

fòrum Beijing+30

Una mirada cap el futur
en la igualtat de gènere



**Intel·ligència Artificial responsable
i biaixos de gènere**

Intel·ligència Artificial responsable i biaixos de gènere

Meritxell Bautista Quiñones, presidenta executiva de Manix Capital i consellera de l'ODEE

Esther Nin Camps, directora de l'assessoria jurídica internacional, Banc de Sabadell i cosellera de l'ODEE

Marta Rodríguez Martín, directora d'organització, persones i cultura, Mercabarna i cosellera de l'ODEE

Montse Serra Valldeoriola, vicepresidenta de Serveis i Solucions d'HP i consellera de l'ODEE

Jordina Torrents Barrena, HP Senior AI Manager and Principal AI Architect

Iolanda Montserrat Siles, HP Personal Systems Marketing Segment Lead

Karina Gibert Oliveras, degana del Col·legi oficial d'enginyeria informàtica de Catalunya i catedràtica d'intel·ligència artificial i ciència de les dades de la UPC, donesCOEINF

Edita ODEE, Cambra de Comerç, Setembre 2025



ÍNDEX

1. Introducció	7
2. Per què la IA té biaixos de gènere?	8
3. Impactes negatius de la IA en qüestions de gènere	9
4. Marc jurídic europeu	10
5. Propostes per mitigar els biaixos de gènere en IA: actuacions, eines i metodologies	13
5.1. Tractament del biaix a les dades	
5.2. Detecció de biaixos en els models	
5.3. Detecció de biaixos en els resultats	
5.4. Propostes de l'àmbit jurídic	
5.5. Actuacions addicionals	
6. L'àmbit de la comunicació i premsa	18
7. L'àmbit dels departaments de gestió de persones	22
8. Conclusions finals	24

1. Introducció

La intel·ligència artificial (en endavant, IA) està transformant múltiples aspectes de la societat, des de l'economia fins a la sanitat, passant per l'educació i l'entreteniment.

En un món cada cop més digitalitzat, la IA s'ha convertit en una eina indispensable per a les empreses. Les seves capacitats per analitzar gran volum de dades i prendre decisions informades han revolucionat múltiples sectors empresarials.

No obstant això, aquesta tecnologia no està exempta de problemes. Darrera d'aquesta tecnologia s'amaguen riscos latents. Un dels més destacats és la perpetuació i ampliació de biaixos de gènere que poden tenir impactes significatius en l'equitat i la inclusió.

Els biaixos en la IA, arrelats en les mateixes dades de partida i en els algoritmes utilitzats, poden perpetuar desigualtats, discriminacions i presa de decisions injustes. Malgrat la creixent importància de la IA, encara hi ha moltes empreses que desconeixen la profunditat del funcionament d'aquests mecanismes que generen biaixos i les conseqüències que poden comportar per al seu negoci i per a la seva reputació.

La presència de biaixos en el marc empresarial, fruit de la utilització de la IA, poden portar conseqüències significatives per a les empreses, tals com:

- **Pèrdua de reputació:** associar la marca amb decisions injustes o discriminatòries pot generar una crisi de confiança
- **Danys en imatge de marca:** els biaixos poden afectar negativament a la lleialtat i al compromís per part d'empleats, clients i proveïdors.
- **Riscos legals:** les empreses poden arribar a ser sancionades legalment si els seus sistemes o usos del IA discriminen a determinats grups socials.
- **Decisions empresarials errònies:** els biaixos poden portar a conclusions incorrectes i a decisions estratègiques equivocades.

Per això, és important analitzar la problemàtica i identificar bones pràctiques. Aquest informe té com a objectiu proporcionar un marc de referència per a que les empreses entenguin els riscos associats als biaixos de gènere en la IA i adoptin pràctiques responsables i ètiques en el desenvolupament i l'ús d'aquesta tecnologia. Oferint guies pràctiques, metodologies i eines per ajudar a les empreses a identificar, mitigar i prevenir biaixos en els seus models de IA. Ja sigui perquè estan iniciant o implementant un procés de transformació digital, per millorar la seva productivitat o bé perquè busquin ampliar la seva oferta de productes i serveis actuals.



2. Per què la IA té biaixos de gènere?

Falta de consciència

Existeix un buit de coneixement sobre què són i quins són els biaixos a la IA, com s'originen i com afecten als resultats a l'hora d'utilitzar-la.

Pressió per la presa de decisions i obtenir resultats immediats

La urgència del dia a dia sovint eclipsa la necessitat d'una avaluació rigorosa de les dades i els algoritmes a utilitzar. La detecció i mitigació dels biaixos en cadascun d'ells afegixen una capa de complexitat i temps addicional en el procés de desenvolupament. Entre el 70% i el 90% dels prototips funcionals no arriben mai a producció pel seu cost i temps associat, que malauradament impacta en els biaixos dels resultats d'aquesta IA.

Cada capa de responsabilitat té un cost

La implementació de sistemes basats en IA requereix una planificació exhaustiva que consideri des de l'inici els costos associats a garantir la qualitat, l'ètica i la sostenibilitat de les dades i del model. Sovint manca incorporar una partida pressupostària que cobreixi des del disseny, el desenvolupament de les proteccions que evitaran els biaixos, la producció i el manteniment dels models de IA per tal d'evitar funcionaments esbiaixats, imprevistos o retards en la implementació causats per la detecció tardana dels biaixos quan el procés ja està avançat. Aquesta falta de previsió pot generar retards significatius i costos addicionals, obligant a reentrenar el model completament.

Limitacions en les eines i experiència en l'ús

Moltes empreses manquen de les eines i metodologies necessàries per detectar, mesurar i mitigar els biaixos en el seu ús de la IA. Hi ha un buit d'experts en gestió de la responsabilitat i l'ètica en els projectes de IA. Són projectes que requereixen d'un alt coneixement i sensibilitat per aquests aspectes, formació tècnica, però també ètica i habilitats tècniques per solucionar-les.

Absència de protocols

La falta de protocols i processos establerts per avaluar i controlar els resultats dels models de la IA augmenta el risc de perpetuar els biaixos, que poden provenir de les dades o dels models o directament dels propis resultats finals de les aplicacions de IA. La teoria del mostreig i el disseny d'experiments aporten principis que es coneixen des del segle passat que podrien contribuir fortament a disminuir els biaixos provocats per una descompensació en les dades que alimenten els models de IA en la fase d'entrenament. Malgrat tot, això no és suficient i cal sempre desenvolupar protocols de validació molt exhaustius que garanteixin el comportament no esbiaixat del sistema/model. Addicionalment, hi ha reptes importants relacionats amb la manca de tècniques solvents de verificació formal del software

per garantir que l'algorisme no s'ha dissenyat imprimint un biaix indesitjat i que actualment només es pot mitigar amb testos de validació intensiva del sistema abans de posar-los en producció. Quan es parla de models de IA molt concrets que prediuen un paràmetre específic, és senzill fer aquests testos de validació. Quan es parla de grans sistemes basats en IA més complexos no és tan senzill degut a l'alt nombre de característiques que influeixen en els resultats de la IA i que no permeten la validació exhaustiva, ja que la casuística és de naturalesa altament combinatòria.

3. Impactes negatius de la IA en qüestions de gènere

1. Prediccions i recomanacions que desfavoreixen les dones quan les dades d'entrenament estan esbiaixades

La majoria dels algoritmes de IA es formen amb grans conjunts de dades que, sovint, reflecteixen els biaixos existents en la societat. Per exemple, si un conjunt de dades conté més exemples d'homes que de dones en rols tècnics, l'algoritme pot perpetuar aquesta desigualtat.

2. Vulneració en els drets fonamentals (drets d'igualtat en totes les seves manifestacions: d'accés, dret al treball, drets socials...)

L'automatització impulsada per la IA està transformant el mercat laboral. No obstant això, els biaixos de gènere en els algoritmes de contractació poden portar a la subrepresentació de dones en sectors tecnològicament avançats. Un exemple conegut és el sistema de contractació d'Amazon que mostrava preferència per candidats masculins a causa de biaixos en les dades històriques de contractació.

3. Perpetuació d'estereotips de gènere

Els assistents virtuals com Siri, Alexa i Google Assistant solen tenir veus femenines i es programen per ser submises i servicials, reforçant estereotips sobre el rol de la dona en la societat.

4. Major risc de falses atribucions per a les dones, incloent processos legals

Els biaixos de gènere en els sistemes de visió per computador, inclosos els algorismes de reconeixement facial que poden utilitzar fins i tot les forces de l'ordre, han mostrat majors taxes d'error en identificar dones, especialment dones de color, a causa de conjunts de dades no representatius i mal equilibrats. Això pot tenir greus implicacions en àrees com la seguretat i la vigilància.

4. Marc jurídic europeu

La Unió Europea va aprovar, el 21 de maig de 2024, el Reglament d'Intel·ligència Artificial (RIA)¹ per promoure una intel·ligència artificial segura i fiable (trustworthy) i centrada en l'ésser humà, assegurant un nivell alt de protecció a la salut, a la seguretat i als drets fonamentals; és aquí on queda inclosa, doncs, la igualtat entre sexes.

Aquest marc regulador introdueix un sistema d'avaluació dels riscos associats a l'ús d'una intel·ligència artificial, n'identifica quatre, acompanyats de mesures d'exigència creixent en funció del nivell de risc. La regulació inclou diverses disposicions clau per abordar el risc del biaix de gènere en la IA a través, entre d'altres però bàsicament, de la protecció al dret fonamental de la igualtat.

Anteriorment però, la Unió Europea va publicar les Directrius ètiques per una IA segura i fiable², elaborades per un grup independent d'experts IA d'alt nivell, creat per la Comissió Europea al juny de 2018. És un valuós text on es presenta un codi ètic basat en 7 eixos, un dels quals recull de forma expressa, i entre d'altres, la diversitat, la no discriminació, l'equitat, que impactarà directament en un altre dirigit al benestar –en particular el benestar social–, com a requisits ètics per comptar amb una IA segura i fiable; i s'identifica de forma expressa el risc dels biaixos per raó de sexe.

L'Observatori Català d'Ètica de la Intel·ligència Artificial va llançar, l'any 2023, un segell (model PIO) per qualificar el nivell ètic d'una aplicació intel·ligent a partir d'un constructe que avalua tots els eixos del model ètic europeu per la IA, i aporta un resultat en codi de colors on cada dimensió es representa amb un color que gradua entre el verd i el vermell indicant la intensitat ètica d'aquella dimensió. Es tracta d'un qüestionari gratuït i el resultat és un segell, molt visual que, ràpidament, dona la perspectiva de quins aspectes de l'aplicació estan més o menys madurs en matèria d'ètica. Certament una bona part de les recomanacions ètiques han estat recollides en la nova regulació europea, però no totes. Això significa que no sempre tenir una aplicació auditada i que compleixi la llei en vigor significarà tenir una aplicació èticament segura.

Es podria concloure que, en l'estadi actual, comptem amb una norma i unes directrius ètiques que ens faciliten un marc en el que el biaix per raó de gènere es veu identificat i es vol evitar:

1. Requisits per als sistemes de IA d'alt risc

- **Qualitat de les dades.** Els sistemes de IA d'alt risc han de ser desenvolupats utilitzant dades d'alta qualitat, representatives i sense biaixos. Les empreses han de demostrar que les dades utilitzades són adequades per al context específic de l'ús de la IA. En concret, provar que les dades que s'utilitzaran per entrenar la IA no estan esbiaixades en gènere serà fonamental per garantir models intel·ligents equitatius.

¹ Reglament (UE) 2024/1689 del Parlament Europeu i del Consell, de 13 de juny de 2024, pel qual s'estableixen normes harmonitzades en matèria d'intel·ligència artificial i pel qual es modifiquen els Reglaments (CE) núm. 300/2008, (UE) núm. 167/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 i les Directives 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (Reglament d'Intel·ligència Artificial).

² Directrius ètiques per a una intel·ligència artificial fiable (Ethics Guidelines for Trustworthy Artificial Intelligence), elaborades pel Grup d'Experts d'Alt Nivell sobre la IA de la Comissió Europea i publicades a l'abril del 2019.

- **Supervisió humana.** Els sistemes de IA han de ser dissenyats per permetre la intervenció humana quan sigui necessari, garantint que les decisions automàtiques poden ser revisades i corregides si es detecten errors o biaixos. En particular, tots els sistemes basats en IA que involucrin decisions que afecten a persones requereixen aquesta supervisió humana. I no només això sinó que serà molt important formar bé a aquests supervisors de la IA i que tinguin clar que poden qüestionar-ne els resultats si no els hi encaixen i rebutjar-los.

2. Avaluació i gestió de riscos

- **Identificació de riscos.** Abans de desplegar un sistema de IA, les empreses han d'identificar i avaluar els riscos potencials associats, inclosos els riscos de biaix de gènere.
- **Mitigació de riscos.** Desenvolupar estratègies efectives per mitigar els riscos identificats, com la formació en igualtat, diversitat i inclusió per als desenvolupadors de IA, l'auditoria regular dels algorismes, i formar equips paritaris on la mirada femenina estigui ben representada per evitar justament dissenys on les necessitats i drets de les dones es puguin veure vulnerats

3. Directrius ètiques

- **Respecte per la igualtat, la diversitat i la no discriminació.** Els desenvolupadors de IA han de seguir principis ètics que respectin la igualtat i evitin discriminació per raó de gènere, així com respecte a la diversitat i evitin la discriminació per raó de raça o altres característiques protegides. La formació i constitució d'equips paritaris ja esmentada contribuirà a la igualtat.
- **Transparència i responsabilitat.** Els sistemes de IA han de ser transparents en els seus processos de decisió, permetent la traçabilitat i la comprensió de com es prenen les decisions i s'ha d'informar quan hi ha intervingut una IA i la persona destinatària d'aquesta decisió tindrà sempre dret a rebre una argumentació sobre com la IA ha calculat aquella decisió (per això és fonamental utilitzar algorismes de la IA que siguin explicables). A més, les empreses han de ser responsables dels impactes de la seva IA en la societat.

4. Inclusió de la perspectiva de gènere en el desenvolupament de la IA

- **Equips paritaris i diversos.** Promoure la paritat en els equips de desenvolupament de la IA per assegurar que es consideri la mirada femenina, així com promoure la pluralitat de perspectives i experiències, reduint així el risc de biaixos
- **Formació i sensibilització.** Proporcionar formació específica sobre la igualtat de gènere i la identificació de biaixos a les persones desenvolupadores, responsables i que utilitzen la IA. És també fonamental formar sobre la naturalesa imperfecta dels models en general i els de IA en particular, i donar criteris a les persones que estaran a càrrec d'utilitzar aquests sistemes per qüestionar el resultat d'una IA en un cas concret sempre que sigui necessari i, sobretot, quan hi hagi risc de vulnerar injustament drets fonamentals.

Aquestes mesures reflecteixen el compromís de la UE per assegurar que el desenvolupament i la implementació de la IA respectin els drets fonamentals i promoguin una societat més justa i equitativa. En especial, s'espera que el RIA serveixi com a model per a altres jurisdiccions, promovent una adopció global d'aquests principis.

Podem acabar dient que el camí cap a la igualtat real dona/home, i tot el que implica, encara no s'ha acabat, però la IA no pot ser un obstacle ni, molt menys, fer-nos retrocedir. Al contrari, és la gran ocasió per reduir les desigualtats. Però és necessari un marc ètic i legal complert i ambiciós, més enllà dels bons propòsits. Caldrà també tenir en compte que el marc legal, de desenvolupament molt més lent que la tecnologia, per si sol no serà suficient, sinó que haurà de venir precedit pel marc ètic, aquest sí, d'evolució més dinàmica i que s'haurà d'anar adaptant al ritme que la tecnologia requereix. Donat que l'ètica és de compliment voluntari, és molt important operar un canvi cultural generalitzat on tothom des del seu rol es comprometi en aquests valors ètics mentre es desenvolupa el marc legal corresponent. El RIA i les Directrius ètiques són un primer pas.

La IA no és una finalitat, és un mitjà molt prometedor per afavorir la prosperitat de la humanitat, però tota la humanitat i no hauria de deixar ningú enrere.

Conclusions

- El marc regulatori europeu és sòlid.
- La llei refereix la necessitat de garantir que les dades i els algorismes no estiguin esbiaixats i tracta globalment totes les possibles fonts de biaix, entre les que menciona el biaix per raó de sexe.
- A través dels mecanismes per a la protecció expressa dels drets fonamentals s'entén protegit el dret a la igualtat i, per tant, la igualtat dona/home, evitant biaixos per raó de sexe.
- El marc ètic contempla explícitament el dret a la igualtat i esbossa alguns mecanismes per evitar els biaixos.
- No comptem amb marcs jurídics i ètics similars en altres jurisdiccions concretes, malgrat els pronunciaments de l'ONU.
- Es fa necessari el compromís de tots els actors (desenvolupadors, distribuïdors, usuaris, etc.) amb el codi ètic, més enllà del que marqui la llei, en prevenció del desajust temporal que es produeix entre el ritme tecnològic i els processos legislatius.
- Existència de diferents sensibilitats respecte a la igualtat dona/home en les diferents jurisdiccions que, arribat el cas, es reflectiran al seu marc jurídic.

5. Propostes per mitigar els biaixos de gènere en IA: actuacions, eines i metodologies

Detectar i mitigar els biaixos de gènere en sistemes d'IA requereix accions específiques així com l'ús d'eines i metodologies dedicades que aborden les diferents fases del cicle de vida del sistema. A continuació, detallem algunes eines de codi lliure i metodologies clau en cadascuna d'elles:

1. Diversificació de dades des del disseny

Assegurar que els conjunts de dades que es recopilen o construeixen per entrenar algoritmes de la IA siguin representatius de les dones.

- **Auditories de dades:** realitzar auditories exhaustives del conjunt de dades utilitzades per entrenar algoritmes de la IA siguin diversos i representatius de totes les demografies. Això inclou identificar i quantificar la presència de biaixos de gènere.
- **Auditories d'algoritmes:** implementar auditories regulars d'algoritmes per identificar i corregir biaixos. Aquestes auditories han de ser realitzades per equips diversos que incloguin dones i persones de diferents orígens.

2. Transparència i rendició de comptes

Fomentar la transparència en el desenvolupament de IA, fent que les organitzacions publiquin informes sobre el disseny, implementació i resultats dels seus algoritmes. A més, establir mecanismes de rendició de comptes per a les empreses que utilitzin IA de manera esbiaixada.

3. Educació i formació

Invertir en l'educació i formació en IA amb enfocament en la inclusió de gènere. Això inclou incentivar a més dones a participar en camps STEM (Ciència, Tecnologia, Enginyeria i Matemàtiques) i proporcionar formació sobre la identificació i correcció de biaixos en IA.

4. Sensibilització en l'ús responsable, ètic i deontològic de la IA

És fonamental que tots els actors de la cadena de valor de la IA assumeixin uns valors ètics que garantitzin l'ús responsable i equitatiu de la IA més enllà del que marquin les regulacions, i sobretot, mentre les regulacions no arriben.

5. Regulació i polítiques públiques

Desenvolupar polítiques i marcs reguladors que obliguin a les empreses a considerar l'impacte de gènere en els seus sistemes de la IA. Això pot incloure directrius específiques per a l'avaluació i mitigació de biaixos.

5.1. Tractament del biaix a les dades

Recollida de dades representatives

Cal dissenyar la composició de les dades d'entrenament assegurant que són representatives de la diversitat de la població/fenomen d'on provenen. Recordem que la teoria del mostreig i del disseny d'experiments pot aportar un bon marc metodològic per abordar aquestes fases. Això pot implicar la recopilació de noves dades o l'augment de les dades existents per equilibrar la representació de dones i altres grups subrepresentats. És fonamental garantir la representació de dones i minories.

Etiquetatge correcte de les dades

Assegurar que les dades estiguin correctament etiquetades per reflectir adequadament les identitats de gènere. Això inclou l'ús d'etiquetes inclusives i la correcció d'etiquetes errònies.

Auditories de dades

Revisar com s'han recopilat i etiquetat les dades per identificar biaixos en la representació de gènere (p. ex., sobrerrepresentació d'homes en llocs d'alt rang).

Detecció del biaix

Visualitzar les dades i fer l'anàlisi exploratòria per identificar patrons, desigualtats i desequilibris de les dades. (p. ex., comparació de proporcions de gènere en categories clau).

Neteja i preprocessament adequats

Eliminar *outliers*, valors que falten, corregir les inconsistències i normalitzar la terminologia. D'especial interès és l'emascarament d'informació sensible que puguin induir biaixos de gènere, com noms o pronoms, abans d'entrenar els models.

Correcció del biaix

Cal actuar considerant els següents aspectes:

- **Balanceig de les mostres d'entrenament**, submostreig (*undersampling*) o remostreig (*oversampling*) per ajustar l'equilibri de les dades i evitar que el model s'esbiaixi cap a la categoria majoritària.
- **Normalització de dades**: Aplicar tècniques de normalització per equilibrar les distribucions de diferents gèneres. Per exemple, el submostreig de la majoria o el sobremostreig de la minoria per aconseguir una distribució equitativa.

- **Reweighting:** Assigna pesos als exemples en funció del grup de gènere per equilibrar el conjunt de dades.
- **Data Augmentation:** Genera dades addicionals per al grup subrepresentat.
- **Learning Fair Representations:** Transformen les dades perquè siguin estadísticament indistingibles entre grups abans de passar-les al model.
- **Generació de dades sintètiques:** Crear dades artificials per augmentar la representativitat de les mateixes i reduir els biaixos en les dades originals.

Eines a considerar en aquesta fase:

- **Fairlearn** (<https://fairlearn.org/>): Aquesta biblioteca de Python ajuda a avaluar i mitigar biaixos en models d'aprenentatge automàtic. Ofereix mètriques per identificar biaixos segons grups demogràfics i eines per ajustar la paritat.
- **AI Fairness 360** (<https://aif360.res.ibm.com/>), desenvolupada per IBM, o Fairness Indicators de Google (<https://research.google/blog/fairness-indicators-scalable-infrastructure-for-fair-ml-systems/>) proporcionen un conjunt d'algorismes i mètriques per avaluar biaixos en dades i models. És útil per visualitzar la distribució desigual d'atributs com el gènere. Inclou algorismes específicament desenvolupats per la detecció de biaixos, com per exemple: *optimized pre-processing*, *reweighting*, *adversarial debiasing*, *reject option classification*, *disparate impact remover*, *learning fair representations*, *prejudice remover*, *meta fair classifier*, *calibrated equalized odds post-processing*, *equalized odds post-processing*, etc.
- **Google What-If Tool** (<https://pair-code.github.io/what-if-tool/>): Una eina que permet explorar els resultats dels models de ML canviant les entrades de manera interactiva, la qual cosa pot revelar biaixos no esperats.
- **DataSheets for Datasets:** Aquesta metodologia recomana documentar detalls sobre la creació, curació i ús del conjunt de dades, per ajudar a identificar possibles biaixos des de l'origen.
- **Fair Clustering:** Divideix les dades en grups assegurant que els grups són equilibrats pel que fa a la representació de gènere.

5.2. Detecció de biaixos en els models

Eines per a l'anàlisi de models (p. ex. Tècniques d'explicabilitat) ajuden a veure amb quins principis actua l'algoritme. En el cas de les xarxes neuronals, es poden visualitzar els pesos de les connexions de la xarxa entrenada, que tenen un rang que mostren la probabilitat de que hagi punts/regions amb possibles problemàtiques.

Aquestes mètriques i algorismes permeten una visió global i profunda dels biaixos en els models, tant en el moment de la seva creació com durant el desplegament. La combinació de diverses tècniques pot ajudar a fer que el sistema no perpetui les desigualtats.

- **Algorismes d'aprenentatge justos:** utilitzar algorismes dissenyats específicament per mitigar biaixos.
- **Regularització d'igualtat:** aplicar tècniques per penalitzar models que discriminin per gènere i assegurar que les taxes d'error en les prediccions siguin equivalents entre gèneres.

- **Tècniques d'explicabilitat:** prioritzar l'ús de models que siguin interpretables sempre que sigui possible, com els arbres de decisió, per entendre com i per què la IA elabora les decisions. Quan sigui necessari l'ús de models opacs, utilitzar tècniques com LIME o SHAP per entendre com el model elabora decisions. L'atribut d'explicabilitat és el que permet a una IA argumentar els seus propis resultats i aporta molta llum sobre l'existència de possibles biaixos.
- **Mètriques d'equitat:** utilitzar mètriques específiques per avaluar l'equitat del model.
- **Avaluació intensiva del model incloent l'anàlisi de sensibilitat:** avaluar com varien les prediccions del model al modificar lleugerament les dades d'entrada.

Metodologies a considerar en aquesta fase. Són mètodes que modifiquen l'entrenament del model per mitigar el biaix.

- **Mètriques d'equitat:** utilitzar mesures com ara:
 - **Paritat demogràfica:** el percentatge d'individus seleccionats pel model (positius) hauria de ser igual per a tots els grups.
 - **Equal Opportunity:** les taxes de certs i falsos positius han de ser iguals per a tots els grups.
 - **False Positive / Negative Rate:** analitzar si certs gèneres tenen més errors específics.
- **Comparació de resultats creuats:** comparar la precisió, sensibilitat i altres mètriques clau segregant les dades per gènere.
- **Proves A/B:** realitzar proves A/B per avaluar com diferents versions d'un model afecten a diferents gèneres i ajustar els models en conseqüència.
- **Adversarial Debiasing:** utilitzar una xarxa adversària per minimitzar la capacitat del model de predir el gènere a partir de les dades, forçant-lo a ser imparcial.
- **Fairness Constraints:** incloure restriccions d'equitat, com garantir la paritat demogràfica, a la funció de pèrdua durant l'entrenament.
- **Regularització específica:** Afegir termes de regularització que penalitzin els biaixos de gènere (p. ex., a través de la TPR o FPR).

Eines a considerar en aquesta fase:

- **SHAP (SHapley Additive ExPlanations):** Proporcionen explicacions sobre les decisions dels models, ajudant a identificar si un atribut com el gènere té un impacte desproporcionat.
- **LIME (Local Interpretable Model-Agnostic Explanation):** Explica les prediccions d'un model a nivell local, permetent veure si el gènere influeix en casos específics.
- **Counterfactual Fairness:** Comprova si un model faria la mateixa predicció si l'atribut de gènere fos diferent.
- **Bayesian Fairness Models:** Utilitza inferència bayesiana per modelar la distribució de biaixos en funció dels atributs de les dades.
- **Model Cards for Model Reporting** (<https://arxiv.org/abs/1810.03993>): Una guia per documentar models de IA, incloent-hi l'anàlisi de possibles biaixos i recomanacions d'ús responsable.

5.3. Detecció de biaixos en els resultats

- **Monitoreig continu:** implementar sistemes de monitoreig per detectar canvis en el rendiment del model al llarg del temps i identificar possibles desviacions. Per evitar la degeneració del sistema pot ser interessant bloquejar-ne la capacitat d'aprenentatge adaptatiu en continu i treballar per versions, realitzant la fase de retroalimentació en *back-office* de forma periòdica, verificant que el model millora i no esbiaixa abans de posar-lo en producció.
- **Auditories regulars:** realitzar auditories periòdiques per avaluar l'equitat i la transparència del model.
- **Retroalimentació humana:** implicació d'experts humans en el procés de resultats, per identificar biaixos.
- **Transparència:** tenint en compte que la IA no hauria de tenir agència és molt important que pugui explicar com funciona el model per augmentar la confiança d'ús i facilitar la detecció de biaixos. Caldrà mantenir una documentació clara i detallada de com es recopilen les dades, com es processen, com es produeix l'entrenament i com s'ajusten els models.

Metodologies a considerar en aquesta fase. Aquests mètodes actuen sobre les prediccions del model després de l'entrenament:

- **Testing per grups d'usuari:** Provar el sistema amb dades que incloïen diverses representacions del gènere (incloent-hi gèneres no binaris).
- **Avaluació post-deplegament:** Monitoritzar com interactuen els usuaris amb el sistema per detectar patrons discriminatoris emergents.
- **Equalized Odds Post-Processing:** Ajusta les probabilitats del model per garantir que compleix les mètriques d'oportunitats igualades.
- **Calibration Across Groups:** Ajusta els llindars de decisió del model perquè siguin consistents entre els grups.

5.4. Propostes de l'àmbit jurídic

- Protecció jurídic-regulatòria directe del dret a la igualtat dona/home, inspirant-se en els principis ètics de la IA ja existents a nivell UE.
- Aportar al legislador UE escenaris concrets en els que la no protecció directa del dret a la igualtat dona/home deriva en biaixos que perjudiquen la meitat de la ciutadania.
- Exportar al món la necessitat de compartir un codi ètic comú en els valors que protegeixen els drets fonamentals.
- Regular la IA globalment, a partir d'una regulació que tingui presents els drets humans
- Articular elements de canvi cultural que promoguin el compromís amb els valors ètics, més enllà del compliment estricte de la legislació vigent.
- Articular un marc global internacional, un mínim consens regulatori, estandarditzat i transparent que trobi l'equilibri entre innovació i respecte als drets humans, posant el focus en evitar els biaixos i, en especial per la seva representativitat i rellevància, aquells que poden afectar a la meitat de la població mundial.

5.5. Actuacions addicionals

Més enllà de les dades y els models, una IA responsable implica també prendre actuacions fonamentals en altres aspectes com la cultura y la transformació dels processos de les pròpies empreses, posant èmfasi en les següents àrees:

- **Diversitat en els equips:** fomentar equips de desenvolupament i auditoria diversos que incloguin persones de diferents gèneres, antecedents i perspectives per reduir, identificar i mitigar biaixos inconscients
- **Perspectiva interseccional:** considerar altres característiques (com raça, edat o situació socioeconòmica) i entendre com poden interactuar amb el gènere per amplificar els biaixos.
- **Principis ètics:** establir principis ètics i guies justes per al desenvolupament i l'ús de la IA dins de cada empresa.
- **Col·laboració amb experts:** consultar a experts en la matèria i en l'ètica de la IA per garantir desenvolupaments i validacions justes i equitatives.
- **Desenvolupament iteratiu:** Revisar contínuament les dades i el model per corregir biaixos a mesura que es detecten.
- **Feedback continu:** Establir canals de retroalimentació contínua amb comunitats diverses per rebre informació i millorar contínuament els models.
- **Transparència i rendició de comptes:** Utilitzar eines de documentació com *DataSheets* i *Model Cards* per promoure un ús ètic i conscient dels models.
- **Polítiques i normatives:** desenvolupar i adherir-se a polítiques organitzacionals i normatives governamentals que promoguin l'equitat de gènere en el desenvolupament i desplegament de IA.

6. L'àmbit de la comunicació i premsa

1. Representació en les notícies

Exemple:

Quan es cobreixen notícies sobre ciència i tecnologia, és més probable que els mitjans de comunicació citin homes com a experts que dones. Tot i que hi ha moltes dones qualificades en aquests camps, sovint se'ls dona menys visibilitat.

Impacte:

Això perpetua la percepció que els homes són més experts en ciència i tecnologia que les dones, desincentivant les noies joves a interessar-se per aquestes àrees.

Solució:

Promoure la diversitat d'experts en els mitjans de comunicació, buscant activament veus femenines i donant-les visibilitat igualitària en els reportatges sobre ciència i tecnologia.

2. Fotografia i imatges

Exemple:

En articles de premsa sobre negocis i lideratge, les imatges que s'utilitzen solen mostrar més homes en posicions de poder (com ara consellers delegats, directius, etc.) que dones.

Impacte:

Aquesta representació desequilibrada pot reforça l'estereotip que els homes són més aptes per a rols de lideratge i que les dones són menys visibles en aquests àmbits.

Solució:

Seleccionar imatges que representin una diversitat de gènere en les posicions de poder, assegurant-se que les dones líders també siguin visibles en els articles de negocis i lideratge.

3. Titulars i descripcions

Exemple:

Quan una dona aconsegueix un èxit notable, els titulars de les notícies sovint es centren en la seva vida personal o la seva aparença. Per exemple, "Mare de dos fills guanya el premi de ciència" en lloc de simplement destacar el seu èxit professional.

Impacte:

Aquests titulars poden desviar l'atenció dels mèrits professionals de les dones i reforçar l'èmfasi en els seus rols tradicionals de gènere, minimitzant els seus èxits.

Solució:

Redactar titulars amb noms i cognoms i descripcions que es centrin en els assoliments professionals de les dones, igual que es faria amb els homes, evitant referències innecessàries a la seva vida personal o aparença.

4. Biaixos en el llenguatge

Exemple:

Els articles de premsa sovint utilitzen termes diferents per descriure homes i dones. Per exemple, un home pot ser descrit com "determinant" o "assertiu", mentre que una dona amb les mateixes qualitats pot ser etiquetada com "agressiva" o "dominant".

Impacte:

Aquest tipus de llenguatge pot perpetuar estereotips de gènere negatius i afectar la percepció pública de les dones que assumeixen rols de lideratge.

Solució:

Utilitzar un llenguatge neutral i equitatiu quan es descriuen les qualitats de les persones, independentment del seu gènere, i revisar els articles per evitar terminologia esbiaixada.

5. Cobertura esportiva

Exemple:

En la cobertura esportiva, les dones esportistes sovint reben menys atenció mediàtica que els homes. Quan es cobreixen les dones, el focus pot estar més en la seva aparença física o vida personal que en els seus assoliments esportius.

Impacte:

Això pot reduir el reconeixement dels èxits de les dones esportistes i perpetuar la idea que els esports masculins són més importants o interessants que els femenins.

Solució:

Dedicar més espai i recursos a la cobertura dels esports femenins, centrant-se en els assoliments i les capacitats de les esportistes de la mateixa manera que es fa amb els homes.

6. Estereotips en la publicitat

Exemple:

Els anuncis publicitaris sovint retraten les dones en rols domèstics o relacionats amb la cura personal, mentre que els homes apareixen en contextos professionals o d'aventura. Per exemple, un anunci de productes de neteja mostrarà una dona fent les tasques de la llar, mentre que un anunci de cotxes mostrarà un home conduint.

Impacte:

Aquests estereotips poden influir en les expectatives de rols de gènere i perpetuar la idea que certes activitats són "més adequades" per a un gènere que per a un altre.

Solució:

Crear anuncis publicitaris que mostrin homes i dones en una varietat de rols, trencant els estereotips tradicionals de gènere i promovent la igualtat.

7. Cobertura de personatges públics**Exemple:**

Quan es parla de dones polítiques o figures públiques, sovint es fa referència al seu aspecte físic, la seva roba o la seva vida familiar, en comptes de centrar-se en les seves polítiques o assoliments. Per exemple, un article sobre una política podria destacar el seu vestit en lloc de les seves propostes polítiques.

Impacte:

Aquest tipus de cobertura desvia l'atenció dels assoliments i capacitats professionals de les dones, minimitzant el seu impacte i contribucions.

Solució:

Centrar els reportatges sobre figures públiques en les seves contribucions professionals i polítiques, evitant comentaris sobre l'aparença física o la vida personal tret que sigui rellevant per a la notícia.

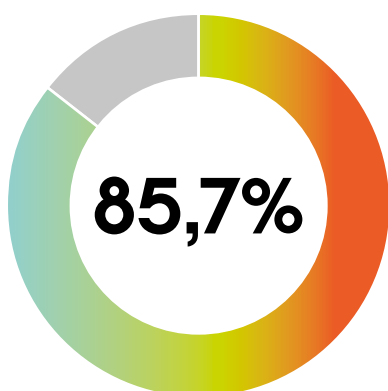
Aquests exemples i solucions mostren com es poden identificar i abordar els biaixos de gènere en la comunicació i la premsa per fomentar una representació més equitativa i justa de les dones.

7. L'àmbit dels departaments de gestió de persones

La IA cada vegada és més present en la gestió del dia a dia de les empreses. Un dels departaments on més comencen a utilitzar-se eines de IA és en els de gestió de persones.

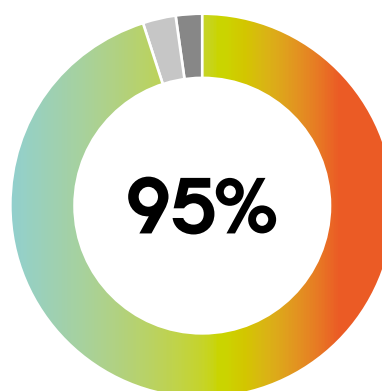
Per tal de poder copsar quina és la situació i veure com i en quin processos s'està utilitzant s'ha realitzat una enquesta que han respost una cinquantena d'empreses pimes i grans empreses que ens aporten les següents dades:

Percentatge d'ús d'IA a la feina



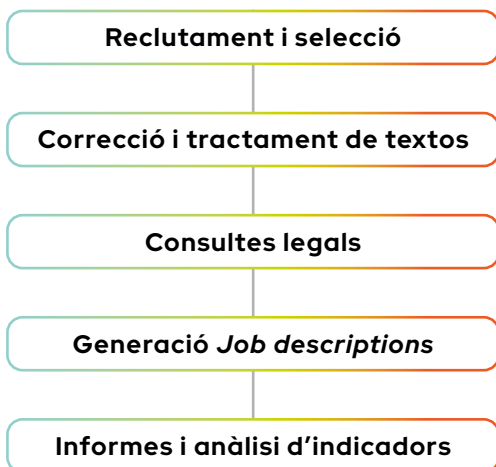
■ Si ■ No

Eines d'IA més utilitzades

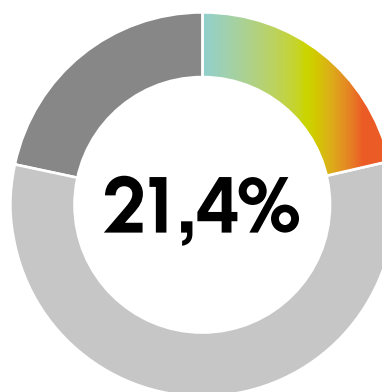


■ Responen ChatGPT o Copilot
■ Utilitzen Chatbot, Power BI, Deepseek o People Analytics
■ No utilitzen cap eina

En quins processos l'utilitzes?



Percentatges de biaixos identificats



■ Si ■ No ■ De vegades

Exemples de biaixos

Respostes amb llenguatge no inclusiu

Les respostes sovint no utilitzen llenguatge inclusiu

El llenguatge no contempla diversitat cultural ni discapacitat

Registre retributiu esbiaixat

Preferència per candidats masculins

Quan es demanen els millors CV, sempre apareixen homes

ChatGPT i llenguatge inclusiu

ChatGPT no acostuma a ser inclusiu en el seu llenguatge

Biaixos a les imatges

Mètodes de correcció

Correcció immediata

Modificació del text amb llenguatge inclusiu

Correcció manual

Revisió constant de les respostes

Adaptació de les respostes

Supervisió humana

Equip humà revisant i corregint

Revisió posterior

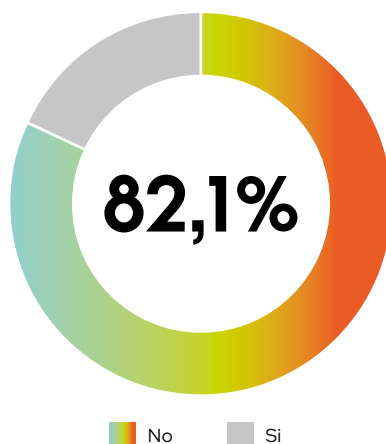
Auditoria de biaixos

Revisió del text després de la resposta

Preguntes addicionals

Fer més preguntes al chat per assegurar la correcció

Empreses que no tenen polítiques d'ús intern de la IA



Conclusions de l'enquesta

- Les empreses estan utilitzant la IA de forma intensiva tot i que majoritàriament no tenen establerta una política d'ús intern amb els riscos que això comporta.
- En l'àmbit de les gestió de persones els processos on més s'utilitza la IA són en la selecció, gestió administrativa i en la part més analítica.
- Aparentment no sembla que s'identifiquin amb facilitat els biaixos de gènere i en cas de detectar-se més enllà de la correcció manual, tampoc es té gaire coneixement d'eines que ho mitiguin de forma més automàtica.



8. Conclusions finals

El desenvolupament i ús d'una IA responsable és un repte multidisciplinari, que combina aspectes tècnics, ètics, de negoci i socials.

La complexitat dels sistemes de IA i la varietat dels possibles biaixos a banda del de gènere, fa que no existeixi una solució universal per a totes les empreses.

A nivell empresarial, serà necessari dissenyar estratègies a mida, que considerin els reptes específics de l'estat de cada empresa (indústria, negoci, funcions, usuaris, etc.), així com el potencial impacte i ús de la IA i els recursos disponibles.

En qualsevol cas, la clau està en l'exercici responsable de la professió i la implementació d'un procés iteratiu i continu que permeti millorar aquestes estratègies i accions a mitjà i llarg termini, que incloguin formació dels equips que utilitzen les eines de IA, marc d'ús /regulació interna clara i coneixement de les eines que permeten mitigar/identificar els biaixos en processos concrets de l'empresa i que són a la vegada crítics.

No obstant això, sí que hi ha alguns principis generals que són d'aplicació universal i poden contribuir a mitigar la construcció i desplegament de les IAs esbiaixades, i especialment amb biaixos de gènere.

L'impacte de la IA en la perpetuació de baixos de gènere és significatiu, però no insuperable. A través de la diversificació de dades, auditories d'algoritmes, transparència, educació, compromís ètic, regulació i equips paritaris és possible mitigar aquests biaixos i promoure una IA més equitativa i inclusiva. Les mesures proposades no només beneficien a les dones, sinó que també milloren la qualitat i justícia dels sistemes de IA per a tota la societat.

Descarrega't l'informe



fòrum Beijing+30

#BeijingBCN

forumigualtatbeijingbcn.cat



Entitats organitzadores:

odee
observatori dona empresa economia

 **Cambra**
de Comerç de Barcelona

 **Generalitat**
de Catalunya