Digeat N.7 - 19 Settembre 2025

L’Intelligenza Artificiale è quello che mangia

Di Pierangelo Felici

Abstract

I problemi di qualità legati all’utilizzo di fonti non affidabili per l’addestramento dei modelli di Intelligenza Artificiale sono spesso sottovalutati. E i rischi non si limitano agli aspetti di accuratezza o bias, ma possono avere gravi implicazioni anche in termini di protezione dei dati personali (data protection).

Indice

Introduzione

Stiamo vivendo un periodo nel quale si parla tanto, forse troppo, di **Intelligenza Artificiale** (IA o AI). Da un lato, c’è chi la teme, adombrando rivoluzioni negative, più o meno marcate, dall’altro, chi la idolatra, pensando che possa risolvere qualunque problema.

> Pochi – è questa l’impressione – hanno una posizione equilibrata. Sembra che se ne riconoscano le potenzialità, attuali ma soprattutto future, senza tenere però presente che si tratta solo di uno strumento. Con molti limiti. Uno dei quali, la qualità dell’informazione che la “nutre” e la modella.

L’importanza della qualità dei dati non può e non deve essere sottovalutata. I sistemi di IA, in particolare quelli basati sull’apprendimento automatico (machine learning) e sull’apprendimento profondo (deep learning), dipendono in modo sostanziale dai dati con cui vengono addestrati. **Più i dati sono completi, coerenti e accurati, più le prestazioni del modello risultano affidabili e generalizzabili.** Tuttavia, quando le fonti utilizzate per addestrare questi sistemi non sono attendibili o sono affette da errori, disinformazione o *bias*, i problemi emergono e possono generare conseguenze significative.

L’importanza della qualità delle fonti

Il processo di addestramento di un modello di IA comporta l'assimilazione di grandi quantità di informazioni, spesso raccolte da Internet, da archivi digitali, database aziendali, articoli, social media e molte altre fonti. Quando tali dati provengono da fonti autorevoli, verificabili, o comunque esatte, i modelli tendono a sviluppare comportamenti e risposte coerenti con la realtà con sorprendente velocità. Tuttavia, l'affidabilità delle fonti non è sempre garantita, specialmente nel caso di scraping automatico del web o di popolamento di informazioni derivante dall'uso sin troppo disinvolto da parte degli utenti. In tal caso, il confine tra contenuto esatto e disinformazione è spesso labile.

Le fonti non affidabili possono provenire da diverse direzioni: siti web contenenti fake news, blog non verificati con contenuti imprecisi, commenti anonimi sui social, articoli pseudoscientifici, contenuti generati automaticamente da altri sistemi di IA (che potrebbero, di conseguenza, amplificare errori preesistenti) e via dicendo. L'inclusione di tali fonti nel corpus di addestramento può compromettere seriamente l'integrità del modello.

> **Uno dei principali rischi associati all'utilizzo di fonti inaffidabili è l'assimilazione di informazioni errate o fuorvianti da parte del sistema**. Un modello che ha appreso contenuti falsi o imprecisi può fornire risposte non corrette, replicare falsità o perpetuare teorie infondate.

Ad esempio, un chatbot addestrato su testi tratti da siti complottisti potrebbe inconsapevolmente riportare informazioni false su vaccini o altri argomenti sanitari, sul cambiamento climatico o anche su eventi storici. Se non sono stati applicati adeguati filtri di qualità o meccanismi (umani) di controllo, l'IA potrebbe sembrare convinta delle proprie affermazioni, insistendo e inducendo in errore gli utenti.

Questo problema diventa particolarmente pericoloso nei settori critici, come la medicina, il diritto, l'istruzione o la finanza, dove decisioni errate possono avere gravi ripercussioni. Un sistema diagnostico che fornisce suggerimenti clinici basati su dati non del tutto attendibili può mettere a rischio la salute del paziente. Allo stesso modo, un assistente legale automatizzato che trae informazioni da forum amatoriali o blog non ufficiali può generare documenti errati o fuorvianti.

Le fonti inaffidabili non solo possono contenere errori, ma possono anche riflettere pregiudizi culturali, stereotipi di genere, razziali o politici. Quando un sistema di IA assimila dati distorti, rischia di riprodurre e amplificare tali *bias* nel proprio comportamento. È noto, ad esempio, che modelli linguistici di grandi dimensioni possono riflettere disuguaglianze di genere, discriminazioni razziali o pregiudizi ideologici, se queste componenti sono presenti nei dati di addestramento.

La difficoltà sta nel fatto che tali distorsioni sono spesso sottili e sistemiche. Non si tratta solo di affermazioni esplicitamente offensive, ma anche di associazioni implicite e frequenze sbilanciate: se la maggior parte degli articoli su ruoli di leadership menziona uomini, ad esempio, mentre le donne vengono associate più frequentemente a ruoli di supporto, il modello potrebbe imparare – e perpetuare – questi schemi senza che sia immediatamente evidente, anche solo nel modo di esprimersi o nei giudizi che induce.

I modelli generativi, come quelli utilizzati per produrre testi, immagini o suoni, sono particolarmente vulnerabili alla propagazione degli errori. Se addestrati su dati contaminati, questi modelli possono "imparare" a generare contenuti scorretti, ingannevoli o addirittura dannosi.

> Un rischio ulteriore è l'effetto catena: quando i contenuti prodotti da IA contaminata vengono pubblicati online, potrebbero a loro volta essere raccolti da altri modelli in fase di addestramento. Questo ciclo di sub-alimentazione può portare a una crescente degradazione della qualità dell'informazione, in quello che alcuni ricercatori chiamano il fenomeno del **"model collapse"**, cioè l'indebolimento progressivo della capacità dei modelli di distinguere il vero dal falso.

Mitigare il problema?

Per contrastare i problemi derivanti da fonti inaffidabili, è essenziale adottare strategie di filtraggio e valutazione dei **dati a monte del processo** di addestramento. Alcune delle pratiche più efficaci possono includere, ad esempio:

1. **Cura manuale delle fonti**: selezionare attivamente informazioni basate su testi provenienti da fonti certificate (pubblicazioni scientifiche, enciclopedie, documentazione tecnica, archivi governativi, ecc.).
2. **Filtri automatici di qualità**: utilizzare algoritmi per identificare e rimuovere contenuti duplicati, inconsistenti, offensivi o incoerenti.
3. **Valutazione umana a posteriori**: affiancare l'automazione con revisioni umane per valutare l'affidabilità e la neutralità dei risultati prodotti dal modello.
4. **Tecniche di debiasing**: applicare metodi per ridurre la riproduzione dei bias nei dati, attraverso riequilibrio dei dataset o modifiche ai parametri di addestramento.
5. **Audit e trasparenza**: fornire documentazione aperta e tracciabilità sulle fonti utilizzate, in modo da permettere valutazioni indipendenti.

Considerazioni etiche e responsabilità

> La questione delle fonti non affidabili è strettamente connessa anche all'etica dell'IA. Le aziende e le istituzioni che sviluppano modelli devono assumersi la responsabilità di garantire che i loro sistemi non diventino vettori di disinformazione o strumenti di discriminazione.

Inoltre, è fondamentale che gli utenti siano consapevoli dei limiti dei modelli e del potenziale rischio di errore, specialmente quando si tratta di sistemi a scopo generale. **La trasparenza sulle fonti e sui processi di addestramento può aiutare a costruire un rapporto di fiducia tra tecnologia e società**.

Implicazioni sulla protezione dei dati personali

Un altro aspetto critico, spesso trascurato, nell'utilizzo di fonti non affidabili per l'addestramento dei sistemi di IA, riguarda la **protezione dei dati personali**. Molte fonti online non autorevoli – come forum, social network, blog amatoriali o siti non regolamentati – contengono **informazioni personali pubblicate in modo non conforme** alle normative sulla data protection, oppure **dati particolari** che dovrebbero essere esclusi dal processo di raccolta e trattamento.

Quando i modelli di IA vengono addestrati su dati raccolti in modo indiscriminato, rischiano di assimilare informazioni personali riconoscibili, come nomi, indirizzi email, numeri di telefono e, ancora peggio, informazioni sanitarie, opinioni politiche o dati biometrici. Questo può rappresentare una **violazione del Regolamento Generale sulla Protezione dei Dati (GDPR)** e di altre normative internazionali che impongono il consenso per l'uso di dati personali e il rispetto dei principi di minimizzazione e finalità.

> L'utilizzo di fonti non affidabili, infatti, **rende difficile stabilire l'origine dei dati**, verificarne la liceità e garantire i diritti degli interessati, come il diritto all'oblio o alla rettifica. Inoltre, i modelli addestrati su questi dati possono accidentalmente

> “rigenerare” informazioni personali nei loro output, specialmente se interrogati in modo mirato.

Anche quando i dati sono stati apparentemente anonimizzati, **l’aggregazione di fonti multiple e l’elevata capacità dei modelli di IA di trovare correlazioni indirette possono consentire la reidentificazione** degli individui. Questo è un rischio concreto soprattutto con modelli di grandi dimensioni (LLM), che possono ricostruire informazioni sensibili attraverso combinazioni di dati appresi.

Conclusioni

L’efficacia di un sistema di intelligenza artificiale dipende in larga parte dalla qualità dei dati con cui è stato addestrato. **L’utilizzo di fonti non affidabili compromette la precisione, l’eticità e l’utilità reale di tali sistemi, con effetti potenzialmente dannosi per individui e collettività**. In un’epoca in cui l’IA assume un ruolo sempre più centrale nella nostra vita quotidiana, è imprescindibile promuovere pratiche di addestramento trasparenti, responsabili e basate su fonti di comprovata affidabilità. Solo così sarà possibile costruire tecnologie realmente al servizio del bene comune.

> E, soprattutto, è fondamentale diffondere consapevolezza. Perché chiunque usi sistemi di IA sia conscio delle potenzialità ma anche consapevole dei rischi, tutt’altro che trascurabili, che rischia di generare e amplificare.