Profice

AI ACADEMY

Applicare l’Intelligenza Artificiale nello sviluppo software

AI ACADEMY

AI bias, fairness ed etica applicata
07/07/2025

INTRODUZIONE DELL'ISTRUTTORE

Tamas Szakacs

Formazione

- Laureato come programmatore matematico
- MBA in management

Principali esperienze di lavoro

- Amministratore di sistemi UNIX
- Oracle DBA
- Sviluppatore di Java, Python e di Oracle PL/SQL
- Architetto (solution, enterprise, security, data)
- Ricercatore tecnologico e interdisciplinare di IA

Dedicato alla formazione continua

- Teorie, modelli, framework IA
- Ricerche IA
- Strategie aziendali
- Trasformazione digitale
- Formazione professionale

email: tamas.szakacs@proficegroup.it

MOTIVI E RIASSUNTO DEL CORSO

L'**Intelligenza Artificiale (AI)** è oggi il motore dell'innovazione in ogni settore, grazie alla sua capacità di analizzare dati, automatizzare processi e generare nuove soluzioni. Questo corso offre una panoramica completa e pratica sullo sviluppo di applicazioni AI moderne, guidando i partecipanti dall'ideazione al rilascio in produzione.

Attraverso una **combinazione di teoria chiara ed esercitazioni pratiche**, saranno affrontate le tecniche e gli strumenti più attuali: **machine learning, deep learning, reti neurali, Large Language Models (LLM), Transformers, Retrieval Augmented Generation (RAG)** e progettazione di agenti AI.
Le competenze acquisite saranno applicate in progetti concreti, dallo sviluppo di chatbot all'integrazione di modelli generativi, fino al deploy di soluzioni AI in ambienti reali e collaborativi.

Il percorso è pensato per chi vuole imparare a progettare, valutare e integrare sistemi AI di nuova generazione, con particolare attenzione alle best practice di programmazione, collaborazione in team, sicurezza, valutazione delle performance ed etica dell'AI.

DURATA: 17 GIORNI

OBIETTIVI

Il percorso formativo è progettato per **giovani consulenti junior**, con una conoscenza base di programmazione, che stanno iniziando un percorso professionale nel settore AI.

L’obiettivo centrale è fornire una panoramica pratica, completa e operativa sull’intelligenza artificiale moderna, guidando ogni partecipante attraverso tutte le fasi fondamentali.

OBIETTIVI

- Allineare conoscenze AI, ML, DL di tutti i partecipanti
- Saper usare e orchestrare modelli LLM (closed e open-weight)
- Costruire pipeline RAG complete (retrieval-augmented generation)
- Progettare agenti AI semplici con strumenti moderni (LangChain, tool calling)
- Capire principi di valutazione, robustezza e sicurezza dei sistemi GenA
- Migliorare la produttività come sviluppatori usando tool GenAI-driven
- Padroneggiare best practice di sviluppo, versioning e deploy AI
- Introdurre i fondamenti di Graph Data Science e Knowledge Graph
- Ottenere capacità di valutazione dei modelli e metriche
- Comprensione dell'etica e dei bias nei modelli di intelligenza artificiale
- Approfondire le normative di riferimento: AI Act, compliance e governance AI

Il corso è **estremamente pratico** (circa il 40% del tempo in esercitazioni hands-on, notebook, challenge e hackathon), con l'utilizzo di Google Colab, GitHub, e tutti gli strumenti necessari per lavorare su progetti reali e simulati.

STRUTTURA DELLE GIORNATE – PROGRAMMA BREVE

Tutte le giornate sono di 8 ore (9:00-17:00), con 1 ora di pausa suddivisa (mezz'ora pranzo, due pause da 15 min durante la mattina e il pomeriggio).

La progettazione sintetica delle giornate:

Giorno	Tema	Breve descrizione
1	Git & Python clean-code	Collaborazione su progetti reali, versionamento, codice pulito e testato
2	Machine Learning Supervised	Modelli supervisionati per predizione e classificazione
3	Machine Learning Unsupervised	Clustering, riduzione dimensionale, scoperta di pattern
4	Prompt Engineering avanzato	Scrivere e valutare prompt efficaci per modelli generativi
5	LLM via API (multi-vendor)	Uso pratico di modelli LLM via API, autenticazione, deployment
6	Come costruire un RAG	Pipeline end-to-end per Retrieval-Augmented Generation
7	Tool-calling & Agent design	Progettare agenti AI che usano strumenti esterni
8	Hackathon: Agentic RAG	Challenge pratica: chatbot agentico RAG in team

STRUTTURA DELLE GIORNATE – PROGRAMMA BREVE

Tutte le giornate sono di 8 ore (9:00-17:00), con 1 ora di pausa suddivisa (mezz'ora pranzo, due pause da 15 min durante la mattina e il pomeriggio).

La progettazione sintetica delle giornate:

Giorno	Tema	Breve descrizione
9	Hackathon: Rapid Prototyping	Da prototipo a web-app con Streamlit e GitHub
10	AI Productivity Tools	Workflow con IDE AI-powered, automazione e refactoring assistito
11	Docker & HF Spaces Deploy	Deployment di app GenAI containerizzate o su HuggingFace Spaces
12	AI Act & ISO 42001 Compliance	Fondamenti di compliance e governance AI
13	Knowledge Base & Graph Data Science	Introduzione a Knowledge Graph e query con Neo4j
14	Model evaluation & osservabilità	Metriche avanzate, explainability, strumenti di valutazione
15	AI bias, fairness ed etica applicata	Analisi dei rischi, metriche e mitigazione dei bias
16-17	Project Work & Challenge finale	Lavoro a gruppi, POC/POD, presentazione e votazione progetti

METODOLOGIA DEL CORSO

1. Approccio introduttivo ma avanzato

Il corso è introduttivo nei concetti base dell'AI applicata allo sviluppo, ma affronta anche tecnologie, modelli e soluzioni avanzate per garantire un apprendimento completo.

2. Linguaggio adattato

Il linguaggio utilizzato è chiaro e adattato agli studenti, con spiegazioni dettagliate dei termini tecnici per favorirne la comprensione e l'apprendimento graduale.

3. Esercizi pratici

Gli esercizi pratici sono interamente svolti online tramite piattaforme come Google Colab o notebook Python, eliminando la necessità di installare software sul proprio computer.

4. Supporto interattivo

È possibile porre domande in qualsiasi momento durante le lezioni o successivamente via email per garantire una piena comprensione del materiale trattato.

NOTA

Il corso segue un **approccio laboratoriale**: ogni giornata combina sessioni teoriche chiare e concrete con molte attività pratiche supervisionate, per sviluppare *competenze reali* immediatamente applicabili.

I partecipanti lavoreranno spesso in gruppo, useranno notebook in Colab e versioneranno codice su GitHub, vivendo una vera simulazione del lavoro in azienda AI.

Nessun prerequisito avanzato richiesto: si partirà dagli strumenti e flussi fondamentali, con una crescita graduale verso le tecniche più attuali e richieste dal mercato.

ORARIO TIPICO DELLE GIORNATE

Orario	Attività	Dettaglio
09:00 – 09:30	Teoria introduttiva	Concetti chiave, schema della giornata
09:30 – 10:30	Live coding + esercizio guidato	Esempio pratico, notebook Colab
10:30 – 10:45	*Pausa breve*	
10:45 – 11:30	Approfondimento teorico	Tecniche, best practice
11:30 – 12:30	Esercizio hands-on individuale	Sviluppo o completamento di codice
12:30 – 13:00	Discussione soluzioni + Q&A	Condivisione e correzione
13:00 – 14:00	*Pausa pranzo*	
13:30 – 14:15	Teoria avanzata / nuovi tools	Nuovi strumenti, pattern, demo
14:15 – 15:30	Esercizio a gruppi / challenge	Lavoro di squadra su task reale
15:30 – 15:45	*Pausa breve*	
15:45 – 16:30	Sommario teorico e pratico	
16:30 – 17:00	Discussioni, feedback	Riepilogo, best practice, domande aperte

DOMANDE?

Cominciamo!

OBIETTIVI DELLA GIORNATA

Obiettivi della giornata

- Riconoscere le principali forme di bias nei dati, negli algoritmi e nelle interazioni uomo-macchina.
- Analizzare le metriche di fairness più usate per valutare la non-discriminazione nei modelli AI.
- Saper utilizzare tool per l'audit dei dataset e la rilevazione del bias.
- Applicare tecniche pratiche di debiasing: oversampling, reweighting, approcci adversarial.
- Analizzare casi di discriminazione algoritmica (es. recruiting, credit scoring).
- Applicare esercizi su dataset reali (COMPAS, Adult Income, HuggingFace datasets) per valutare e mitigare il bias.
- Conoscere le principali linee guida etiche (OCSE, UNESCO, EU AI Act).

ESERCIZIO: VALUTAZIONE METRICA E FEEDBACK UMANO

Obiettivo:
Rendere capace l'agente RAG (NER + GPT su dati aziendali) che i suoi risultati vengano valutati automaticamente tramite una metrica (es. F1-score, ROUGE, ecc.)
(Opzionale) I risultati possono essere valutati opzionalmente tramite feedback esplicito dell'utente (voto, commento). L'interazione successiva deve considerare questa valutazione per dare la risposta..

Step operativi:

1. **Generazione risposta:**
   - L'agente riceve una domanda (query) e produce una risposta sfruttando i documenti e la knowledge base.
2. **Valutazione automatica:**
   - Per ogni risposta, calcolare una metrica di qualità rispetto alla risposta attesa, ad esempio:
     - F1-score, se è classificazione.
     - ROUGE o BERTScore, se è risposta testuale.
   - Visualizzare e registrare il valore calcolato nella chat.

ESERCIZIO: VALUTAZIONE METRICA E FEEDBACK UMANO

Step operativi:

3. **Feedback utente (opzionale):**
   - L’interfaccia deve permettere all’utente di valutare la risposta (ad es. punteggio 1-5, o “utile/non utile”, oppure un commento testuale).
   - Salvare questo feedback insieme ai risultati automatici.
4. **Analisi e miglioramento (opzionale):**
   - Dopo almeno 10 risposte, mostrare statistiche aggregate (es. media dei punteggi automatici e dei feedback umani).
   - Permettere all’utente di rivedere alcune risposte con bassa valutazione, e correggere (ad es. riformulare prompt o migliorare l’agente).

PERCHÉ BIAS E FAIRNESS SONO TEMI CALDI IN AI OGGI?

Profice

Il bias e fairness in AI

- L'AI è sempre più usata per decidere chi ottiene un lavoro, un mutuo, o anche solo cosa vediamo online.
- Se i dati sono "sporchi" o sbilanciati, i modelli rischiano di prendere decisioni ingiuste (anche senza volerlo!).
- Una decisione automatica sbagliata può creare problemi seri: esclusioni, errori, discriminazioni.
- Oggi le aziende (e i developer!) devono stare attente: evitare bias non è solo una questione etica, ma anche di legge e reputazione.
- In pratica: AI affidabile = AI attenta al bias. Serve per la fiducia, la qualità e per rispettare le regole (EU AI Act e simili).

Da dove nasce il bias nei dati?

- **Genere**
  Esempio: Un modello HR "preferisce" candidati da una specifica zona geografica.
- **Etnia**
  Dati raccolti soprattutto su un gruppo etnico.
- **Età**
  Modelli che "favoriscono" giovani o anziani a seconda dei dati a disposizione.
- **Regione/lingua**
  Training su dati di una zona/language, e il modello non funziona bene altrove.
- **Condizioni socio-economiche**
  Dati da gruppi più abbienti, e l'AI capisce poco le esigenze di altri.
- **Altri fattori**
  Disabilità, orientamento, status familiare… tutto ciò che non è rappresentato correttamente può creare bias.

BIAS E FAIRNESS IN AI

Bias in AI

Il bias nell'intelligenza artificiale è una distorsione che porta il modello a favorire o penalizzare certi gruppi o individui, spesso in modo involontario. Questa distorsione può nascere dai dati di addestramento, dalle scelte progettuali o da come viene utilizzato il sistema. Il risultato è che le decisioni dell'AI non sono neutre, ma riflettono – e talvolta amplificano – squilibri o pregiudizi presenti nei dati o nella società.

Fairness in AI

La fairness (equità) in AI significa che un sistema prende decisioni giuste e imparziali, trattando tutti gli utenti o i gruppi allo stesso modo, senza favoritismi o discriminazioni ingiustificate. Raggiungere la fairness richiede attenzione sia nella progettazione che nella valutazione del modello, utilizzando dati rappresentativi e metriche specifiche per garantire risultati equi e affidabili per tutti.

BIAS NEI DATI

Bias nei dati: esempi e fonti

Anche i dati che usiamo per addestrare i modelli possono portare dentro pregiudizi. Vediamo come nasce il bias già nella fase di raccolta e selezione dei dataset.

Bias nei dati: esempi e fonti

- **Dataset sbilanciati**
  Se il dataset contiene molti più esempi di un gruppo rispetto ad altri, il modello imparerà a “preferire” il gruppo più rappresentato.
  Esempio*:* 90% dati di uomini, 10% di donne → risultati poco affidabili per le donne.
- **Errori di raccolta**
  Dati raccolti solo in certi luoghi, orari o condizioni possono escludere gruppi importanti.
  Esempio*:* Sondaggio online che esclude chi non usa internet, dati sanitari presi solo da ospedali cittadini.
- **Dati storici con pregiudizi**
  Se i dati riflettono decisioni passate già influenzate da bias umani, l’AI li replica.
  Esempio*:* Vecchi record di assunzione che penalizzavano alcune etnie o generi.

Conclusione:

Per valutare sistemi generativi (es. chatbot, summarization, traduzione automatica) servono metriche più avanzate e specifiche, capaci di cogliere la qualità linguistica e la vicinanza semantica tra output e riferimento.

BIAS NEGLI ALGORITMI

Bias nei dati: esempi e fonti

- **Dataset sbilanciati**
  Se il dataset contiene molti più esempi di un gruppo rispetto ad altri, il modello imparerà a “preferire” il gruppo più rappresentato.
  Esempio*:* 90% dati di uomini, 10% di donne → risultati poco affidabili per le donne.
- **Errori di raccolta**
  Dati raccolti solo in certi luoghi, orari o condizioni possono escludere gruppi importanti.
  Esempio*:* Sondaggio online che esclude chi non usa internet, dati sanitari presi solo da ospedali cittadini.
- **Dati storici con pregiudizi**
  Se i dati riflettono decisioni passate già influenzate da bias umani, l’AI li replica.
  Esempio*:* Vecchi record di assunzione che penalizzavano alcune etnie o generi.

In pratica:
Un dataset “difettoso” genera modelli poco equi, anche se l’algoritmo è corretto.

ALTRI FONTI DI BIAS

Bias nell’interazione

- **UX (User Experience):**
  Un’interfaccia poco inclusiva o poco chiara può portare alcuni utenti a fare errori, influenzando i dati raccolti.
- **Feedback loop:**
  Se il modello si aggiorna continuamente con i dati prodotti dagli utenti, può rafforzare i bias già presenti (“effetto eco”).
- **Label leakage:**
  Quando le etichette o le risposte dell’AI diventano visibili agli utenti, questi possono copiarle o adattarsi, falsando i dati futuri.

In sintesi:
Il modo in cui l’AI interagisce con le persone può creare o amplificare bias nel tempo.

BIAS IN RAG

Bias nei sistemi RAG

- I sistemi RAG combinano un modello generativo (es. LLM) con un motore di recupero di informazioni (retriever).
- Il bias può nascere **nei documenti recuperati**: se la knowledge base è sbilanciata o incompleta, l'output dell'AI rifletterà quei limiti.
- Anche il **retriever** può "preferire" certi contenuti, escludendo fonti meno rappresentate o minoritarie.
- Il modello generativo può amplificare i bias già presenti nei dati recuperati o nei prompt.

Esempio:
Se il sistema RAG per un chatbot HR recupera solo documenti con storie di manager, le risposte AI tenderanno a rappresentare solo quel gruppo.

In pratica:
La qualità e la varietà delle fonti usate nel retrieval sono fondamentali per ridurre il rischio di bias nei risultati.

ESEMPI REALI DI BIAS: CASI FAMOSI

Amazon Recruiting
Sistema AI scartava automaticamente CV di donne per ruoli tecnici
Il modello aveva ‘imparato’ dai dati storici, tutti maschili.

COMPAS (USA)
Valutazione rischio recidiva più alta per afroamericani
Algoritmo accusato di essere più severo con alcuni gruppi.

Riconoscimento facciale
Errori maggiori su donne e persone con pelle scura
(MIT/IBM: fino a 34% di errore contro 1% su volti bianchi maschili)

Annunci lavoro su Facebook
Ruoli STEM mostrati più spesso agli uomini
Targeting automatico riproduce stereotipi.

L’intelligenza artificiale è potente, ma riflette chi siamo e i dati che le diamo.

DOMANDE?

PAUSA

ANALISI DEL BIAS NEL DATASET – ADULT INCOME

Cos'è:
Dataset pubblico sui redditi negli Stati Uniti, usato per studiare se si possono prevedere guadagni sopra/sotto i 50K$.

Gruppi sensibili
- Genere (sex), etnia (race), età.

Come si analizza il bias:
- Conta quanti esempi ci sono per ogni gruppo (es: uomini/donne, etnie diverse)
- Calcola la percentuale di ">50K" per ogni gruppo
- Cerca squilibri: se solo il 10% delle donne ha income >50K contro il 30% degli uomini, il modello può imparare a "favorire" un gruppo

Strumenti:
pandas, matplotlib per i grafici

Domande da porsi:
- Tutti i gruppi sono rappresentati abbastanza?
- I risultati sono equilibrati tra i gruppi?
- Dove potrebbe nascere il bias se addestriamo un modello su questi dati?

ANALISI DEL BIAS NEL DATASET – ADULT INCOME

Scarica il dataset
Adult Data Set – UCI: https://archive.ics.uci.edu/dataset/2/adult

Analizza le colonne sensibili:

```
# Colonne, come definite nel dataset UCI
columns = [
    "age", "workclass", "fnlwgt", "education", "education-num",
    "marital-status", "occupation", "relationship", "race", "sex",
    "capital-gain", "capital-loss", "hours-per-week", "native-country", "income"
]

# Scarica e carica il dataset
df = pd.read_csv(url, header=None, names=columns, na_values=" ?", skipinitialspace=True)

print(df["sex"].value_counts())
print(df["race"].value_counts())
```

DOMANDA PER LA DISCUSSIONE

Domande – Bias nel dataset

- Quanti dati ci sono per ciascun gruppo (es. uomini/donne, diverse etnie)?
- Noti differenze marcate nelle percentuali di income alto (>50K) tra i gruppi?
- Secondo te, queste differenze riflettono la realtà o potrebbero essere dovute a come sono stati raccolti i dati?
- Se un modello si addestra su questi dati, quali rischi di bias vedi?
- Ci sono gruppi sottorappresentati o assenti del tutto? Che impatto può avere?
- Come potresti rendere il dataset più “fair”?
- Pensi che il modello AI riuscirebbe a essere imparziale partendo da questi dati?
- Quali soluzioni pratiche proporresti per mitigare il bias nei dati o nel modello?

FAIRNESS IN AI: PERCHÉ È COMPLICATA?

Fairness in AI

La fairness è la capacità di un sistema di intelligenza artificiale di prendere decisioni eque, senza creare vantaggi o svantaggi ingiustificati tra individui o gruppi diversi.
Richiede attenzione nella scelta dei dati, nelle metriche di valutazione e nel contesto di utilizzo, per garantire risultati affidabili e il più possibile imparziali.

- **Tanti tipi di fairness:**
  Gruppi, individui, opportunità, risultati… quale scegliere?
- **Un equilibrio difficile:**
  Aiutare un gruppo può danneggiarne un altro: non si può ottimizzare tutto insieme.
- **Metriche in conflitto:**
  Spesso le diverse metriche di fairness danno indicazioni opposte sulla stessa situazione.
- **Cambia secondo il contesto:**
  Sanità, credito, lavoro: cosa è “giusto” dipende dal settore e dalle conseguenze.
- **Anche la società conta:**
  Le aspettative su cos’è equo variano tra paesi, aziende, culture.

DOMANDA PER LA DISCUSSIONE

Secondo voi, un’AI che porta in sé dei bias e non è sempre capace di fairness, potrebbe imparare e migliorare da sola nel tempo?

Cosa servirebbe perché questo succeda davvero?

DOMANDA PER LA DISCUSSIONE

Secondo voi, un’AI che porta in sé dei bias e non è sempre capace di fairness, potrebbe imparare e migliorare da sola nel tempo?

Cosa servirebbe perché questo succeda davvero?

I bias sono spesso difficili da identificare e la fairness non è facile da applicare.

DOMANDA PER LA DISCUSSIONE

Secondo voi, un’AI che porta in sé dei bias e non è sempre capace di fairness, potrebbe imparare e migliorare da sola nel tempo?

Cosa servirebbe perché questo succeda davvero?

I bias sono spesso difficili da identificare e la fairness non è facile da applicare.

Non sempre possiamo cambiare l’apprendimento del modello.

DOMANDA PER LA DISCUSSIONE

Secondo voi, un'AI che porta in sé dei bias e non è sempre capace di fairness, potrebbe imparare e migliorare da sola nel tempo?

Cosa servirebbe perché questo succeda davvero?

I bias sono spesso difficili da identificare e la fairness non è facile da applicare.

Non sempre possiamo cambiare l'apprendimento del modello.

Con tecniche avversariali possiamo controllare e modificare le risposte.

DOMANDA PER LA DISCUSSIONE

Secondo voi, un'AI che porta in sé dei bias e non è sempre capace di fairness, potrebbe imparare e migliorare da sola nel tempo?

Cosa servirebbe perché questo succeda davvero?

I bias sono spesso difficili da identificare e la fairness non è facile da applicare.

Non sempre possiamo cambiare l'apprendimento del modello.

Con tecniche avversariali possiamo controllare e modificare le risposte.

Con il fine-tuning possiamo modificare il comportamento del modello e ridurre (o correggere) eventuali bias, rendendo le sue risposte più eque e bilanciate.

DEMOGRAPHIC PARITY – DEFINIZIONE E FORMULA

Definizione:
Un modello soddisfa la demographic parity se la probabilità di ricevere una certa predizione positiva (es. “income >50K”) è la stessa per tutti i gruppi protetti (es. uomini e donne, etnie diverse).

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

Dove:

- $\hat{Y}$ = 1 = predizione positiva
- A = attributo protetto (es. sesso, etnia)

In pratica:
Il modello non deve “favorire” un gruppo rispetto a un altro solo per appartenenza al gruppo.

DEMOGRAPHIC PARITY – DEFINIZIONE E FORMULA

Cosa significa Demographic Parity?
Se un modello rispetta la demographic parity, la percentuale di persone che ricevono la predizione positiva (ad esempio: “viene approvato il prestito” o “income >50K”) è uguale in tutti i gruppi, indipendentemente da genere, etnia o altra caratteristica protetta.
Non importa se le persone sono diverse tra loro: il modello assegna lo stesso tasso di “successo” a ciascun gruppo.

Esempio:
Se il 25% degli uomini riceve “income >50K”, anche il 25% delle donne dovrebbe riceverlo.

Limite:
Demographic parity non considera se le differenze tra i gruppi siano dovute a fattori reali nei dati. Si concentra solo sul risultato “finale” uguale per tutti, anche se la distribuzione di caratteristiche tra i gruppi è diversa.

EQUALIZED ODDS – DEFINIZIONE

Definizione:
Equalized Odds richiede che il modello abbia le stesse probabilità di fare errori (falsi positivi e falsi negativi) per ogni gruppo protetto, dati i veri valori.

In pratica:
La probabilità di una predizione corretta o errata deve essere uguale tra i gruppi, a parità di verità sottostante.

Differenza con Demographic Parity (DP)

- **DP** guarda solo alla percentuale complessiva di predizioni positive per ciascun gruppo, senza considerare se la predizione è corretta o meno.
- **Equalized Odds** invece tiene conto della correttezza: richiede che il modello sia equo sia nei veri positivi che nei veri negativi per ogni gruppo.
- **Esempio pratico:**
  - DP: Stessa percentuale di prestiti approvati in ogni gruppo.
  - EO: Stessa percentuale di approvazioni corrette e rifiuti corretti in ogni gruppo.

ALTRE METRICHE DI FAIRNESS

Predictive parity

Il modello soddisfa la predictive parity se la probabilità che una predizione positiva sia davvero corretta è uguale tra tutti i gruppi.

Esempio: se il modello dice “income >50K”, questa predizione dovrebbe essere giusta nella stessa percentuale per ogni gruppo.

Individual fairness

Richiede che persone simili ricevano decisioni simili dal modello, indipendentemente dal gruppo a cui appartengono.

Esempio*:* due candidati con profili quasi identici dovrebbero avere la stessa probabilità di ottenere un prestito, anche se appartengono a gruppi diversi.

LIMITI DELLE METRICHE DI FAIRNESS

- **Tradeoff tra metriche:**
  Migliorare una metrica di fairness (es. demographic parity) può peggiorarne un'altra (es. equalized odds). Non si possono spesso ottimizzare tutte insieme.
- **Impossibility theorem:**
  È stato dimostrato che, in molti casi, non è possibile soddisfare contemporaneamente tutte le metriche di fairness, a meno che i gruppi siano identici nei dati di partenza.
- **Cosa significa in pratica?**
  Spesso bisogna scegliere quali metriche sono più importanti per il proprio caso d'uso, spiegare questa scelta e accettare dei compromessi.

In sintesi:
Non esiste una fairness "perfetta": ogni scelta ha vantaggi e svantaggi.

TOOL PER ANALIZZARE BIAS E FAIRNESS NEI MODELLI AI

- **Fairlearn**
  Libreria Python per analizzare, visualizzare e mitigare bias nei modelli di machine learning.
  Offre metriche di fairness e strumenti per produrre report dettagliati.
- **AIF360 (AI Fairness 360)**
  Toolkit open-source sviluppato da IBM con tantissime metriche di fairness e tecniche di debiasing.
  Supporta dataset classici come Adult Income, COMPAS, German Credit.
- **What-If Tool (Google)**
  Interfaccia visuale per TensorFlow e scikit-learn che permette di esplorare interattivamente dati, predizioni e fairness.
- **Responsibly**
  Toolkit più leggero per valutare e spiegare metriche di fairness in modelli di classificazione.
- **Shap, LIME**
  Strumenti per l'interpretabilità dei modelli che aiutano a capire l'origine del bias nelle predizioni.

ESERCIZIO: VISUALIZZARE LE METRICHE DI FAIRNESS

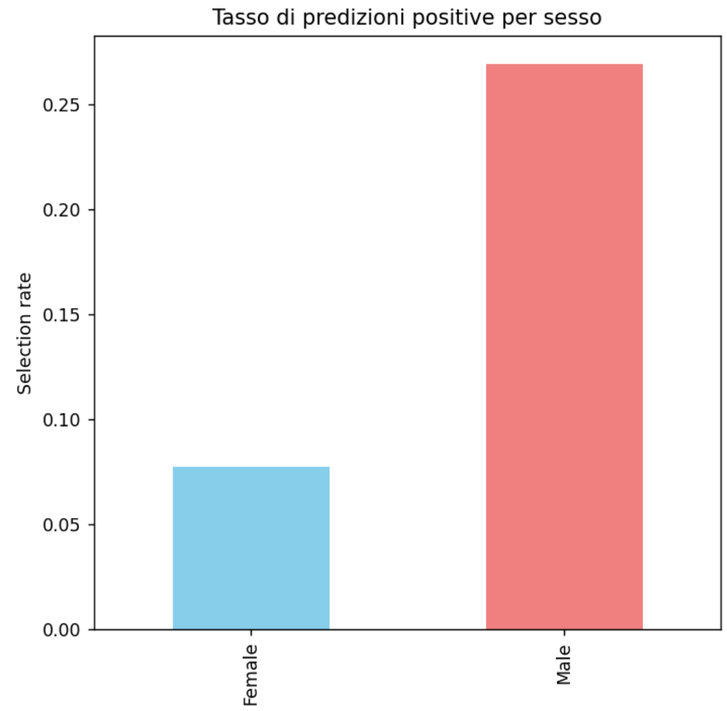
- Calcoliamo la percentuale di predizioni positive (>50K) per uomini e donne:
  Esempio:
  Uomini: 30%
  Donne: 10%
- Usiamo un grafico a barre per confrontare i gruppi:
- (Mostrare un grafico: asse X = sesso, asse Y = % predizione positiva)
- Calcoliamo la precisione del modello in ogni gruppo:
  Esempio:
  Precision uomini: 75%
  Precision donne: 68%
- Confrontiamo le differenze e discutiamo:
  - Il modello rispetta la demographic parity?
  - Ci sono gap di accuratezza tra i gruppi?

ESERCIZIO: FAIRLEARN SU ADULT INCOME

Cosa mostra Fairlearn?

- **Misura** quanto il modello “amplifica”, “riduce” o “trasforma” lo sbilancio già presente nei dati.
- Permette di confrontare il tasso di predizioni positive tra i gruppi protetti dopo che il modello ha imparato dai dati.
- Se la differenza di demographic parity è grande, il modello **sta trasferendo (o aumentando) il bias** nei risultati.
- In classe: aiuta a vedere concretamente che un modello “corretto” può comunque essere poco equo.

```
from fairlearn.metrics import MetricFrame
mf = MetricFrame(
    metrics=selection_rate,
    y_true=y_test,
    y_pred=y_pred,
    sensitive_features=sex_test
)
print("\nDemographic Parity):")
print(mf.by_group)
```



COSA AIUTA A TROVARE I BIAS?

- **Analisi esplorativa dei dati**
  Guardare le distribuzioni delle variabili sensibili (es. genere, etnia) e delle etichette di output.
- **Metriche di fairness**
  Usare strumenti come Fairlearn, AIF360 per calcolare differenze tra gruppi (demographic parity, equalized odds, ecc.).
- **Visualizzazioni**
  Grafici a barre, heatmap e tabelle che confrontano i risultati tra diversi gruppi.
- **Audit dei dati**
  Verificare se alcuni gruppi sono sottorappresentati o hanno risultati molto diversi.
- **Simulazioni e test avversariali**
  Provare input specifici per vedere se il modello tratta i gruppi in modo diverso.

DATASET AUDITING: COS'È E PERCHÉ SERVE

Cos'è:
È il processo di analizzare sistematicamente un dataset per identificare squilibri, errori o potenziali bias (es. gruppi sottorappresentati, dati mancanti, etichette errate).

Perché serve:
Un audit accurato permette di:

- Scoprire problemi prima di addestrare i modelli
- Prevenire che il bias dei dati si trasferisca nel modello
- Documentare la qualità e l'equità del dataset (utile anche per la compliance normativa)
- Migliorare la trasparenza e la fiducia nell'intero processo AI

Auditare i dati è il primo passo per costruire sistemi AI più equi e affidabili.

AUDIT PRATICO: COME LEGGERE I REPORT DI FAIRNESS

- **Guarda i tassi di predizione positiva**
  Sono simili tra i gruppi? Se no, c'è rischio di bias.
- **Analizza le metriche chiave**
  Demographic parity, equalized odds, predictive parity…
  Valori molto diversi indicano squilibri.
- **Cerca gap e anomalie**
  Differenze marcate o "zero" predizioni per un gruppo segnalano problemi.
- **Osserva anche le performance globali**
  Un modello può essere accurato in media ma ingiusto per certi gruppi.
- **Non fermarti al numero:**
  Rifletti sulle cause: dipende dai dati, dal modello o da entrambi?

Usa i report di fairness per individuare dove intervenire e quali azioni proporre (es. bilanciamento dati, debiasing).

APPROCCI GENERALI AL DEBIASING

- **Preprocessing dei dati**
  Bilanciare i dati prima dell'addestramento (oversampling, undersampling, reweighting).
- **Modifica del modello**
  Aggiungere vincoli o penalità di fairness durante il training, oppure usare modelli specifici progettati per la fairness.
- **Postprocessing delle predizioni**
  Correggere le predizioni del modello dopo l'addestramento per ridurre il bias (es. changing thresholds, equalizing outcomes tra gruppi).
- **Feedback continuo**
  Monitorare le prestazioni e la fairness nel tempo, aggiornando il modello quando emergono nuovi bias.

Il debiasing può intervenire prima, durante o dopo l'addestramento del modello. Spesso serve combinare più strategie.

OVERSAMPLING E UNDERSAMPLING

Oversampling
Si aumentano gli esempi del gruppo minoritario, copiandoli o generandone di simili, così che tutti i gruppi abbiano lo stesso peso.
Esempio: Se ci sono poche donne nel dataset, si duplicano le loro righe finché sono pari agli uomini.

Undersampling
Si riducono gli esempi del gruppo maggioritario, eliminandone alcuni per equilibrare le classi.
Esempio: Se ci sono troppi uomini, se ne tengono solo quanti sono le donne, scartando gli altri.

Obiettivo:
Evitare che il modello “ignori” il gruppo meno rappresentato, rendendo la previsione più equa.

REWEIGHTING: COME FUNZIONA E CASI D'USO

Cos'è:
Ogni esempio del dataset viene "pesato" in modo diverso: i casi dei gruppi minoritari (o svantaggiati) ricevono un peso maggiore nell'addestramento, quelli dei gruppi maggioritari un peso minore.

Come funziona:
Il modello "impara" dando più importanza agli esempi poco rappresentati, così da non trascurarli nelle predizioni.

Casi d'uso:
- Quando non vuoi duplicare o eliminare dati (come con oversampling/undersampling)
- Per dataset con grandi squilibri tra classi o gruppi sensibili
- In problemi dove i dati sono costosi o difficili da raccogliere per alcuni gruppi

Vantaggio:
Corregge il bias senza modificare direttamente la composizione del dataset.

APPROCCI ADVERSARIALI: PANORAMICA

Cos'è:
Si usa una "rete avversaria" che cerca di capire da quale gruppo proviene un esempio, mentre il modello principale cerca di fare predizioni senza rivelare questa informazione.

Come funziona:
Il modello principale viene penalizzato se l'avversario riesce a indovinare il gruppo sensibile (es. genere, etnia) dalle sue predizioni.
L'obiettivo è che le predizioni diventino "neutre" rispetto ai gruppi sensibili.

Quando si usa:
- Quando si vuole che il modello "dimentichi" informazioni sui gruppi protetti
- Per problemi complessi dove i bias sono sottili e difficili da rimuovere con semplici bilanciamenti

Risultato:
Il modello finale è meno influenzato dal gruppo sensibile e più equo nelle sue decisioni.

PIPELINE DI DEBIASING – SCHEMA

1. **Audit dei dati**
   Analisi delle distribuzioni e ricerca di bias
2. **Preprocessing**
   Bilanciamento (oversampling, undersampling, reweighting)
   Cleaning dei dati sensibili
3. **Addestramento modello**
   Modello standard o con vincoli/penalità di fairness
   Possibile uso di tecniche adversariali
4. **Valutazione fairness**
   Calcolo delle metriche di fairness sui risultati (es. Demographic Parity, Equalized Odds)
5. **Postprocessing**
   Correzione soglie o risultati per migliorare l'equità finale
6. **Monitoraggio continuo**
   Controllo periodico della fairness su nuovi dati

ESERCIZIO PRATICO: AUDIT E DEBIASING SU ADULT INCOME

Audit

- Carica il dataset Adult Income.
- Analizza la distribuzione di “income >50K” per genere ed etnia.
- Calcola il tasso di predizioni positive e la differenza di Demographic Parity tra i gruppi.

Debiasing

- Applica una tecnica (**oversampling**, undersampling o reweighting) per ridurre il bias di genere.
- Allena di nuovo il modello e ripeti la valutazione della fairness.
- Confronta i risultati: la differenza tra i gruppi si è ridotta?

```
# Oversampling della classe minoritaria nei dati di training
from imblearn.over_sampling import RandomOverSampler

ros = RandomOverSampler(random_state=42)
X_res, y_res = ros.fit_resample(X_train, y_train)

print("Distribuzione classi nel training set DOPO oversampling:")
print(y_res.value_counts())
```

DOMANDE?

PAUSA

ESERCIZIO: NER/GPT + AVVERSARIALE

Scenario applicativo:

- Un sistema NER anonimizza i dati nei documenti.
- Un modello GPT processa questi documenti e gestisce la chat (es. via Langchain).

Obiettivo:

- Prima di inviare ogni risposta del GPT all'utente, utilizzare un ulteriore controllo automatico (può essere lo stesso GPT, oppure un modello dedicato) per valutare la presenza di bias e la qualità in termini di fairness.

Come implementare:

- Dopo che GPT genera la risposta, passa il testo a un valutatore (prompt "critico" oppure un classificatore separato).
- Il valutatore avversariale controlla se la risposta contiene espressioni discriminatorie, squilibri, stereotipi, o se tratta i gruppi in modo imparziale.
- Se viene rilevato bias o rischio, il sistema può:
  - rigenerare la risposta,
  - aggiungere un avviso per l'utente,
  - oppure loggare il caso per miglioramenti futuri.

ESERCIZIO: NER/GPT + AVVERSARIALE

Spunti pratici:

- Progetta prompt specifici per chiedere a GPT “Vedi bias o mancanza di fairness in questa risposta? Spiega.”
- Allenare o fine-tuning di un classificatore su esempi di risposte biased/non-biased.
- Definisci delle checklist di fairness per guidare la valutazione automatica (ad es: “il testo include stereotipi di genere/etnia?”, “tutte le categorie sono rappresentate equamente?”).

DOMANDE DI SUPPORTO – BIAS, FAIRNESS E AVVERSARIO

1. **Cosa intendiamo per bias in una risposta AI, anche dopo l'anonimizzazione?**
   - Bias può emergere solo dal testo, o anche dalla selezione delle informazioni?
2. **In quali casi una risposta del GPT potrebbe risultare non "fair", anche se i dati personali sono stati rimossi?**
   - Pensa a stereotipi, generalizzazioni, esclusioni di gruppi, ecc.
3. **Che tipo di prompt o regola potrebbe aiutare il valutatore a identificare bias evidenti?**
   - Puoi scrivere un esempio di prompt?
4. **Come definiresti "fairness" in questo contesto?**
   - Vuol dire solo trattare tutti uguale, o anche rappresentare tutte le categorie in modo accurato?
5. **Quali sono i rischi di falsi positivi o negativi nella valutazione del bias?**
   - Il valutatore potrebbe "vedere bias" dove non c'è, o non rilevare bias reale?
6. **Come puoi migliorare il valutatore col tempo?**
   - Loggare i casi discussi, usare feedback umano, creare set di esempi per training?
7. **Hai esempi pratici di risposte GPT da sottoporre al valutatore?**
   - Cosa succede se la risposta menziona solo un genere, una nazionalità, o un punto di vista?

ETICA E NORMATIVE

INTRODUZIONE – ETICA NELL'INTELLIGENZA ARTIFICIALE

L'etica nell'intelligenza artificiale

L'intelligenza artificiale sta entrando in tanti aspetti della nostra vita e delle nostre decisioni.
Per questo è fondamentale porsi domande etiche:

- L'AI è davvero giusta ed equa per tutti?
- Come possiamo prevenire rischi, discriminazioni o usi scorretti?
- Chi è responsabile delle scelte fatte da un sistema AI?

L'etica in AI serve a garantire che queste tecnologie siano al servizio delle persone, rispettino i diritti fondamentali e siano utilizzate in modo trasparente e responsabile.

Etica e AI: uno sguardo filosofico

L'etica nasce dalla filosofia (filosofia morale) e si occupa di capire cosa è giusto o sbagliato, buono o dannoso per le persone e la società.
Applicare questi principi all'AI significa riflettere su scelte, responsabilità e conseguenze delle decisioni prese dalle macchine.

LINEE GUIDA ETICHE: OCSE, UNESCO, EU AI ACT

OCSE* (OECD AI Principles)

- Sviluppo e utilizzo responsabile dell'AI, trasparenza, sicurezza, inclusività.
- Promuove sistemi affidabili, rispetto dei diritti umani e monitoraggio continuo.

UNESCO

- Raccomanda AI "centrata sull'uomo": tutela diritti, diversità culturale e inclusione.
- Sottolinea il controllo umano, la responsabilità e la non discriminazione.

EU AI Act

- Prima proposta di legge europea specifica sull'AI.
- Richiede valutazione dei rischi, trasparenza, limiti a usi "ad alto rischio", protezione contro discriminazioni e bias.
- Impone regole precise su dati, spiegabilità, audit e responsabilità.

Queste linee guida puntano a rendere l'AI più equa, trasparente e affidabile, con attenzione ai diritti e ai rischi sociali.

*** OCSE** sta per **Organizzazione per la Cooperazione e lo Sviluppo Economico**.
In inglese: **OECD – Organisation for Economic Co-operation and Development**.

EU AI ACT: PARTE ETICA E IMPATTI OPERATIVI

Perché servono le normative?
Per garantire che l'AI sia sicura, equa, trasparente e rispettosa dei diritti.

EU AI Act – Parte etica:

- Impone l'identificazione e la gestione dei rischi di bias e discriminazione.
- Chiede trasparenza sulle decisioni e sulle fonti dei dati.
- Obbliga a spiegare il funzionamento dei sistemi AI (spiegabilità).
- Promuove la supervisione umana e la responsabilità.

Impatti operativi per le aziende:

- Devono auditare e documentare dati e processi AI.
- Necessità di testare la fairness e mitigare bias.
- Maggiori controlli su sistemi "ad alto rischio".
- Sanzioni in caso di non conformità.

Le regole non sono solo un obbligo, ma un'opportunità per rendere l'AI più affidabile, trasparente e accettata nella società.

RESPONSABILITÀ UMANA VS ALGORITMICA

Responsabilità umana:
Le persone (sviluppatori, aziende, operatori) restano responsabili delle decisioni prese dai sistemi AI.
Sono loro a scegliere i dati, i modelli, le soglie e l'uso finale.

Responsabilità algoritmica:
Un sistema AI può prendere decisioni automatiche, ma non può essere ritenuto responsabile come un essere umano.
Tuttavia, è importante tracciare e spiegare le sue scelte (accountability).

Sfida attuale:
Trovare il giusto equilibrio tra controllo umano e autonomia dei sistemi AI, per garantire trasparenza, sicurezza e rispetto dei diritti.

Anche con AI avanzata, la responsabilità ultima resta nelle mani umane.

COME INTEGRARE PRINCIPI ETICI NEI PROGETTI AI

- **Definisci linee guida etiche**
  Adotta principi come trasparenza, equità, sicurezza e rispetto della privacy già nella fase di progettazione.
- **Audita e bilancia i dati**
  Analizza i dati per individuare e correggere bias prima dell'addestramento.
- **Valuta l'impatto sociale**
  Considera chi potrebbe essere svantaggiato dalle decisioni dell'AI e adotta misure per proteggere i gruppi vulnerabili.
- **Spiegabilità e trasparenza**
  Progetta sistemi che possano spiegare le proprie decisioni, sia agli utenti che agli auditor.
- **Supervisione e miglioramento continuo**
  Mantieni un controllo umano sul sistema e aggiorna regolarmente i modelli e le policy etiche.

L'etica non va solo dichiarata, ma integrata e verificata in ogni fase del progetto AI.

CASO STUDIO: NER + GPT SU DOCUMENTI AZIENDALI

Scenario:
Dopo l'anonimizzazione con NER, il modello GPT elabora i documenti aziendali e risponde alle richieste degli utenti interni.

Il problema:
Anche con i dati personali rimossi, le risposte di GPT possono comunque contenere **bias** o risultare non "fair" verso determinati gruppi di dipendenti (ad esempio, potrebbero privilegiare ruoli, sedi o categorie specifiche presenti nei dati storici o nel linguaggio usato).

Domande per l'analisi:
- Dopo l'anonimizzazione, quali tipi di bias possono comunque emergere nelle risposte GPT?
- Il GPT potrebbe utilizzare informazioni residue (contesto, linguaggio, esempi ricorrenti) che favoriscono o penalizzano certi gruppi o ruoli aziendali?
- In quali casi una risposta GPT potrebbe risultare non "fair" per categorie specifiche di dipendenti (es. junior/senior, sede, ruolo)?
- Come il modello di valutazione può riconoscere e segnalare queste situazioni?
- Che controlli aggiuntivi o prompt potresti progettare per rafforzare la fairness delle risposte?

GRAZIE PER L'ATTENZIONE