



# Microcredențial de IA etică

CARTEA

CU1 | Ce este părtinirea algoritmică?

Numărul proiectului:  
2022-1-ES01-KA220-HED-000085257



# Cum să utilizați acest Flipbook?

Acest document este interactiv. De-a lungul documentului, veți găsi linkuri către informații suplimentare.



Buton care vă duce la începutul documentului. Această pictogramă apare în colțul din dreapta sus al paginilor.



Ori de câte ori vedeți această săgeată, înseamnă că aveți un **text color interactiv** pe care trebuie să faceți clic, care are asociat un link extern.

**DECLINARE DE RESPONSABILITATE:** Vă rugăm să rețineți că nu putem garanta disponibilitatea continuă a conținutului extern, cum ar fi videoclipurile, deoarece acestea pot fi modificate sau eliminate de către autorii sau platformele gazdă.

# Index

Faceți clic pe meniu

**01. Introducere**

**02. Conținutul cursului și rezultatele așteptate**

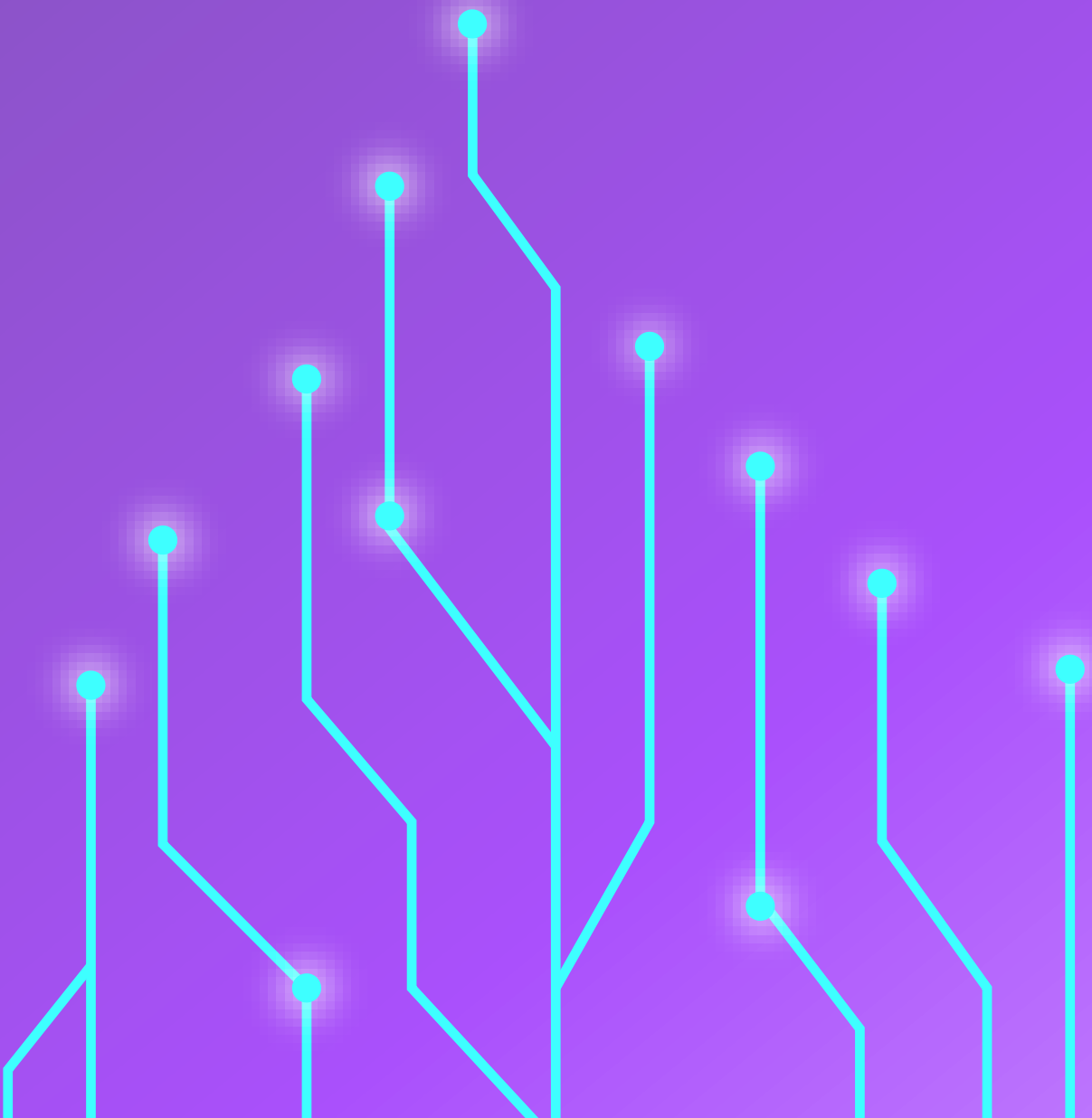
**03. Ce este părtinirea algoritmică?**

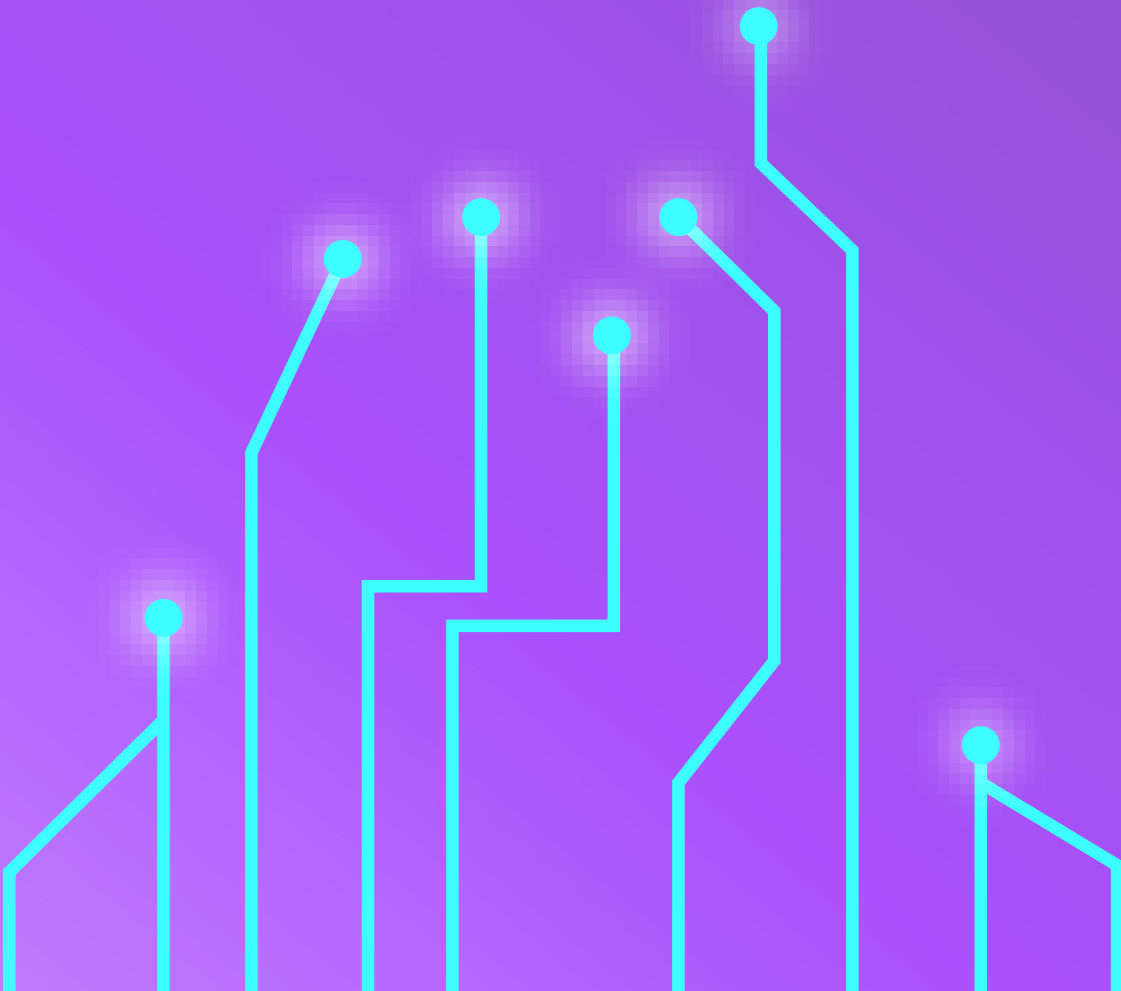
**04. Definirea prejudecăților algoritmice**

**05. Înțelegerea prejudecăților în sistemele AI**

# 01. Introducere

CU1 | Ce este părtinirea algoritmică?





## 01. Introducere

În peisajul în evoluție rapidă al inteligenței artificiale (AI), asigurarea dezvoltării și utilizării etice a tehnologiilor AI este extrem de importantă. Această broșură servește drept ghid cuprinzător pentru microcredențialul Ethical AI, concentrându-se pe șase unități de competență concepute pentru a vă dota cu cunoștințele și abilitățile necesare pentru a naviga în complexitatea etică a AI.

Pe măsură ce porniți în această călătorie, veți explora șase unități de competențe distincte, fiecare abordând aspecte cruciale ale dezvoltării și implementării etice a IA. De la înțelegerea prejudecăților algoritmice la promovarea transparenței și respectarea drepturilor omului, aceste unități de competență sunt concepute pentru a vă oferi instrumentele necesare pentru a face față provocărilor etice inerente tehnologiilor IA.

Pe parcursul acestei broșuri, veți studia următoarele Unități de competență (de acum înainte CU):

- CU1 - Ce este părtinirea algoritmică?
- CU2 - Non-maleficiență
- CU3 - Răspundere
- CU4 - Transparență
- CU5 - Drepturile omului și corectitudinea
- CU6 - Etica IA, o abordare practică



Fiecare unitate vă va oferi o înțelegere mai profundă a principiilor și practicilor etice cheie în IA, împreună cu perspective practice și exemple din lumea reală pentru a vă consolida învățarea.

Fie că sunteți un student adult, un profesionist sau un entuziast al IA, această broșură oferă o resursă valoroasă pentru a vă extinde cunoștințele și expertiza în IA etică. Vă invităm să porniți în această călătorie alături de noi, pe măsură ce explorăm dimensiunile etice ale IA și lucrăm la crearea unui viitor mai responsabil și mai echitabil.

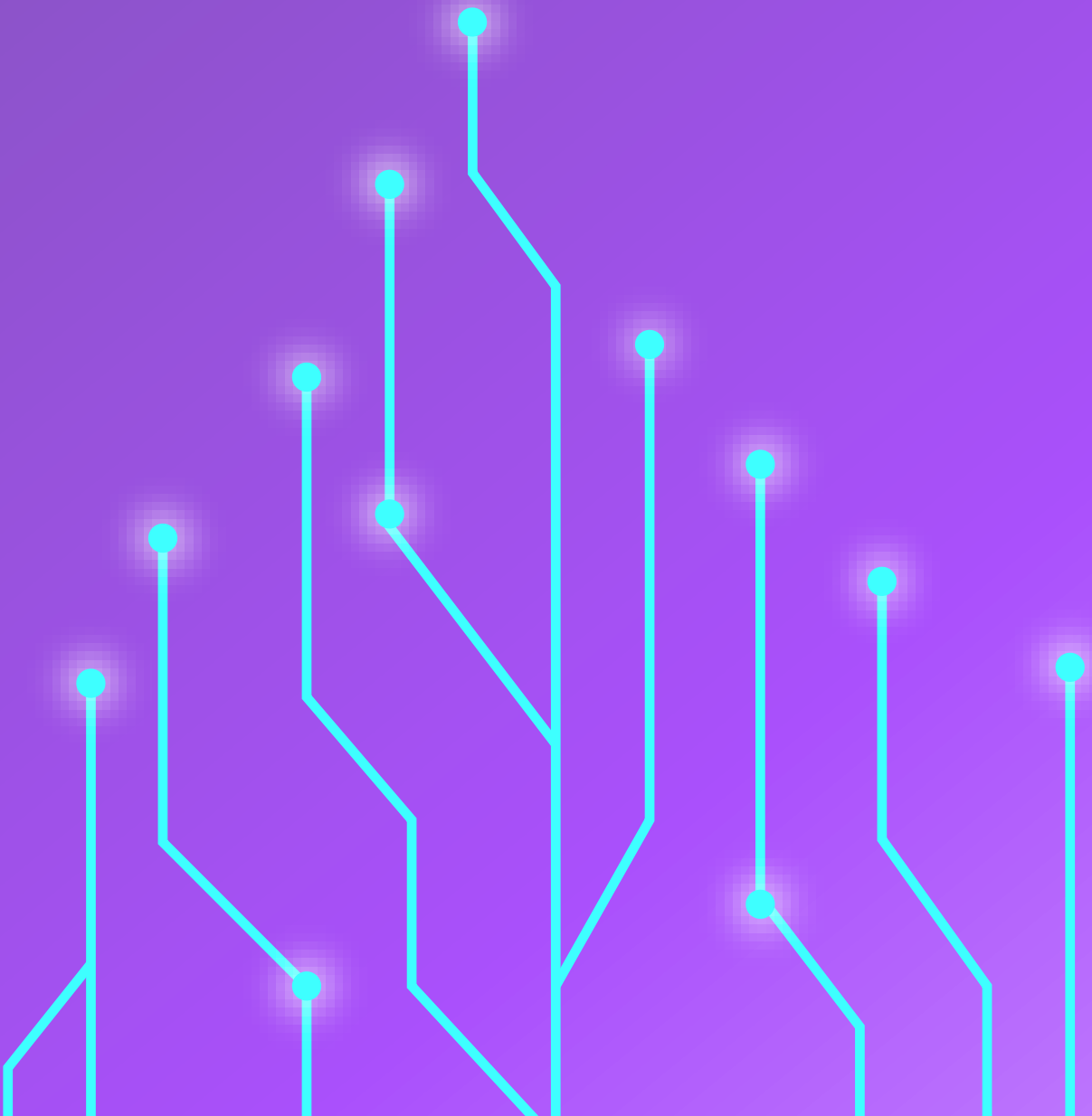
Vă mulțumim pentru că ați ales această broșură ca ghid pentru dezvoltarea și practica etică a IA.

Să începem împreună această călătorie transformatoare!

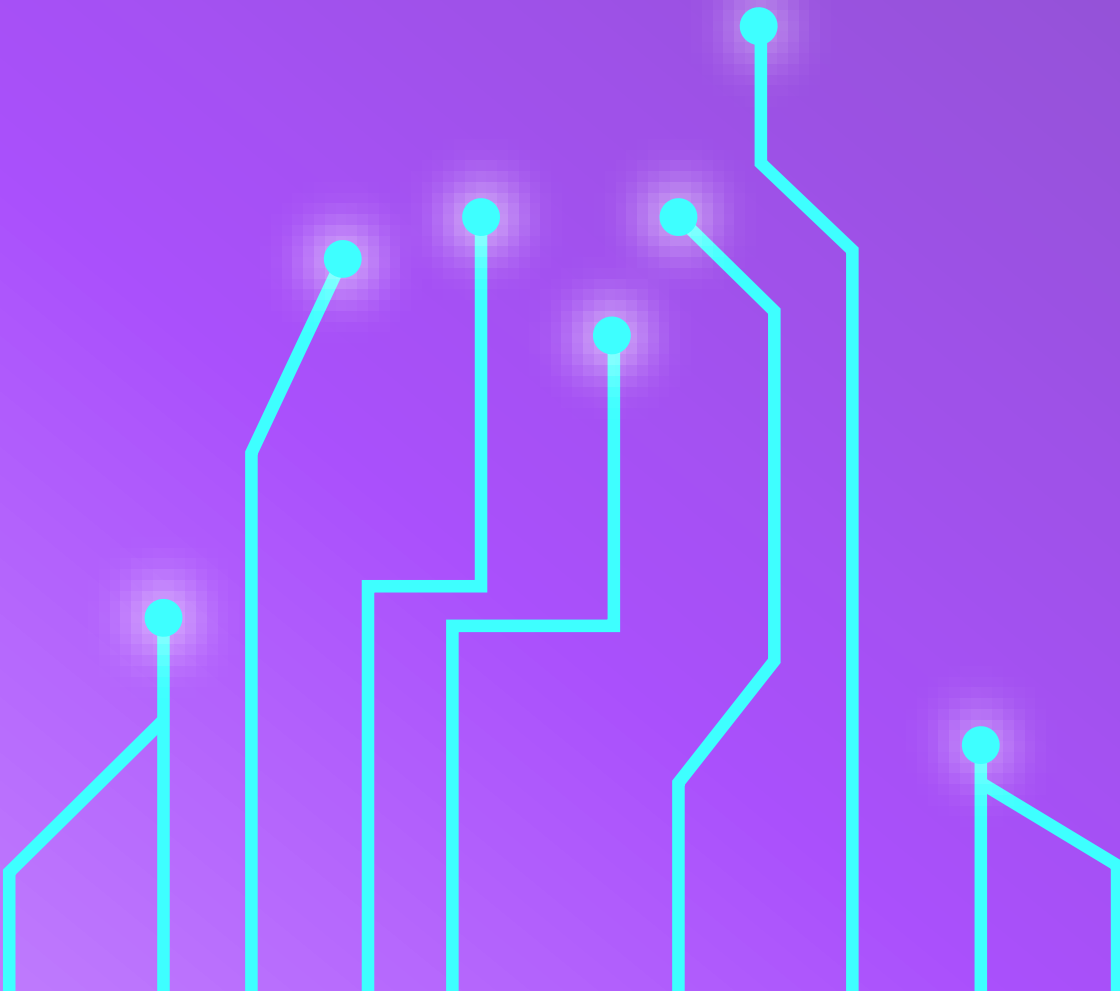
*Echipa proiectului CHARLIE*

# 02. Conținutul cursului și rezultatele așteptate

CU1 | Ce este părtinirea algoritmică?







## 02. Conținutul cursului și rezultatele așteptate

Microcredențialul "Ethical AI Microcredential" la nivelul EQF4 este conceput pentru a atinge următoarele rezultate:

1. Stabilirea unei înțelegeri fundamentale a prejudecăților algoritmice, explorând originile și implicațiile acestora pentru indivizi și societate.
  - Aprofundați definiția, sursele și manifestările prejudecăților algoritmice.
  - Analizați ramificațiile societale și individuale ale algoritmilor părtinitori.
2. Cultivarea conștientizării și aplicării principiului etic al non-maleficenței în dezvoltarea IA.
  - Evaluați riscurile și daunele asociate algoritmilor părtinitori.
  - Elaborarea de strategii de atenuare a riscurilor și de promovare a dezvoltării etice a IA.
3. Să aprecieze importanța responsabilității în sistemele AI, examinând cadrele juridice și etice pertinente.
  - Investigarea rolurilor diferitelor părți interesate în ceea ce privește responsabilitatea IA.
  - Aflați cele mai bune practici pentru încurajarea responsabilității în dezvoltarea IA.



4. Descoperiți conceptul de **transparență în sistemele AI** și rolul său esențial în luarea deciziilor algoritmice.
  - Explorarea metodologiilor și instrumentelor de sporire a transparenței în IA.
  - Înțelegerea provocărilor și constrângerilor legate de redarea algoritmilor complecși într-un mod mai ușor de înțeles.
5. Explorați **intersecția dintre inteligența artificială, drepturile omului și echitate**, precum și implicațiile acestora pentru dezvoltarea etică a inteligenței artificiale.
  - Evaluarea impactului algoritmilor părtinitori asupra drepturilor omului, inclusiv asupra nediscriminării, vieții private și libertății de exprimare.
  - Elaborarea de strategii pentru a asigura corectitudinea și echitatea în dezvoltarea și implementarea IA.
6. Aplicarea principiilor etice în dezvoltarea și implementarea IA prin **abordări practice și scenarii din lumea reală**.
  - Examinarea diverselor cadre și orientări etice și aplicarea lor la sistemele AI.
  - Să înțeleagă importanța implicării părților interesate, a colaborării interdisciplinare și a proceselor etice de dezvoltare a IA.

La finalizarea acestui curs, participanții vor avea o înțelegere holistică a prejudecăților algoritmice, a impactului lor specific sectorului, precum și a instrumentelor și strategiilor de abordare a acestora. Aceste cunoștințe permit profesioniștilor și cadrelor universitare/studentilor din domeniile bazate pe algoritmi să contribuie la rezultate mai echitabile și mai corecte într-o lume bazată pe date.

Cursul microcredențial este structurat în jurul a **6 unități de competență (CU)**, fiecare fiind concepută pentru a dota participanții cu cunoștințele și abilitățile necesare pentru a face față provocărilor și oportunităților din domeniul bias-ului algoritmic.

**CU1 - Ce este părtinirea algoritmică?** În această unitate, elevii vor explora conceptul de părtinire algoritmică și diferitele sale manifestări. Aceasta acoperă definiția, cauzele și implicațiile societale ale acesteia. Elevii vor analiza originile prejudecăților, sursele din cadrul algoritmilor și impactul potențial asupra indivizilor și societății.

**CU2 - Non-maleficență:** Această unitate analizează principiul etic al non-maleficienței, acordând prioritate evitării daunelor în dezvoltarea și implementarea IA. Participanții vor explora riscurile și prejudiciile inerente legate de algoritmi părtinitori, descoperind în același timp strategii de atenuare a acestor riscuri și de promovare a practicilor etice în domeniul IA.



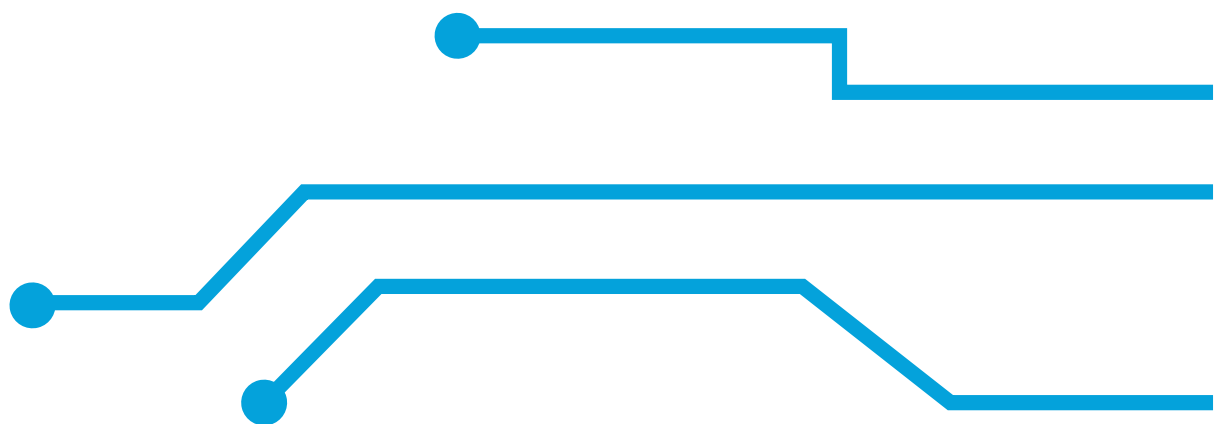
**CU3 - Răspundere:** În această unitate, studenții pătrund în domeniul critic al responsabilității în cadrul dezvoltării și utilizării IA. Expunând necesitatea trasării unor linii clare de responsabilitate, participanții explorează cadrele juridice și etice care reglementează responsabilitatea. În plus, curriculumul analizează rolurile diverselor părți interesate și analizează cele mai bune practici care asigură responsabilitatea în cadrul eforturilor de dezvoltare a IA.

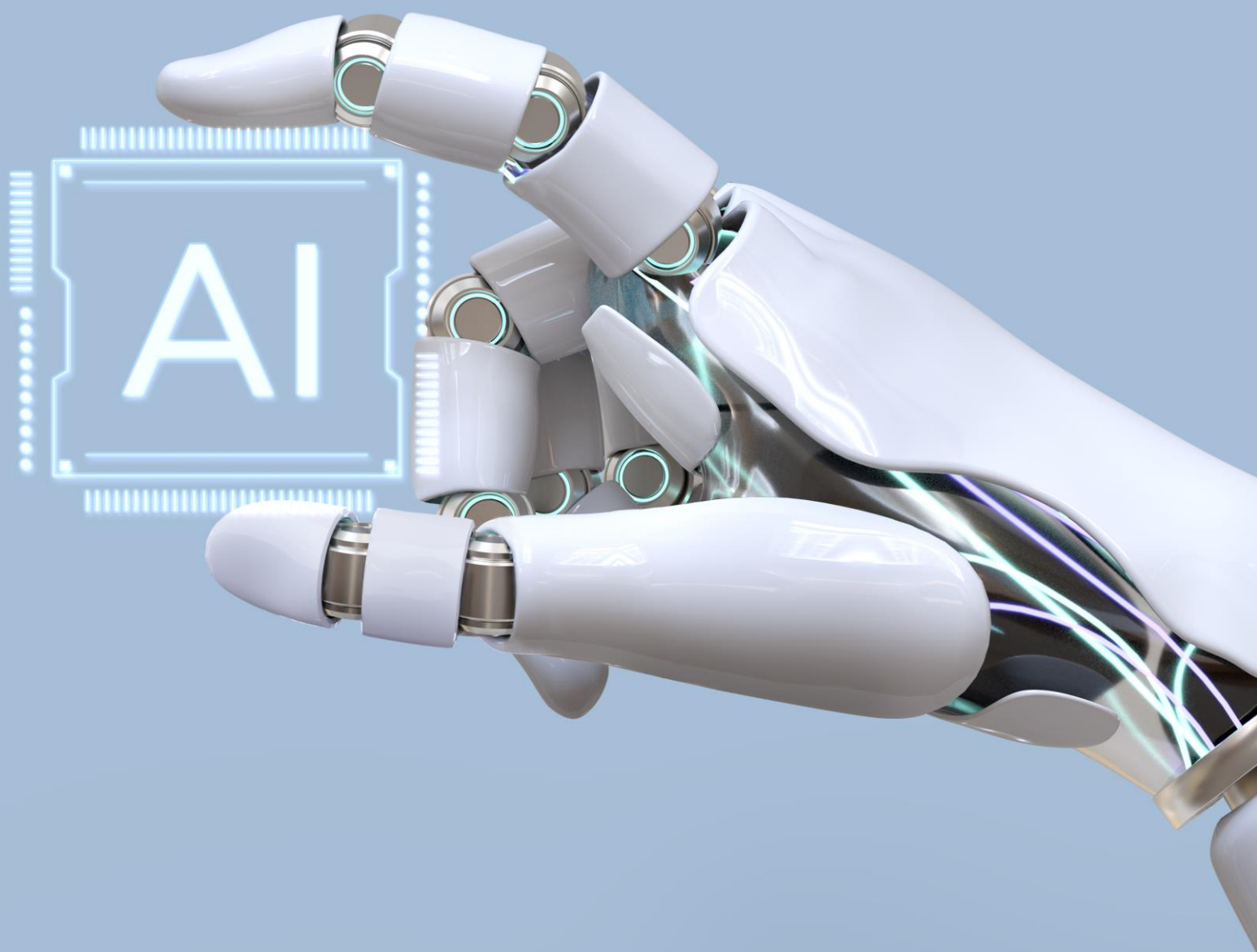
**CU4 - Transparență:** Această unitate pune în lumină semnificația transparenței în cadrul sistemelor AI, subliniind valorile deschiderii, comunicării și explicabilității în luarea deciziilor algoritmice. Participanții se vor angaja în tehnici și resurse menite să sporească transparența în inteligența artificială, luptând în același timp cu provocările și constrângerile inerente pentru a face algoritmi complicați inteligibili.

**CU5 - Drepturile omului și echitatea:** În cadrul unității Drepturile omului și echitatea, elevii vor explora intersecția dintre inteligența artificială, drepturile omului și echitate. Ei vor examina modul în care algoritmi părtinitori pot afecta drepturile omului, inclusiv dreptul la nediscriminare, confidențialitate și libertatea de exprimare. Elevii vor învăța, de asemenea, despre strategiile de asigurare a corectitudinii și echității în dezvoltarea și implementarea IA.

**CU6 - Etica IA, o abordare practică:** Această unitate pune accentul pe aplicarea pragmatică a principiilor etice pe parcursul dezvoltării și implementării AI. Participanții analizează diverse cadre și orientări etice, dobândind o perspectivă asupra aplicării lor în lumea reală în scenariile IA. În plus, unitatea subliniază importanța implicării părților interesate, a colaborării interdisciplinare și a integrării proceselor etice de dezvoltare a IA pentru promovarea inovării responsabile în domeniul IA.

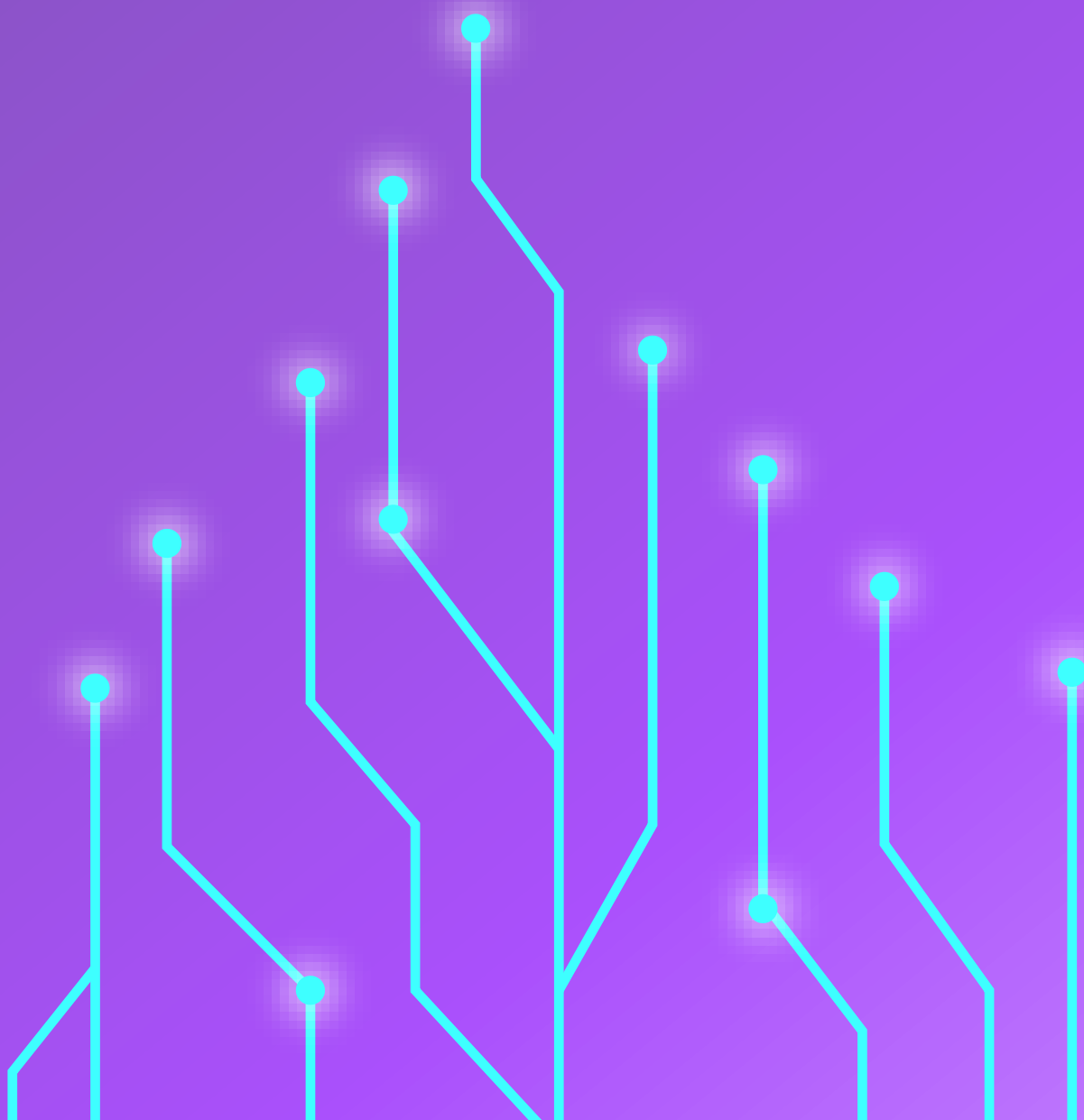
Secțiunea următoare detaliază conținutul fiecărei unități de competență.



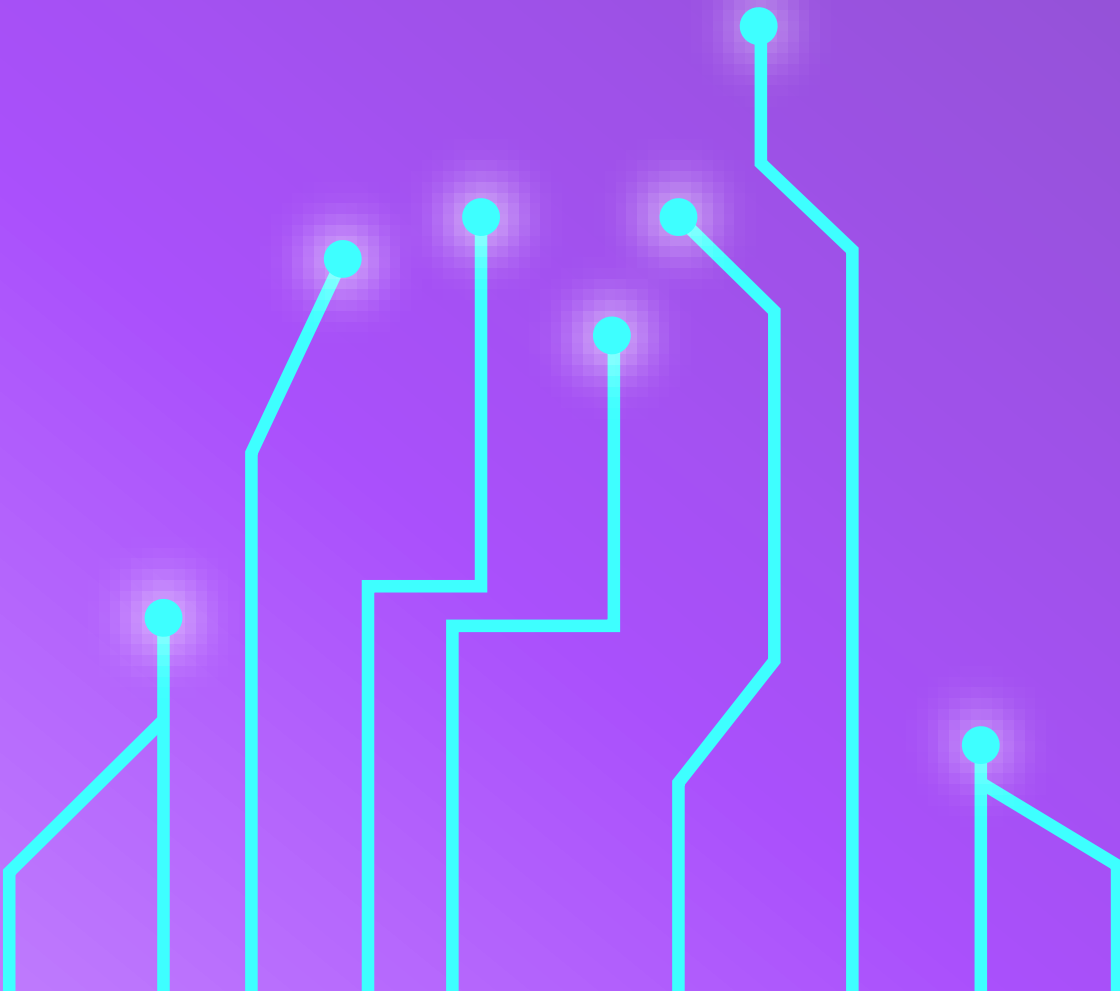


# 03. Ce este părtinirea algoritmică?

CU1 | Ce este părtinirea algoritmică?







### 03. Ce este părtinirea algoritmică?

Algoritmii sunt utilizați pentru a lua decizii importante. Cu toate acestea, ele pot fi uneori părtinitoare și nedrepte față de anumite grupuri de persoane. Acest lucru este cunoscut sub numele de părtinire algoritmică.

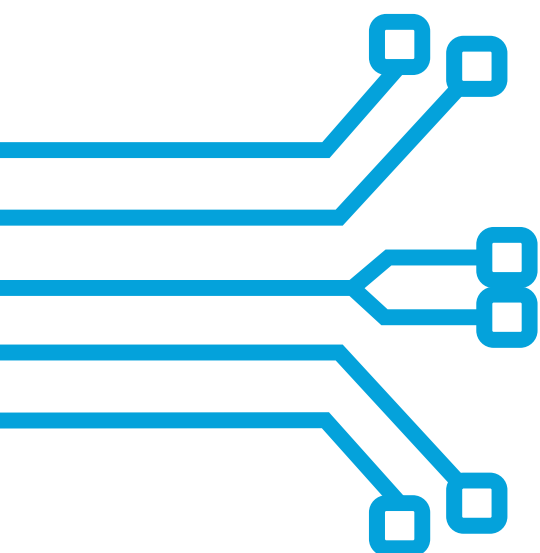
În această unitate de competențe, elevii vor învăța despre prejudecățile algoritmice, diferitele lor forme și cum să le identifice. Ei vor explora, de asemenea, motivele care stau la baza prejudecăților din algoritmi, inclusiv impactul prejudecăților umane asupra procesului decizional. În plus, elevii vor examina potențialele consecințe ale algoritmilor părtinitori asupra indivizilor și societății, care pot duce la discriminare și tratament inechitabil. La sfârșitul acestei unități, elevii vor avea o mai bună înțelegere a prejudecăților algoritmice și a modului de abordare a acestora în activitatea lor viitoare.

Rezultatele cunoștințelor pentru această unitate includ:

- **Definirea prejudecăților algoritmice:** Elevii vor învăța despre prejudecățile algoritmice și cauzele acestora, inclusiv colectarea părtinitoare a datelor, datele de formare distorsionate și procesul decizional uman. Aceste cunoștințe îi vor ajuta să înțeleagă modul în care părtinirea poate afecta aplicațiile AI, cum ar fi sistemele de recunoaștere facială care identifică greșit anumite grupuri.

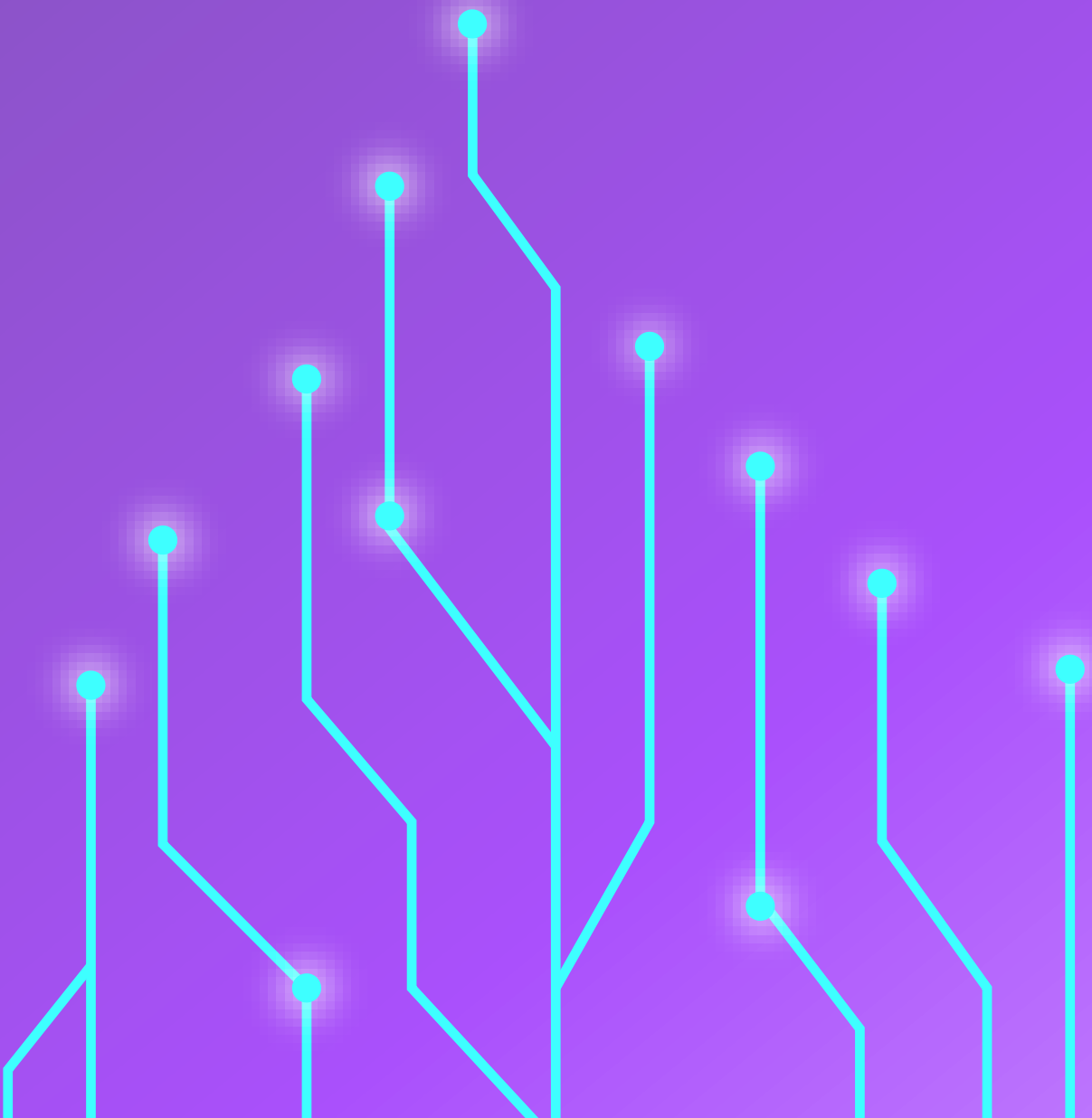


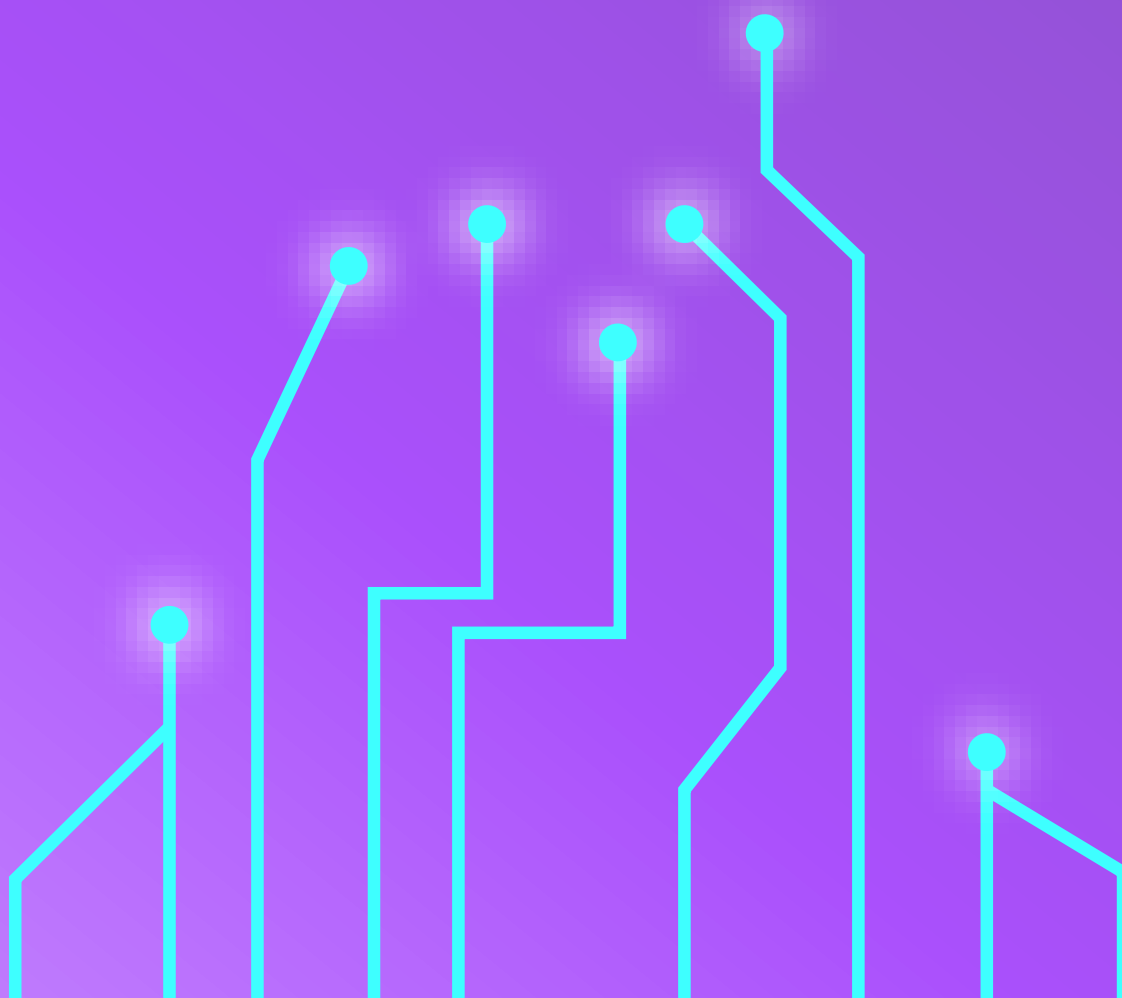
- **Identificarea tipurilor de prejudecăți algoritmice:** Elevii vor învăța despre prejudecățile algoritmice, inclusiv prejudecățile bazate pe date, bazate pe modele și bazate pe oameni. Ei vor înțelege modul în care aceste prejudecăți pot cauza nedreptate în sistemele AI. De exemplu, prejudecățile determinate de date pot rezulta din date de instruire nereprezentative, ceea ce duce la predicții părtinitoare în domenii precum scoringul creditelor sau selectarea candidaților la un loc de muncă.
- **Implicațiile prejudecăților algoritmice în lumea reală:** În cadrul acestui curs, studenții vor învăța despre consecințele prejudecăților algoritmice în diferite sectoare, cum ar fi sănătatea, finanțele și justiția penală. Ei vor înțelege necesitatea de a minimiza prejudecățile algoritmice în sistemele AI pentru a promova corectitudinea și echitatea. Vor fi discutate exemple de sisteme AI părtinitoare care conduc la rezultate negative în domeniul sănătății și al justiției penale.



# 04. Definirea prejudecăților algoritmice

CU1 | Ce este părtinirea algoritmică?





## 04. Definirea prejudecăților algoritmice

Prejudecarea algoritmică este un aspect critic al IA care a atras atenția în ultimii ani. Înțelegerea acestuia este esențială pentru oricine este implicat în dezvoltarea, implementarea sau reglementarea IA. Să definim ce este părtinirea algoritmică și de ce este esențial să o studiem.

### > Ce este prejudecata algoritmică?

Prejudecățile algoritmice se referă la erori sistematice sau nedreptăți în rezultatele sistemelor AI din cauza diferiților factori, cum ar fi datele părtinitoare, algoritmi defectuoși sau deciziile umane. Aceste prejudecăți pot conduce la un tratament discriminatoriu sau nedrept al persoanelor sau grupurilor, perpetuând inegalitățile sociale existente și consolidând stereotipurile.

### De ce să studiem prejudecățile algoritmice?

Pentru a explora diferitele forme, cauze și consecințe ale prejudecăților algoritmice, este important să înțelegem mai întâi definiția și semnificația acestora. Cu aceste cunoștințe, ne putem dota cu instrumentele necesare pentru a identifica, atenua și preveni părtinirea algoritmică în sistemele AI.



- 1. Implicații etice:** Prejudecățile algoritmice pot duce la un tratament incorect al persoanelor în funcție de rasă, sex, vârstă sau alte caracteristici protejate, încălcând principiile de corectitudine și echitate.
- 2. Impactul social:** Sistemele AI părtinitoare pot exacerba inegalitățile societale și discriminarea, afectând accesul la oportunități, resurse și servicii pentru comunitățile marginalizate.
- 3. Preocupări juridice și de reglementare:** Pe măsură ce tehnologiile AI devin din ce în ce mai răspândite, există o atenție tot mai mare din partea legiuitorilor și a organismelor de reglementare pentru a aborda prejudecățile algoritmice, pentru a asigura conformitatea cu legile anti-discriminare și pentru a proteja drepturile persoanelor.
- 4. Reputație și încredere:** Organizațiile care utilizează sisteme AI părtinitoare riscă să afecteze reputația și să piardă încrederea publicului, ceea ce poate avea consecințe semnificative asupra imaginii lor de marcă și a credibilității pe piață.



## ➤ **Factori care contribuie la rezultate părtinitoare**

Mai mulți factori interconectați contribuie la apariția sistemelor AI părtinitoare, subminându-le fiabilitatea, corectitudinea și eficiența. În această secțiune vom explora unii dintre cei mai comuni factori care contribuie la rezultate părtinitoare în sistemele AI.

- **Date tendențioase:** Datele părtinitoare utilizate pentru instruirea sistemelor de inteligență artificială generează o părtinire algoritmică, care poate conduce la rezultate discriminatorii. Pentru a atenua această situație, trebuie luate în considerare cu atenție colectarea și preprocesarea datelor, inclusiv eșantionarea reprezentativă, algoritmi de detectare și atenuare a părtinirilor și creșterea diversității datelor.
- **Algoritmi defectuoși:** Sistemele AI pot avea rezultate părtinitoare din cauza algoritmilor defectuoși, a alegerilor de proiectare, a arhitecturilor de model, a procedurilor de optimizare sau a variabilelor de intrare. Tehnicile de învățare automată bazată pe corectitudine, transparența algoritmică și interpretabilitatea pot contribui la atenuarea acestor prejudecăți.
- **Biasuri umane:** Prejudecățile din sistemele AI pot rezulta din influențele inconștiente ale dezvoltatorilor, cercetătorilor de date și factorilor de decizie. Pentru a evita aceste prejudecăți, echipele de dezvoltare a IA ar trebui să se concentreze pe diversitate, orientări etice și mecanisme de responsabilitate.



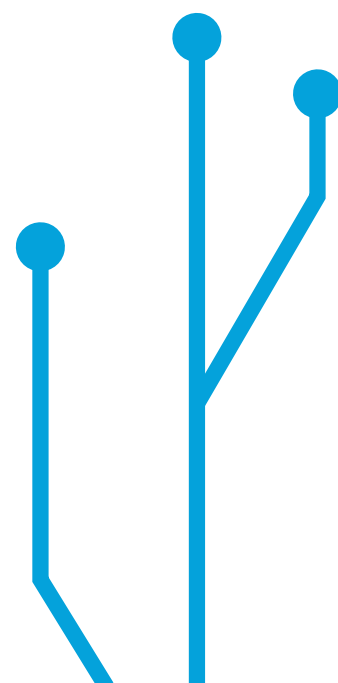


## > Exemple de sisteme părtinitoare

Sistemele AI pot avea prejudecăți, ceea ce duce la rezultate inechitabile. Mai jos sunt prezentate câteva exemple rapide din lumea reală de sisteme AI cu prejudecăți frecvente, care evidențiază consecințele potențiale ale prejudecăților algoritmice. Le vom explora mai în profunzime în secțiunile ulterioare ale acestui curs.

- **Algoritmi de recunoaștere facială:** Tehnologia de recunoaștere facială poate avea prejudecăți care perpetuează disparitățile rasiale sau de gen, conducând la arestări sau supravegheri greșite ale anumitor grupuri. Abordarea acestor prejudecăți este esențială pentru asigurarea corectitudinii și echității în sistemele AI și pentru restabilirea încrederii publice.
- **Algoritmi predictivi de poliție:** Algoritmii de poliție predictivă pot perpetua prejudecățile prezente în datele istorice privind criminalitatea, ceea ce duce la o poliție excesivă a anumitor comunități sau grupuri demografice. Algoritmii părtinitori pot exacerba disparitățile existente în practicile de aplicare a legii și pot ridica probleme legate de echitate, responsabilitate și potențialul de rezultate discriminatorii în sistemele de justiție penală.

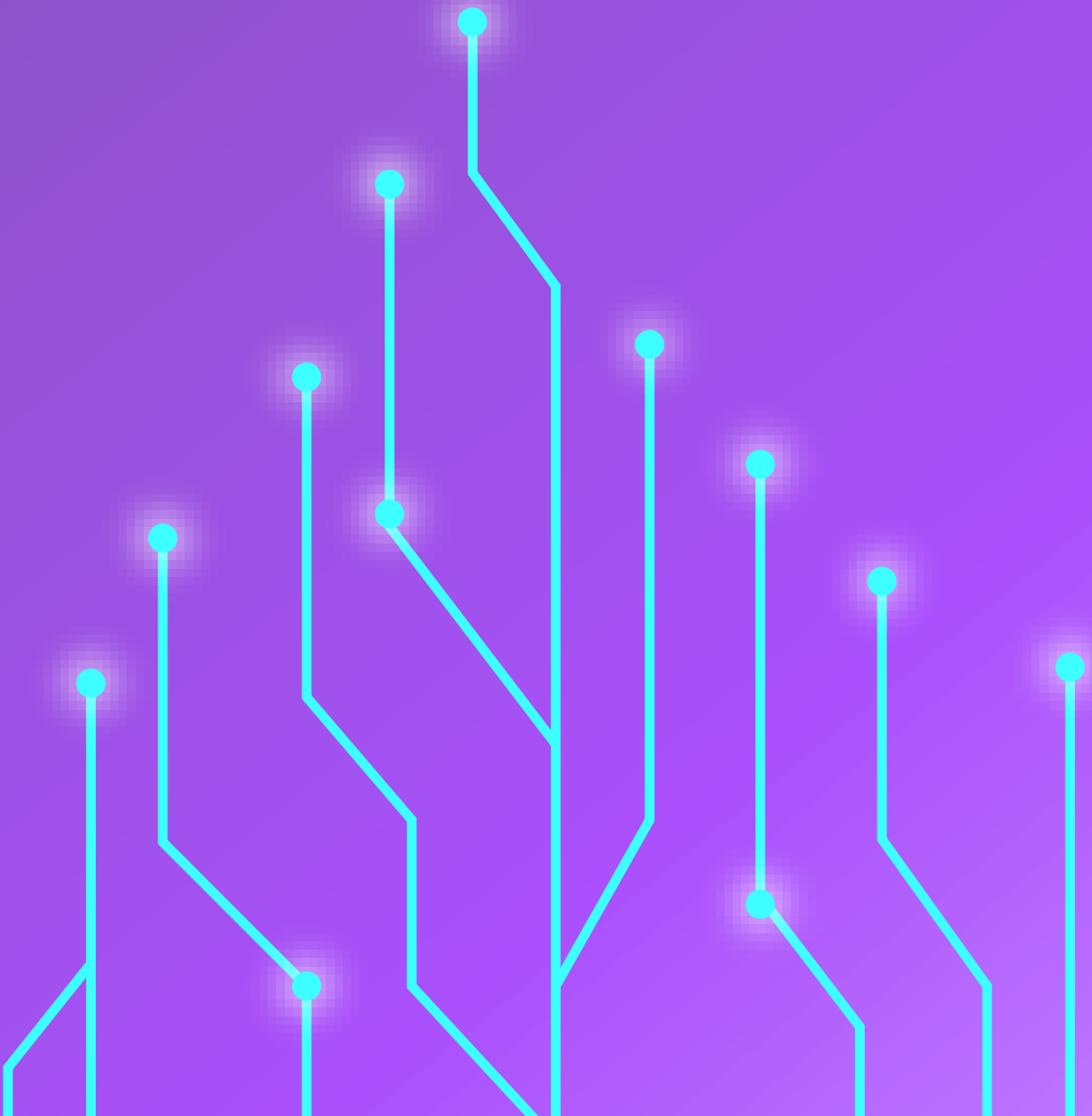
- **Sisteme automatizate de angajare:** Sistemele automatizate de angajare pot perpetua prejudecățile, conducând la practici discriminatorii și limitând diversitatea forței de muncă. Algoritmii părtinitori pot învăța modele de părtinire din datele istorice, ceea ce duce la tratamentul preferențial al anumitor grupuri demografice. Auditul și atenuarea prejudecăților sunt esențiale pentru a asigura corectitudinea, echitatea și responsabilitatea în procesele de recrutare bazate pe inteligența artificială.

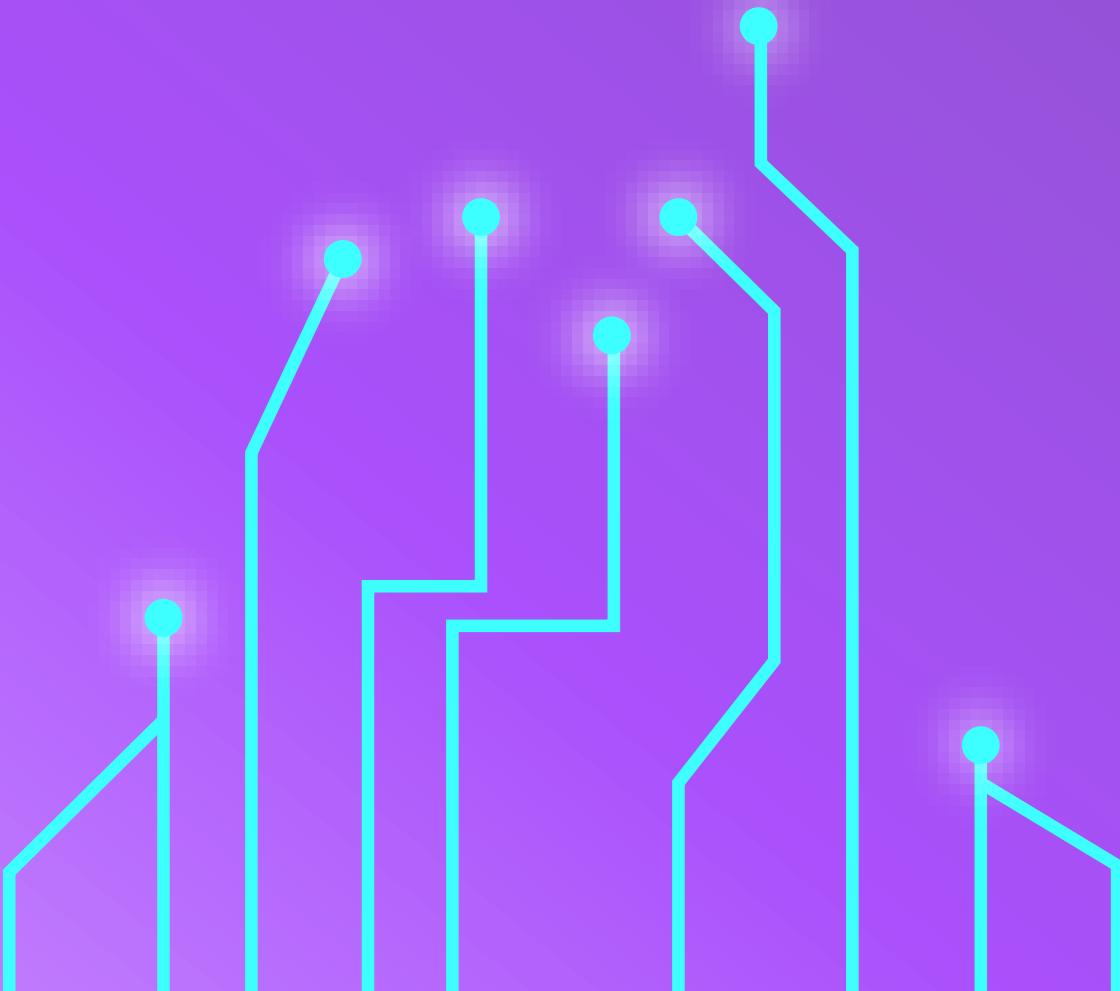




# 05. Înțelegerea prejudecăților în sistemele AI

CU1 | Ce este părtinirea algoritmică?





## 05. Înțelegerea prejudecăților în sistemele AI

În această secțiune vom explora trei tipuri de prejudecăți: bazate pe date, bazate pe modele și bazate pe oameni.

Aceste prejudecăți pot afecta acuratețea și fiabilitatea sistemelor AI, iar înțelegerea lor este primul pas în prevenirea lor.

### > Prejudecăți bazate pe date

Ce este prejudecata generată de date?

Prejudecățile determinate de date se referă la prejudecățile care rezultă din caracteristicile sau distribuția datelor de formare utilizate pentru a dezvolta modele de învățare automată.

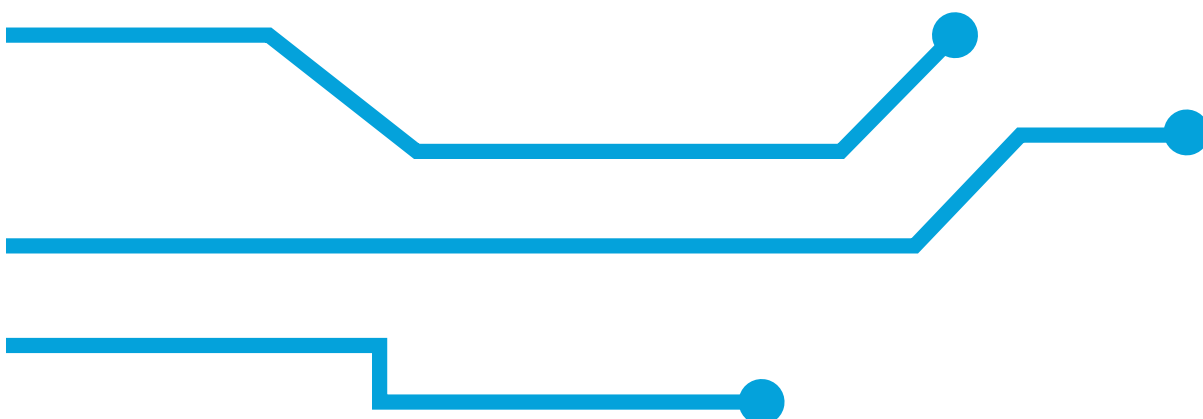
Datele de instruire părtinitoare pot reflecta inegalități istorice, prejudecăți sociale sau discriminare sistemică, ceea ce duce la reprezentări distorsionate ale anumitor grupuri demografice sau la subreprezentarea altora.

Înțelegerea prejudecăților generate de date este esențială pentru a recunoaște modul în care datele de instruire prejudiciate pot perpetua și exacerba stereotipurile, inegalitățile și practicile discriminatorii existente în sistemele de inteligență artificială.



## Cauzele prejudecăților generate de date

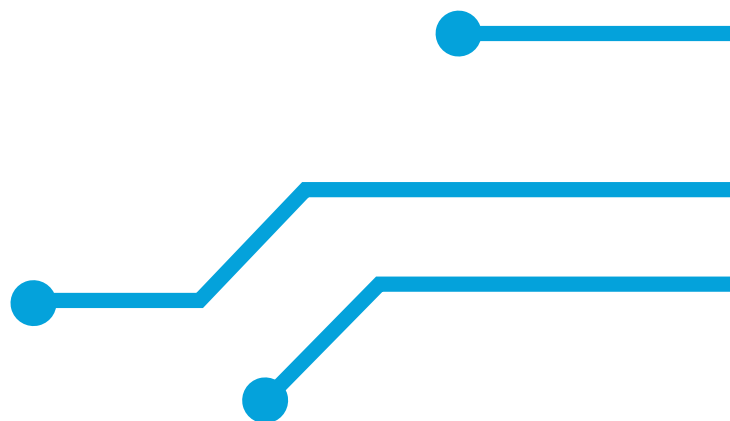
- 1. Eșantionare incompletă sau tendențioasă:** Seturile de date de instruire pot fi lipsite de diversitate sau nu reușesc să reprezinte în mod adecvat anumite grupuri demografice, ceea ce conduce la reprezentări distorsionate și predicții părtinitoare ale modelului.
- 2. Prejudecăți istorice:** Datele de instruire pot reflecta inegalități istorice sau prejudecăți sistemice prezente în societate, perpetuând rezultate discriminatorii în sistemele AI.
- 3. Biasuri de etichetare:** Practicile de etichetare părtinitoare sau subiective pot introduce prejudecăți în datele de formare, influențând predicțiile modelului și consolidând stereotipurile existente.





## Exemple de prejudecăți generate de date

- 1. Recunoaștere facială distorsionată:** Algoritmii de recunoaștere facială antrenați pe seturi de date dezechilibrate pot prezenta prejudecăți rasiale sau de gen, ducând la o identificare greșită și la discriminarea anumitor grupuri demografice.
- 2. Prejudecăți de gen în modelele lingvistice:** Modelele lingvistice antrenate pe corpusuri de texte tendențioase pot genera un limbaj discriminatoriu sau stereotip de gen, reflectând și perpetuând prejudecățile sociale.
- 3. Prejudecăți rasiale în poliția predictivă:** Algoritmii de poliție predictivă instruiți pe baza datelor tendențioase privind criminalitatea pot viza în mod disproporționat comunitățile minoritare, exacerband disparitățile rasiale în aplicarea legii.







## Impactul prejudecăților generate de date

- 1. Întărirea stereotipurilor:** Datele de formare tendențioase pot consolida stereotipurile și prejudecățile existente, perpetuând discriminarea și inegalitatea în sistemele AI.
- 2. Amplificarea inegalităților:** Prejudecățile bazate pe date pot exacerba inegalitățile și disparitățile existente, conducând la un tratament inechitabil și la rezultate discriminatorii pentru grupurile marginalizate.
- 3. Erodarea încrederii:** Sistemele AI părtinitoare subminează încrederea în tehnologie, exacerbând preocupările legate de corectitudine, responsabilitate și transparență.

Prejudecățile generate de date reprezintă o provocare semnificativă pentru sistemele de inteligență artificială corecte și echitabile. Prin înțelegerea cauzelor și consecințelor sale, părțile interesate pot lua măsuri proactive pentru a atenua prejudecățile din datele de formare și pentru a promova incluziunea în IA.

## > Prejudecăți determinate de model

Ce este prejudecata generată de model?

Prejudecățile determinate de modele se referă la prejudecățile care rezultă din proiectarea, structura sau optimizarea modelelor de învățare automată, conducând la rezultate discriminatorii sau predicții distorsionate.

Cauzele prejudecăților determinate de model

- 1. Prejudecăți în selectarea caracteristicilor:** Caracteristicile modelului selectate în timpul procesului de modelare pot codifica în mod neintenționat prejudecățile prezente în datele de formare, ceea ce duce la predicții părtinitoare sau la rezultate discriminatorii.
- 2. Complexitatea algoritmică:** Algoritmii complecși de învățare automată pot capta și consolida prejudecățile subtile prezente în datele de instruire, amplificând impactul acestora asupra predicțiilor modelului.
- 3. Obiective de optimizare:** Obiectivele de optimizare definite în timpul procesului de formare a modelului pot prioritiza în mod neintenționat anumite rezultate în detrimentul altora, ceea ce conduce la predicții părtinitoare sau inechitabile.



## Exemple de prejudecăți determinate de model

- 1. Prejudecăți de gen în algoritmi de angajare:** Algoritmii automatizați de angajare pot favoriza în mod involuntar candidații de sex masculin în detrimentul candidaților de sex feminin din cauza selecției părtinitoare a caracteristicilor sau a obiectivelor de optimizare, perpetuând disparitățile de gen în cadrul forței de muncă.
- 2. Prejudecăți rasiale în algoritmi de condamnare:** Algoritmii predictivi de stabilire a sentințelor utilizați în sistemele de justiție penală pot recomanda în mod disproporționat sentințe mai aspre pentru inculpații aparținând minorităților, amplificând disparitățile rasiale în ceea ce privește ratele de încarcerare.
- 3. Prejudecăți socioeconomice în modelele de aprobare a împrumuturilor:** Modelele de învățare automată utilizate pentru aprobarea împrumuturilor pot refuza sistematic acordarea de împrumuturi persoanelor din comunitățile marginalizate, exacerband inegalitățile socioeconomice în ceea ce privește accesul la serviciile financiare.



## Impactul prejudecăților generate de modele

- 1. Perpetuarea discriminării:** Prejudecățile bazate pe modele pot perpetua și consolida discriminarea și inegalitățile existente în societate, ducând la un tratament incorect și la rezultate părtinitoare pentru grupurile marginalizate.
- 2. Lipsa responsabilității:** Modelele AI părtinitoare pot fi lipsite de transparență și responsabilitate, ceea ce face dificilă identificarea și abordarea practicilor discriminatorii în sistemele AI.
- 3. Implicații etice:** Prejudecățile determinate de modele ridică probleme etice legate de echitate, justiție și drepturile omului, subliniind necesitatea unor orientări și reglementări etice care să guverneze dezvoltarea și implementarea inteligenței artificiale.

Prejudecățile generate de modele reprezintă provocări semnificative pentru dezvoltarea și implementarea unor sisteme de inteligență artificială echitabile și responsabile. Prin înțelegerea mecanismelor și a implicațiilor prejudecăților generate de modele, părțile interesate pot pune în aplicare strategii de atenuare a prejudecăților și de promovare a corectitudinii și echității în tehnologiile IA.





## > Prejudecăți cauzate de om

Ce este prejudecata generată de om?

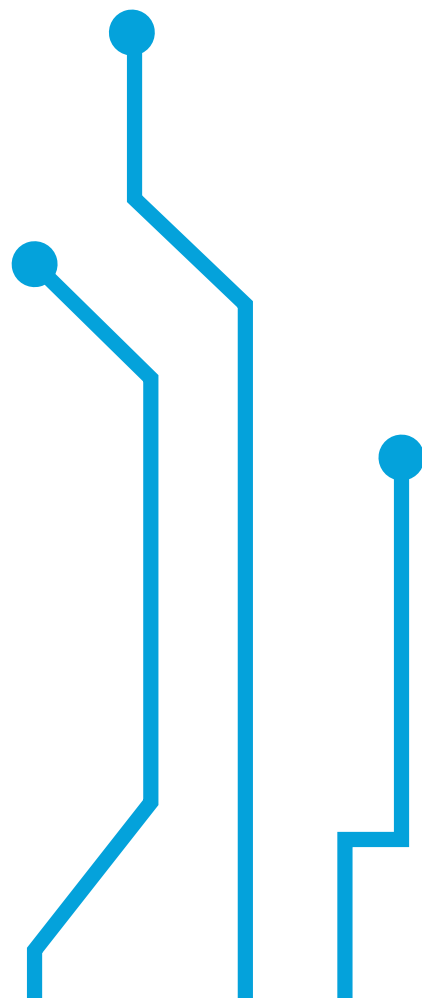
Prejudcățile provocate de om în IA se referă la prejudcățile care rezultă din deciziile, acțiunile sau judecățile persoanelor implicate în dezvoltare și implementare. Acestea pot proveni din prejudcăți cognitive, influențe culturale și prejudcăți societale, conducând la rezultate părtinitoare sau practici discriminatorii.

Cauzele prejudcăților provocate de om

- 1. Prejudcăți în colectarea datelor:** Prejudcățile de colectare a datelor, cum ar fi prejudcățile de eșantionare sau de selecție, pot duce la date de instruire părtinitoare și la predicții distorsionate ale modelului.
- 2. Prejudcăți de proiectare algoritmică:** Algoritmii inteligenței artificiale pot fi părtinitori din cauza alegerilor designerilor și dezvoltatorilor umani, perpetuând rezultate părtinitoare în sistemele de inteligență artificială.
- 3. Prejudcăți de interpretare și implementare:** Interpreții umani și factorii de decizie pot prezenta prejudcăți atunci când implementează sisteme AI, ceea ce poate duce la practici discriminatorii și tratamente inechitabile.

## Exemple de prejudecăți provocate de om

- 1. Prejudecăți în sistemele de recunoaștere facială:** Prejudecățile umane în colectarea datelor de formare și în proiectarea algoritmică pot duce la prejudecăți rasiale sau de gen în sistemele de recunoaștere facială, rezultând în identificarea eronată sau subreprezentarea anumitor grupuri demografice.
- 2. Echitate în algoritmi de angajare:** Prejudecățile din procesele decizionale umane, cum ar fi selectarea CV-urilor sau evaluarea interviurilor, pot perpetua disparitățile de gen sau rasiale în rezultatele angajărilor, chiar și atunci când se utilizează algoritmi de angajare pe bază de inteligență artificială.





## Impactul prejudecăților provocate de om

- 1. Exacerbarea inegalităților existente:** Prejudecățile cauzate de om în IA pot exagera inegalitățile și disparitățile existente în societate. Colectarea, proiectarea algoritmică și interpretarea părtinitoare a datelor pot conduce la un tratament inechitabil pentru grupurile marginalizate, perpetuând discriminarea și împiedicând progresul social.
- 2. Erodarea încrederii și a încrederii publicului:** Sistemele AI afectate de prejudecăți umane pot eroda încrederea publicului în tehnologie. Pot apărea îngrijorări cu privire la corectitudine, transparență și responsabilitate, împiedicând adoptarea și acceptarea inteligenței artificiale în diverse sectoare.
- 3. Reducerea eficacității sistemelor AI:** Prejudecățile provocate de om pot submina eficacitatea sistemelor AI. Datele de formare părtinitoare sau interpretările părtinitoare ale oamenilor pot conduce la predicții inexacte, recomandări eronate și rezultate suboptimale, împiedicând beneficiile potențiale ale AI.

Prejudecățile umane reprezintă o provocare semnificativă pentru crearea unor sisteme AI corecte și responsabile. Prin înțelegerea și atenuarea prejudecăților algoritmice, părțile interesate pot crea sisteme AI mai fiabile și mai transparente.



# CharTe



Cofinanțat de  
Uniunea Europeană

Finanțat de Uniunea Europeană. Punctele de vedere și opiniile exprimate aparțin, însă, exclusiv autorului (autorilor) și nu reflectă neapărat punctele de vedere și opiniile Uniunii Europene sau ale Agenției Executive Europene pentru Educație și Cultură (EACEA). Nici Uniunea Europeană și nici EACEA nu pot fi considerate răspunzătoare pentru acestea.



Universitat  
de les Illes Balears



helixconnect



2022-1-ES01-KA220-HED-000085257

CC BY 4.0