



# Microcredențial de IA etică

CARTEA

**CU2 | Non-maleficență**

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**

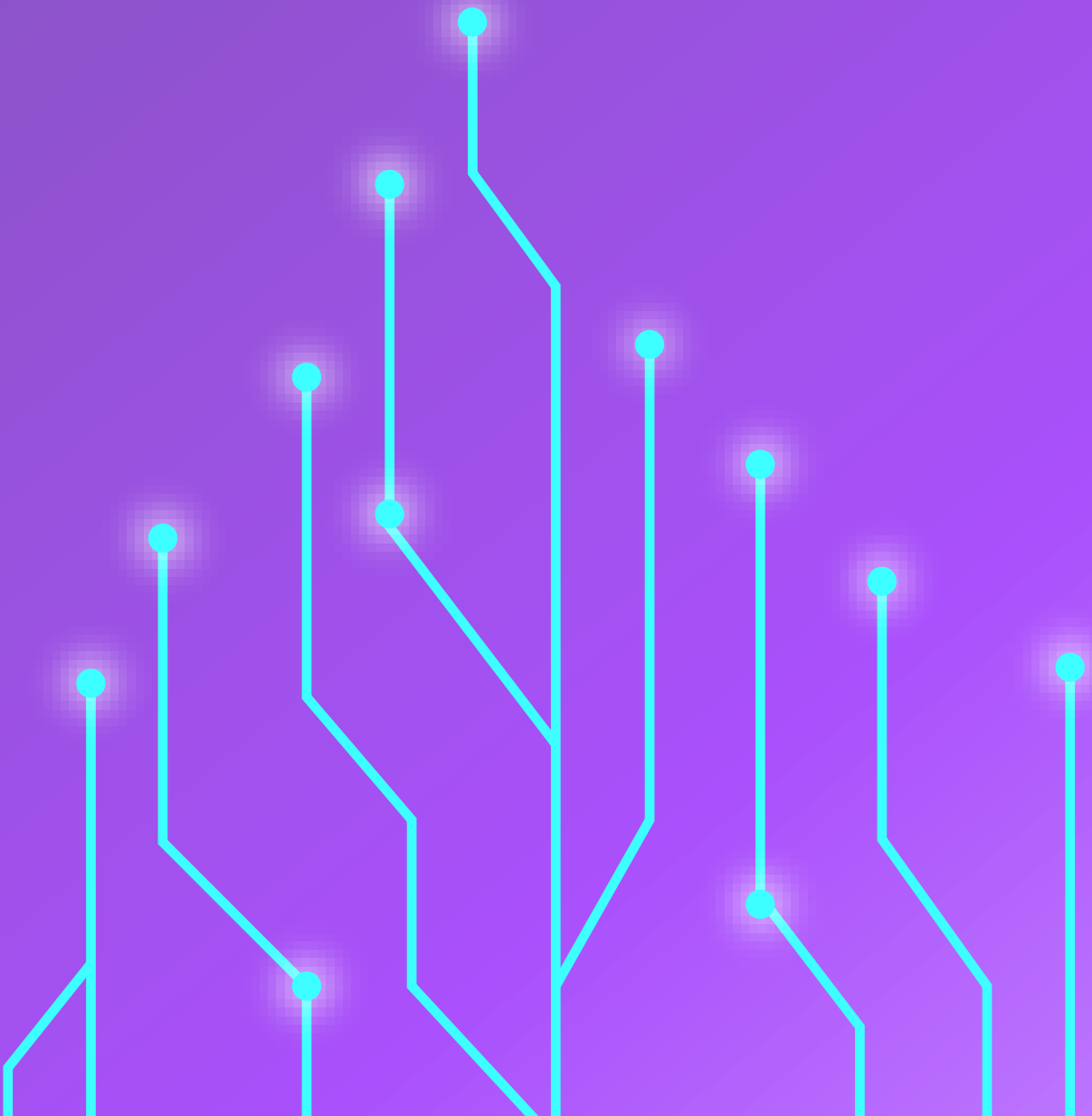
**02. Non-maleficență**

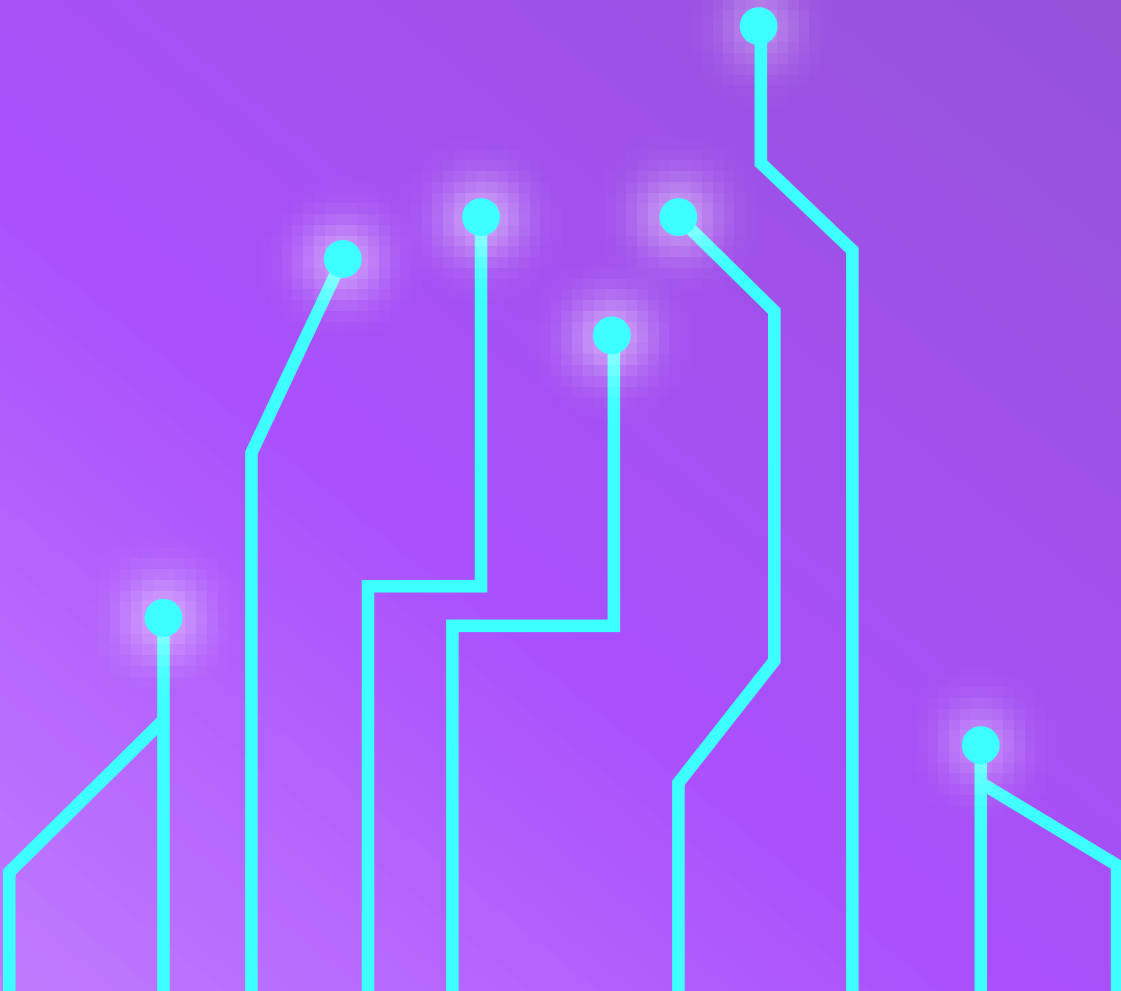
**03. Posibile prejudicii cauzate de  
inteligența artificială părtinitoare**

**04. Strategii pentru a face sistemele AI mai puțin  
dăunătoare**

# 01. Introducere

CU2 | Non-maleficență





## 01. Introducere

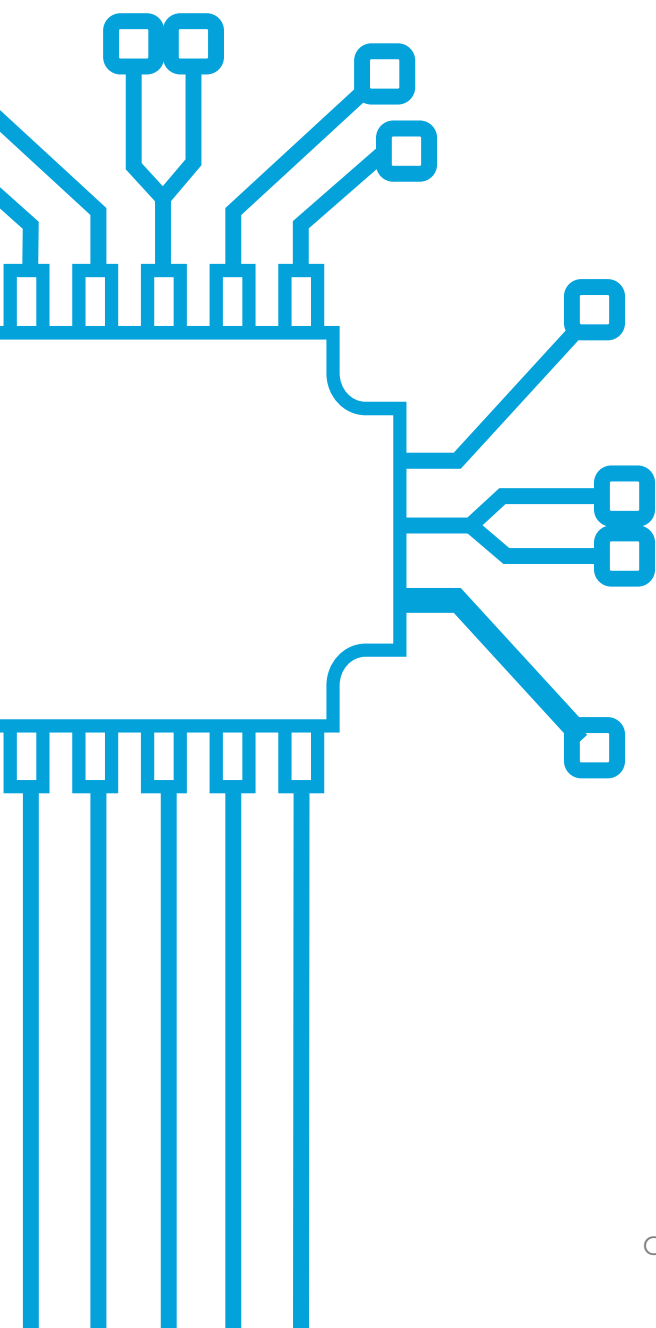
În cadrul acestei unități de competență, studenții vor dobândi cunoștințe fundamentale privind conceptul de non-maleficență în IA, responsabilitățile dezvoltatorilor și utilizatorilor de IA în asigurarea unor sisteme de IA etice cu daune minime și în recunoașterea implicațiilor din lumea reală, apreciind adoptarea și punerea în aplicare a mecanismelor care promovează responsabilitatea în sistemele de IA.

Rezultatele cunoștințelor pentru această unitate de competență includ:

- **Principiul non-maleficenței:** studenții prezintă conceptul de bază al non-maleficenței, subliniind importanța evitării daunelor în crearea și utilizarea sistemelor AI și modul în care această idee contribuie la dezvoltarea responsabilă a AI.
- **Posibile prejudicii cauzate de inteligența artificială părtinitoare:** elevii vor recunoaște diferitele moduri în care sistemele de inteligență artificială părtinitoare pot cauza prejudicii, cum ar fi discriminarea sau încălcarea vieții private, și vor folosi exemple din lumea reală pentru a ilustra importanța abordării părtinirii algoritmice.

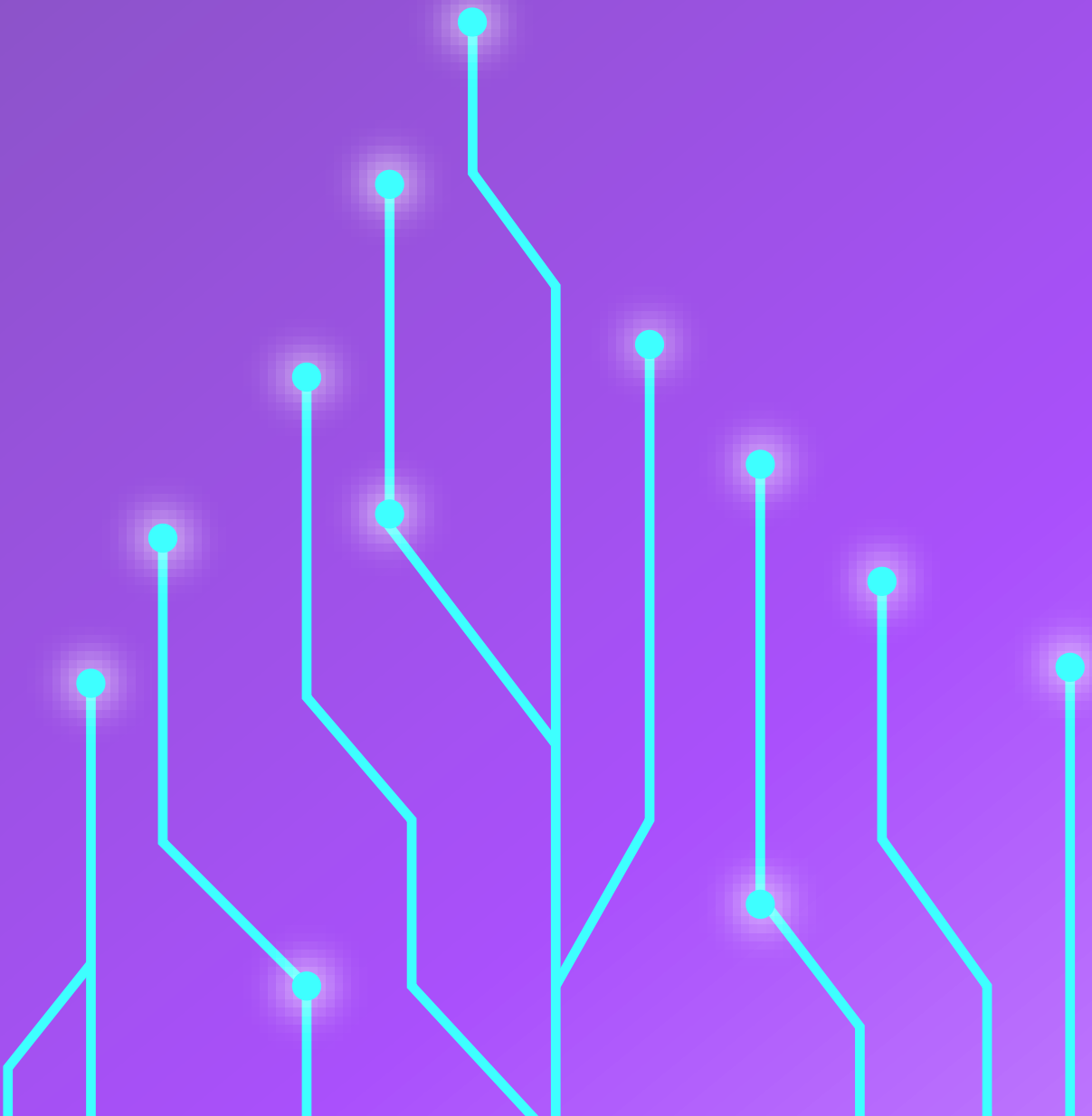


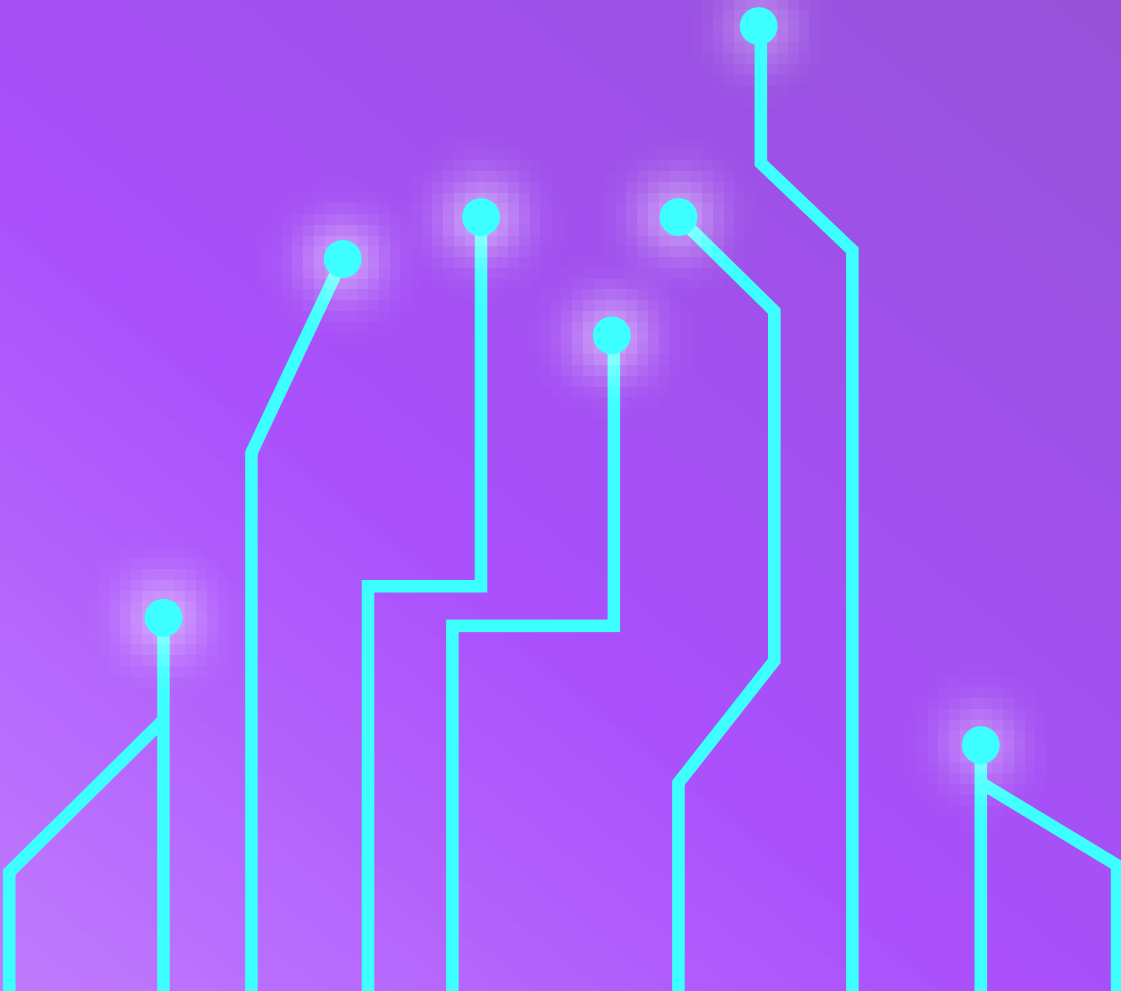
- **Strategii pentru a face sistemele AI mai puțin dăunătoare:** Elevii se vor familiariza cu strategii simple care pot face sistemele AI mai puțin dăunătoare, inclusiv promovarea echității, responsabilității și transparenței în dezvoltarea AI și încurajarea colaborării cu experți din diverse domenii.



# 02. Non-maleficență

CU2 | Non-maleficență





## 02. Non-maleficență

În această secțiune, vom prezenta principiul non-maleficenței și relevanța sa pentru tehnologiile IA și big data. Principiul non-maleficenței, adesea rezumat la "a nu face rău", este o piatră de temelie a procesului decizional etic în diverse domenii, inclusiv medicina, tehnologia și cercetarea. În contextul IA și al big data, non-maleficența subliniază importanța acordării de prioritate siguranței și bunăstării persoanelor și a societății în momentul dezvoltării și implementării acestor tehnologii.

### > Ce este non-maleficența?

Non-maleficența, derivată din expresia latină "*primum non nocere*" care înseamnă "în primul rând, să nu faci rău", este un principiu etic fundamental care îi ghidează pe profesioniști în prevenirea prejudiciilor aduse altora. Acesta pune accentul pe obligația morală de a evita provocarea de daune, fie ele fizice, psihologice sau societale, prin acțiunile sau deciziile proprii. În contextul IA și al big data, non-maleficența impune dezvoltatorilor, cercetătorilor și responsabililor politici să ia în considerare riscurile și consecințele potențiale ale tehnologiilor IA și să ia măsuri proactive pentru a preveni daunele.





## > De ce este importantă non-maleficența?

Non-maleficența este deosebit de importantă în domeniul IA și al big data datorită impactului semnificativ pe care aceste tehnologii îl pot avea asupra indivizilor și societății. Sistemele de inteligență artificială sunt utilizate din ce în ce mai mult în procesele decizionale esențiale, cum ar fi diagnosticarea în domeniul asistenței medicale, acordarea de împrumuturi financiare și pronunțarea sentințelor penale.

Asigurarea faptului că aceste sisteme acordă prioritate considerentelor etice și nu cauzează prejudicii este esențială pentru menținerea încrederii publice, prevenirea discriminării și susținerea valorilor societale precum echitatea și justiția.

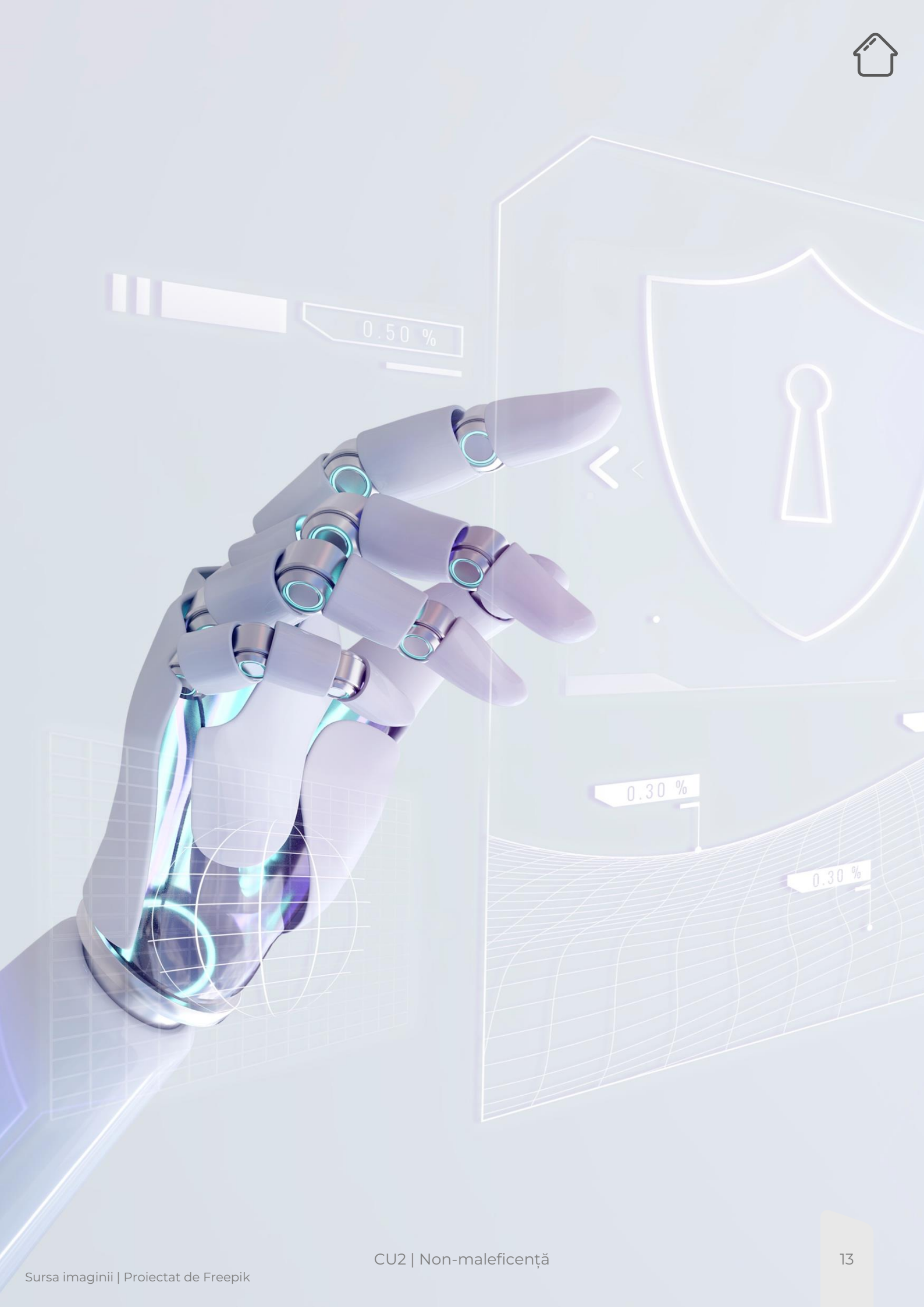
## > Principiile non-maleficenței

Non-maleficența impune persoanelor și organizațiilor implicate în dezvoltarea inteligenței artificiale să identifice și să atenueze în mod activ potențialele prejudicii pe care le pot cauza sistemele de inteligență artificială. Aceasta implică luarea în considerare nu numai a impactului imediat al tehnologiilor AI, ci și a consecințelor pe termen lung și a efectelor neintenționate ale acestora. Non-maleficența încurajează o abordare proactivă a eticii, în care dezvoltatorii anticipează și abordează riscurile potențiale înainte ca acestea să se materializeze.

## > Aplicație în dezvoltarea IA

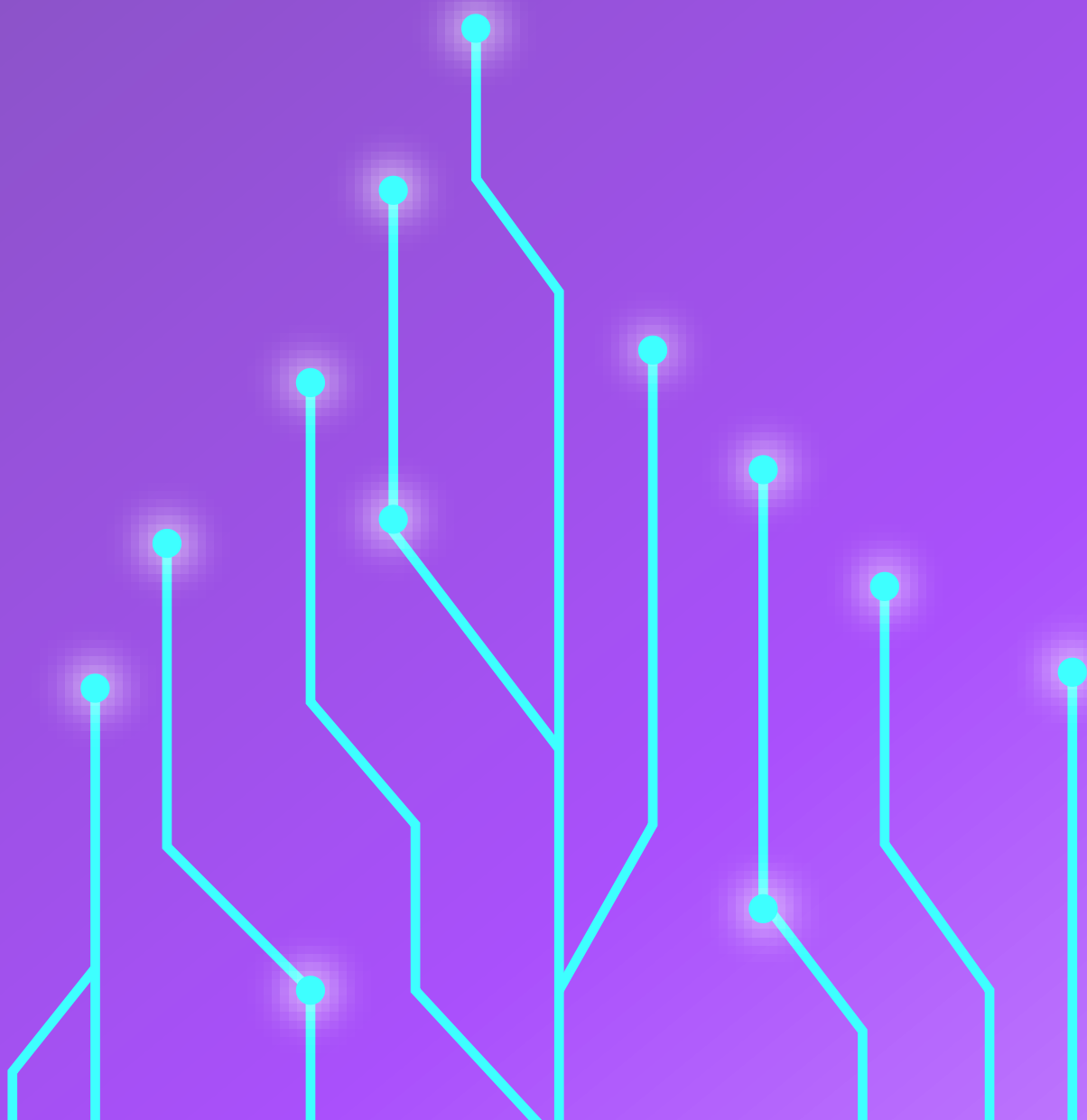
În contextul dezvoltării IA, non-maleficența se manifestă prin diverse practici menite să minimizeze daunele și să promoveze utilizarea etică. Acestea includ proceduri riguroase de testare și validare pentru a identifica și corecta prejudecățile din algoritmi AI, documentarea transparentă a proceselor decizionale ale sistemelor AI pentru a spori responsabilitatea, precum și monitorizarea și evaluarea continuă a implementărilor AI pentru a se asigura că acestea sunt conforme cu standardele etice și cu valorile societății.

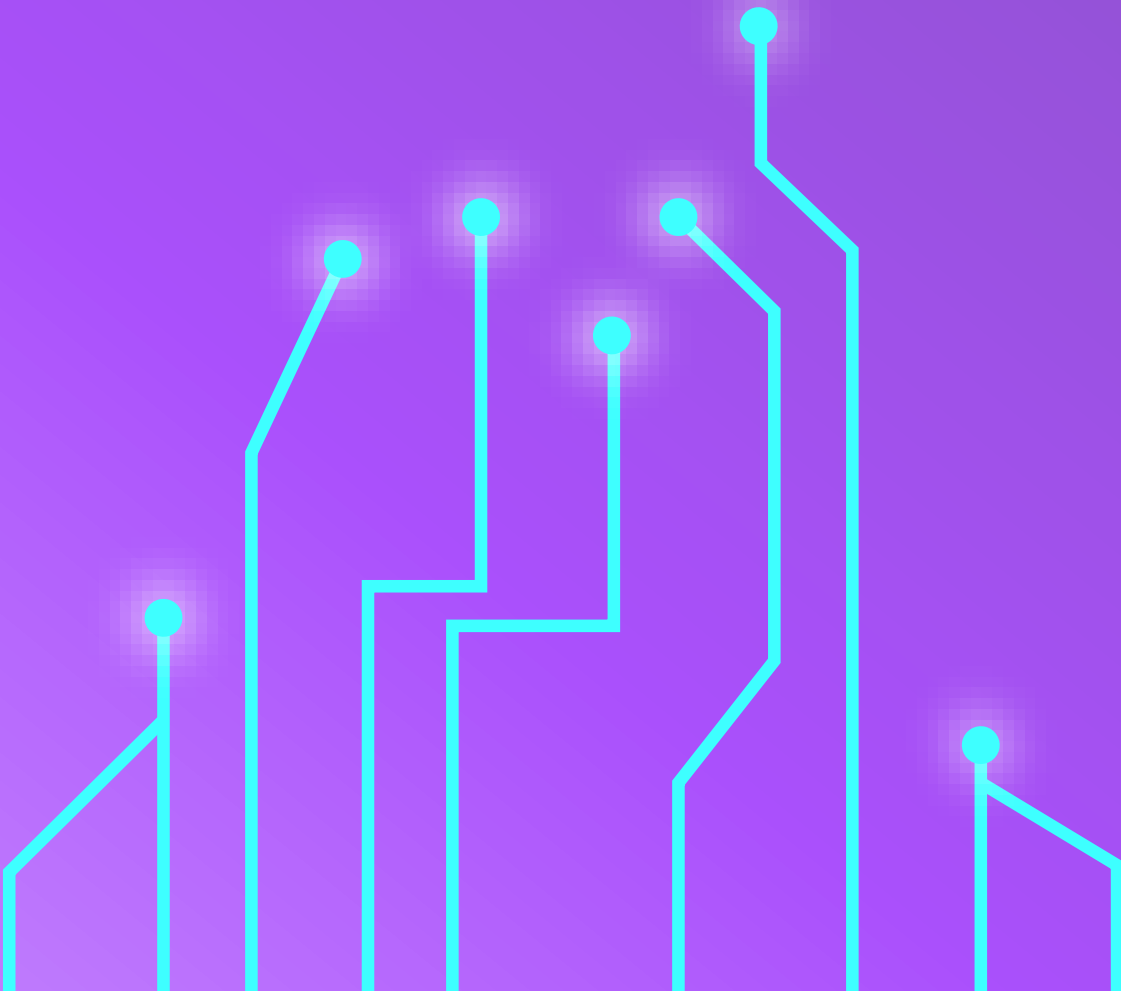




# 03. Posibile prejudicii cauzate de inteligența artificială părtinitoare

CU2 | Non-maleficență





### 03. Posibile prejucicii cauzate de inteligența artificială părtinitoare

În această secțiune, vom explora diferitele moduri în care sistemele AI părtinitoare pot cauza prejucicii, de la discriminare la încălcarea vieții private. Înțelegerea acestor prejucicii potențiale este esențială pentru a recunoaște importanța abordării prejudecăților algoritmice și a promovării unor practici responsabile de dezvoltare a IA.

#### > **Recunoașterea efectelor nocive**

Sistemele AI părtinitoare au potențialul de a perpetua și exacerba inegalitățile și nedreptățile existente în societate. Imaginați-vă o lume în care un algoritm vă refuză pe nedrept un împrumut din cauza codului poștal sau un sistem de recunoaștere facială vă identifică greșit ca infractor din cauza prejudecăților rasiale. Acestea sunt doar câteva dintre pericolele potențiale reprezentate de inteligența artificială părtinitoare. Mai jos, vom analiza zece dintre cele mai frecvente scenarii dăunătoare care pot apărea din cauza sistemelor AI părtinitoare.

- 1. Rezultate discriminatorii:** Algoritmii AI părtinitori pot conduce la rezultate discriminatorii, în care anumite persoane sau grupuri sunt tratate inechitabil pe baza unor caracteristici precum rasa, sexul sau statutul socioeconomic. Acest lucru poate duce la disparități în diverse domenii, inclusiv ocuparea forței de muncă, educație și justiție penală.

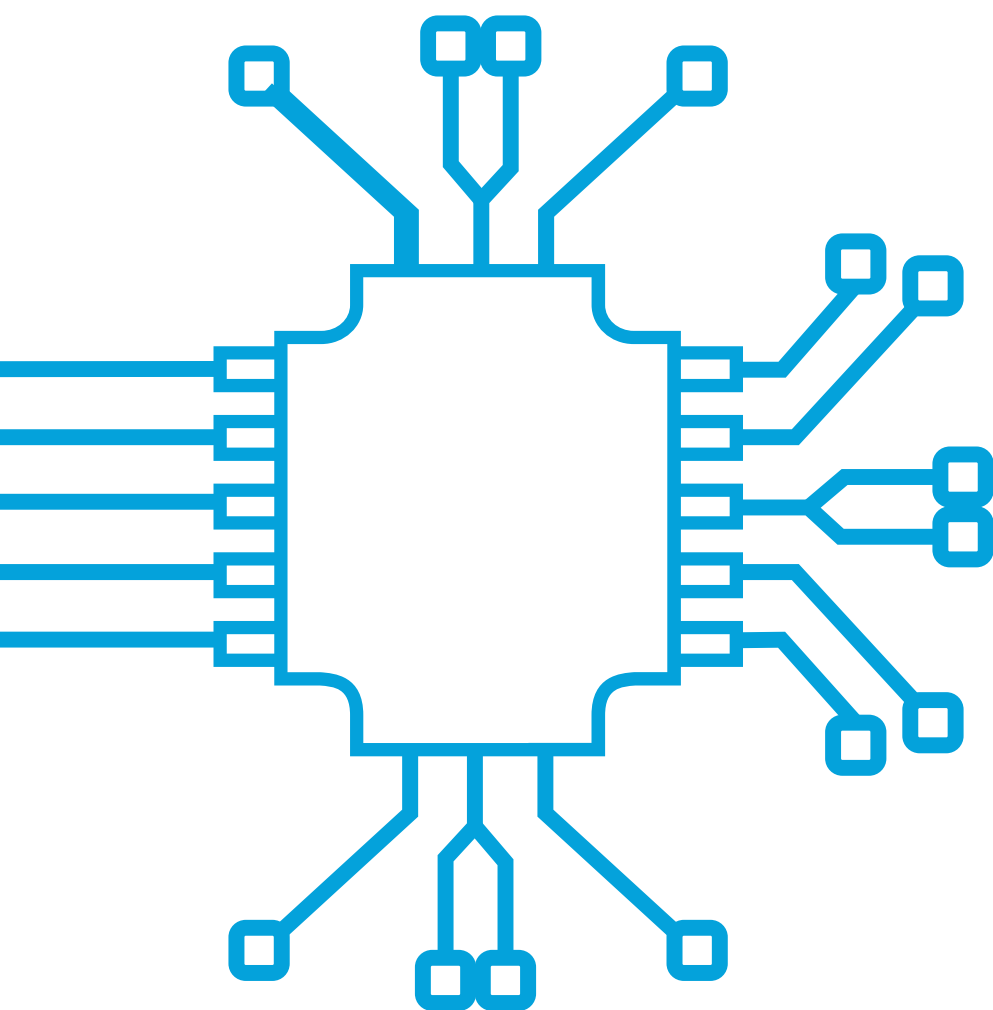


- 2. Încălcarea vieții private:** Sistemele AI tendențioase pot încălca dreptul la viață privată al persoanelor prin luarea de decizii bazate pe date personale sensibile fără consimțământul acestora. De exemplu, tehnologia de recunoaștere facială implementată în spațiile publice poate supune persoanele la supraveghere și urmărire nejustificate, ceea ce ridică probleme legate de încălcarea vieții private și a libertăților civile.
- 3. Consolidarea stereotipurilor:** Algoritmii AI părtinitori pot perpetua și consolida stereotipurile și prejudecățile dăunătoare prezente în societate. Acest lucru poate duce la marginalizarea și stigmatizarea anumitor grupuri, exacerband inegalitățile existente și inhibând progresul social.
- 4. Luarea de decizii inexacte:** Prejudecățile din datele de instruire sau algoritmii defectuoși pot duce la luarea unor decizii inexacte sau eronate de către sistemele AI. Acest lucru poate avea consecințe grave, în special în domenii critice precum diagnosticarea medicală, creditarea financiară și pronunțarea sentințelor penale, unde deciziile incorecte pot dăuna persoanelor și comunităților.
- 5. Lipsa responsabilității:** Sistemele AI părtinitoare pot fi lipsite de mecanisme de transparență și responsabilitate, ceea ce face dificilă identificarea și rectificarea cazurilor de părtinire. Acest lucru poate submina încrederea în tehnologiile AI și poate împiedica eforturile de a aborda în mod eficient părtinirea algoritmică.

- 6. Diversitate și incluziune limitate:** Algoritmii AI părtinitori pot perpetua inegalitățile existente prin favorizarea anumitor grupuri demografice în detrimentul altora. Acest lucru poate contribui la o lipsă de diversitate și incluziune în dezvoltarea și implementarea AI, limitând reprezentarea și perspectivele reflectate în sistemele AI și exacerband inegalitățile sociale.
- 7. Impactul negativ asupra inovării:** Algoritmii AI părtinitori pot împiedica inovarea și progresul prin perpetuarea practicilor învechite sau discriminatorii și prin limitarea oportunităților de creativitate și explorare. Abordarea prejudecăților în IA este esențială pentru promovarea unui mediu care încurajează diversitatea de gândire și promovează inovarea în beneficiul întregii societăți.
- 8. Pierderea încrederii și a încrederii:** Instinctele de părtinire în sistemele AI pot eroda încrederea publică în tehnologie și în capacitatea acesteia de a servi binele comun. Acest lucru poate duce la scepticism, rezistență și reticență în adoptarea soluțiilor AI, împiedicând potențialul acestora de a avea un impact pozitiv asupra societății.
- 9. Preocupări juridice și etice:** Sistemele AI părtinitoare pot ridica probleme juridice și etice legate de corectitudine, responsabilitate și transparență. Abordarea acestor preocupări necesită cadre de reglementare solide, orientări etice și practici responsabile de dezvoltare a IA, pentru a garanta că tehnologiile IA se aliniază valorilor societale și respectă drepturile fundamentale.



**10. Implicații sociale și economice:** Impactul omniprezent al AI părtinitoare se extinde dincolo de cazurile individuale de discriminare și are implicații sociale și economice mai ample. Sistemele de inteligență artificială părtinitoare pot exacerba inegalitățile existente, pot mări decalajul digital și pot perpetua nedreptățile sociale, reprezentând provocări semnificative pentru construirea unei societăți corecte și echitabile.



## > Exemple din lumea reală

Folosind exemple din lumea reală, vom ilustra importanța abordării prejudecăților algoritmice și impactul potențial al acestora asupra persoanelor și comunităților. Aceste exemple vor evidenția cazuri în care sistemele AI părtinitoare au condus la consecințe dăunătoare, cum ar fi arestări eronate, tratament inechitabil în deciziile de angajare sau creditare și perpetuarea stereotipurilor și prejudecăților.

- **EXEMPLU #1 - Algoritmul Amazon a discriminat femeile**

Instrumentul de recrutare AI al Amazon a avut ca scop găsirea celor mai bune talente în domeniul tehnologiei, dar a ajuns să filtreze femeile. De ce? Algoritmul, antrenat pe baza CV-urilor anterioare (majoritatea ale bărbaților), a favorizat cuvintele-cheie utilizate de bărbați și le-a penalizat pe cele asociate cu femeile. Acest lucru evidențiază o provocare majoră a inteligenței artificiale: datele părtinitoare conduc la algoritmi părtinitori. La fel ca un student care se bazează pe manuale greșite, inteligența artificială moștenește prejudecățile din datele sale de formare. Citiți mai multe în:

<https://www.reuters.com/article/idUSKCN1MK0AG/>





- **EXEMPLUL nr. 2 - Prejudecăți rasiale algoritmice în predicția ratei de recidivă penală**

Imaginați-vă un instrument care prezice cine comite infracțiuni. În SUA, COMPAS face exact acest lucru, dar cu o întorsătură rasială. Studiile arată că inculpații de culoare sunt etichetați cu risc ridicat mult mai des decât inculpații albi cu antecedente similare. De ce această părtinire? COMPAS reflectă inegalitățile sociale deja prezente în datele privind arestările. Aceste prejudecăți fac ca persoanele să fie reținute înainte de proces sau să primească sentințe mai aspre, ceea ce afectează în mod nedrept persoanele de culoare. Cazul COMPAS subliniază necesitatea unor verificări atente ale inteligenței artificiale utilizate în sistemele de justiție pentru a asigura echitatea pentru toți. Citiți mai multe în: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

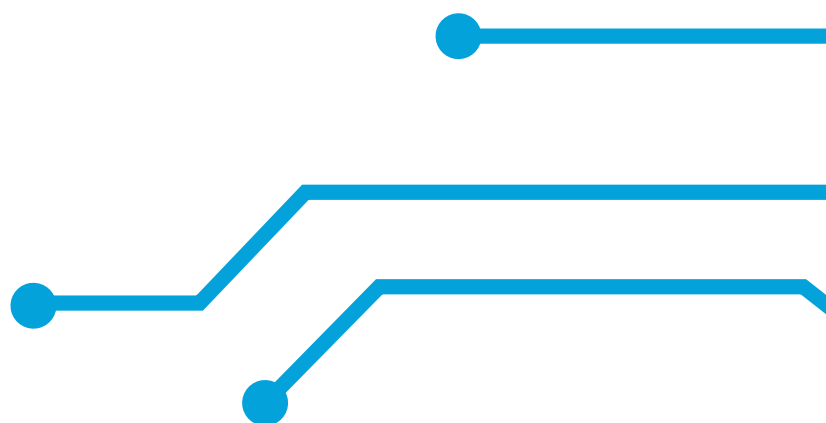


- **EXEMPLU #3 - Algoritmul de asistență medicală din SUA a subestimat nevoile pacienților de culoare**

Luați în considerare un sistem de sănătate care favorizează pacienții care cheltuiesc mai mult. Din păcate, acest lucru a afectat pacienții de culoare din SUA. Un algoritm conceput pentru a-i identifica pe cei care au nevoie de îngrijiri suplimentare a omis mulți pacienți de culoare din cauza prejudecăților. De ce? Sistemul s-a bazat pe datele privind cheltuielile medicale anterioare, care nu reflectă accesul limitat al pacienților de culoare la îngrijiri preventive, din cauza disparităților economice. Astfel, pacienții de culoare au fost clasificați ca fiind mai sănătoși și nu au beneficiat de îngrijiri critice. Repararea algoritmului ar putea ajuta mult mai mulți pacienți de culoare. Acest caz evidențiază necesitatea unei inteligențe artificiale corecte în domeniul asistenței medicale pentru a se asigura că toată lumea primește tratamentul de care are nevoie.

Citiți mai multe în:

<https://www.theguardian.com/society/2019/oct/25/healthcare-algorithm-racial-biases-optum>

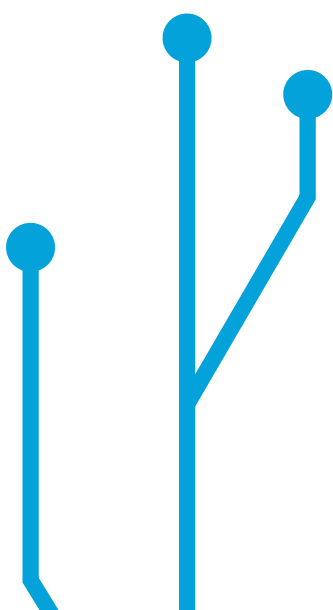




- **EXEMPLUL #4 - ChatBot a transmis mesaje discriminatorii**

Chatbotul Tay de la Microsoft a fost conceput pentru a învăța din conversațiile ocazionale. Lansat pe Twitter, acesta a început rapid să lanseze mesaje rasiste și ofensatoare. De ce? Pentru că "trolii" îl bombardau pe Tay cu conținut instigator la ură, pe care acesta îl absorbea și îl imita. Acest incident evidențiază o provocare majoră a interacțiunii IA cu lumea reală. Social media poate fi un loc toxic, iar inteligența artificială expusă la aceasta poate învăța negativitatea. Tay este un avertisment: proiectarea AI pentru interacțiunea online necesită luarea în considerare a contextului social și a potențialului de utilizare abuzivă.

Citiți mai multe în: <https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>



- **EXEMPLUL #5 - Sistem de recunoaștere facială distorsionat**

Imaginați-vă că vă sărbătoriți ziua de naștere cu o excursie la cumpărături, doar pentru a fi acuzat de furt de către un sistem de recunoaștere facială! Acest lucru i s-a întâmplat unei femei Māori din Noua Zeelandă. Tehnologia, concepută pentru a prinde hoții din magazine, a identificat-o în mod eronat și i-a provocat o suferință semnificativă. Acest caz scoate în evidență pericolele recunoașterii faciale părtinitoare. Studiile arată că aceste sisteme pot identifica greșit persoanele, în special femeile și persoanele de culoare. Pe măsură ce tehnologia de recunoaștere facială devine tot mai răspândită, este esențial să se asigure corectitudinea și să se prevină astfel de incidente. Citiți mai multe în:

<https://www.1news.co.nz/2024/04/22/rotorua-mother-wrongly-identified-by-supermarket-as-a-thief/>

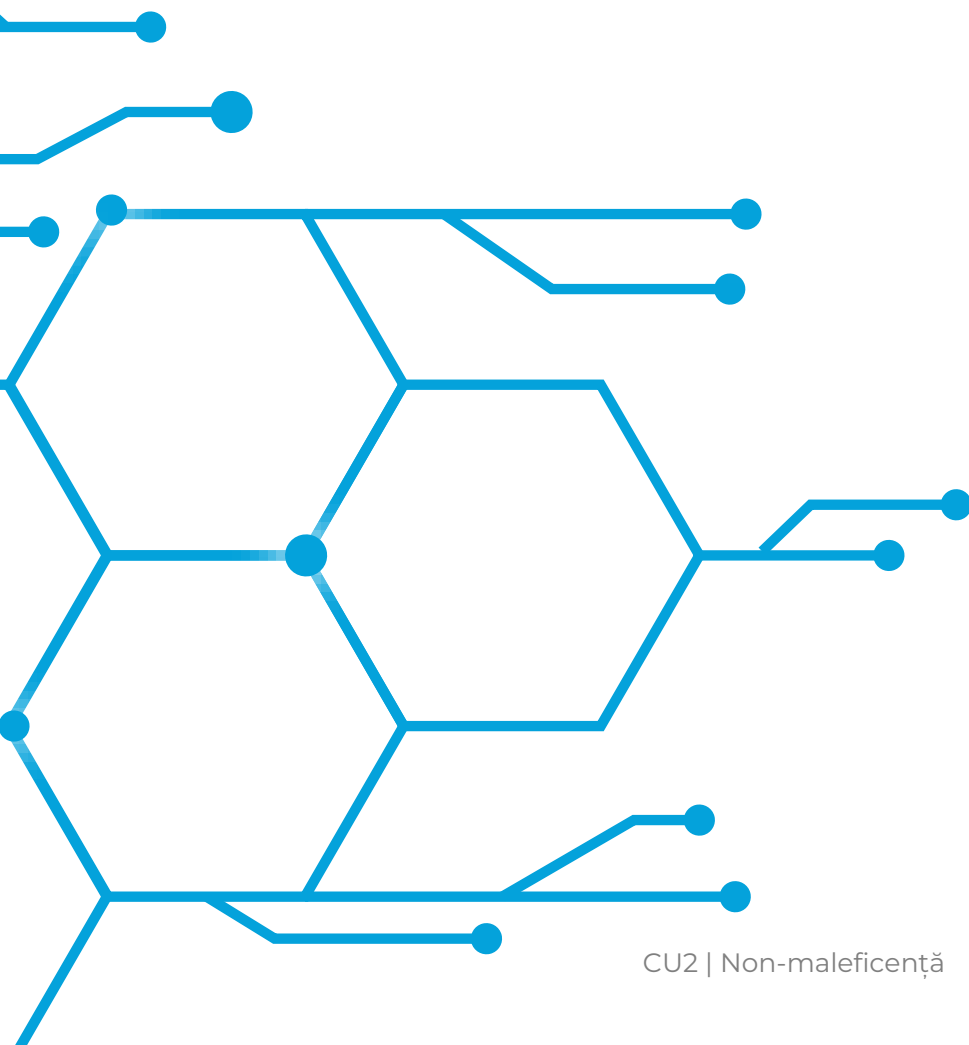




- **EXEMPLUL nr. 6 - AI generatoare de text care fabrică fapte**

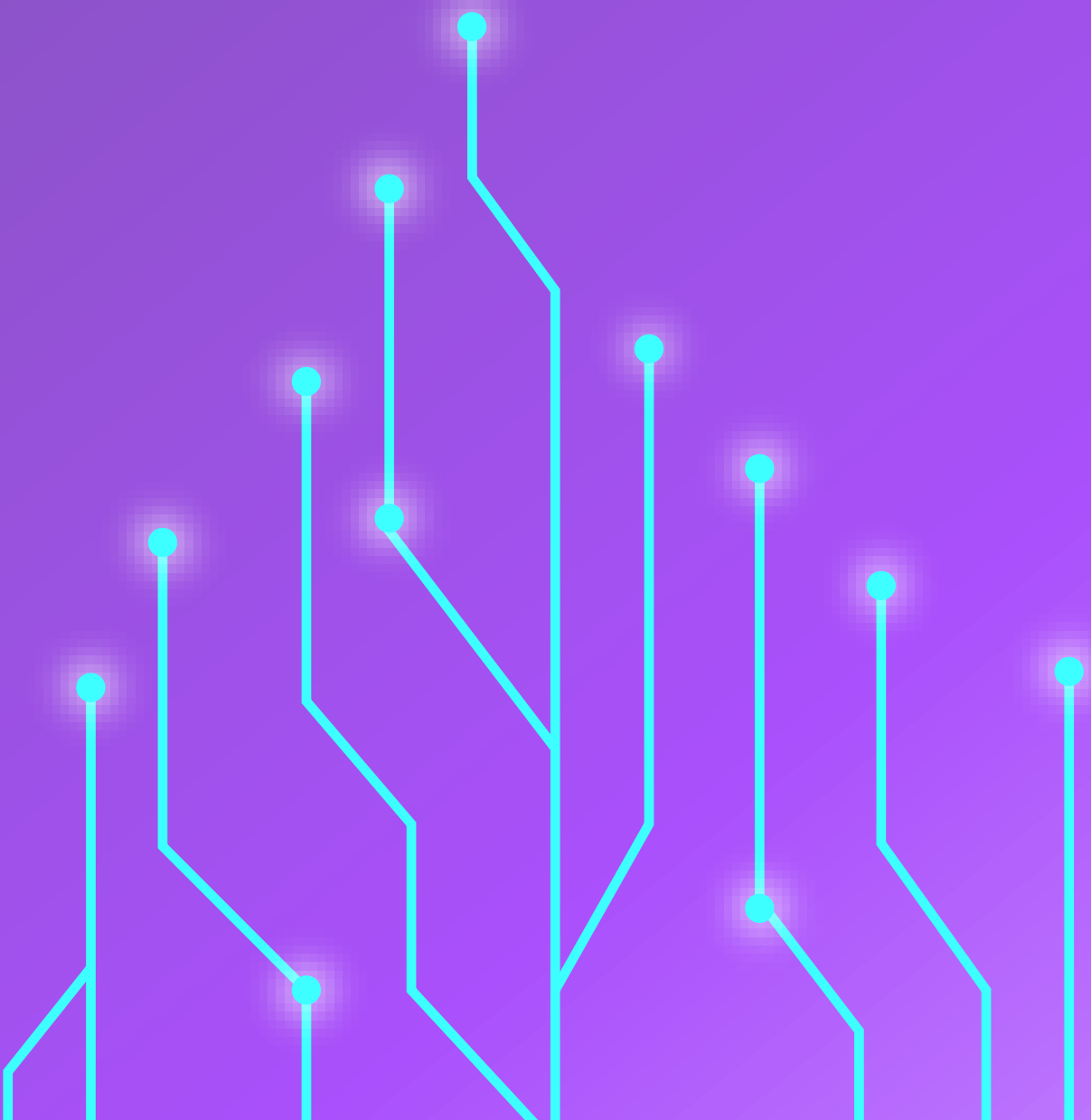
Reputația unui profesor de drept a fost pătată de un chatbot AI. ChatGPT a fabricat o plângere de hărțuire sexuală împotriva sa, completată cu un articol de știri fals. Acest caz expune un risc major al IA: generarea de dezinformări dăunătoare. Profesorul s-a confruntat cu daune reputaționale, în ciuda faptului că minciuna a fost dezvăluită. Pe măsură ce inteligența artificială devine din ce în ce mai frecventă, asigurarea unor informații factuale și stabilirea responsabilității pentru falsurile generate de inteligența artificială sunt aspecte esențiale. Citiți mai multe în:

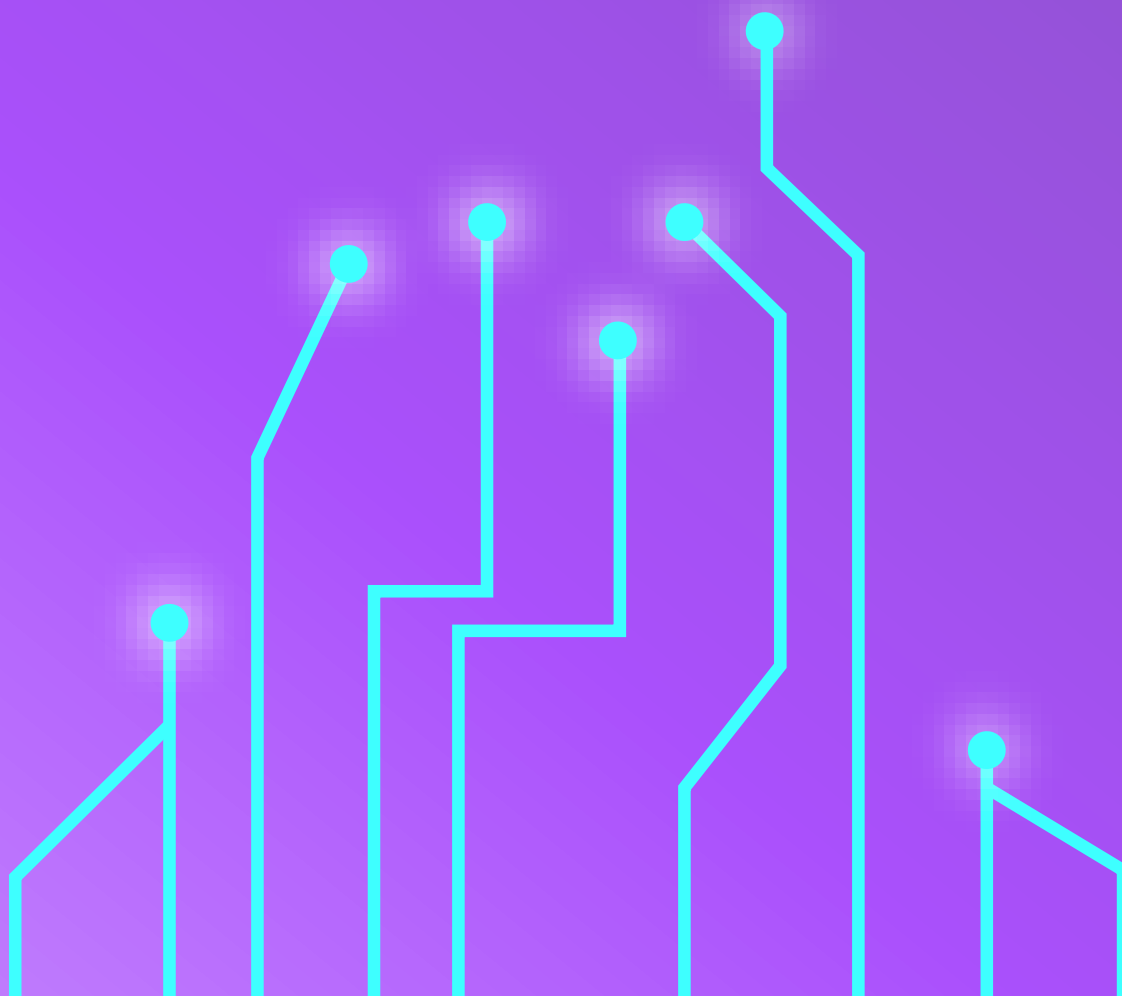
<https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/>



# 04. Strategii pentru a face sistemele AI mai puțin dăunătoare

CU2 | Non-maleficență





## 04. Strategii pentru a face sistemele AI mai puțin dăunătoare

În această secțiune, vom prezenta strategii menite să facă sistemele AI mai puțin dăunătoare prin promovarea echității, responsabilității și transparenței în dezvoltarea și implementarea acestora. Aceste strategii permit dezvoltatorilor, responsabililor politici și părților interesate să abordeze în mod proactiv prejudecățile algoritmice și să atenueze potențialele consecințe negative ale acestora.

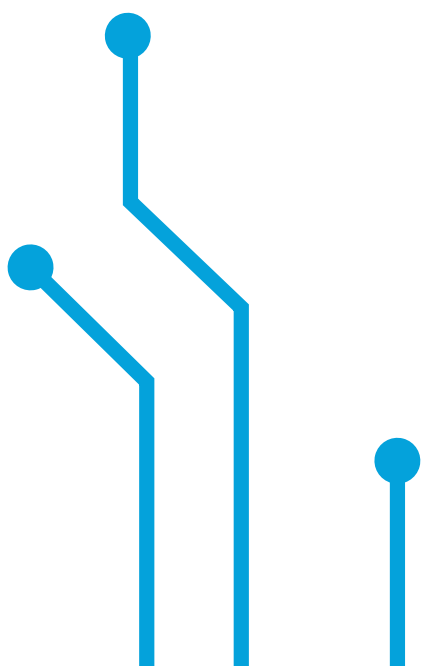
### > **Promovarea echității**

O strategie-cheie pentru atenuarea efectelor nocive ale sistemelor AI părtinitoare este promovarea echității în procesele decizionale algoritmice. Aceasta implică asigurarea faptului că modelele AI sunt antrenate pe seturi de date diverse și reprezentative, fără prejudecăți discriminatorii. În plus, pot fi utilizate tehnici de învățare automată care țin seama de echitate pentru a identifica și a atenua prejudecățile din predicțiile algoritmice, promovând astfel rezultate echitabile pentru toate persoanele.



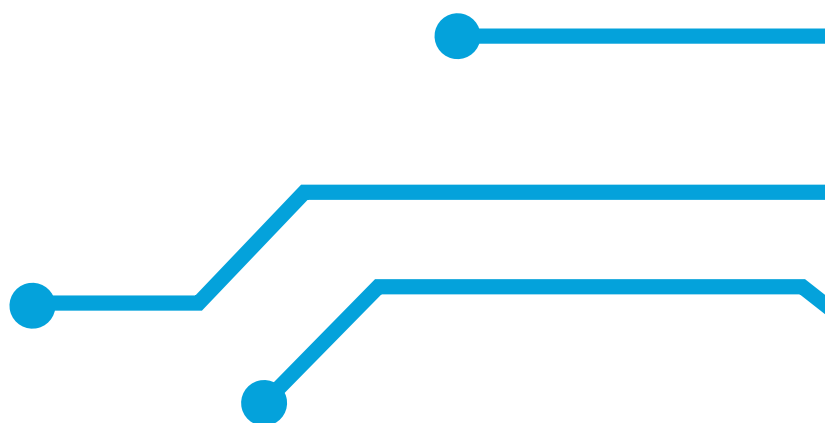
## > **Sporirea responsabilității**

Un alt aspect important al reducerii daunelor provocate de sistemele AI părtinitoare este sporirea responsabilității în rândul dezvoltatorilor, organizațiilor și factorilor de decizie politică. Aceasta include punerea în aplicare a orientărilor etice și a celor mai bune practici pentru dezvoltarea inteligenței artificiale, cum ar fi realizarea unor evaluări de impact aprofundate pentru a identifica riscurile și prejudiciile potențiale. În plus, stabilirea unor mecanisme clare de responsabilizare și a unor cadre de supraveghere poate contribui la responsabilizarea persoanelor și organizațiilor cu privire la implicațiile etice ale implementării AI. În următoarea unitate a acestui curs, vom explora mai în detaliu conceptul de responsabilitate.



## > Încurajarea transparenței

Transparența este esențială pentru ca sistemele de inteligență artificială să fie mai puțin dăunătoare, prin promovarea responsabilității și a încrederii între părțile interesate. Documentarea transparentă a algoritmilor AI și a proceselor decizionale permite examinarea și validarea externă, asigurând identificarea și soluționarea în timp util a prejudecăților și erorilor. În plus, încurajarea dialogului deschis și a colaborării între dezvoltatorii de inteligență artificială, cercetători și comunitățile afectate poate facilita o mai mare transparență și înțelegere a implicațiilor etice ale tehnologiilor de inteligență artificială. Unitatea 4 a acestui curs va aprofunda conceptul de transparență, deoarece acesta este unul dintre cele mai fundamentale aspecte pentru asigurarea unei IA responsabile.





## > **Asigurarea confidențialității**

Sistemele de inteligență artificială sunt instrumente puternice, dar confortul lor nu ar trebui să se facă în detrimentul vieții private. Această strategie se concentrează pe protejarea informațiilor dvs. personale. Dezvoltatorii ar trebui să colecteze și să utilizeze cât mai puține date posibil, în special detalii sensibile. Măsurile de securitate trebuie să fie de top pentru a păstra informațiile în siguranță. Sistemele AI ar trebui, de asemenea, construite pentru a respecta legile și reglementările privind confidențialitatea, inclusiv Regulamentul general privind protecția datelor (GDPR) din Europa, care acordă persoanelor fizice un control semnificativ asupra datelor lor personale.

## > **Prioritizarea siguranței**

Când vine vorba de inteligența artificială, siguranța ar trebui să fie prioritatea absolută. Aceasta înseamnă ca sistemele AI să fie supuse unor procese riguroase de testare și validare înainte de a fi lansate în lumea reală. Scopul este de a identifica și remedia orice riscuri sau probleme potențiale care ar putea provoca daune. Prin asigurarea funcționării fiabile și sigure a sistemelor AI, putem proteja persoanele și societatea în ansamblu.



# Charlie



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



ENGAGING PEOPLE



INNOVATION TRAINING CENTER



AARHUS UNIVERSITY



VAMK UNIVERSITY OF APPLIED SCIENCES

helixconnect



2022-1-ES01-KA220-HED-000085257