

Comprendere il bias algoritmico

Identificare e mitigare i bias nei dati e negli algoritmi dell'IA per promuovere equità e equità.



Obiettivi di apprendimento **In poche parole**



- Capire cos'è il bias algoritmico e come ha origine. Identificare diversi tipi di bias nei sistemi di dati e IA. Riconosci i rischi associati all'uso dell'IA nei contesti educativi. Analizza l'impatto del pregiudizio su equità e inclusione. Applica strategie di base per mitigare i pregiudizi nella loro pratica didattica.



Concetti chiave e definizioni

*I CONCETTI ESSENZIALI PER IDENTIFICARE E AFFRONTARE
I BIAS NEI SISTEMI IA*

Bias algoritmico

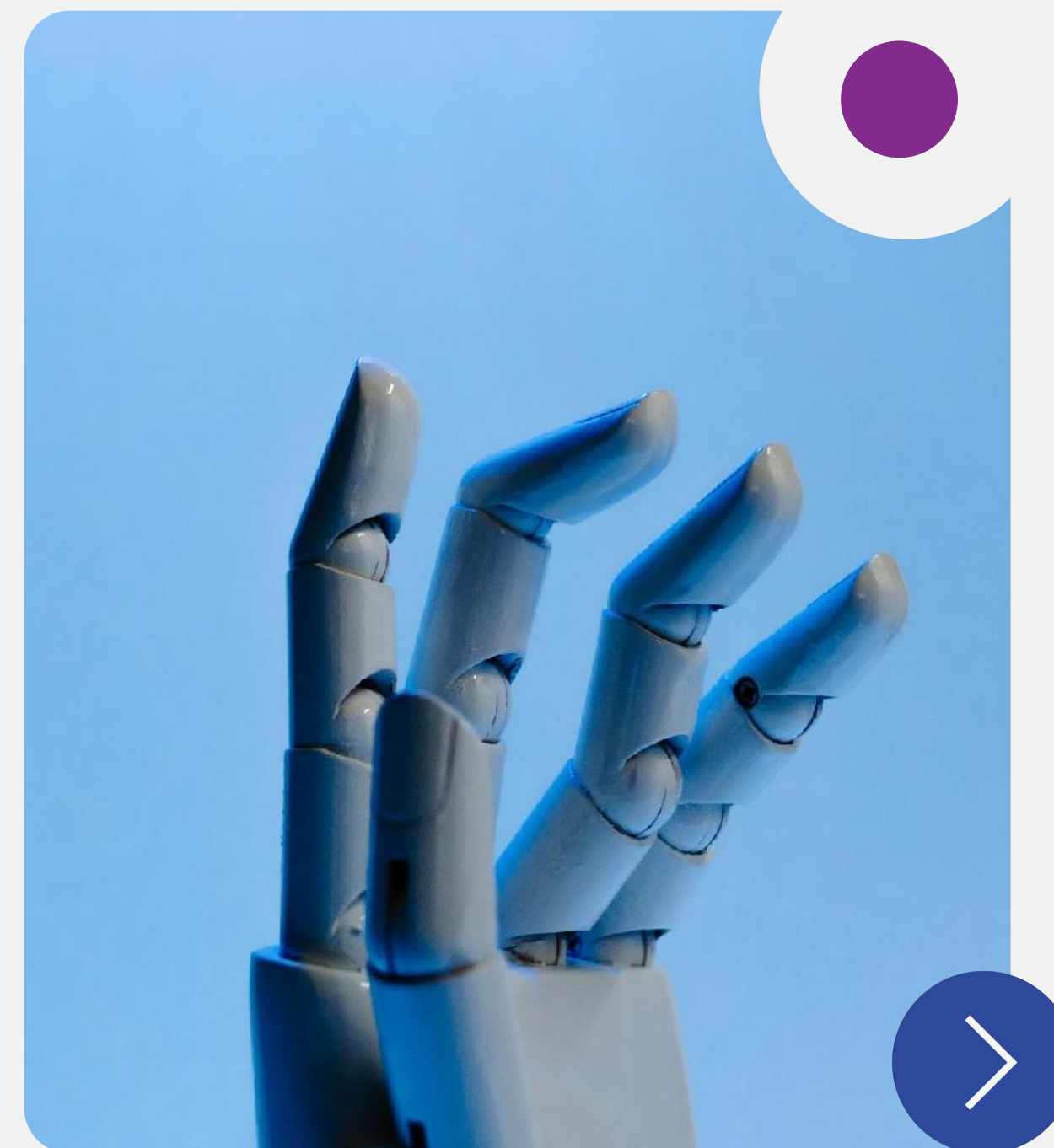
Il bias algoritmico si verifica quando un sistema IA produce risultati sistematicamente ingiusti a causa di dati distorti, scelte di design o assunzioni implicite

Bias di rappresentanza

Il bias di rappresentazione sorge quando alcuni gruppi sono sottorappresentati o sovrarappresentati nei dati di addestramento.

Orientamento storico

Il bias storico si verifica quando i dataset riflettono disuguaglianze passate che vengono riprodotte o amplificate dai sistemi di IA



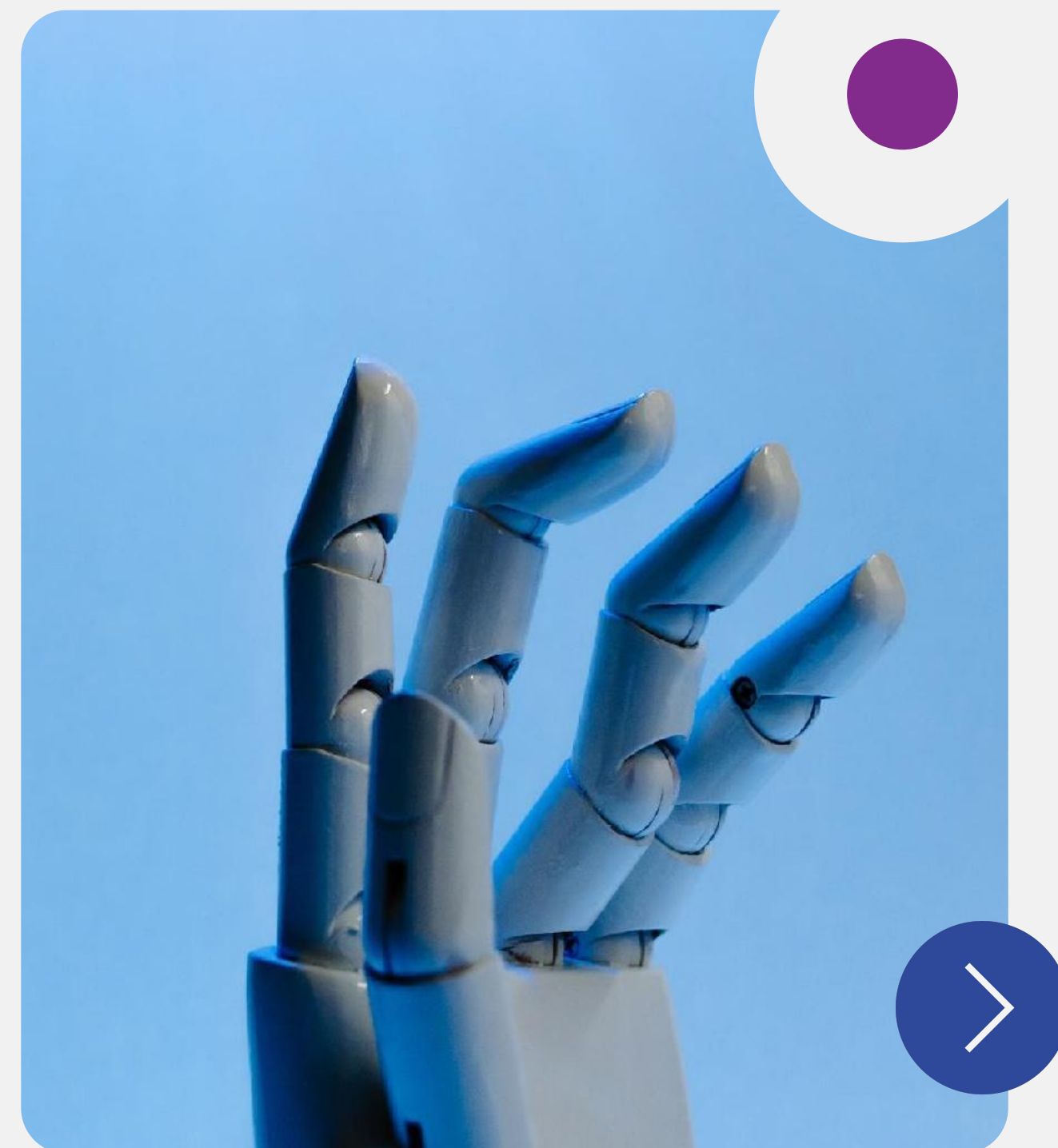
Cos'è il bias algoritmico?

Come emergono le distorsioni nei dati e negli algoritmi dell'IA.

- Il bias algoritmico si verifica quando un sistema IA produce risultati sistematicamente ingiusti o diseguali a causa dei dati su cui è stato addestrato, delle scelte di design fatte o del modo in cui interpreta le informazioni.

Non è sempre intenzionale, ma può riprodurre o amplificare disuguaglianze esistenti.

Il bias algoritmico non è solo una questione tecnica, ma anche un riflesso di più ampie disuguaglianze sociali radicate nei dati e nelle pratiche istituzionali.



Principali tipi di bias



Orientamento storico

Quando i dati di addestramento riflettono disuguaglianze passate che si incorporano nel sistema di IA.

Bias di misurazione

Quando i dati utilizzati non catturano accuratamente ciò che intende valutare.

Bias di aggregazione

Quando un modello viene applicato a gruppi diversi senza considerare le differenze.

Bias di progettazione o implementazione

Introdotta attraverso scelte umane nella definizione di variabili, categorie o criteri di valutazione.

Bias di rappresentanza

Quando certi gruppi sono sottorappresentati o sovrarappresentati nei dati di addestramento.

Bias di conferma

Quando sviluppatori o progettisti selezionano dati o funzionalità che rafforzano le assunzioni esistenti.



Rischi di dati legati al bias algoritmico

Come i dati possano portare a risultati ingiusti nei sistemi di IA educativi.

Il bias algoritmico spesso ha origine proprio nei dati. Quando i dataset sono incompleti, sbilanciati o strutturati in modo scadente, i sistemi di IA possono generare risultati distorti o ingiusti.

Rischi chiave legati ai dati:

- **Sotto-rappresentazione di alcuni gruppi di studenti** → Le esperienze e i bisogni di alcuni studenti potrebbero non essere adeguatamente riflessi.
Dataset sbilanciati → Il sistema potrebbe ottenere risultati migliori per i gruppi di maggioranza rispetto alle minoranze.
Uso di variabili proxy (indicatori indiretti) → Variabili come il codice postale o il tipo di scuola possono riflettere involontariamente differenze socio-economiche o demografiche.
Dati di bassa qualità o incoerenti → Informazioni inaccurate o obsolete possono distorcere previsioni e raccomandazioni.
L'eccessiva dipendenza dai dati quantitativi → le realtà educative complesse possono essere ridotte a metriche semplificate.

Se i dati sono distorti, anche un algoritmo tecnicamente valido può produrre risultati diseguali.

L'impatto del bias algoritmico sull'equità educativa



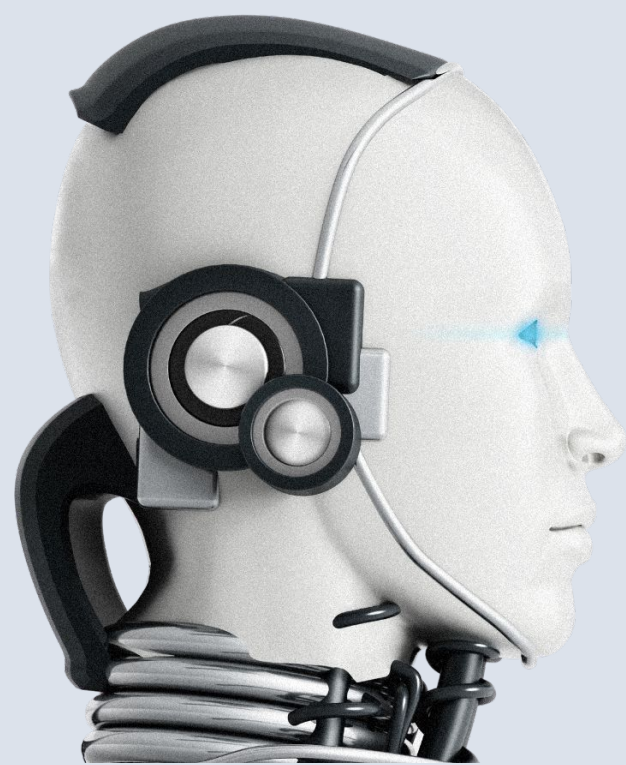
Come i sistemi di IA possono influenzare l'uguaglianza di opportunità nell'istruzione superiore

Quando i sistemi di IA incorporano il bias, possono influenzare gruppi diversi di studenti in modo diseguale, anche senza un intento esplicito. I potenziali impatti nei contesti educativi sono:

- **Risultati di valutazione ingiusti o sproporzionati.** Alcuni profili di studenti possono costantemente ricevere voti o raccomandazioni meno favorevoli.
Rafforzamento delle disuguaglianze preesistenti. I modelli storici possono essere riprodotti nell'analisi predittiva o nelle raccomandazioni dei percorsi accademici.
Limitazione delle opportunità accademiche. Sistemi di guida o monitoraggio possono influenzare le scelte future e il progresso degli studenti.
Perdita di fiducia nella tecnologia educativa. La percepita ingiustizia può minare la fiducia negli strumenti digitali e nelle pratiche istituzionali.
Impatto sull'inclusione e sulla diversità. Gruppi specifici possono essere sistematicamente svantaggiati.

Un sistema predittivo progettato per identificare gli "studenti a rischio" potrebbe segnalare in modo sproporzionato studenti provenienti da contesti svantaggiati se i dati storici sull'abbandono scolastico riflettono disuguaglianze strutturali esistenti.



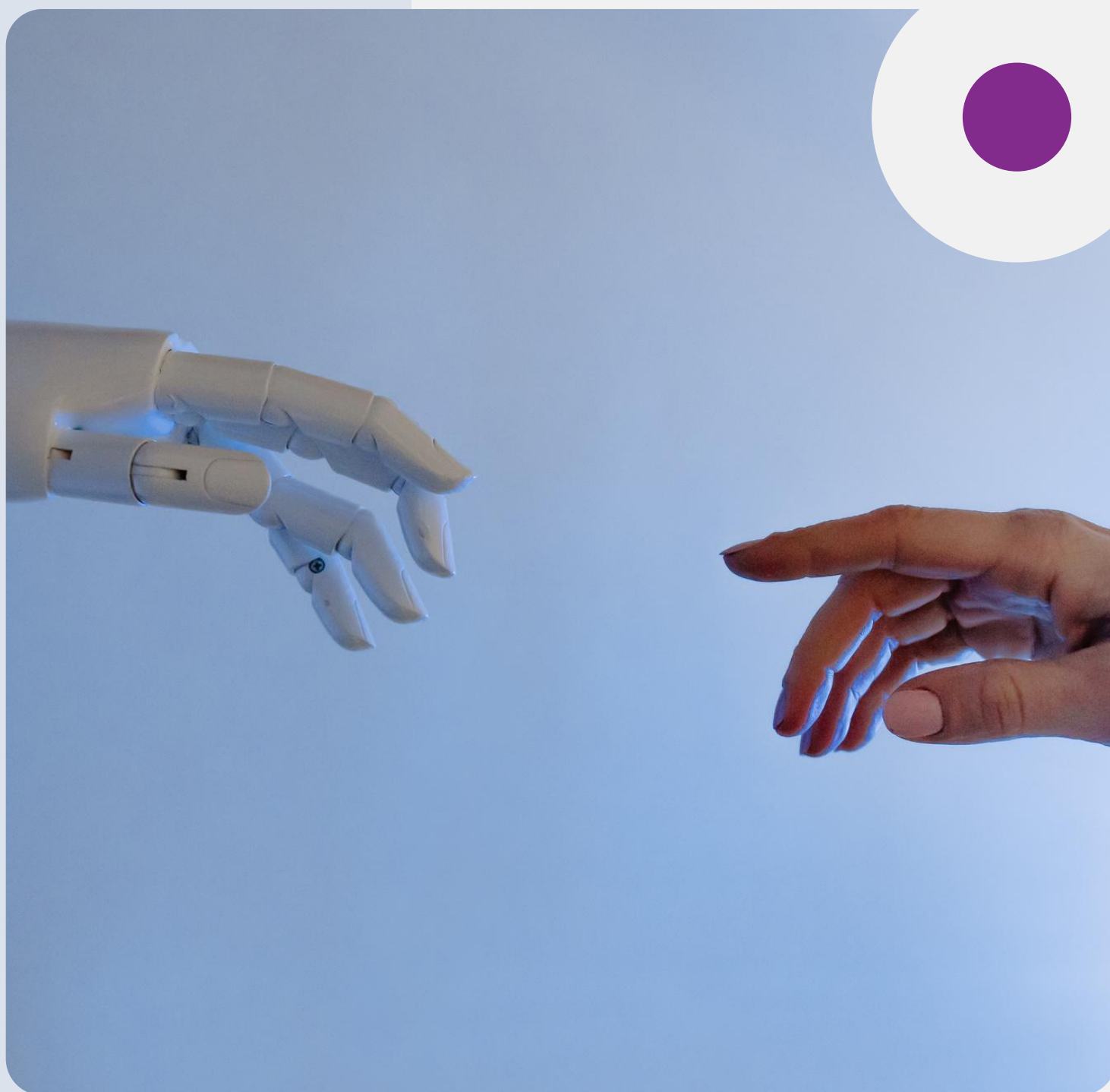


Azioni pratiche per mitigare i bias algoritmici e promuovere l'equità nell'IA educativa

Sebbene il pregiudizio non possa sempre essere eliminato completamente, può essere identificato e ridotto attraverso decisioni consapevoli nella pratica didattica.

- Chiedersi se i dati utilizzati rappresentino adeguatamente tutti i profili degli studenti.
Mantenere una supervisione umana attiva e rivedere le raccomandazioni generate dall'IA prima di prendere decisioni accademiche.
Promuovi la trasparenza spiegando agli studenti come vengono utilizzati gli strumenti di IA e quali sono i loro limiti.
Osserva se certi gruppi sono regolarmente influenzati.
Rivedi la rappresentatività dei dati considerando se il sistema è stato addestrato su dataset diversi o se alcuni gruppi potrebbero essere sottorappresentati.
Confronta se il sistema produce risultati sistematicamente diversi in base a genere, background, lingua o altri fattori rilevanti.
Ascolta le prospettive degli studenti per capire se un gruppo percepisce un trattamento diseguale.





Riesci a identificare il pregiudizio?

Applica ciò che hai imparato

Il bias algoritmico non è sempre evidente. Spesso appare in modo sottile negli strumenti educativi e nei sistemi decisionali. In questa attività, analizzerai scenari educativi reali e identificherai il tipo di pregiudizio coinvolto.

Metti in pratica le tue conoscenze e metti alla prova il tuo occhio critico.

Avvia l'attività

Riferimenti

- Barocas, S., Hardt, M., & Narayanan, A. (2019). Equità e apprendimento automatico: limiti e opportunità.
<https://fairmlbook.org>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). Un sondaggio su bias e equità nel machine learning. *Indagini di calcolo ACM*, 54(6), 1–35.
<https://doi.org/10.1145/3457607>
- UNESCO. (2021). Raccomandazione sull'Etica dell'Intelligenza Artificiale.
<https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- Commissione Europea. (2022). Linee guida etiche sull'uso dell'intelligenza artificiale e dei dati nell'insegnamento e nell'apprendimento per gli educatori.
<https://education.ec.europa.eu/news/ethical-guidelines-on-the-use-of-artificial-intelligence-and-data-in-teaching-and-learning-for-educators>

Promuovere l'equità nell'istruzione richiede non solo buone intenzioni, ma anche una consapevolezza fondamentale di come la tecnologia plasmi le opportunità.

Il consorzio:



<https://trustai.unibit.bg/>

 Trust AI project

Disclaimer: Finanziato dall'Unione Europea. Le opinioni espresse sono però solo quelle dell'autore o degli autori e non riflettono necessariamente quelle dell'Unione Europea o dell'Agenzia Nazionale (NA). Né l'Unione Europea né l'NA possono essere ritenute responsabili per queste.

Numero di riferimento del progetto: 2024-1-BG01-KA220-HED-000249185



Co-funded by
the European Union

