Dimensione Critica
Limiti e Potenzialità dell'Intelligenza Artificiale
Dati, Spiegabilità e Uso Consapevole dell'AI

Giovanni Della Lunga
giovanni.dellalunga@unibo.it

A lezione di Intelligenza Artificiale

Siena - Giugno 2025

Indice

L'Importanza dei Dati

Data and Information Literacy nell'era dell'IA

Una delle principali conseguenze della digitalizzazione è la **proliferazione incontrollabile dei dati** (Borgman, 2016).

Il processo di infrastrutturazione digitale delle attività informative, cognitive, educative e comunicative sta producendo un fenomeno senza precedenti:

≪**La generazione di enormi quantità di dati generati dall'azione umana all'interno delle piattaforme**≫ (Van Dijck, 2014).

Due concetti chiave:

- **Piattaformizzazione**: penetrazione delle piattaforme in infrastrutture, processi economici e culturali (Poell, Nieborg, Van Dijck, 2019).
- **Datafication**: trasformazione di pratiche storicamente non quantificabili in dati (Van Dijck, 2014).

La raccolta dei dati e l'importanza della consapevolezza

Il processo di datafication include non solo dati volontari (profilazione), ma anche:

- **Metadati comportamentali**: raccolti tramite app, plug-in, sensori, tracker, dispositivi mobili.
- Questi dati trasformano **ogni interazione umana** in un flusso informativo digitale.
- **Diffusione incontrollata**: i dati vengono spesso condivisi con attori esterni in modo imprevedibile.

"I dati sono il nuovo petrolio" (*The Economist*, 2017)

Riflessione critica:

- I recenti sviluppi dell'IA sono stati resi possibili da questa disponibilità massiva di dati.
- È **cruciale riflettere sul concetto di dato** per comprendere il funzionamento dei sistemi di IA e le implicazioni sociali.

La raccolta dei dati e l'importanza della consapevolezza

Come la condivisione di dati inconsapevole può creare problemi di sicurezza nazionale ad una superpotenza

Nel 2018, l'app di fitness Strava ha pubblicato una "heatmap" globale che mostrava le attività aggregate degli utenti, come corse e pedalate, tracciate tramite GPS. Sebbene l'intento fosse quello di fornire una visualizzazione delle rotte più popolari, è emerso che in aree isolate, come zone di conflitto o deserti, le tracce lasciate da militari in servizio erano chiaramente visibili. Questo ha permesso di identificare la posizione di basi militari segrete, rivelando anche i percorsi abituali dei soldati durante l'allenamento .

La raccolta dei dati e l'importanza della consapevolezza

Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

- Latest: Strava suggests military users 'opt out' of heatmap as row deepens

A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap

Ripensare il concetto di dato

Non possiamo ignorare i problemi legati all'acquisizione e all'uso dei dati personali.

Secondo Borgman (2016), i dati:

- non sono oggetti naturali;
- sono **rappresentazioni** di osservazioni, oggetti, entità;
- variano nel tempo, nel contesto e secondo l'osservatore.

≪**I dati esistono in un contesto, che ne influenza il significato insieme alla prospettiva dell'osservatore.**≫

Il dato non è neutro → può contenere stereotipi e pregiudizi.

Conclusione: è necessario promuovere una **critical data literacy education** per affrontare le sfide della *data society*.

Data literacy e cittadinanza digitale

Si è passati da una visione tecnico-statistica a una concezione della **data literacy** come prerequisito per la **partecipazione proattiva alla società digitale** (Carmi et al., 2020).

Tre livelli di data citizenship:

1. **Data thinking**: lettura, raccolta e comprensione critica dei dati;
2. **Data doing**: azioni concrete come la cancellazione e l'uso etico dei dati;
3. **Data participation**: attivismo civico e promozione della cultura dei dati.

Bhargava & D'Ignazio (2015):

- Competenza tecnico-matematica e critica;
- Capacità di leggere, creare e interpretare dati;
- Comprensione della realtà rappresentata nei dati;
- Narrazione pubblica basata sui dati.

Critical data literacy: definizione operativa

Selwyn & Pangrazio (2018) definiscono la *critical data literacy* come la capacità critica di gestione dei dati personali. Essa include:

- **Identificazione dei dati**: comprendere il tipo di dato (ceduto volontariamente o estratto automaticamente);
- **Comprensione dei dati**: sapere come vengono trattati e processati;
- **Riflessività sui dati**: analizzare le implicazioni legate al riuso;
- **Uso critico dei dati**: leggere Termini di servizio, configurare la privacy, ecc.;
- **Uso tattico dei dati**: impiego strategico nella prospettiva dell'attivismo civico.

Questa forma di alfabetizzazione è essenziale per affrontare i rischi dell'IA e della data society.

Bias e Altri Rischi

Bias e rischi cognitivi nei sistemi di IA

Bias di conferma → pregiudizio cognitivo per cui le persone tendono a interpretare nuove informazioni secondo convinzioni preesistenti.

Implicazioni per l'IA:

- I sistemi di IA possono rafforzare bias già presenti nei dati;
- Il rischio è che le decisioni automatizzate riflettano pregiudizi umani;
- Particolarmente critico in contesti ad alto impatto sociale:
  - Giustizia penale
  - Assistenza sanitaria
  - Finanza, alloggi, occupazione

È necessaria una **consapevolezza critica** per contrastare questi effetti sistemici.

Conseguenze critiche dell'IA generativa

Possibili rischi e implicazioni emerse dall'uso dell'IA generativa:

- **Conseguenze non intenzionali**: risultati imprevisti che richiedono il coinvolgimento di più attori per una valutazione attenta;
- **Atrofia mentale**: perdita di efficacia delle capacità cognitive dovuta alla delega del pensiero ai sistemi automatizzati;
- **Allucinazioni**: output insensati o fuorvianti generati dal modello; portano a informazioni errate o dannose;
- **Protezione della proprietà intellettuale**: il contenuto generato dall'IA solleva dubbi su copyright, uso non autorizzato e divulgazione accidentale di dati sensibili;
- **Erosione della fiducia**: rischio di manipolazione sociale, diffusione di notizie false e indebolimento dell'autorevolezza delle fonti.

Tutti questi elementi sottolineano la necessità di una **information literacy** critica e partecipata.

Conseguenze critiche dell'IA generativa

- Superare il **bias di conferma** $\rightarrow$ ricerca strategica, non conferma automatica;
- Contrastare l'**atrofia mentale** $\rightarrow$ esercizio del pensiero critico e flessibilità cognitiva;
- Ridurre le **allucinazioni** $\rightarrow$ confronto attivo con fonti e comunità;
- Promuovere **competenze critiche** $\rightarrow$ riflessione, confronto, valutazione.

Opacità algoritmica e responsabilità decisionale

Un modello è buono quanto lo sono i dati con cui è addestrato.

Con l'aumento della complessità dei modelli di IA:

- Diminuisce la trasparenza nel processo decisionale;
- È sempre più difficile comprendere il **come** e il **perché** di una decisione;
- Diventa arduo garantire **responsabilità decisionale** e possibilità di **interventi correttivi** (O'Neil, 2016).

Ciò richiede:

- una **postura critica** verso l'interpretabilità dei modelli;
- riflessioni etiche su potere, controllo e governance dell'IA;
- alfabetizzazione algoritmica come forma di cittadinanza.

XAI: eXplicable Artificial Intelligence

Explainability e responsabilità clinica

Le applicazioni di IA come "black box":

- Producono risultati accurati, ma spesso non interpretabili;
- Critiche in contesti ad alto impatto umano: medicina, diritto, educazione.

L'explainability (Ribeiro et al., 2016):

- Capacità di comprendere e spiegare il processo decisionale di un modello;
- Essenziale per evitare errori diagnostici, discriminazioni, stress e danni sistemici;
- Anche un modello preciso può essere rischioso se non comprensibile dagli utenti (es. medici).

Conclusione: la spiegabilità è *non solo desiderabile, ma essenziale* per un uso etico e responsabile dell'IA.

Trade-off tra precisione e spiegabilità

Una delle sfide principali dell'explainability è bilanciare:

- **Precisione**: maggiore nei modelli complessi, ma meno interpretabili;
- **Spiegabilità**: più alta nei modelli semplici, ma con minor accuratezza.

Il contesto è fondamentale:

- In medicina o diritto $\rightarrow$ meglio modelli spiegabili;
- In meteorologia o altri ambiti $\rightarrow$ può prevalere la precisione;
- Diversi utenti (ingegneri, medici, cittadini) hanno **diverse esigenze esplicative**.

Conclusione:

- Non esiste una "taglia unica" per l'explainability;
- Deve essere adattata all'utente;
- Aiuta a comprendere decisioni, prevenire errori e scoprire bias.

Modelli spiegabili e trasparenza decisionale

Trustworthy AI e linee guida europee (AI HLEG, 2019):

- Richiedono tracciabilità delle decisioni automatizzate;
- È essenziale documentare: dati, etichettatura, algoritmo e logica decisionale.

Explainability e metodi di trasparenza:

- **Metodi intrinseci**:
  - Alberi decisionali, regressione lineare/logistica, modelli GLM;
  - Interpretabili "nativamente", la relazione input/output è chiara.
- **Metodi post hoc**:
  - Spiegazioni applicate dopo l'elaborazione (es. LIME, SHAP).

Conclusione: un modello intrinsecamente spiegabile è una base solida per un'IA trasparente e responsabile.

Spiegabilità post hoc: LIME e SHAP

I metodi post hoc sono utilizzati per spiegare modelli complessi dopo l'addestramento.

Due tecniche principali:

- **LIME** (*Local Interpretable Model-agnostic Explanations*) Fornisce spiegazioni locali, indicando l'impatto di ciascuna feature sulla decisione;
- **SHAP** (*SHapley Additive exPlanations*) Usa la teoria dei giochi per valutare l'importanza di ogni variabile in modo consistente e globale (Lundberg & Lee, 2017).

Obiettivo: anche modelli opachi (es. reti neurali) diventano parzialmente interpretabili grazie a queste tecniche.

Inquiry Based Learning e Uso Critico dell'AI

ChatGPT e il pensiero critico: l'Inquiry Based Learning

IA generativa e accesso all'informazione:

- LLM come ChatGPT offrono nuove opportunità educative;
- Ma pongono sfide alla formazione del **pensiero critico**.

Rischi evidenziati:

- Dipendenza da risposte automatizzate;
- Diminuzione della capacità di analizzare, valutare, contestualizzare;
- Rischi educativi: copia-incolla, assenza di rielaborazione, plagio intellettuale (Ranieri, 2022).

Soluzione proposta: usare **ChatGPT come palestra** per sviluppare pensiero critico, secondo l'approccio **Inquiry Based Learning (IBL)**.

Il potere delle domande: Prompt design educativo

Non è l'output a fare la differenza, ma l'interazione:

- Esiste un "modo e modo" di porre domande a ChatGPT;
- Un prompt generico produce risposte generiche;
- Un prompt ben costruito stimola analisi, verifica, confronto.

Esempio educativo:

- Chiedere una valutazione delle fonti;
- Richiedere il punteggio di affidabilità con motivazioni;
- Simulare un dialogo argomentativo.

Conclusione: ChatGPT può diventare una **palestra per la formulazione di domande**, allenando lo **spirito critico** e la capacità di riflessione.

Inquiry Based Learning: la struttura in 5 fasi

L'IBL (Inquiry Based Learning) è un approccio didattico che:

- Collega esplorazione e metacognizione;
- Si basa sulla scoperta guidata, l'ipotesi, la verifica;
- È adatto per integrare ChatGPT in una didattica attiva e critica.

Inquiry Cycle (Pedaste et al., 2015):

1. **Orientamento**: introduzione al problema;
2. **Concettualizzazione**: generazione di domande e ipotesi;
3. **Scoperta**: esplorazione e analisi;
4. **Conclusione**: sintesi e soluzione proposta;
5. **Discussione**: valutazione e confronto finale.

Prima fase operativa: *Familiarizzare con ChatGPT*.

IBL con ChatGPT: le prime tre fasi operative

Fase 1 – Familiarizzazione:

- Studenti in coppie interagiscono con ChatGPT;
- Provano prompt, valutano le risposte, migliorano l'interazione.

Fase 2 – Generare un testo:

- Suddivisione in gruppi;
- Ogni gruppo elabora un prompt per risolvere un problema;
- Si genera un testo con ChatGPT e si confrontano gli output.

Fase 3 – Mettere alla prova ChatGPT:

- Discussione plenaria;
- Analisi di qualità, affidabilità e coerenza;
- Riflessione su prompt e ruolo dell'utente.