



**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 (XVIII)

Biasul algoritmic și gândirea critică (J.3)

**Motto: ”Dacă nu ești psihiatru, nu-ți pierde vremea cu idioții: sunt prea proști ca să plătească unor
neprofesioniști pentru companie.” Stanislaw Jerzy Lec¹**

Fără doar și poate, gândirea critică este esențială pentru fiecare dintre noi, indiferent dacă profesăm sau nu în domeniul juridic....dar este cu atât mai importantă pentru juriști! Analizând motto-ul din acest material putem să înțelegem că e în puterea noastră să evoluăm, să învățăm, să analizăm, să recunoaștem ceea ce este greșit, și tot în puterea noastră să schimbăm ceea ce nu este corect, benefic, evoluând prin învățare! Toate acestea nu se realizează decât printr-o gândire critică/autocritică.

În condițiile în care inteligența artificială (AI) tinde să monopolizeze întreaga noastră viață actuală, dar mai ales cea viitoare, devine esențială înțelegerea legăturii dintre manipularea sau părtinirea AI cum mai este ea numită și gândirea critică (umană).

Potrivit Encyclopediei Britannica, una dintre cele mai prestigioase enciclopedii din lume², *părtinirea* este o tendință de a crede că unele persoane, idei etc. sunt mai bune decât altele, ceea ce de obicei duce la tratarea nedreaptă a unor persoane. Părtinirea face parte din experiența umană. (devenită prejudecată) Nu poți pur și simplu să spui: "Oh, nu, nu am niciodată părtinire."

¹ <https://rightwords.ro/citate/daca-nu-esti-psihiatru-nu-ti-pierde-vremea-cu-idiotii-233259-> Stanislaw Jerzy Lec-
poet polonez

² publicată pentru prima dată în 1768. În prezent, oferă gratuit atât o enciclopedie, cât și un dicționar complet de limba engleză (The Britannica Dictionary).



De fapt, domeniul părtinirii în experiența umană este adesea denumit un punct orb, adică prejudecata — fie că este cognitivă (de exemplu, biasul de confirmare) sau social (de exemplu, prejudecățile rasiale), operează în afara conștientizării noastre, influențându-ne judecățile și deciziile fără ca noi să ne dăm seama.

În condițiile în care AI este creată de om, pare corect ca și judecata AI să aibă părtiniri (bias), prejudecăți?!?

Este necesară definirea termenilor.

Biasul algoritmic reprezintă tendința unui sistem informatic sau de inteligență artificială de a genera rezultate sistematice nedrepte, cauzate de erori în design, de datele folosite pentru antrenare (care reflectă prejudecăți umane) sau de parametrii implicați ai modelului și se manifestă frecvent în diverse domenii, generând consecințe sociale grave. (conform definițiilor confirmate chiar de AI)

Cauzele principale fiind **”date de antrenament viciate”**, întrucât algoritmi învață din date istorice, care includ și reflectă prejudecăți trecute; fiind deci firesc ca inteligența artificială să le reproducă și chiar să le amplifice. **”Proiectarea defectuoasă”** adică erorile umane pot apărea la stabilirea parametrilor și a variabilelor pe care algoritmul le consideră importante, efectul fiind firesc suportat de către toți cei care utilizează AI, și mai grav, consideră informația ca fiind corectă!

Tipurile comune ale bias-ului algoritmic se clasifică ca și în cazul prejudecăților umane în părtinire de gen și rasială. (și nu numai) În cazul părtinirii de gen, algoritmi favorizează bărbații în procesele de recrutare sau subvaluează capacitatea femeilor în anumite roluri (situații comune cu prejudecățile umane reglementate).

Prejudecățile rasiale sunt mai ușor de demonstrat prin recunoașterea facială cu o rată de eroare mai mare pentru persoanele de culoare sau algoritmi medicali care alocă resurse mai puține minorităților, care deja produc victime.

Prejudecata în materie de inteligență artificială este cunoscută și sub numele de *”distorsiune AI”* și se referă la discriminarea sistematică încorporată în sistemele AI care pot consolida prejudecățile existente și pot amplifica discriminarea, prejudecățile și stereotipurile.³

³ Ce este distorsiunea AI? 30 octombrie 2024- <https://www.sap.com/romania/resources/what-is-ai-bias-> care lucrează cu Agenții AI Joule și asistenți Joule



Conform Wikipedia: *"Primele programe de calculator au fost concepute să imite raționamentul și deducțiile umane și erau considerate funcționale atunci când reproduceau cu succes și consecvent acea logică umană. În cartea sa din 1976, **Computer Power and Human Reason**, pionierul inteligenței artificiale **Joseph Weizenbaum** a sugerat că părtinirea poate apărea atât din datele folosite într-un program, cât și din modul în care acesta este programat.*

*Un exemplu timpuriu de părtinire algoritmică a dus la faptul că până la 60 de femei și minorități etnice au fost refuzate în fiecare an la Școala de Medicină a Spitalului St. George's între 1982 și 1986, pe baza implementării unui nou sistem de evaluare computerizată, care a refuzat accesul femeilor și bărbaților cu "nume care sună străin" pe baza tendințelor istorice de admitere."*⁴

Simplificând, **părtinirea algoritmică se referă la erori sistematice și recurente într-un sistem informatic care duce la rezultate nedrepte sau distorsionate, adesea favorizând un grup în detrimentul altuia în moduri care se abat de la scopul inițial al algoritmului.**

Este destul de interesant, totuși, că **mulți algoritmi își declară ca obiective eliminarea părtinirii umane.** Totuși, este important să înțelegem că IA nu este lipsită de prejudecăți. Din contră, IA poate avea prejudecăți și, în multe cazuri, poate chiar extinde prejudecățile umane!

Dificultatea practică este majoră, întrucât, prejudecățile umane sunt reglementate cu titlul exemplificativ, tocmai datorită faptului că omul, plin de imaginație, poate oricând inova în zona discriminărilor și prejudecăților pe care AI, la rândul său, le poate dezvolta, și multiplica!

Riscul major în cazul bias-ului algoritmic va conduce la discriminări și prejudecăți grave față de oameni, care până sunt demonstrate legal că sunt erori ale sistemului AI, oamenii, în realitatea zilelor noastre, vor fi victime ale acestor grave nedreptăți!

Complexitatea bias în dimensiunea algoritmică oferă o listă cuprinzătoare, deși nu exhaustivă, care este realizată după modelul prejudecăților reglementate deja la nivelul societăților care valorizează drepturile și libertățile fundamentale ale omului.

⁴ Citat în materialul publicat în 24.martie 2025: "Înțelegerea și atenuarea părtinirii algoritmice: un ghid cuprinzător"

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



Deja se scriu articole și se fac analize și studii care clasifică spațiul bias din domeniul AI în **bias sistemic, bias umane și biasuri statistice/computaționale (algoritmice)**.⁵

Deși biasurile statistice și computaționale reprezintă erori sistematice care denaturează rezultatele unui studiu sau predicțiile unui algoritm, acestea apar atunci când datele folosite nu sunt reprezentative, iar modelele matematice favorizează sau discriminează anumite grupuri.

În acest moment intervine nevoia stringentă a intervenției gândirii umane critice, care după o analiză critică eficientă, să descopere prejudecățile și să nu permită algoritmului să aplice automat un criteriu discriminatoriu care va afecta omul supus analizei algoritmului respectiv!

Biasul Computațional (Algoritmice) apare în inteligența artificială și în procesul ”machine learning (vs deep learning)⁶”, adesea ca o consecință a datelor istorice viciate pe care se antrenează modelul.

Distincția între învățarea automată (mecanică) și învățarea profundă, (conștientă) face diferența între judecata automată realizată prin algoritmi AI și gândirea critică automată. Astfel:

- Biasul de date (Data Bias) este prezent dacă ”datele de antrenament” conțin prejudecăți umane (ex: discriminare de gen sau rasă în CV-uri), algoritmul va învăța și va amplifica aceste stereotipuri.
- Biasul de asociere este situația în care algoritmi asociază greșit anumite concepte; de exemplu, un motor de căutare vizual poate asocia automat femeile cu activitățile casnice.
- Biasul de buclă (Feedback Loop) este algoritmul care creează un efect de bulgăre de zăpadă. Recomandările video sau de știri bazate exclusiv pe istoricul utilizatorului îl limitează la conținut similar, blocându-l în propriile sale preferințe.

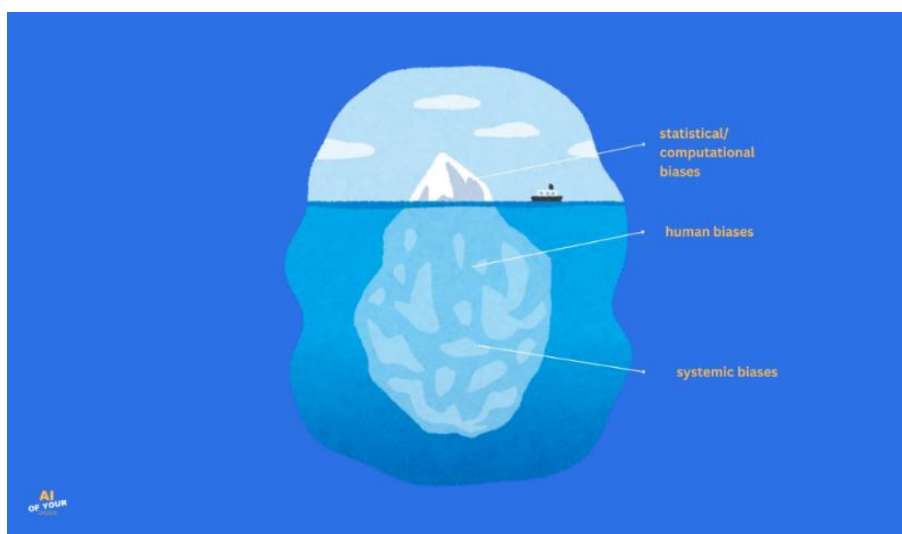
⁵ a se vedea și <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf>

articolul NIST: ”Către un standard pentru identificarea și gestionarea părtinirii în inteligența artificială”

⁶ Învățarea automată este un subset larg al inteligenței artificiale (este un grup sau o colecție de elemente care fac parte dintr-o altă mulțime mai mare); în care **algoritmii învață tipare din date pentru a face predicții, bazându-se în mare măsură pe selecția caracteristicilor ghidată de om. Învățarea profundă este subsetul specializat al învățării automate care utilizează rețele neuronale artificiale multistratificate pentru a extrage automat caracteristici și a gestiona date nestructurate.**



De ce este periculos bias-ul algoritmic, ...este semnificativ ilustrat în imaginea de mai jos, tocmai pentru că AI depășește gândirea umană, accentuând prejudecățile umane astfel încât judecățile AI va afecta direct viața și viitorul umanității, dacă omul nu intervine rapid și eficient asupra fenomenului care devine îngrijorător!



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

Mai jos sunt prezentate tipuri comune de bias algoritmic care afectează direct relația AI cu omul:

1. **Părtinirea de gen** apare atunci când un gen este favorizat în detrimentul altuia sau exclus din rezultate. (inclusiv discriminarea pozitivă). De exemplu, un algoritm de angajare care prioritizează bărbații în detrimentul femeilor pentru anumite roluri sau care nu recunoaște persoanele transgender⁷.
2. **Prejudecățile rasiale** apar atunci când un algoritm favorizează sau discriminează disproporționat anumite grupuri rasiale. Un exemplu este IA în sistemul de justiție penală, care poate prezice inexact rate mai mari de recidivă pentru anumite demografii rasiale.

⁷ persoană cu o identitate de gen diferită de genul atribuit la naștere și care dorește să-și exprime identitatea de gen în alt mod, conform https://eige.europa.eu/publications-resources/thesaurus/terms/1285?language_content_entity=ro a se vedea și <https://ro.wikipedia.org/wiki/Transgen>



3. **Părtinire de vârstă** se manifestă atunci când un algoritm favorizează anumite grupe de vârstă în detrimentul altora, adesea ca urmare a unei reprezentări dezechilibrate a datelor. De exemplu, AI-ul pentru recrutare ar putea prioritiza candidații mai tineri în detrimentul candidaților mai în vârstă și mai experimentați.
4. **Prejudecățile socio-economice** apar atunci când algoritmi reprezintă disproporționat indivizi cu avantaje economice față de cei din medii cu venituri mici. Acest lucru poate fi observat în sistemele de recunoaștere facială bazate pe AI sau în sistemele de monitorizare la locul de muncă, care presupun accesul la anumite tehnologii.
5. **Biasul de confirmare** se realizează atunci când un algoritm întărește credințe preexistente prin furnizarea continuă a conținutului aliniat cu comportamentul lor anterior. Acest lucru este comun în sistemele de recomandare pentru rețelele sociale de pe platforme precum YouTube, TikTok, Facebook și Instagram.
6. **Părtinirea de reprezentare** apare atunci când anumite grupuri demografice sunt subreprezentate în datele de antrenament ale unui algoritm, ceea ce duce la rezultate inexacte sau nedrepte.
7. **Deriva conceptului** apare atunci când un algoritm nu reușește să se adapteze la evenimente neprevăzute sau circumstanțe schimbătoare. De exemplu, un asistent AI care prezice comportamentul de cumpărare poate continua să sugereze produse dintr-o locație anterioară, după ce un utilizator se mută într-un oraș nou.
8. **Părtinirea confidențialității** apare atunci când persoanele care renunță la colectarea datelor sunt excluse din luarea deciziilor algoritmice. De exemplu, o inteligență artificială pentru cumpărături online care personalizează recomandările poate dezavantaja utilizatorii care refuză urmărirea datelor.
9. **Prejudecăți legate de dizabilitate** se întâmplă atunci când un algoritm nu ia în calcul utilizatorii cu dizabilități fizice sau mentale. De exemplu, detectarea emoțiilor condusă de AI poate interpreta greșit expresiile persoanelor cu afecțiuni neurodivergente, ducând la rezultate inexacte la locul de muncă, săli de clasă sau vehicule.



10. **Părtinirea lingvistică** apare atunci când un sistem AI nu reușește să țină cont de diversitatea lingvistică, dezavantajând vorbitorii non-nativi. Un exemplu este un algoritm automatizat de notare care penalizează elevii care scriu în engleză ca a doua limbă.
11. **Părtinirea regională** apare atunci când anumite zone geografice sunt reprezentate nedrept în date. De exemplu, utilizarea codurilor poștale pentru a determina aprobările împrumuturilor fără a lua în considerare disparitățile economice, istorice poate dezavantaja comunitățile marginalizate.
12. **Biasul recent**⁸ se manifestă atunci când un algoritm prioritizează datele recente în detrimentul contextului istoric, ceea ce poate duce la predicții inexacte. Un exemplu este un sistem AI care prognozează tendințele pieței de capital, doar pe baza evenimentelor recente, ignorând tiparele pe termen lung.

Biasul (părtinirea) recentă, ca eroare cognitivă prin care supraevaluăm importanța ultimelor informații, evenimente sau experiențe primite, afectează grav deciziile corecte, excluzând analiza și gândirea critică. De exemplu în sfera investițiilor și a finanțelor, se manifestă în cazul investitorilor care tind să cumpere acțiuni care au crescut abrupt recent și să vândă în panică la cea mai mică scădere din ultima perioadă, uitând de istoricul pe termen lung al pieței. (și iată cum printr-un algoritm, la un click de calculator se creează crize inexistente, prin manipularea piețelor financiare)

La locul de muncă, algoritmul se manifestă în evaluările de performanță, atunci când managerii au tendința de a puncta mai bine angajații care au avut un proiect de succes în ultima lună, ignorând rezultatele bune, dar mai vechi din restul anului.

În sport se manifestă atunci când un sportiv sau o echipă sunt adesea judecați exclusiv după ultimul meci jucat, fanii și comentatorii minimalizând performanțele anterioare constante.

⁸ **Biasul de recentă** (recency bias) părtinirea recentă- este o eroare cognitivă prin care supraevaluăm importanța ultimelor informații, evenimente sau experiențe primite, ignorând datele mai vechi sau contextul de ansamblu. Mintea noastră tinde să rețină și să prioritizeze ceea ce s-a întâmplat recent, deoarece este mai proaspăt în memorie. A se vedea și <https://www.psychologies.ro/dezvoltare-personala-cunoaste-te-2/cele-mai-comune-tipuri-de-bias-cognitiv-si-cum-sa-le-recunosti-2212496-> 17 oct 2025



Pentru a contracara acest bias, omul, managerul, cel care are de luat decizii poate folosi **jurnale de activitate**, **stabilirea unor criterii obiective** (în loc să te bazezi pe impresii) și **analiza datelor istorice** pe termen lung înainte de a lua o decizie.

Alte exemple de părtiniri algoritmice și prejudecăți care afectează viața și deciziile umane și non umane, recomandate de algoritmi sunt relatate în cartea **Design Justice: Community-Led Practices to Build the Worlds We Need**, autor **Sasha Costanza-Chock- Design Justice**⁹, evidențiază modul în care sistemele de securitate aeroportuare, în special scanerile de unde milimetrice, codifică prejudecățile de gen, prin impunerea unei înțelegeri rigide și binare a genului.

Tehnologia de scanare cere ofițerilor să selecteze "Bărbat" sau "Femeie", iar orice abatere de la modelul binar, cis-normativ al sistemului, declanșează un avertisment de anomalie, supunând persoanele transgender și nonbinare unei supravegheri și disconfort suplimentare.

Această părtinire nu este doar tehnică, ci profund înrădăcinată în sistemele sociotehnice, reflectând și consolidând presupuneri cisnormative mai largi în designul algoritmic. Experiența depășește genul, intersectându-se cu prejudecățile rasiale și ale dizabilității, deoarece femeile de culoare, călătorii musulmani și persoanele cu dizabilități se confruntă, de asemenea, cu poveri disproporționate din cauza acestor protocoale de securitate defectuoase.¹⁰

Cazul de angajare Strategieon¹¹ ilustrează părtinirea algoritmică în recrutare, unde un sistem AI (PARiS) a discriminat neintenționat un candidat calificat, Hara, din cauza prejudecății de reprezentare și a prejudecății indirecte privind dizabilitatea.

Inteligența artificială a fost antrenată pe foști angajați—majoritatea foști militari—care aveau o corelație puternică cu experiențele sportive. Ca urmare, a învățat să asocieze experiența sportivă cu a fi o "potrivire

⁹ Sasha Costanza-Chock - Design Justice- <https://direct.mit.edu/books/oa-monograph/4605/Design-JusticeCommunity-Led-Practices-to-Build-the-> O explorare a modului în care designul ar putea fi condus de comunități marginalizate, ar putea demonta inegalitatea structurală și ar putea promova eliberarea colectivă și supraviețuirea ecologică.

¹⁰ Conform - Costanza-Chock, Design Justice, pp. 1-5- <https://mitpress.mit.edu/9780262043458/design-justice/>

¹¹ <https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/12/Princeton-AI-Ethics-Case-Study-5.pdf>



bună", ceea ce a dus la excluderea candidaților precum Hara, care nu aveau istoric de participare sportivă din cauza utilizării scaunului cu roțile pe viață.

Acest lucru demonstrează cum datele părtinitoare privind instruirea pot întări practicile nedrepte de angajare, chiar și atunci când atribute protejate, cum ar fi statutul de dizabilitate, nu sunt luate în considerare explicit.

Cazul evidențiază riscurile de prejudecată în automatizare, unde echipele umane de HR, inițial sceptice, se încred tot mai mult în AI, fără a-i pune la îndoială deciziile. Acest studiu de caz subliniază provocările etice ale angajărilor bazate pe AI și necesitatea unor strategii de reducere a prejudecăților pentru a asigura echitatea și diversitatea în recrutare ¹².

Strategii de atenuare a prejudecății AI

Strategiile de atenuare a prejudecăților sunt esențiale pentru viața omului. Analizând diferitele tipuri de bias, inclusiv prin efectul disfuncționalităților din procesul de programare umană al algoritmului inițial, se constată că pot apărea disfuncționalități urmare deciziilor luate de AI în timpul selecției datelor, selecției caracteristicilor sau altor alegeri ale celor care implementează aceste sisteme.

În cazul problemelor sociotehnice, codul și datele singure sunt insuficiente, deoarece provocările reale depășesc performanța sistemului și a modului în care este folosit sistemul.

Întrebările cheie nu sunt doar despre eficiență, ci și despre putere: Cine are de câștigat? Cine este rănit? Cine ia deciziile?...dar și Cine răspunde pentru daunele produse omului?

Concluzii

Cea mai bună strategie de reducere părtinitoare algoritmică este să fie acceptate în echipe perspective și viziuni diferite asupra lumii, să fie create spații sigure pentru a spune adevărul, să se stabilească sisteme de avertizare de integritate și să existe o asumare a responsabilității personale pentru a conștientiza posibila discriminare. Dacă are loc discriminarea, este important să existe și asumarea de către un OM a responsabilității pentru aceasta.

¹² Studiu de caz privind etica IA de la Princeton, 2023-

<https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/12/Princeton-AI-Ethics-Case-Study-5.pdf>



Cofinanțat de
Uniunea Europeană



Programul Educație și Ocupare 2021-2027
Cod apel: PEO/71/PEO_P7/OP4/ESO4.5/PEO_A49

Numele proiectului: DreptCompAct - COMPetențe ACTuale pentru studenți la DREPT
Cod proiect: 312957



<https://rightwords.ro/citate/adevarul-este-rareori-pur-si-niciodata-simplu--566>