



Microcredențial de IA etică

CARTEA

CU4 | Transparență

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. Importanța transparenței în sistemele AI

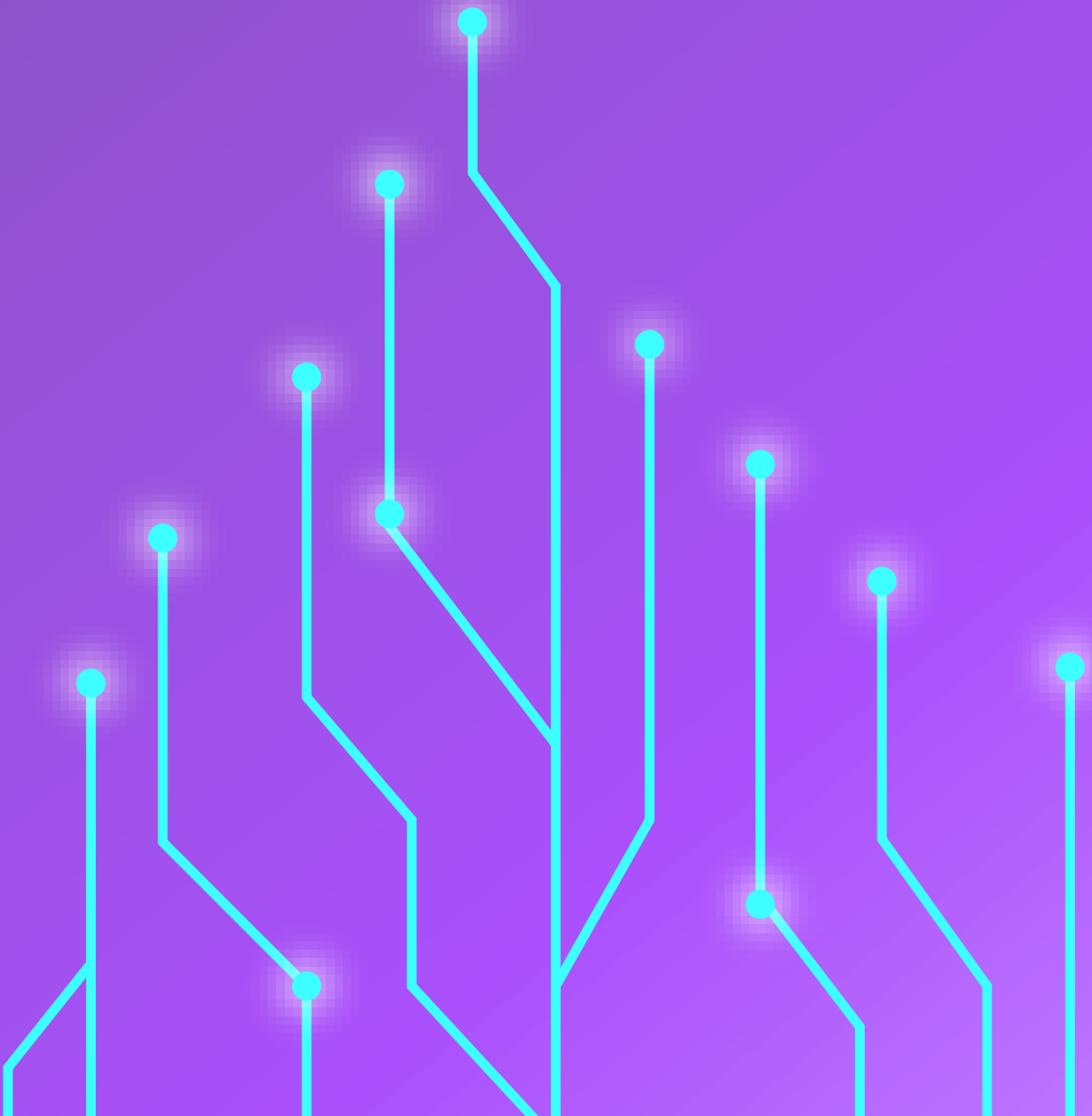
**03. Relația dintre transparență și părtinirea
algoritmică**

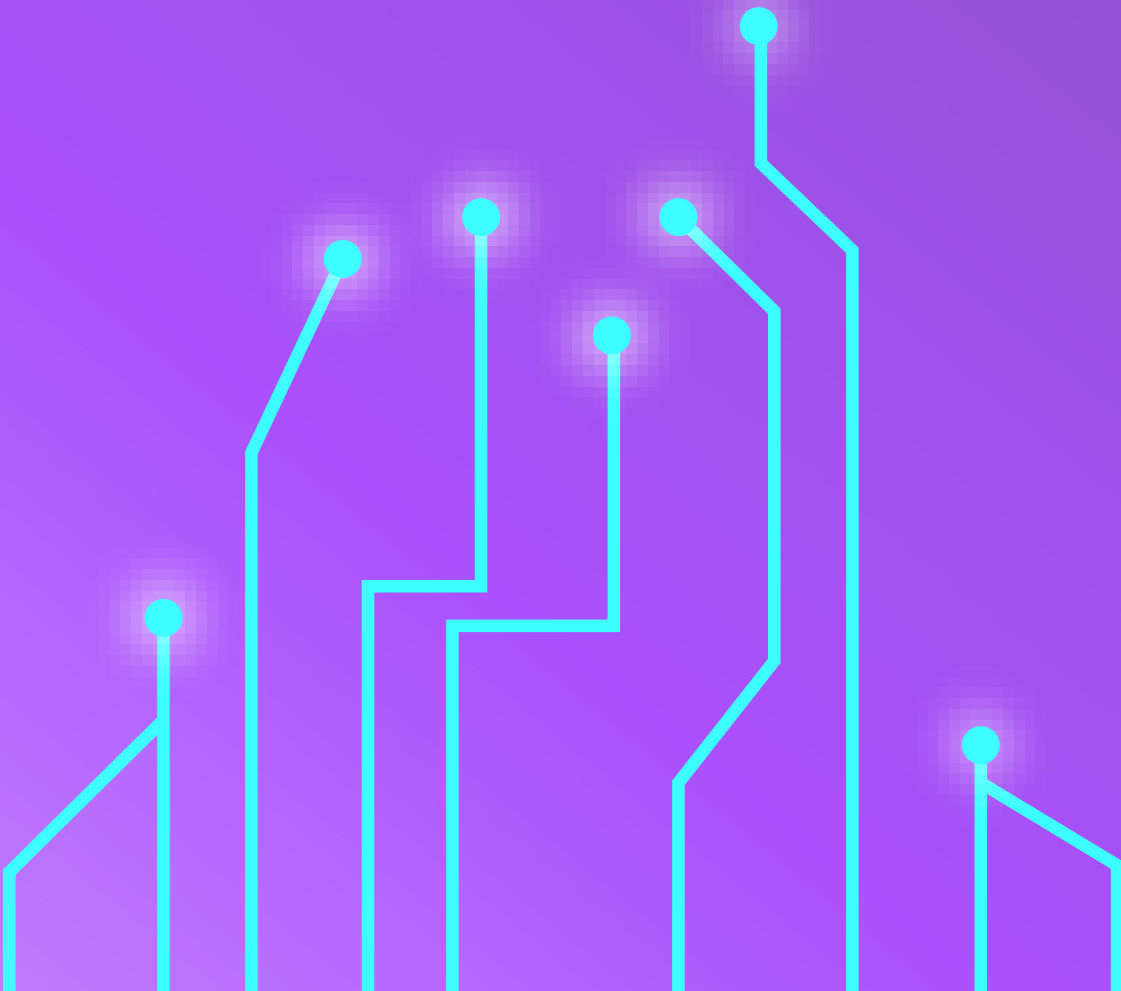
**04. Strategii pentru promovarea transparenței
în sistemele AI**

05. Concluzii

01. Introducere

CU4 | Transparență





01. Introducere

În această unitate de competență, cursanții vor dobândi cunoștințe cu privire la importanța transparenței în sistemele AI, concentrându-se pe înțelegerea conceptelor de bază, relația dintre transparență și prejudecățile algoritmice și relevanța strategiilor pentru a se asigura că sistemele AI sunt inteligibile, explicabile și accesibile părților interesate, recunoscând implicațiile din lumea reală, apreciind cât de importante pot fi modelele interpretabile, documentația clară și comunicarea eficientă în promovarea unei culturi a transparenței, atenuând prejudecățile algoritmice.

Rezultatele cunoștințelor pentru acest curs includ:

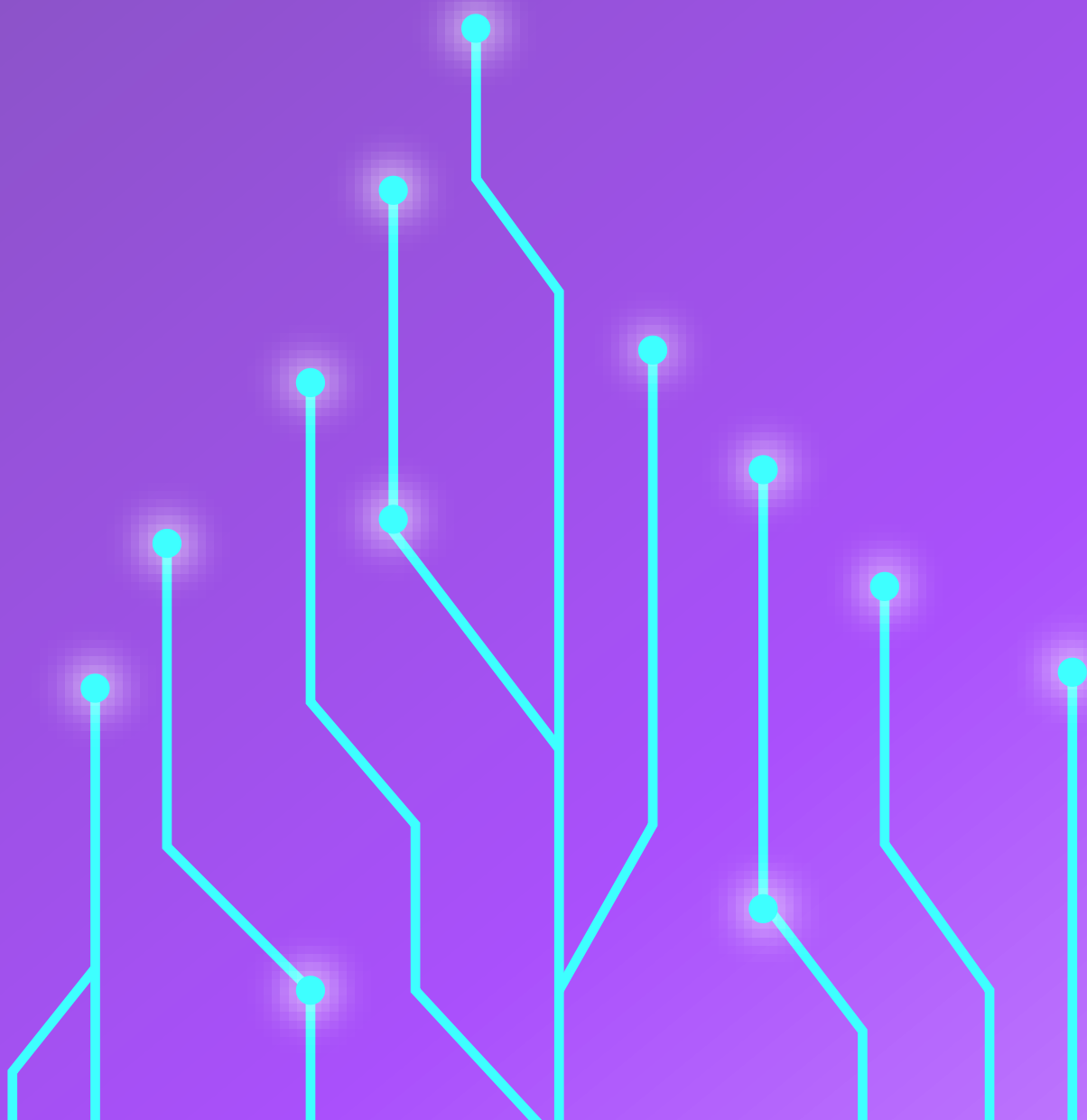
- **Importanța transparenței în sistemele AI** și relevanța acesteia în asigurarea faptului că sistemele AI sunt inteligibile, explicabile și accesibile părților interesate. Vom identifica beneficiile și vom aprecia importanța sistemelor AI transparente pentru construirea încrederii și facilitarea înțelegerii părților interesate. Ca exemplu: Un model AI conceput pentru a detecta cancerul, chiar dacă este greșit doar cu 1%, ar putea amenința o viață. În astfel de cazuri, AI și oamenii trebuie să lucreze împreună, iar sarcina devine mult mai ușoară atunci când modelul AI poate explica cum a ajuns la o anumită decizie. Transparența face din inteligența artificială un jucător de echipă.

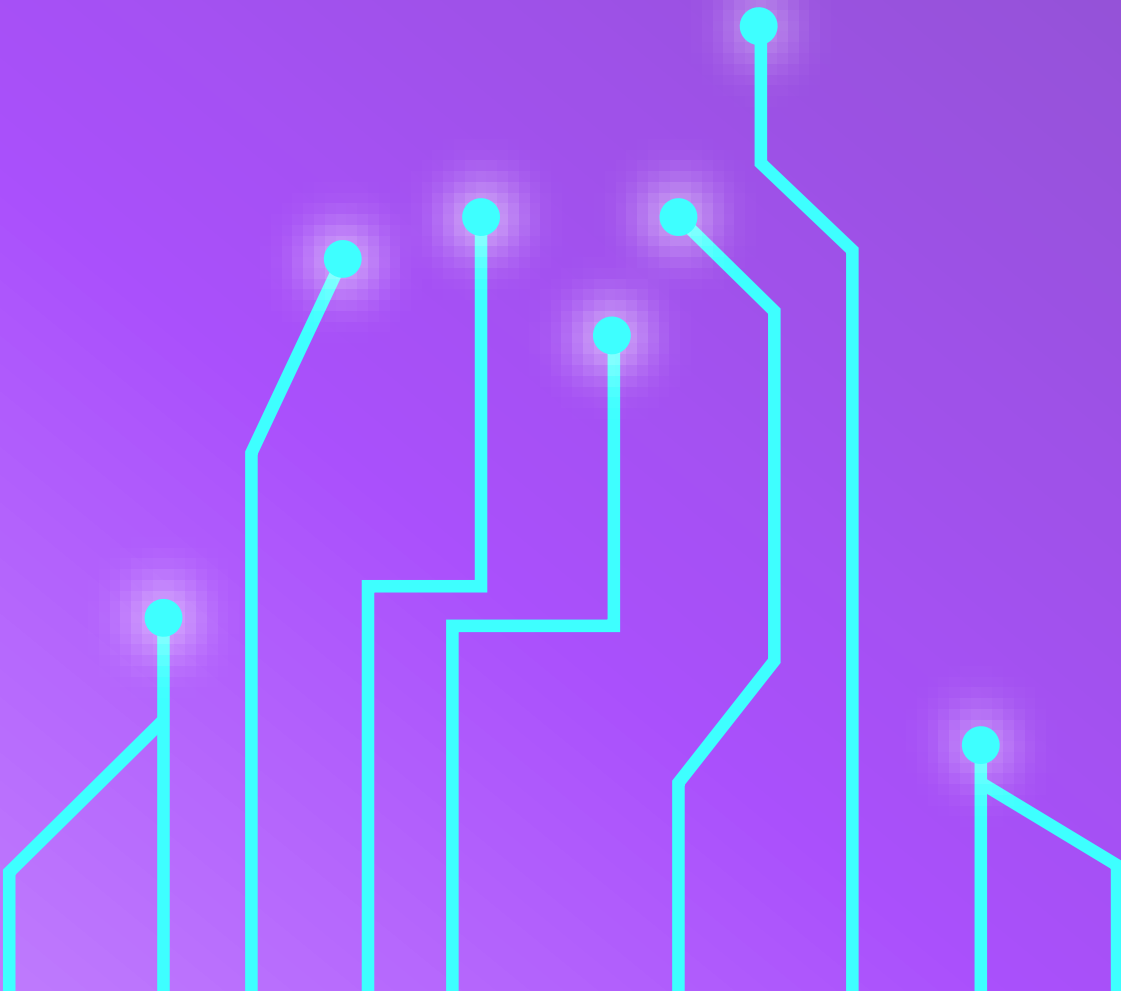


- **Relația dintre transparență și părtinirea algoritmică** pentru a găsi legătura dintre transparență și părtinirea algoritmică, recunoscând pericolele opaciei și modul în care o transparență sporită poate ajuta la identificarea, prevenirea și atenuarea rezultatelor părtinitoare în sistemele AI. Vom recunoaște importanța transparenței în abordarea și atenuarea părtinirilor algoritmice. Ca exemplu: destul de des, algoritmi AI sunt opaci în sensul că astfel de explicații nu sunt disponibile pentru toate părțile interesate. Această opacitate poate avea diferite surse. Uneori, instituțiile sau corporațiile nu reușesc să comunice atunci când se bazează pe sistemele AI sau pe modul în care funcționează aceste sisteme.
- **Strategii pentru promovarea transparenței în sistemele AI**, cum ar fi utilizarea de modele interpretabile, furnizarea de documentație clară și comunicarea proceselor decizionale ale aplicațiilor AI. Vom explica cât de importante sunt aceste strategii pentru promovarea unei culturi a transparenței și atenuarea prejudecăților algoritmice. De exemplu, IA poate afecta diverse părți interesate, cum ar fi utilizatorii, clienții, angajații, managerii, autoritățile de reglementare sau societatea. Pentru a asigura transparența și responsabilitatea, trebuie să vă implicați și să împuterniciți părțile interesate din domeniul IA pe tot parcursul ciclului de viață al SI.

02. Importanța transparenței în sistemele AI

CU4 | Transparență





02. Importanța transparenței în sistemele AI

Transparența este un principiu fundamental în dezvoltarea și implementarea sistemelor de inteligență artificială (AI).

Transparența în IA se referă la deschiderea și accesibilitatea sistemelor de IA, permițând părților interesate să înțeleagă cum funcționează algoritmi, de ce sunt luate anumite decizii și ce factori influențează rezultatele acestora. Aceasta cuprinde diverse aspecte, inclusiv disponibilitatea informațiilor despre sursele de date, modelele algoritmice, procesele decizionale și potențialele prejudecăți. Sistemele de IA transparente permit părților interesate, inclusiv utilizatorilor, dezvoltatorilor, factorilor de decizie politică și publicului larg, să analizeze și să conteste rezultatele algoritmilor, promovând încrederea și responsabilitatea.

Unul dintre principalele avantaje ale sistemelor AI transparente este inteligibilitatea acestora. Atunci când algoritmi AI sunt transparenți, părțile interesate pot înțelege cum funcționează și de ce produc anumite rezultate. Această înțelegere permite utilizatorilor să aibă încredere în tehnologiile AI și să ia decizii în cunoștință de cauză cu privire la utilizarea acestora. De exemplu, în contextul unui model AI de diagnostic medical, transparența permite profesioniștilor din domeniul sănătății să înțeleagă modul în care modelul a ajuns la diagnosticul său, permițându-le să valideze acuratețea și fiabilitatea acestuia înainte de a lua decizii de tratament.



În plus, transparența facilitează explicabilitatea, care este esențială pentru a garanta că sistemele AI pot oferi explicații interpretabile pentru deciziile și acțiunile lor. Inteligența artificială explicabilă permite părților interesate să înțeleagă logica din spatele rezultatelor algoritmice și să identifice și să corecteze prejudecățile sau erorile. De exemplu, în cazul unui sistem AI de aprobare a împrumuturilor, transparența și explicabilitatea permit solicitanților de împrumuturi să înțeleagă de ce cererea lor a fost aprobată sau respinsă, oferind informații despre procesul decizional și căi de atac în cazul în care consideră că decizia a fost părtinitoare sau nedreaptă.

În plus, transparența îmbunătățește accesibilitatea sistemelor AI, făcându-le mai incluzive și mai echitabile. Atunci când algoritmiile inteligenței artificiale sunt transparente, părțile interesate din diverse medii și niveluri de expertiză pot accesa și interpreta informații privind funcționarea și rezultatele acestora. Această accesibilitate garantează că tehnologiile IA nu sunt doar ușor de înțeles, ci și utilizabile de către o gamă largă de utilizatori, inclusiv de către cei cu handicap sau cu cunoștințe tehnice limitate. De exemplu, în dezvoltarea de instrumente de accesibilitate bazate pe IA pentru persoanele cu handicap, transparența permite utilizatorilor să înțeleagă cum funcționează instrumentele și cum pot beneficia de ele.

Un exemplu ilustrativ al importanței transparenței în sistemele AI este dezvoltarea de modele AI pentru diagnosticarea medicală, cum ar fi detectarea cancerului. Chiar dacă un model AI este extrem de precis, cu o rată de succes de 99%, marja de eroare rămasă de 1% ar putea avea consecințe amenințătoare pentru viața pacienților. În astfel de scenarii critice, transparența devine esențială pentru a se asigura că profesioniștii din domeniul sănătății pot înțelege modul în care modelul AI a ajuns la diagnosticul său și pot verifica acuratețea acestuia înainte de a lua decizii de tratament. Prin furnizarea de explicații transparente cu privire la procesul său decizional, modelul AI devine un instrument valoros pentru profesioniștii din domeniul sănătății, sporind capacitatea acestora de a diagnostica și trata pacienții în mod eficient.

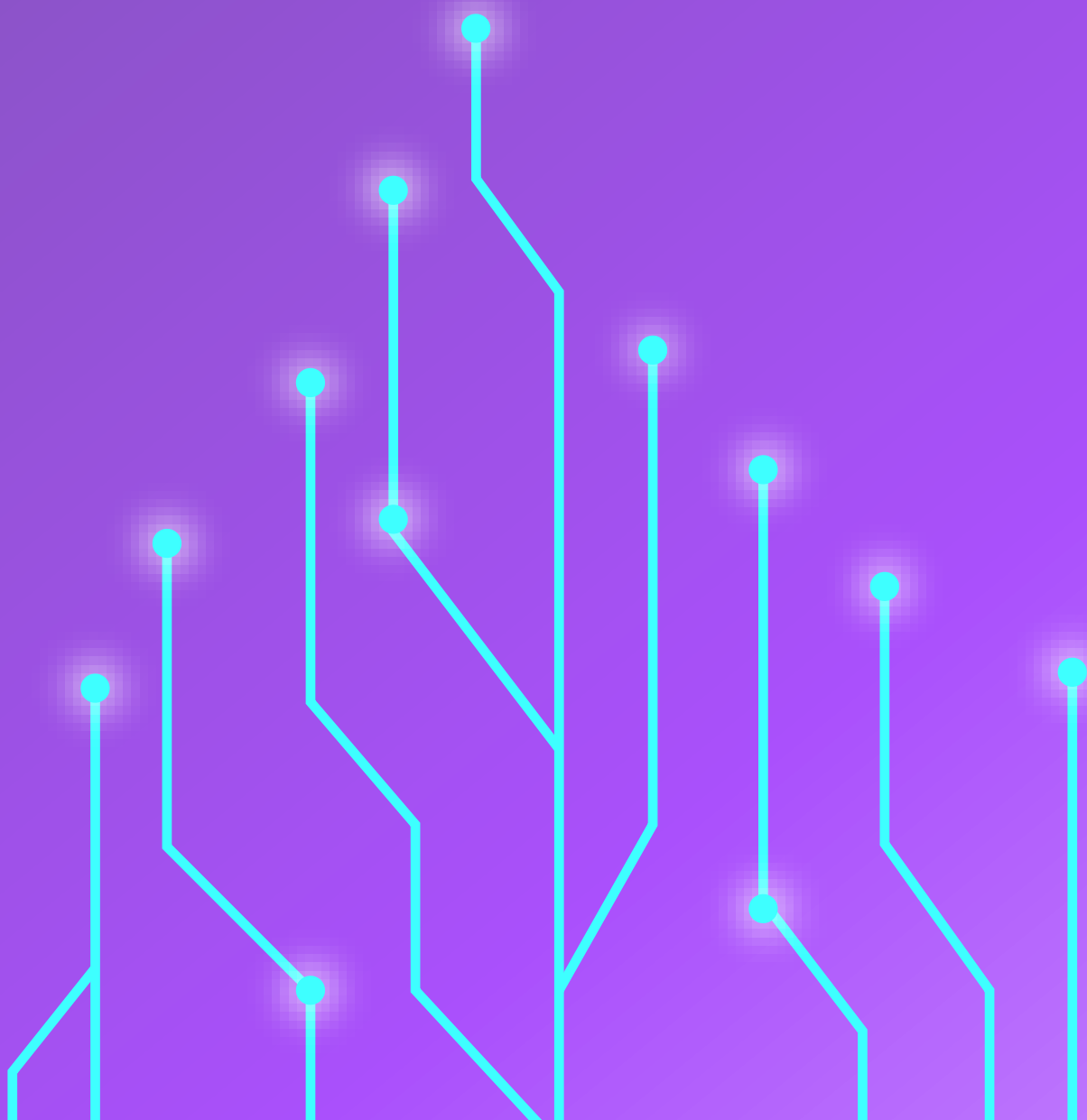
După cum am citit deja în această broșură, prejudecățile algoritmice se referă la erori sistematice sau nedreptate în algoritmiile inteligenței artificiale care conduc la rezultate discriminatorii pentru anumite persoane sau grupuri. Aceste prejudecăți pot proveni din diverse surse, inclusiv date de formare părtinitoare, concepție algoritmică greșită sau prejudecăți umane codificate în sistem. Consecințele prejudecăților algoritmice pot fi de mare amploare, perpetuând inegalitățile, consolidând stereotipurile și subminând încrederea în sistemele AI.

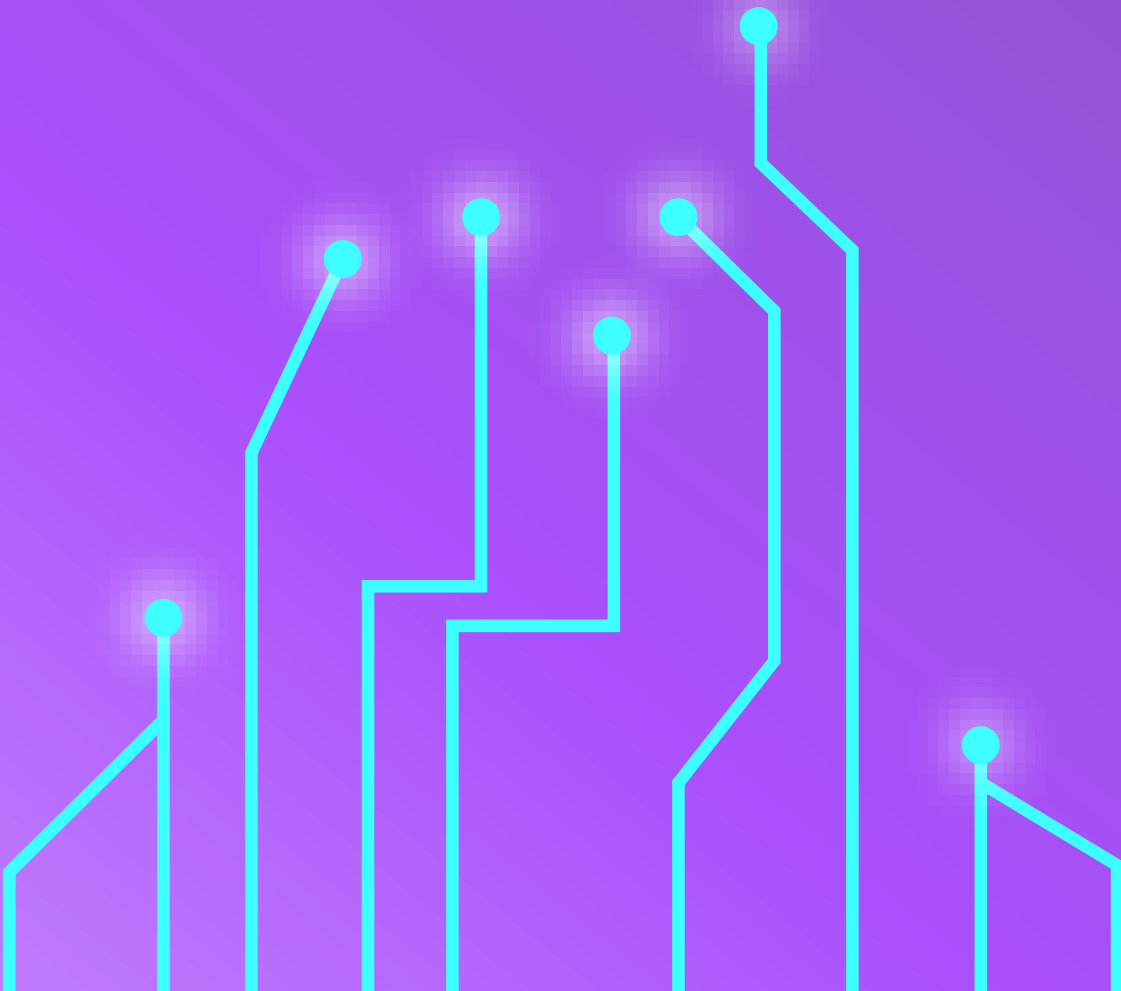




03. Relația dintre transparență și prejudecățile algoritmice

CU4 | Transparență





03. Relația dintre transparență și prejudecățile algoritmice

Opacitatea, sau lipsa de transparență, exacerbează riscurile asociate cu părtinirea algoritmică.

Destul de des, algoritmiile inteligenței artificiale sunt opaci, ceea ce înseamnă că explicațiile privind deciziile și acțiunile lor nu sunt ușor accesibile tuturor părților interesate. Această opacitate poate proveni din diverse surse, inclusiv secretul instituțional, confidențialitatea corporativă sau complexitatea tehnică. Atunci când părțile interesate nu au acces la informații despre sistemele de inteligență artificială, acestea nu sunt în măsură să evalueze corectitudinea, fiabilitatea sau implicațiile etice ale rezultatelor algoritmice, ceea ce conduce la o lipsă de responsabilitate și la prejudicii potențiale.

Transparența este un antidot esențial la opacitatea sistemelor de IA, permițând părților interesate să analizeze și să conteste deciziile algoritmice, reducând astfel riscurile de părtinire algoritmică. Prin creșterea transparenței, dezvoltatorii și practicienii AI pot oferi părților interesate informații despre modul în care funcționează sistemele AI, de ce sunt luate anumite decizii și ce factori influențează rezultatele acestora. Sistemele de IA transparente permit părților interesate să identifice și să abordeze prejudecățile, să valideze acuratețea algoritmilor și să îi tragă la răspundere pe dezvoltatori pentru utilizarea etică și echitabilă a tehnologiilor de IA.

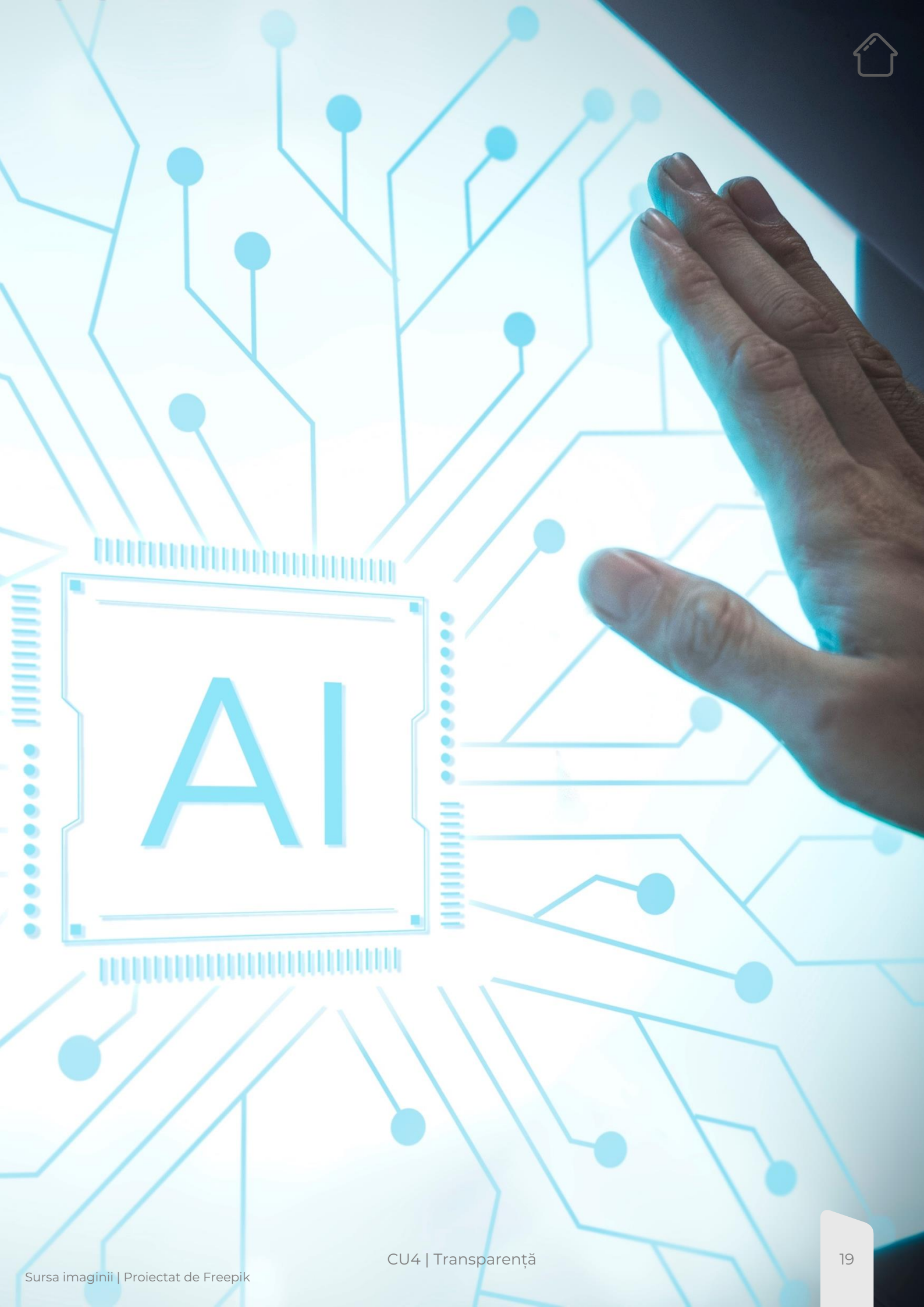


Unul dintre principalele beneficii ale transparenței în abordarea prejudecăților algoritmice este capacitatea de a detecta și de a atenua rezultatele părtinitoare. Atunci când algoritmi AI sunt transparenți, părțile interesate pot examina procesul decizional și pot identifica cazurile în care pot fi prezente prejudecăți. De exemplu, în contextul unui sistem AI de angajare, transparența permite părților interesate să evalueze dacă sistemul discriminează pe nedrept anumite grupuri demografice în procesul de selecție. Prin identificarea rezultatelor părtinitoare, părțile interesate pot lua măsuri corective pentru a atenua prejudiciile cauzate de părtinirea algoritmică și pentru a promova corectitudinea și echitatea.

În plus, transparența facilitează responsabilitatea și încrederea în sistemele IA. Atunci când părțile interesate au acces la informații despre algoritmi AI, acestea pot trage la răspundere dezvoltatorii și practicienii pentru utilizarea etică și echitabilă a tehnologiilor AI. Sistemele de IA transparente creează încredere în rândul utilizatorilor, al autorităților de reglementare și al publicului larg, stimulând încrederea în fiabilitatea și corectitudinea rezultatelor algoritmice. De exemplu, în cazul implementării sistemelor de IA în domeniul justiției penale sau al asistenței medicale, transparența permite părților interesate să înțeleagă modul în care sunt luate deciziile și să se asigure că aceste decizii sunt conforme cu principiile etice și cu standardele juridice.

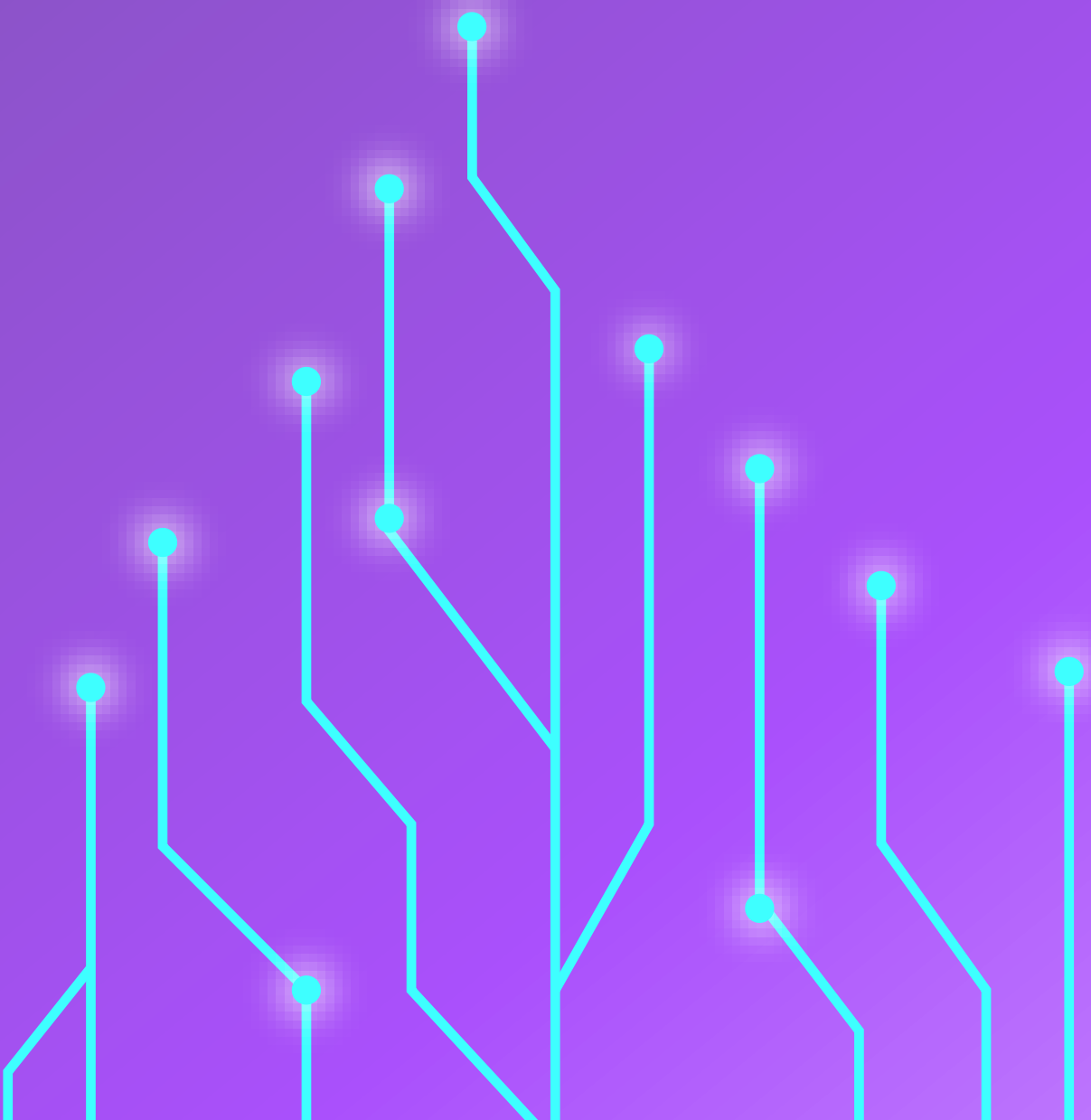
Transparența joacă un rol esențial în abordarea și atenuarea prejudecăților algoritmice din cadrul sistemelor AI. Prin creșterea transparenței, părțile interesate pot detecta și atenua rezultatele părtinitoare, pot promova responsabilitatea și pot consolida încrederea în tehnologiile IA. Pe măsură ce inteligența artificială continuă să evolueze și să se integreze tot mai mult în diverse aspecte ale societății, transparența rămâne esențială pentru a garanta că sistemele de inteligență artificială sunt dezvoltate și implementate într-un mod care respectă standardele etice și promovează corectitudinea și echitatea. Printr-o înțelegere cuprinzătoare a relației dintre transparență și părtinirea algoritmică, cursanții pot contribui la dezvoltarea responsabilă și etică a tehnologiilor IA, creând astfel un viitor mai echitabil și mai favorabil incluziunii.

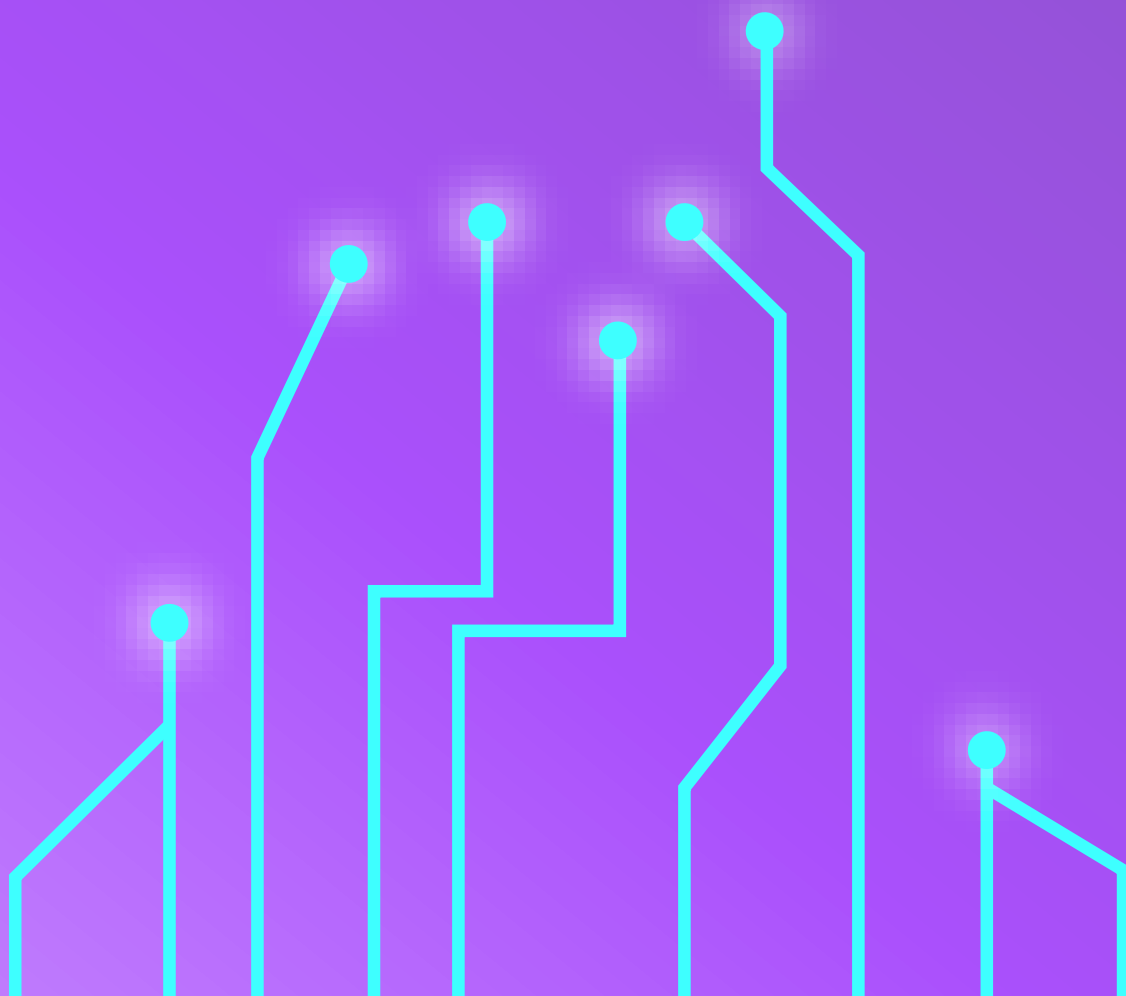




04. Strategii de promovare a transparenței în sistemele IA

CU4 | Transparență





04. Strategii de promovare a transparenței în sistemele IA

Există diverse strategii de promovare a transparenței în sistemele de inteligență artificială (AI), cum ar fi utilizarea de modele interpretabile, furnizarea de documente clare și comunicarea proceselor decizionale.

> Modele interpretabile

Modelele interpretabile reprezintă o strategie-cheie pentru promovarea transparenței în sistemele AI. Acestea sunt modele de învățare automată care produc rezultate care sunt ușor de înțeles și de interpretat de către oameni. Iată câteva exemple:

- **Regresia liniară:** Regresia liniară este un model simplu și interpretabil utilizat în mod obișnuit pentru a prezice rezultate numerice. Acesta funcționează prin ajustarea unei linii drepte la punctele de date, facilitând interpretarea relației dintre variabilele de intrare și rezultat.
- **Arbori decizionali:** Arborii decizionali sunt modele ierarhice care iau decizii pe baza unei serii de afirmații if-then. Fiecare nod din arbore reprezintă o decizie bazată pe o caracteristică a datelor, facilitând urmărirea logicii din spatele predicțiilor modelului.



- **Regresia logistică:** Regresia logistică este un model statistic utilizat pentru sarcinile de clasificare binară. Acesta calculează probabilitatea ca o instanță să aparțină unei anumite clase pe baza caracteristicilor sale de intrare, ceea ce îl face interpretabil și ușor de înțeles.
- **Modele bazate pe reguli:** Modelele bazate pe reguli, cum ar fi arborii de clasificare și regresie (CART) sau regulile de decizie, transformă direct caracteristicile de intrare în reguli de decizie. Aceste reguli sunt ușor de interpretat și pot oferi informații despre modul în care modelul face predicții.
- **Modele aditive generalizate (GAM):** GAM-urile sunt modele flexibile care pot surprinde relații complexe între variabilele de intrare și variabila țintă, menținând în același timp interpretabilitatea. Ele utilizează funcții netede pentru a reprezenta relația dintre fiecare variabilă de intrare și ieșire, permițând o interpretare ușoară a predicțiilor modelului.

Spre deosebire de modelele complexe de tip cutie neagră, modelele interpretabile permit părților interesate să înțeleagă modul în care algoritmi AI iau decizii și factorii care influențează rezultatele acestora. Prin utilizarea modelelor interpretabile, dezvoltatorii de AI pot spori transparența și responsabilitatea, permițând părților interesate să valideze rezultatele algoritmilor și să identifice eventualele prejudecăți sau erori.

De exemplu, în contextul unui sistem AI de scoring al creditelor, utilizarea modelelor interpretabile permite părților interesate să înțeleagă factorii care contribuie la deciziile de creditare, cum ar fi venitul, istoricul creditelor și nivelul datoriilor, promovând astfel transparența și corectitudinea practicilor de creditare.



> Documentație clară

O documentație clară reprezintă o altă strategie esențială pentru promovarea transparenței în sistemele de inteligență artificială.

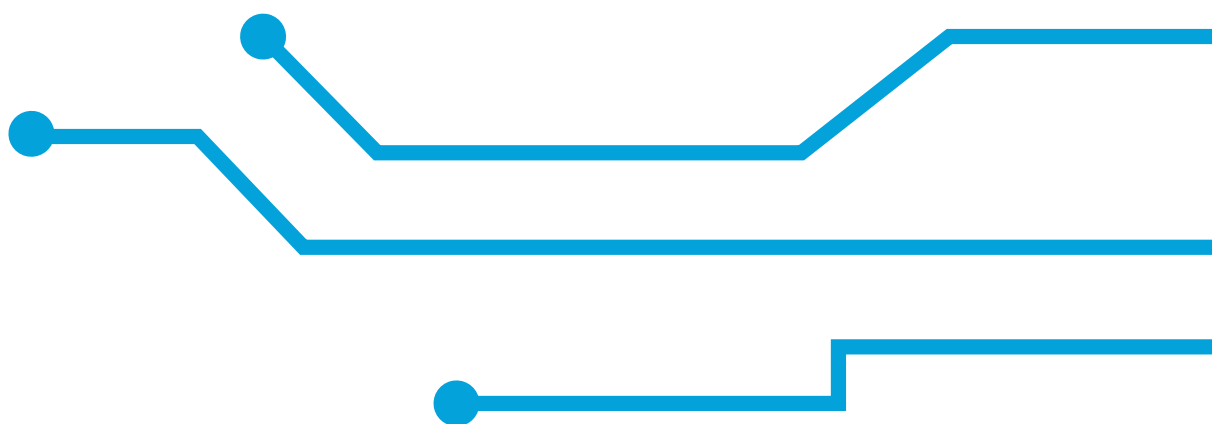
Documentația oferă părților interesate informații cu privire la proiectarea, dezvoltarea și implementarea algoritmilor AI, inclusiv sursele de date, tehnicile de preprocesare, arhitectura modelelor și parametrii de evaluare.

Prin documentarea cuprinzătoare a sistemelor AI, dezvoltatorii pot spori transparența și responsabilitatea, permițând părților interesate să înțeleagă procesele și ipotezele care stau la baza tehnologiilor AI. De exemplu, în dezvoltarea unui sistem AI de întreținere predictivă pentru echipamente industriale, documentația clară permite părților interesate să evalueze fiabilitatea și acuratețea modelelor predictive, să înțeleagă recomandările de întreținere și să verifice conformitatea cu standardele de siguranță.

> Comunicarea eficientă a procesului decizional

Comunicarea eficientă a proceselor decizionale este esențială pentru promovarea transparenței în sistemele AI. Comunicarea garantează că părțile interesate sunt informate cu privire la raționamentul, logica și implicațiile deciziilor algoritmice ale IA.

Prin comunicarea clară și transparentă a proceselor decizionale, dezvoltatorii de AI pot consolida încrederea utilizatorilor, autorităților de reglementare și publicului larg. De exemplu, în cazul implementării sistemelor AI pentru diagnosticarea în domeniul sănătății, o comunicare eficientă asigură faptul că profesioniștii din domeniul sănătății și pacienții înțeleg modul în care sunt luate deciziile de diagnosticare, permițându-le să aibă încredere și să verifice acuratețea diagnosticelor generate de AI.

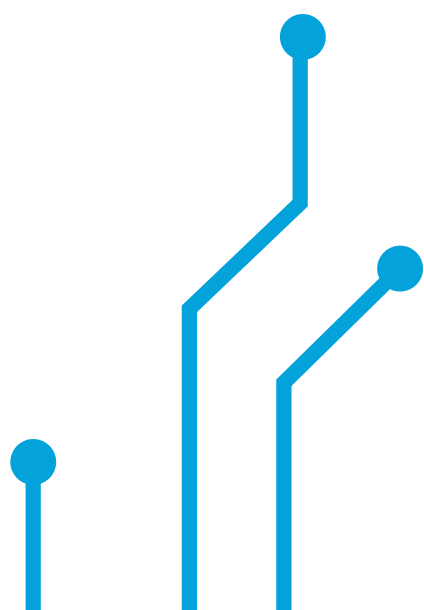




> Implicarea și responsabilizarea părților interesate din domeniul IA

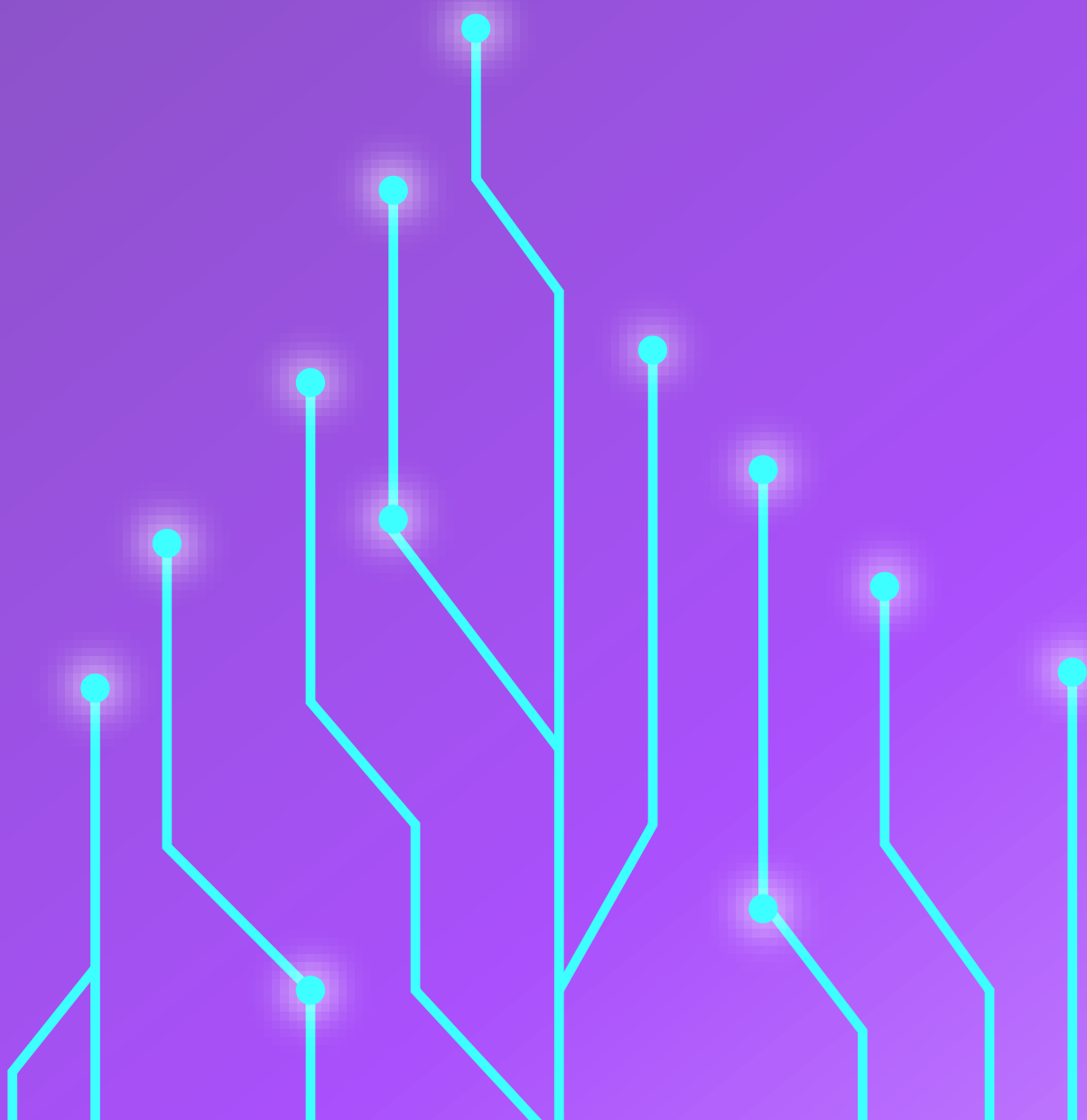
Implicarea și responsabilizarea părților interesate din domeniul IA pe tot parcursul ciclului de viață al IA sunt esențiale pentru asigurarea transparenței și responsabilității. Implicarea părților interesate implică utilizatorii, clienții, angajații, managerii, autoritățile de reglementare și societatea în proiectarea, dezvoltarea, implementarea și evaluarea sistemelor de IA.

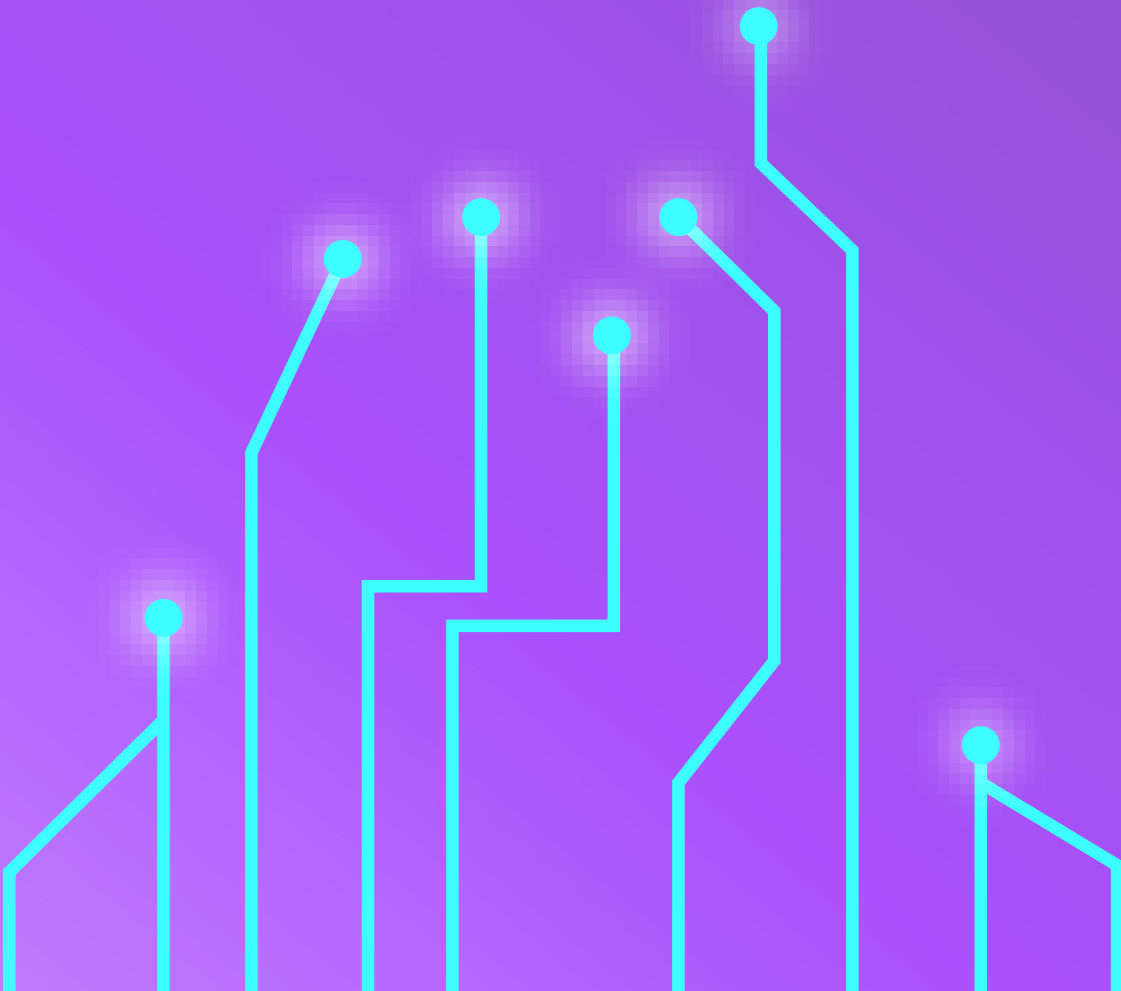
Prin implicarea părților interesate, dezvoltatorii de IA pot obține informații valoroase cu privire la nevoile, preferințele și preocupările acestora, promovând astfel transparența, responsabilitatea și luarea de decizii etice. De exemplu, în dezvoltarea vehiculelor autonome alimentate cu inteligență artificială, implicarea autorităților de reglementare și a societății asigură abordarea aspectelor legate de siguranță, confidențialitate și etică, sporind transparența și încrederea în tehnologie.



05. Concluzii

CU4 | Transparență

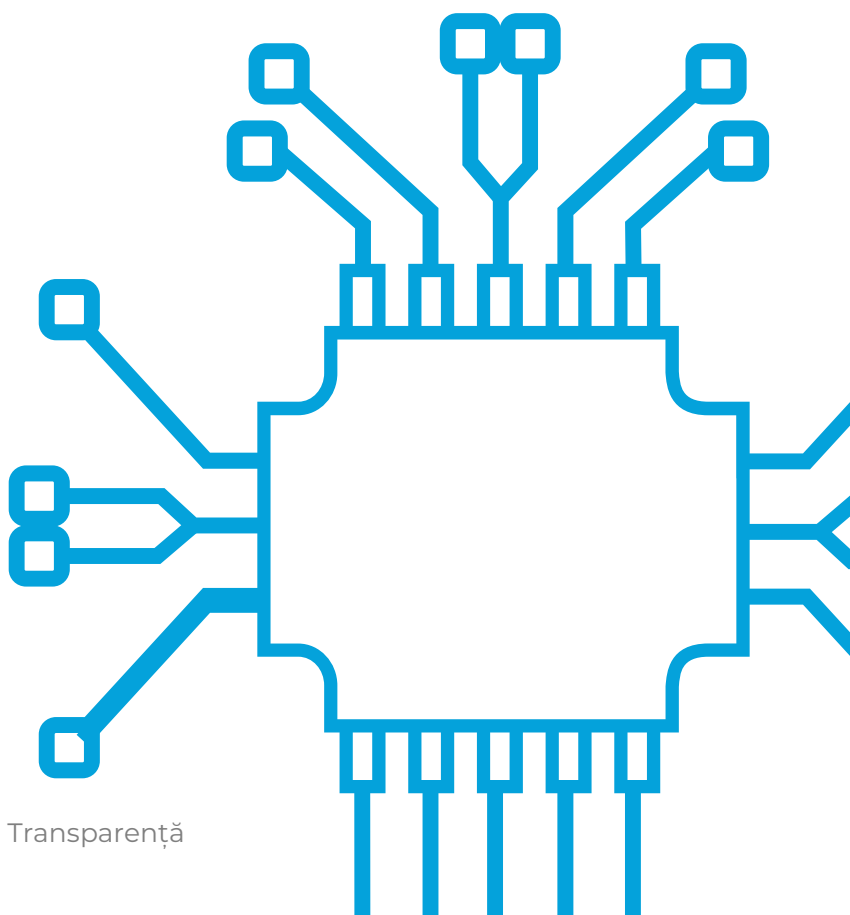


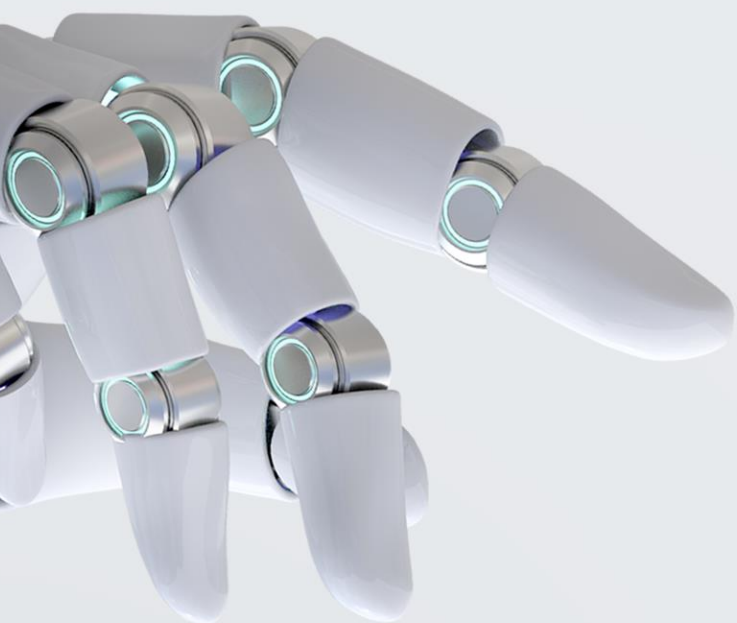


05. Concluzii

În concluzie, importanța transparenței în sistemele de inteligență artificială nu poate fi supraestimată, deoarece aceasta constituie fundamentul pentru construirea încrederii, responsabilizării și atenuării părtinirilor algoritmice. În plus, înțelegerea relației dintre transparență și părtinirea algoritmică evidențiază necesitatea de a aborda opacitatea ca mijloc de identificare, prevenire și atenuare a rezultatelor părtinitoare în sistemele AI.

În cele din urmă, explorarea strategiilor de promovare a transparenței oferă studenților instrumente practice pentru a spori responsabilitatea și a promova încrederea în tehnologiile IA. Printr-o înțelegere cuprinzătoare a acestor concepte, studenții sunt mai bine pregătiți să navigheze printre provocările etice ale dezvoltării și implementării IA, contribuind la avansarea unor sisteme IA responsabile și echitabile.







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



2022-1-ES01-KA220-HED-000085257