume 58, 2020, pp. 82-115;

Blackman R., Ammanath B., *Building transparency into AI projects*, in Harvard Business Review, 2022;

Chakravorti B., *AI's trust problem. Twelve persistent risks of AI that are driving skepticism*, in Harvard Business Review, 2024;

de Bruijn H., Warnier M., Janssen M., *The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making*, in Government Information Quarterly, Volume 39, Issue 2, 2022;

EDPS: TechDispatch 2/2023 – *Explainable Artificial Intelligence*, 2023;

Gillespie N., Lockey S., Curtis C., *Trust in artificial intelligence: A five country study*, The University of Queensland and KPMG Australia, 2021;

Hasan M., A.K.M. Najmul I., Syed I., Kari S., *What influences algorithmic decision-making? A systematic literature review on algorithm aversion*, in Technological Forecasting and Social Change, Volume 175, 2022;

Hassija, V., Chamola, V., Mahapatra, A. *et al.Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence*, Cogn Comput, Volume 16, 2024, pp. 45–74;

Holzinger A., Saranti A., Molnar C., Biecek P., Samek W., *Explainable AI Methods – A Brief Overview*, in Holzinger A., Goebel R. Fong R., Moon T., Müller KR., Samek W. (eds) *xxAI – Beyond Explainable AI. xxAI 2020*. *Lecture Notes in Computer Science,* volume 13200, 2020;

Jussupow E., Benbasat I., Heinzl A., *Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion*, in Proceedings of the 28th European conference on information systems (ECIS), an online AIS conference, June 15–17, 2020;

Love P., Fang W., Matthews J., Porter S., Luo H., Ding L., *Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction*, in Advanced Engineering Informatics, Volume 57, 2023

Periccioli M., *Cos'è l'eXplainable AI e perché non possiamo farne a meno*, in AI4Business, 2020;

Phillips P. J., Hahn C. A., Fontana P. C., Broniatowski D. A., Prxybocki M. A., *Four principles of explainable Artificial Intelligence*, NIST Interagency/Internal report (NISTIR), report number 8312, 2021;

Rai A., *Explainable AI: from black box to glass box*, in J. of the Acad. Sci., volume 48, p. 137–141, 2020;

Schaap, G., Bosse, T. & Hendriks Vettehen, P., *The ABC of algorithmic aversion: not agent, but benefits and control determine the acceptance of automated decision-making*, in AI & Soc, volume 39, 2024, p. 1947–1960;

Speith T., *A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods*, in 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22),