Oltre la “black box” dell’IA: l’Explainable AI come strumento per sviluppare il pensiero critico

DI THEFABLAB
13/10/2026

INTELLIGENZA ARTIFICIALE



Il testo approfondisce il ruolo dell’Explainable AI (XAI), un insieme di tecniche che aiutano a comprendere come l’intelligenza artificiale prende le proprie decisioni. Attraverso esempi pratici e strumenti come LIME, SHAP e Google Teachable Machine, vengono mostrati i limiti degli algoritmi e il rischio che producano risultati basati su correlazioni ingannevoli. L’obiettivo è promuovere un uso consapevole dell’intelligenza artificiale, sviluppando negli studenti il pensiero critico e la capacità di verificare l’affidabilità delle risposte generate.

Dai chatbot basati su IA generativa fino agli algoritmi dei social media, oggi studentesse e studenti hanno **accesso quotidiano all’intelligenza artificiale**. Spesso l’uso di questi strumenti rimane su un livello dinamico e superficiale: in pochi secondi è possibile ottenere un feedback rispetto a qualsiasi richiesta. La tecnologia è insomma comoda e

accessibile, ma questo approccio rischia di perdere di vista un passaggio fondamentale per alimentare lo **spirito critico** necessario a intraprendere un rapporto attivo e propositivo con la tecnologia.

Questo succede anche perché i modelli di IA che siamo abituati a usare funzionano come *black box*, scatole nere: è molto facile ottenere un output, ma è molto difficile risalire al ***perché* è stato prodotto proprio quello specifico output**.

L'**explainable AI (XAI)** è un insieme di **metodi e tecniche che aiutano gli esseri umani a comprendere e verificare** quali passaggi hanno portato l'intelligenza artificiale a produrre un certo output. Il termine fu usato per la prima volta nel **2016** come titolo di un programma di ricerca ufficiale della DARPA, l'agenzia di ricerca del Dipartimento della Difesa degli Stati Uniti. All'epoca l'attenzione era rivolta per lo più all'**AI predittiva**: in quel periodo le istituzioni e i ricercatori, infatti, cominciavano a rendersi conto che gli umani non avrebbero mai potuto fidarsi delle decisioni dell'IA in contesti critici e delicati senza la possibilità di capire come erano state prese.

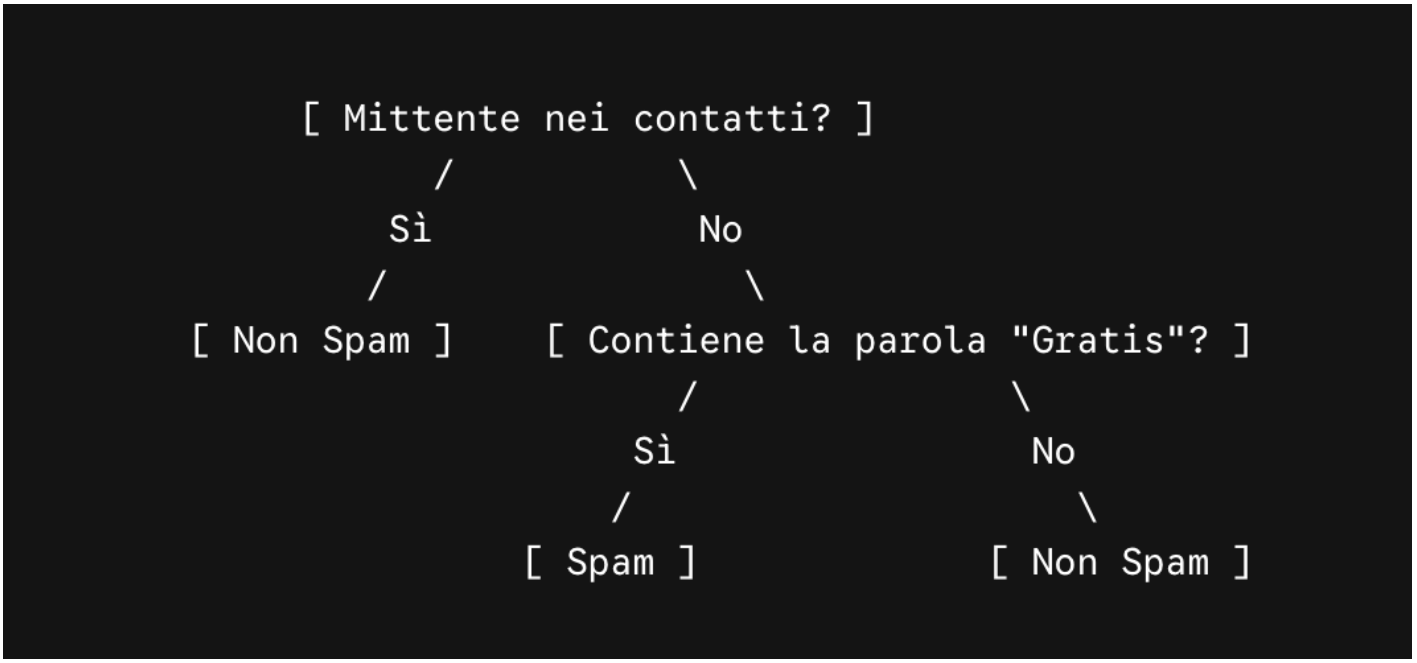
Il concetto si è rafforzato sempre più negli anni successivi. Nel 2018 il GDPR dell'Unione Europea introduce il "**diritto alla spiegazione**": se un cittadino chiede conto a una organizzazione di una decisione presa da un sistema IA, cioè, la risposta dovrà essere in grado di spiegare **come l'algoritmo è arrivato a quella conclusione**. E, con l'AI Act del 2024, viene introdotta la spiegabilità obbligatoria per i sistemi di IA classificati come *ad alto rischio*. Il tema è quindi sempre più centrale ed evolve insieme al nostro uso di questa tecnologia: oggi, per esempio, l'**AI generativa** introduce la necessità di nuove cautele legate a **fonti, violazioni della privacy e possibili *bias***.

Dalla white box alla spiegabilità a posteriori: due approcci alla XAI

Nel panorama informatico attuale esistono due approcci fondamentali alla XAI:

- **il modello della *white box* o dell'IA *interpretable-by-design***. In questo approccio, l'IA è progettata per essere intrinsecamente spiegabile. Gli algoritmi sono immediati da comprendere e il loro studio permette in modo diretto di capire come vengono trattati i dati tra input e output.
- **l'ispezione di modelli *black box* (*post-hoc explainability*)**. In questo caso, si usano strumenti *ad hoc* per applicare tecniche di spiegabilità a posteriori, indagando esternamente le ragioni che hanno portato alla produzione di un output.

Un classico esempio di algoritmo del tipo white box è l'**albero decisionale**. In questo tipo di sistema, il flusso delle decisioni prese dall'algoritmo può essere visualizzato sotto forma di diagramma di flusso in cui ogni nodo rappresenta una condizione e ogni ramo rappresenta un possibile output basato su tale condizione. Le foglie rappresentano poi la classificazione finale.



Ci potremmo chiedere **perché**, se i modelli *white box* esistono e funzionano perfettamente, **la maggior parte degli strumenti IA che usiamo di solito si basano invece su *black box***.

La risposta sta nel **tipo di dati** che abbiamo bisogno di trattare. I modelli *white box* funzionano bene con **dati strutturati**: tabelle, numeri, elenchi di requisiti. Non sono invece in grado di trattare **dati non strutturati** come testi complessi, immagini e tracce audio: tutti quelli che risultano interessanti per l'AI generativa, insomma.

Di conseguenza, se vogliamo **un'IA capace di individuare un'anomalia in una radiografia o di conversare con gli esseri umani**, è necessario ricorrere alla complessità delle **reti neurali profonde** (*Deep Learning*). In questi casi, la struttura interna del modello, composta da miliardi di parametri matematici interconnessi, diventa inaccessibile e *opaca*.

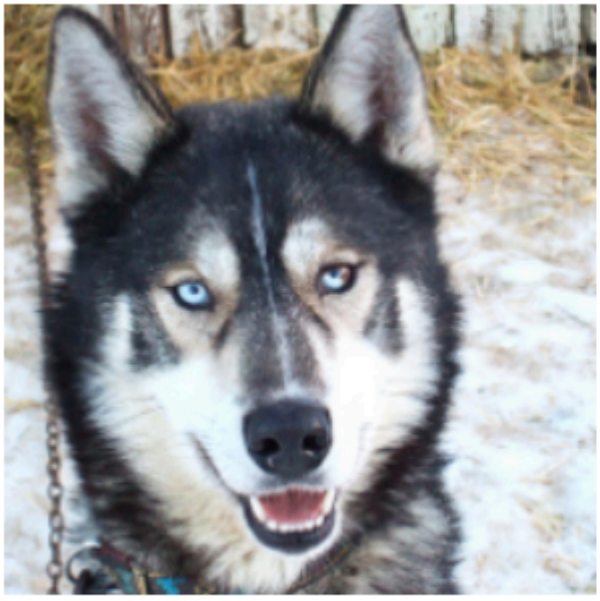
L'ispezione della black box: tecniche e strumenti

Quando le reti neurali e il deep learning hanno cominciato a diventare gli strumenti dominanti nel panorama dell'IA, i ricercatori si sono trovati di fronte a un paradosso: **i**

modelli diventavano via via più accurati ed efficaci, ma al contempo sempre più incomprensibili. Per colmare questo divario, tra il 2016 e il 2017 la ricerca in informatica ha lavorato sullo sviluppo di tecniche di **spiegabilità a posteriori (*post-hoc*)**.

L'idea alla base di questo approccio è trattare l'IA come una scatola chiusa e studiare il suo comportamento dall'esterno, "interrogandola" con domande e variazioni mirate. Questo approccio è stato avviato nel **2016**, quando tre ricercatori dell'Università del Washington pubblicarono lo studio pionieristico che introduceva lo strumento di spiegabilità **LIME** (*Local Interpretable Model-agnostic Explanations*).

Per mostrare l'utilità del loro strumento, gli autori addestrarono un modello di *Deep Learning* per distinguere le immagini di lupi da quelle di cani husky. Il modello sembrava infallibile, con un'accuratezza quasi del 100%. Tuttavia, applicando LIME per ispezionare visivamente le ragioni dietro le decisioni della rete, si scoprì un fatto clamoroso: **l'algoritmo non stava affatto considerando le caratteristiche fisiche degli animali** (come orecchie, occhi o pelo) per eseguire la classificazione, ma si limitava a verificare la presenza della **neve sullo sfondo**. Poiché nel dataset di addestramento tutte le foto di lupi avevano la neve e quelle degli husky no, l'IA aveva imparato una correlazione del tutto fuorviante. Senza una tecnica di XAI *post-hoc*, questo bias non sarebbe emerso per tempo, portando a errori e fraintendimenti.



(a) Husky classified as wolf



(b) Explanation

Oggi l'ispezione delle black box si affida principalmente a tre metodologie:

- **La perturbazione degli input e la rilevanza delle variabili (*Feature Importance*)**
  Strumenti come **LIME** (2016) e **SHAP** (*Shapley Additive exPlanations*, introdotto da Lundberg e Lee nel 2017) funzionano secondo il principio della *perturbazione dell'input* per analizzare e spiegare output specifici. Lo strumento prende una **singola decisione** del modello (ad esempio, la bocciatura di un curriculum vitae) e inizia calcolare i risultati cambiando leggermente i dati forniti in ingresso: cosa succede se rimuoviamo dal CV la parola "laurea"? Cosa cambia invece se modifichiamo l'età o il comune di residenza? Osservando come varia l'output finale a seguito di queste micro-modifiche, lo strumento calcola il "peso" o l'importanza di ciascuna variabile (*Feature Importance*) per il modello. In particolare, SHAP attinge alla rigorosa **teoria dei giochi matematici** elaborata dal Premio Nobel Lloyd Shapley, assegnando a ciascun dato un punteggio esatto che ne misura il contributo marginale al risultato.
- **Le mappe di calore visive (*Saliency Maps*)**
  Nell'ambito dell'analisi di immagini e video, uno degli strumenti più diffusi è **Grad-CAM** (*Gradient-weighted Class Activation Mapping*, sviluppato nel 2016). Questo tool analizza il flusso dei segnali all'interno degli ultimi strati della rete neurale e genera una mappa cromatica, sovrapposta all'immagine originale, che fornisce **una impressione visiva sui dettagli dell'immagine che sono stati maggiormente considerati dall'algoritmo**. I colori spaziano dal rosso/caldo (massima attenzione) al blu/freddo (nessuna rilevanza). Se, per esempio, un modello di IA analizza una radiografia e individua una patologia, la mappa di calore consente al medico di verificare se l'algoritmo ha individuato la reale lesione cellulare o se è stato tratto in inganno da dettagli irrilevanti.
- **Le spiegazioni controfattuali**.
  Per rispondere al "diritto alla spiegazione" previsto dal GDPR, negli ultimi anni sono stati sviluppati diversi strumenti di spiegazione controfattuale, che rispondono cioè alla domanda: "*Qual è il minimo cambiamento necessario nei dati di input per ottenere un risultato opposto?*". Invece di spiegare la complessa matematica del modello, questi sistemi si basano sul principio matematico del minimo cambiamento necessario per fornire all'utente un'indicazione prescrittiva, per esempio: "*Il tuo CV è stato respinto. Se ottenessi una laurea in informatica, il CV verrebbe preso in considerazione.*"

Queste tecniche di spiegabilità sono strumenti preziosi, ma è importante ricordare che rimangono delle approssimazioni: al contrario delle white box, cioè, non offrono una finestra trasparente sul modello. Le mappe di calore, per esempio, ci mostrano dove ha guardato il modello senza spiegarci perché lo ha fatto o perché questo sguardo si è tradotto in una specifica decisione. **Anche la XAI, insomma, va letta con spirito critico**: ogni spiegazione è un'ipotesi da mettere alla prova.

A caccia dell'husky nella neve: un'attività pratica

Per sperimentare la logica delle tecniche di perturbazione dell'input nell'ambito della XAI, è possibile impostare semplici attività in classe basate su **Google Teachable Machine**, uno strumento web gratuito, *no-code* e accessibile direttamente dal browser di qualsiasi computer o tablet.

Sviluppato da Google per scopi didattici, Teachable Machine permette di addestrare in pochi minuti un modello di *Machine Learning* per il riconoscimento di immagini, suoni o posture. Mostrando in tempo reale la percentuale di confidenza dell'algoritmo di fronte a nuovi stimoli, la piattaforma diventa una palestra ideale per analizzare il comportamento delle reti neurali e scoprire eventuali *bias* o sviste logiche.

L'attività, strutturata in stile *reverse engineering*, guida la classe a **spiegare le vere motivazioni dietro le decisioni di un'IA in tre passaggi**:

1. **Addestramento del modello**: Divisi in coppie, studentesse e studenti usano la webcam per addestrare un classificatore visivo a distinguere due oggetti (ad esempio: una borraccia e un astuccio). Inseriscono decine di foto di prova e verificano che l'IA li riconosca con una percentuale di confidenza vicina al 100%.

2. **Perturbazione dell'input**: Gli studenti testano la solidità del modello modificando un solo dettaglio per volta nel contesto o nell'oggetto, annotando quando e come la risposta dell'IA crolla. Per esempio: cosa succede se cambiamo lo sfondo dell'aula o la luminosità? E se copriamo con la mano metà dell'oggetto? L'IA sta riconoscendo davvero la borraccia o sta usando come "scorciatoia" il colore del banco o la forma di una mano?

3. **Discussione e debriefing**: Ricalcando l'aneddoto scientifico dell'husky e della neve, ogni gruppo di lavoro racconta la propria esperienza e la classe riflette su come l'algoritmo possa prendere decisioni apparentemente corrette basandosi su correlazioni del tutto fuorvianti.

Fonti

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144.

Autori

THEFABLAB

TheFablab è attivo dal 2014 nell'ambito dell'educazione tecnologica e della fabbricazione digitale, progettando percorsi formativi per i giovani e le comunità. Collabora a progetti nazionali con aziende, enti pubblici, scuole e istituzioni culturali per promuovere le competenze STEM, l'alfabetizzazione digitale e la cultura scientifica. Attraverso programmi educativi *hands-on*, contribuisce a rafforzare il capitale scientifico delle nuove generazioni e l'approccio critico verso le tecnologie emergenti, sostenendo l'empowerment, l'inclusione e il benessere educativo.