



**Material tematic sesiune clasică/ online de coaching/ consiliere profesională/ orientare în carieră,
individuală/ personalizată și/ sau de grup**

„DreptCompAct- COMPetențe ACTuale pentru studenți la DREPT”, Cod proiect: 312957

Competențe transversale în domeniul juridic. Analiza critică (XX)

Biasul Inteligenței Artificiale și stereotipurile culturale în conexiunea cu gândirea critică (J.5)

Pe măsură ce analizăm bias-ul algoritmic al inteligenței artificiale, constatăm că orice invenție, descoperire, creație nouă creează noi provocări și necesită înțelegerea mecanismului de funcționare și gestionarea efectelor asupra ceea ce deja funcționează, pentru a încerca optimizarea efectelor pozitive și a minimiza eventualele efecte negative.

Din perspectiva gândirii critice, este esențial să analizăm și alte clasificări ale bias-ului, amintindu-ne și definiția acestuia; dar și strategiile de atenuare a prejudecăților care afectează gândirea și influențează deciziile fiecăruia dintre noi. În contextul profesiilor juridice, eliminarea prejudecăților în analiza de către judecător a probelor administrate în dosar, este esențială, dar greu de realizat în practică!

Biasul de IA apare atunci când modelele de IA produc rezultate discriminatorii împotriva anumitor demografii. Există mai multe tipuri de biasuri care pot pătrunde în sistemele de IA și produce rezultate inexacte. Alte biasuri de IA, decât cele prezentate în materialele anterioare¹ sunt:²

Biasul Stereotipic care se referă la fenomenul în care rezultatele unui model de IA constau în stereotipuri sau noțiuni percepute despre o anumită demografie.

Biasul Rasial în IA apare atunci când rezultatul unui model de IA este discriminatoriu și injust față de o persoană sau un grup pe baza etniei sau rasei.

¹ A se vedea Competențe transversale în domeniul juridic. Analiza critică (XVIII), Biasul algoritmic și gândirea critică (J.3)

² Ce este Biasul de IA? Biasul Inteligenței Artificiale și Stereotipurile Culturale: Efecte, Limitări și Mitigare, Published October 4, 2023- Updated May 22, 2026, <https://www.unite.ai/ro/ai-bias-cultural-stereotypes-effects-limitations-mitigation/>



Biasul Cultural care intervine atunci când rezultatele unui model de IA favorizează o anumită cultură în detrimentul alteia.

În afară de biasuri, alte probleme pot împiedica rezultatele unui sistem de IA, cum ar fi inexactități și halucinații. **Inexactitățile** apar atunci când rezultatele produse de un model de IA sunt inexacte din cauza datelor de antrenare inconsistente. **Halucațiile** apar atunci când modelele de IA produc *rezultate fictive și false care nu se bazează pe date factuale*.³

Câteva strategii de atenuare a prejudecății sunt deja formulate în studii de specialitate:

1. Strategii organizaționale și culturale

Educația și conștientizarea: Educarea celor implicați, atât a angajatorilor, dezvoltatorilor, angajaților, dar și a utilizatorilor cu privire la riscurile IA este crucială. Educația implică creșterea gradului de conștientizare a potențialelor prejudecăți și identificarea unor modalități de a le atenua. Aceasta promovează o cultură a înțelegerii, a managementului proactiv al riscului, dar mai ales al nevoii de formarea unor deprinderi, aptitudini, competențe care să elimine stereotipuri și prejudecăți din educația deja însușită de tinerii și adulții care sunt provocați să se adapteze la cerințe noi în societatea digitalizată; pentru a contracara manipularea informațională care transformă adultul dependent de tehnologie la epocii actuale.

Standarde etice și guvernare: Stabilirea standardelor etice și a cadrelor de guvernare asigură că organizațiile iau în considerare reglementările emergente și riscurile potențiale. Această abordare proactivă ajută la crearea unei culturi axate pe dezvoltarea și implementarea etică a inteligenței artificiale. Aspectele etice fac parte din natura interdisciplinară a tipurilor de bias prezentate mai sus.

Diversitatea gândirii: Integrarea diverselor perspective în toate etapele - dezvoltare, testare și implementare - ajută la identificarea și abordarea potențialelor daune. O echipă diversă poate oferi o gamă mai largă de perspective asupra potențialelor prejudecăți.

2. Transparența și responsabilitatea sunt necesare în eficientizarea utilizării diferitelor tipuri de bias-uri care să servească intereselor speciei umane de progres și de trai în armonie și bună înțelegere.

³ Idem 2



Transparența presupune furnizarea de informații clare despre pașii făcuți pentru identificarea și reducerea riscurilor și ajută părțile interesate să înțeleagă angajamentul organizației față de corectitudine, responsabilitate, implicarea activă în realizarea scopurilor angajatorului, în funcție de obiectul de activitate.

Auditul algoritmului implică efectuarea auditului independent, de către terți, care să evalueze inclusiv sistemele AI în materia prejudecăților, având posibilitatea evaluării inclusive a structurilor de guvernare, ceea ce va asigura transparența. Auditul are rolul de a oferi părților interesate acele asigurări, că inclusiv bias-urile IA sunt verificate, cu propuneri de corectare, poate chiar înlocuire, după caz, dacă încalcă reglementările, regulile etice, drepturile și libertățile fundamentale umane, dar și drepturile și obligațiile părților implicate.

3. Testarea și monitorizarea este un proces continuu, constă în implementarea metricilor pentru bias, acuratețe și alte aspecte tehnice care permit îmbunătățirea continuă. Testarea regulată cu seturi de date reprezentative ajută la identificarea zonelor care necesită îmbunătățiri și asigură un sistem AI mai echitabil.

Evaluarea riscului și impactului AI sunt evaluări care identifică riscuri potențiale în sistemele socio-tehnice, dar și părțile interesate relevante care ar putea fi afectate, ajutând la atenuarea daunelor.

4. Explicabilitatea și Reproducibilitatea se referă la rezultatele explicabile și reproductibile. Astfel, prin aceste procese se asigură faptul că sistemele AI oferă rezultate de înțeles și reproductibile, ajutând organizațiile să ia în considerare și să testeze diverse riscuri, sporind încrederea și responsabilitatea.

5. Tehnicile de eliminare a părtinirii implică tehnici de pre-procesare, tehnici de procesare în procesare, tehnici de post-procesare:

5.1. Tehnici de pre-procesare:

Reponderarea datelor de antrenament ajustează ponderile datelor pentru a asigura o învățare echilibrată între grupuri.

Reeșantionarea datelor echilibrează seturile de date prin tehnici precum supraeșantionarea claselor minoritare.

Tehnica de supraeșantionare a minorităților sintetice (PĂPUȘĂ) generează date sintetice pentru a aborda dezechilibrul de clasă.



Selecția/modificarea caracteristicilor elimină caracteristicile părtinitoare sau creează altele noi pentru a reduce părtinirea.

5.2. Tehnici de procesare în procesare:

Regularizarea limitează învățarea pentru a reduce sensibilitatea la părtinire. (predecesorul prejudecății)

Învățare sensibilă la cost atribuie diferite costuri de clasificare greșită pentru a reduce părtinirea în predicții.

Debiasing adversarial antrenează modelele pentru a face predicții corecte în timp ce adversarul testează prejudecățile.

5.3. Tehnici de post-procesare:

Ajustarea pragului are ca obiect ajustarea pragurilor decizionale pentru a egaliza ratele de eroare între grupuri.

Calibrarea aliniază predicțiile modelului cu rezultatele reale, reducând părtinirea.

6. Abordările de echitate acoperă principiile de bază ale eticii.

Corectitudinea prin conștientizare include explicit atribute sensibile pentru a corecta biasurile setului de date.

Corectitudinea prin lipsa de conștientizare ignoră atributele sensibile, deși acest lucru poate să nu fie eficient, dacă alte caracteristici corelează cu ele.

”Biasul algoritmic este deci, o problemă complexă care reflectă prejudecățile inerente luării deciziilor umane. Pe măsură ce sistemele AI influențează tot mai mult diverse aspecte ale vieții noastre, este esențial să recunoaștem că aceste tehnologii nu sunt în mod inherent neutre. Ele pot perpetua și chiar amplifica prejudecățile existente, dacă nu sunt gestionate cu grijă. Abordarea biasării algoritmice necesită o abordare holistică, care combină strategii organizaționale, transparență, testare continuă și proceduri tehnice de debiasing. Prin promovarea unei culturi a diversității, guvernantei etice și responsabilității, organizațiile pot lucra pentru a crea sisteme AI echitabile.



În cele din urmă, responsabilitatea revine tuturor părților interesate - dezvoltatori, utilizatori și factori de decizie - de a se asigura că tehnologiile AI servesc scopul mai larg al justiției sociale și nu întăresc inegalitățile existente.”

Biasul algoritmic apare când sistemele AI învață din date părtinitoare. A afla cum să identifice și să corecteze aceste distorsiuni în aplicațiile tale, devine obligatoriu pentru a asigura acuratețea informației furnizate de AI.

În același timp, prejudecățile în materie de inteligență artificială este numită ”distorsiunea AI” și se referă la discriminarea sistematică încorporată în sistemele AI, care pot consolida prejudecățile existente și pot amplifica discriminarea, prejudecățile și stereotipurile.⁴

Limitările Modelelor de IA⁵

Deși, IA are potențialul de a revoluționa multe industrii, totuși, dacă scenariile precum cele menționate anterior se multiplică, poate genera o scădere a adopției generale a IA, ceea ce poate duce la oportunități ratate. Asemenea cazuri apar de obicei din cauza limitărilor semnificative ale sistemelor de IA, cum ar fi:

Lipsa creativității deoarece IA poate lua decizii doar pe baza datelor de antrenare furnizate, îi lipsește creativitatea de a gândi în afara cutiei, ceea ce împiedică rezolvarea creativă a problemelor.

Lipsa înțelegerii contextuale presupune că sistemele de IA au dificultăți în a înțelege nuanțele contextuale sau expresiile lingvistice ale unei regiuni, ceea ce duce adesea la erori în rezultate.

Biasul de antrenare semnifică că IA se bazează pe date istorice, care pot conține mostre discriminatorii (prejudecăți). În timpul antrenamentului, modelul poate învăța ușor modele discriminatorii pentru a produce rezultate biasate și injuste.

Posibile metode de reducere a biasului în modelele de IA

Experții estimează că până în 2026, 90% din conținutul online ar putea fi generat sintetic. Prin urmare, este vital să minimizăm rapid problemele prezente în tehnologiile de IA Generativă.⁶

⁴ <https://www.sap.com/romania/resources/what-is-ai-bias-> 30 octombrie 2024

⁵ Ce este Biasul de IA? Biasul Inteligenței Artificiale și Stereotipurile Culturale: Efecte, Limitări și Mitigare Published October 4, 2023- Updated May 22, 2026 <https://www.unite.ai/ro/ai-bias-cultural-stereotypes-effects-limitations-mitigation/>

⁶ Înfruntarea realității? Aplicarea legii și provocarea falsificărilor profunde
[https://www.europol.europa.eu/cms/sites/default/files/documents/Europol Innovation Lab Facing Reality Law Enforcement And The Challenge Of Deepfakes.pdf](https://www.europol.europa.eu/cms/sites/default/files/documents/Europol%20Innovation%20Lab%20Facing%20Reality%20Law%20Enforcement%20And%20The%20Challenge%20Of%20Deepfakes.pdf)



Există mai multe strategii cheie care pot fi implementate pentru a reduce biasul în modelele de IA. Unele dintre acestea sunt:

Asigurarea calității datelor este esențială întrucât ”ingestia” de date complete, exacte și curate într-un model de IA poate ajuta la reducerea biasului și la producerea de rezultate mai precise.

Diversificarea datelor implică introducerea de seturi de date diverse într-un sistem de IA, care poate ajuta la mitigarea biasului, deoarece sistemul de IA devine mai inclusiv în timp. (Inteligența Artificială devine treptat mai inclusivă pe măsură ce modelele lingvistice mari și algoritmi sunt antrenați pe seturi de date mai diverse și sunt ajustate fin pentru a reduce prejudecățile. Totuși, progresul aduce și provocări legate de corectitudine și reprezentare.)

Buclele de feedback și de învățare. Aceste [bucle de feedback și învățare](#) sunt necesare, modelele de IA pot îmbunătăți treptat rezultatele.⁷

Reglementările îmbunătățite la nivel global ale IA sunt esențiale pentru menținerea calității sistemelor de IA la nivel transfrontalier. Prin urmare, organizațiile internaționale trebuie să colaboreze pentru a asigura standardizarea IA.

Adoptarea îmbunătățită a IA implică strategiile de IA responsabilă, care contribuie pozitiv la mitigarea biasului de IA, cultivând echitatea și precizia în sistemele de IA și asigurându-se că acestea servesc o bază de utilizatori diversă, în timp ce se străduiesc să se îmbunătățească în mod constant.

Prin integrarea seturilor de date diverse, a responsabilității etice și a mijloacelor de comunicare deschise, putem asigura că IA este o sursă de schimbare pozitivă la nivel global. (doar dacă omul rămâne în control!)

Impact și soluțiile posibile

⁷ Bucla de feedback AI: menținerea calității modelului de producție în era conținutului generat de IA-

<https://www.unite.ai/ro/the-ai-feedback-loop-maintaining-model-production-quality-in-the-age-of-ai-generated-content/>

O buclă de feedback AI este un proces iterativ în care deciziile și ieșirile unui model de IA sunt colectate și utilizate în mod continuu pentru a îmbunătăți sau reantrena același model, rezultând învățare continuă, dezvoltare și îmbunătățire a modelului. În acest proces, datele de antrenare ale sistemului de IA, parametrii modelului și algoritmi sunt actualizați și îmbunătățiți pe baza intrărilor generate din interiorul sistemului.



Biasul algoritmic subminează încrederea în tehnologie și perpetuează inegalitățile sistemice. Pentru atenuarea efectelor sale, strategiile actuale se concentrează pe diversificarea echipelor de dezvoltare și auditarea transparentă a seturilor de date. În Uniunea Europeană, cadrul legal privind protecția datelor și nediscriminarea impune reguli stricte pentru utilizarea etică a acestor sisteme.

Concluzii privind influența bias-ului algoritmic asupra procesului de luarea deciziilor

În cele mai multe situații primul pas este conștientizarea provocărilor produse de toate formele de bias, pentru a găsi mai apoi soluțiile potrivite pentru corectarea erorilor în algoritmi, erori care afectează procesul de luare a deciziilor, implicând în acest proces toate abilitățile, deprinderile, competențele pentru a eficientiza rezultatul final. O posibilă soluție de reducere a bias-ului algoritmic este propus în link-ul indicat. **Cum poate fi redus biasul algoritmic:**⁸ reducerea se poate realiza prin **diversificarea datelor**, prin asigurarea că ”seturile de antrenament” sunt vaste, echilibrate și reprezentative pentru toate categoriile sociale.

Auditarea algoritmică se realizează prin evaluarea periodică a sistemelor AI pentru a detecta dacă acestea iau decizii discriminatorii și testarea lor înainte de lansare.

Nu în ultimul rând **intervenția umană**, este esențială. Menținerea omului în bucla de decizie (human-in-the-loop) pentru cazurile complexe, unde AI-ul ar putea lua decizii cu impact major asupra vieții indivizilor este soluția obligatoriu de aplicat!

La nivel european și național, există reglementări stricte privind nediscriminarea și protecția drepturilor omului care impun companiilor transparență în utilizarea acestor sisteme. Poți afla mai multe despre cum este reglementată responsabilitatea etică și juridică a inteligenței artificiale citind și analiza publicată de [JURIDICE.ro](https://www.juridice.ro)⁹ sau studiul specializat al Agenției pentru Drepturile Fundamentale a Uniunii Europene de pe FRA.¹⁰ În studiul indicat al FRA, include informații privind:

⁸ Înțelegerea și atenuarea părtinirii algoritmice: un ghid cuprinzător- 24 martie 2025

<https://ro.linkedin.com/pulse/understanding-mitigating-algorithmic-bias-guide-ay%C5%9Feg%C3%BCl-g%C3%BCzel-uyfyf?tl=ro>

⁹ Responsabilitatea juridică și etică în era Inteligenței Artificiale, Laura Adina NOVAC, 28.11.2023,

<https://www.juridice.ro/716065/responsabilitatea-juridica-si-etica-in-era-inteligentei-artificiale.html>

¹⁰ Bias in algorithms - Artificial intelligence and discrimination- 8.12.2022, Inteligența artificială este peste tot și afectează pe toată lumea – de la decizia privind conținutul pe care oamenii îl văd pe rețelele lor de socializare până la determinarea cine va primi beneficii de stat. Tehnologiile IA se bazează de obicei pe algoritmi care fac predicții pentru a sprijini sau chiar automatiza complet procesul decizional., <https://fra.europa.eu/en/publication/2022/bias-algorithm>



- Inteligența artificială și prejudecățile: care este problema?
- Bucle de feedback: cum pot algoritmi să influențeze algoritmi
- Prejudecățile etnice și de gen în detectarea discursului ofensator
- Privind în perspectivă: Concentrarea asupra drepturilor fundamentale în inteligența artificială pentru a atenua prejudecățile și discriminarea

Concluzii privind impactul Biasului de IA asupra Societății

Impactul biasului de IA asupra societății poate fi dăunător! Sistemele de IA biasate pot produce rezultate inexacte, care amplifică prejudecățile existente în societate. Aceste rezultate pot crește discriminarea și încălcarea drepturilor, pot afecta procesele de angajare și pot reduce încrederea în tehnologia de IA.

De asemenea, rezultatele biasate de IA pot duce la previziuni inexacte, care pot avea consecințe grave pentru indivizi nevinovați. De exemplu, în august 2020, Robert McDaniel¹¹ a devenit ținta unui act criminal datorită algoritmului de poliție predictivă al Departamentului de Poliție din Chicago, care l-a etichetat ca „persoană de interes”.

Similar, sistemele de IA pentru sănătate dominate de biasuri pot avea și au rezultate acute asupra pacienților. De exemplu, în 2019, [Science](#)¹² a descoperit că un algoritm medical larg utilizat în SUA era biasat rasial împotriva persoanelor de culoare, ceea ce a dus la faptul că pacienții negri au primit mai puțină îngrijire medicală de înalt risc. Redăm în cele ce urmează sintetizat descoperirea prezentată de Science: **Prejudecățile rasiale în algoritmii de sănătate**¹³

”Sistemul de sănătate din SUA utilizează algoritmi comerciali pentru a ghida deciziile în materie de sănătate. Obermeyer și colab. găsesc dovezi ale prejudecăților rasiale într-un algoritm utilizat pe scară largă, astfel încât pacienții de culoare cărora algoritmul le-a atribuit același nivel de risc sunt mai bolnavi decât pacienții albi (vezi Perspectiva lui Benjamin). Autorii au estimat că această prejudecată rasială reduce numărul de pacienți de culoare identificați pentru îngrijire suplimentară cu mai mult de jumătate. Prejudecățile apar deoarece algoritmul folosește costurile de sănătate ca indicator al nevoilor de sănătate. Se cheltuiesc mai

¹¹ a se vedea și Crime Prediction Keeps Society Stuck in the Past. So long as algorithms are trained on racist historical data and outdated values, there will be no opportunities for change. **Predicția criminalității menține societatea blocată în trecut; Atâta timp cât algoritmii sunt antrenați pe baza datelor istorice rasiste și a valorilor învechite, nu vor exista oportunități de schimbare.** Publicat 2 ian 2022, <https://www.wired.com/story/crime-prediction-racist-history/>

¹² Dissecting racial bias in an algorithm used to manage the health of populations. Analizarea prejudecăților rasiale într-un algoritm folosit pentru gestionarea sănătății populațiilor, <https://www.science.org/doi/10.1126/science.aax2342>

¹³ <https://www.wired.com/story/crime-prediction-racist-history/>



puțini bani pentru pacienții de culoare care au același nivel de nevoie și, prin urmare, algoritmul concluzionează în mod fals, că pacienții de culoare sunt mai sănătoși decât pacienții albi la fel de bolnavi. Reformularea algoritmului astfel încât să nu mai utilizeze costurile ca indicator al nevoilor elimină prejudecățile rasiale în prezicerea ” cine are nevoie de îngrijire suplimentară. ””¹⁴

Rezumatul acestui studiu confirmă ca sursă științifică, verificată și argumentată, în privința prejudecăților considerate adevăruri de către bias-urile utilizate în sistemul de sănătate american, (care produce zilnic victime):” Sistemele de sănătate se bazează pe algoritmi comerciali de predicție pentru a identifica și ajuta pacienții cu nevoi complexe de sănătate. Demonstrăm că un algoritm utilizat pe scară largă, tipic acestei abordări la nivel de industrie și care afectează milioane de pacienți, prezintă o prejudecată rasială semnificativă: la un anumit scor de risc, pacienții de culoare sunt considerabil mai bolnavi decât pacienții albi, după cum o demonstrează semnele de boli necontrolate. Remedierea acestei disparități ar crește procentul de pacienți de culoare care primesc ajutor suplimentar de la 17,7 la 46,5%.

*Prejudecata apare deoarece **algoritmul prezice costurile asistenței medicale mai degrabă decât bolile, dar accesul inegal la îngrijire înseamnă că cheltuim mai puțini bani pentru îngrijirea pacienților de culoare decât a pacienților albi. Astfel, în ciuda faptului că costul asistenței medicale pare a fi un indicator eficient al sănătății prin anumite măsuri de acuratețe predictivă, apar prejudecăți rasiale semnificative. Sugerăm că alegerea unor indicatori convenabili, aparent eficienți, pentru adevărul de teren poate fi o sursă importantă de prejudecată algoritmică în multe contexte.***

Concluzia oferă informații corecte și concrete că algoritmii confirmă prejudecăți care afectează viața și sănătatea poporului american....demonstrat prin cifrele și analizele din studiul analizat, dar poate afecta după același mecanism orice domeniul al oricărui popor de pe Terra care utilizează IA. Prejudecăți asemănătoare create prin algoritmi afectează și deciziile în orice alt domeniu, nu numai în sănătate și justiție!

¹⁴ Science, this issue p. [447](#); see also p. [421](#)