Înțelegerea prejudecăților algoritmice și a strategiilor de atenuare

Ce este prejudecata algoritmică?

Erori sistemice și repetabile în sistemele informatice care conduc la rezultate inechitabile.

Importanța înțelegerii prejudecăților pe măsură ce AI și ML devin tot mai răspândite.

Tipuri de prejudecăți algoritmice

Prejudecăți bazate pe date: Idei nedrepte preluate din date. Exemplu: Sistemul de recrutare AI favorizează candidații de sex masculin.

Prejudecăți algoritmice sau determinate de modele: prejudecăți introduse accidental în modelele AI.

Exemplu: Sistemul de recunoaștere facială recunoaște mai bine fețele cu piele deschisă.

Prejudecăți provocate de om sau de societate: prejudecăți introduse de persoanele care creează sau utilizează IA. Exemplu: Etichete nedrepte în descrierile imaginilor.

Strategii de atenuare

Utilizați date diverse și corecte: Asigurați-vă că datele includ toate tipurile de persoane.

Fiți conștienți de rezultatele părtinitoare: Puneți la îndoială răspunsurile incorecte ale AI.

Puneți întrebări și învățați: Înțelegeți cum funcționează inteligența artificială pentru a depista prejudecățile.

Fiți un susținător al corectitudinii: Ridicați-vă împotriva prejudecăților.

Identificarea prejudecăților

Nevoia de instrumente care să măsoare părtinirea în inteligența artificială, asemănătoare degustării unei prăjituri pentru a verifica dacă rețeta este bună.

Prejudecățile se pot strecura în diferite etape ale realizării AI (proiectare, construire, utilizare).

Pentru a reveni la joc, închideți această fereastră, făcând clic pe X.

Cofinanțat de Uniunea Europeană

Finanțat de Uniunea Europeană. Opiniile și opiniile exprimate sunt exclusiv ale autorilor și nu reflectă neapărat pe cele ale Uniunii Europene sau ale Agenției Executive pentru Educație și Cultură Europeană (EACEA). Nici Uniunea Europeană, nici EACEA nu pot fi considerate responsabile pentru acestea.

2022-1-ES01-KA220-HED-000085257