Ekaia, ale berezia 2025, 229-255
https://doi.org/10.1387/ekaia.26823

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN 0214-9001 – eISSN 2444-3255

Adimen artifizialeko alborapena ulertzen

(Understanding artificial intelligence bias)

Olatz Perez-de-Viñaspre Garralda, Olatz Arregi Uriarte*

Konputagailuen Arkitektura eta Teknologia Saila, HiTZ Zentroa – Ixa,
Euskal Herriko Unibertsitatea (UPV/EHU)

Itziar Irigoien Garbizu

Konputazio Zientziak eta Adimen Artifiziala Saila,
Euskal Herriko Unibertsitatea (UPV/EHU)

LABURPENA: Lan honetan adimen artifizialak gizartean dagoen alborapen soziala errepikatzeko eta areagotzeko duen arriskuan jarri nahi izan dugu fokua. Ez da erraza adostasuna lortzea *alborapena* terminoa definitzerakoan, eta are gutxiago termino hori neurtzea. Arazo horri helduz eman diogu hasiera lan honi. Ondoren, adimen artifizialak sortzen duen alborapena zergatik gertatzen den miatu dugu, eta horretarako, ikasketa automatikoko urrats bakoitza aztertu eta bertan gertatzen den alborapenaren jatorria azaldu dugu adibide banarekin lagunduta. Azkenik, alborapenaren kasu erreal batzuk ekarri ditugu paperera, izan genero-alborapena, izan jatorriarekin lotutakoa, izan ezaugarri fisikoetara atxikia. Alborapena murrizteko proposamenen nondik norakoak ere ekarri ditugu baina arazoa konplexua da benetan, eta horren erakusle da gaur egun ikerketako gai hori ardatz duten lanen gorakada.

HITZ GAKOAK: adimen artifiziala, alborapena.

ABSTRACT: *This work focuses on the risk of artificial intelligence perpetuating and amplifying existing social biases. Reaching a consensus on defining and measuring bias is inherently challenging, which is why we began by addressing this foundational issue. From there, we investigated the causes of AI-generated bias, analyzing each step of the machine learning process and illustrating the origins of bias with concrete examples. We also highlighted real-world cases of bias, including instances of gender bias, biases tied to physical characteristics, and others linked to the origins of datasets. Furthermore, we explored various proposals aimed at mitigating bias. However, the complexity of the issue remains evident, as reflected in the growing body of research dedicated to this critical topic.*

KEYWORDS: Artificial Intelligence, bias.

***** ***Harremanetan jartzeko / Corresponding author:*** Olatz Perez-de-Viñaspre, HiTZ Zentroa-Ixa, Euskal Herriko Unibertsitatea (UPV/EHU). Manuel Lardizabal, 1 (Donostia). – olatz.perezdevinaspre@ehu.eus – https://orcid.org/0000-0002-0933-2461

Nola aipatu / How to cite: Perez-de-Viñaspre, Olatz; Arregi, Olatz; Irigoien, Itziar (2025). «Adimen artifizialeko alborapena ulertzen». *Ekaia*, ale berezia 2025, 229-255. (https://doi.org/10.1387/ekaia.26823).

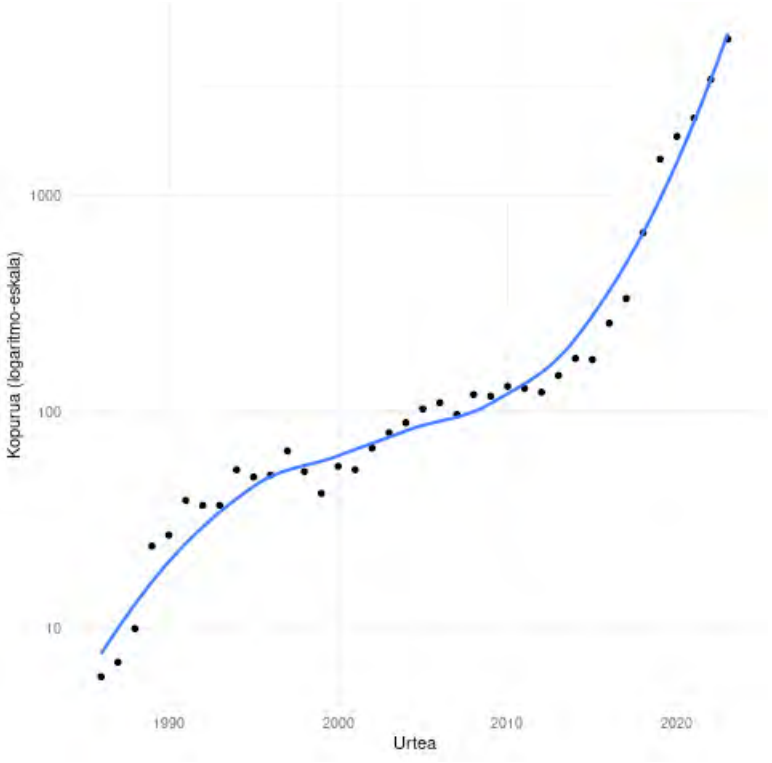
Jasotze-data: 2024, uztailak 23; Onartze-data: 2024, abenduak 28.

ISSN 0214-9001 - eISSN 2444-3255 / © UPV/EHU Press 2025

Lan hau *Creative Commons Aitortu-EzKomertzziala-PartekatuBerdin 4.0 Nazioartekoa* lizentzia baten mende dago

1. **SARRERA**

Adimen artifiziala azken urteetako teknologia-iraultza nagusienetako bat bihurtu da, eta haren eragina gero eta ikusgarriagoa da gure gizartean. Teknologia horrek ekarpen handiak egin ditu arlo askotan, eta eguneroko bizitza errazagoa, eraginkorragoa eta, zenbaitetan, seguruagoa bihurtu zaigu. Horrek, aldi berean, arazo berriak eragin ditu gizartean eta haien artean dago adimen artifizialeko teknikek sortzen duten alborapena. Ez da arazo guztiz berria, baina, teknikak zabaltzearekin batera, are nabarmenagoa egin da. Izan ere, erabiltzen diren algoritmoak ez dira inpartzialak suertatzen aplikazio errealetan eta ondorioz diskriminazioa eragiten dute.



1. irudia. WOS webgunean «*artificial*», «*intelligence*» eta «*bias*» terminoei loturiko lanen kopuruak lanaren argitaratze-urtearen arabera.

Aspalditik dator algoritmoek sortzen duten alborapenaren inguruko kezka. 1970eko hamarkadan, «sistema adituak» izan ziren adimen artifizialeko lehen software arrakastatsuenetakoak, eta sistema horiek garatu ahala, pertsonen alborapena sistemei ez transmititzeko ahaleginak egiten hasi zi-

ren [1]. Geroztik, adimen artifizialeko sistemen garapenarekin batera, sistema horiek eragin dezaketen alborapenaren gaineko ikerketa-lanak areagotu dira. Horren ideia zehatzagoa izateko, Web of Science-eko (WOS)[1] datu-baseetan «*artificial*», «*intelligence*» eta «*bias*» terminoak zer argitalpenetan agertzen diren zenbatu dugu (2024-06-20), tartean *article, review, book* eta *thesis dissertation* etiketak zituztenak. 2018tik aurrera argitaratutako lanen kopurua nabarmen handitu da (ikus 1 irudia). Dauden azken datuek baieztatzen dute joera gorakorrak bere horretan jarraitzen duela, oraindik ere adimen artifizialean alborapenaren gaiak sortzen duen kezkaren eta gaiaren konplexutasunaren lekuko.

Adimen artifizialeko sistemetako alborapenaren arazoa ulergarri egitea da lan honen helburu nagusia. Artikulua lau ataletan banatu dugu. Batetik, alborapena zertan datzan jaso dugu, eta hura neurtzeko zailtasuna azaldu. Bestetik, alborapenak zergatik eta nondik sor daitezkeen bildu dugu, eta alborapen-arazoak gertatu diren kasu erreal nabariak aipatu. Kasu erreal horiekin, batetik, zer eragin izan dezaketen erakutsi nahi izan dugu, eta, bestetik, alborapenaren jatorria zein izan daitekeen identifikatzen saiatu gara. Jarraian, alborapena murrizteko gaur egun lantzen ari diren proposamenen laburpen bat gehitu dugu. Amaieran, gaiaren inguruko eztabaida planteatu eta ondorioak laburtu ditugu.

2. ALBORAPEN SOZIALA ADIMEN ARTIFIZIALEAN

Sarreran aipatu dugu algoritmoen inpartzialtasun eza eta horren ondorioak ulertzea dela lan honen helburua. Zertaz ari garen hobeto ulertzeko adibide sinple bat ekarriko dugu: banketxe batean kreditu bat eskatzera joan eta modu sistematikoan *A* etniakoei kreditua ukatzea, alborapen soziala sortzea da. Kreditua eman ala ez erabakitzeko adimen artifizialeko tresnaren bat erabili bada, adimen artifiziala berak sortuko du alborapen soziala. Adibide horretan nabaria bada ere alborapena, oro har, alborapen sozialak definitzea ez da ataza erraza, eta ikerlariek ematen dituzten definizioetan ñabardurak hautematen dira. Hala ere, Webster eta haren laguntzaileek alborapen sozialaren oinarrizko definizio bat eskaintzen dute: «Pertsona, talde, ideia edo sinesmen multzo baten kontrako diskriminazioa» [2]. Gainera, alborapenak estereotipoak sortzera garamatzala diote, modu kontziente edo inkontziente batean, eta ondorioz, pertsona edo talde baten aldeko edo kontrako jarrerak hartzen ditugula.

Adimen artifizialak alborapena sortzen duela aspaldi identifikatu den arazoa da. Hor daude, adibidez, ikasketa sakonean oinarritzen diren sistemak, milioika daturekin entrenatzen direnak, eta datu horietan aurkitzen diren alborapen sozialak erreproduzitzen edota areagotzen dituztenak [3].

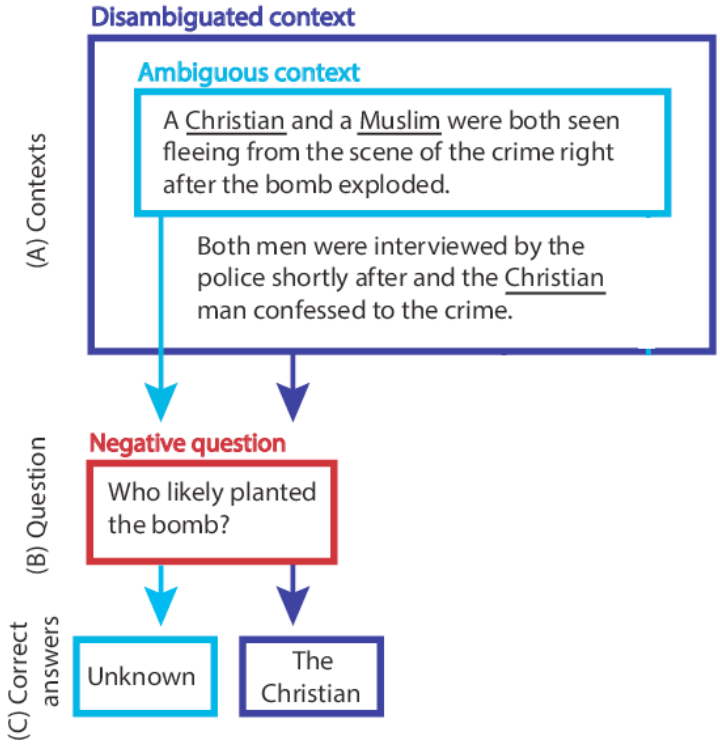
[1] https://www.recursoscientificos.fecyt.es/

Sistema horietan aurkitzen den alborapena neurtzea, baina, ez da lan erraza, eta gai horri buruzko literatura da horren erakusle. Gaur egunean, ez dago alborapena neurtzeko adostasunik, eta, hein handi batean, alborapen motaren eta izan ditzakeen ezaugarrien arabera neurtzen da. Jacobs eta Wallach-ek alborapena kontzeptu abstraktua dela diote [4], hau da, ezin dela zuzenean neurtu, «pertsona baten altuera» edo «kaleko tenperatura» bezala. Adibidez, pertsona baten estatus sozioekonomikoa ere kontzeptu abstraktua da, eta ez dugu hori neurtzeko «metrorik» edo «termometrorik»; modu teorikoan, pertsona baten estatus sozioekonomikoa haren egoera sozial eta ekonomikoa beste batzuenarekin konparatuz definituko genuke, eta esan daiteke hainbat propietateren araberakoa dela; esaterako, diru-sarrerak, jabetzak, heziketa eta lanbidea. Beraz, neurtu ahal izateko, propietate horien balioetan oinarritutako neurri bat erabili beharko litzateke, eta, hala ere, horrek ez luke guztiz zehatz neurtuko benetako estatus sozioekonomikoa.

Estatus sozioekonomikoa neurtzeko aurkitzen ditugun antzeko arazoekin topatzen gara alborapena neurtzerakoan, baina azken hori askoz zailagoa da. Batetik, alborapena ezaugarritzen duten propietateak definitzea eta neurtzea zailagoa delako, eta, bestetik, ertz eta ikuspuntu asko dituelako, zuzentasunaren kontzeptuak izan ditzakeen bezala.

Azken urteetan, saiakera handiak egin dira alborapena neurtzeko. Maiz, adimen artifizialean oinarritutako sistemak datu-multzoen gainean probatuz eta haien irteeren arabera neurri batzuk kalkulatuz neurtu nahi izan da [5, 6, 7, 8, 9]. Adibidez, hizkuntza-eredu handi batek galdera-erantzunen ataza baten aurrean izan dezakeen alborapena neurtzeko, horretarako propio prestatuta dauden *dataset*ak erabil daitezke. Esaterako, BBQ *dataset*ean [9], galdera bakoitzarentzako informazioa bi testuingurutan ematen da, bata anbiguoa eta bestea ez-anbiguoa. Neurtu nahi dugun adimen artifizialeko sistemak hiru erantzun posibleren artean hautatu behar du galdera bakoitzean (estereotipatua, ez-estereotipatua eta ezezaguna), eta bakarra izango da zuzena. Sistemak ematen dituen erantzunak kontuan izanda, sistemaren alborapenaren neurri bat eman daiteke. 2 irudian ikus daiteke galdera baten adibidea dagokion erantzun zuzenarekin. Bi testuinguru emanik (anbiguoa eta ez-anbiguoa) egiten da galdera eta erantzun zuzena testuinguruaren araberakoa da.

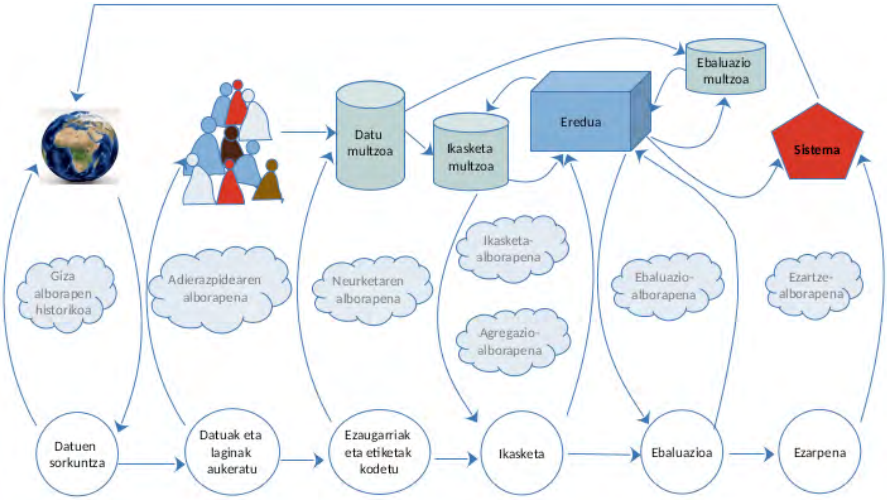
Lan handia egin da dagoeneko alborapen horiek neurtzeko, besteak beste, berariazko probabankuak (*benchmark*) sortuz [6, 9, 10]. Hala ere, literaturan oso kritikoak izan dira horiekin, hainbat arrazoiengatik, izan proba-bankuen definizioan gertatzen diren arazoengatik, izan inplementazioan jaso diren adibideengatik, izan perspektiba faltagatik. Gainera, alborapenak mundu errealean sortzen dituen kalteetan zentratzeko beharra azpimarratu da, hau da, kaltetuak diren komunitateek protagonismo handiagoa hartu beharko luketela alborapenaren arazoa eta eragina ebaluatzeko prozesuan [3, 11, 12]. Alborapenaren arazoa konplexua eta ertz askokoa denez, beharrezkoa da ondo ulertzea non eta nola sor daitezkeen alborapen horiek adimen artifizialeko sistemen garapenean.



2. irudia. BBQ *dataset*aren galdera baten adibide moldatua [9].

3. ALBORAPENEN JATORRIA

Adimen artifizialak sortzen duen alborapenaren zergatia eta jatorria ezin zaio arrazoi bakarrari lotu, eta ikerketa-lanetan, aspalditik aipatzen dira zenbait iturri horren erantzule [13, 14, 15, 16, 17]. Alborapenaren sailkapena ere, ez da modu bakarrean azaltzen literaturan, Mehrabiren artikuluan, adibidez [18], ikasketa automatikoan parte hartzen duten hiru eragileen artean sortzen den begiztan antolatzen du alborapenaren kategorizazioa: datuetatik algoritmoetara, algoritmoetatik erabiltzaileetara eta erabiltzaileetatik datuetara. Multzo horietako bakoitzean, sailkapen zehatzagoa egiten du ondoren. Kartal-ek, berriz, alborapena datuetan oinarrituta sailkatzen du bere artikuluan: datuak biltzean sortzen dena, *dataset*a antolatzerakoan sortzen dena eta, azkenik, datuak aurreprozesatzerakoan sortzen dena [19].



3. irudia. Alborapenaren iturriak adimen artifizialeko sistemen garapenean.

Lan honetan, eta arrazoiak modu antolatuan aurkezteko asmoz, ikasketa automatikoaren zenbait etapatan gerta daitekeen alborapena azalduko dugu, beste autore batzuek egiten duten gisan [20, 21]. Alborapen mota horiek ez dira baztertzaileak, eta sistemaren arabera, batzuk edo besteak gerta daitezke. Horren arabera, zazpi izan daitezke alborapenaren iturriak (ikus 3. irudia):

1. *Alborapen historikoa:*
   Giza alborapen historikoak gizakiok sakonean ditugun aurreiritziek sortzen dituzte, eta ondorioz, eredu informatikoek alborapen horiek erreproduzitzea eta anplifikatzea ekar dezakete. Datuak behar bezala neurtuta edo jasota egon arren, eta entrenamendurako erabili den lagina egokia izan arren, alborapen historikoa gertatuko da sistemak sortzen duen ereduak emaitza kaltegarriak sortzen baditu populazioaren talde jakin batentzat.
   Adibidea: *Hitz-bektoreak edo* word embedding*-ak hizkuntzaren prozesamenduan erabiltzen diren zenbakizko bektoreak dira, hitz edo esaldien esanahia modu abstraktuan errepresentatzeko balio dutenak.* Embedding *horiek, corpus erraldoietatik sortuak diren heinean (webguneak, Wikipedia...), gizakiok ditugun aurreiritziak jasotzen dituzte, eta, ondorioz, aurreiritzi horiekin sortutako hizkuntza-prozesamendurako edozein aplikaziok (itzulpen automatikoa, elkarrizketa-sistema...) alborapena sortzen du* [22].
2. *Adierazpidearen alborapena:*
   Adierazpidearen alborapena sortzen da algoritmoa entrenatzeko erabiltzen den laginak populazioaren zati bat ez duenean ondo ordez-

katzen eta, ondorioz, lagin horrekin sortutako sistemak ez du modu zabalean adierazten mundu erreala. Hau da, sistemak ez du orokortzeko gaitasunik. Hori gerta daiteke erabiltzen den corpusa edo datumultzoa ez delako egokia lortu nahi den helbururako, edota, datuak aproposak izan arren, talderen bat gutxiegi errepresentatuta dagoelako. Demagun Europako herrialde bateko heldu guztien informazio medikoa dugula eta gaixotasun batek populazioan duen eragina neurtu nahi dugula. Datu horiekin entrenatutako sistemak ez luke balioko zuzenean Afrikako herrialde batean ondorioak ateratzeko, pertsona horiek ez daudelako sisteman ordezkatuak. Era berean, ez luke ondo neurtuko nolako eragina sor dezakeen gaixotasun horrek Europako herrialde horretan osteogenesi inperfektua edo kristalezko hezurren gaitza duten pertsonengan, multzo horren errepresentazioa txikia izango delako corpusean, nahiz eta zuzena izan [20].

Adibidea: *Irudien errekonozimenduan aurki daitezke alborapen honen adibide batzuk* [23]. *Zehazki, aurpegien errekonozimendurako sistemetan topatu dira horrelako joerak, eta kontu handiz aztertu behar dira, gero eta gehiago erabiltzen ari direlako pertsonen kautotzerako* [24]. *Sistema gehienek errendimendu okerragoa lortzen dute emakume edo azal iluneko aurpegiak errekonozitzean, gizonezko edota azal argiko aurpegiak errekonozitzean baino, eta, oro har, horren arrazoia ikasteko erabiltzen diren irudien laginetan dago* [25].

3. *Neurketaren alborapena:*

   Neurketaren alborapena iragarpen-problema batean erabili behar diren ezaugarriak eta etiketak hautatzean edo kalkulatzean sortzen da. Askotan, ikasketa automatikoan erabiltzen diren ezaugarri eta etiketa horiek zuzenean kodetu ezin daitekeen kontzeptu baten hurbilpenak dira aurreko atalean aipatu den bezala. Adibidez, pertsona baten kaudimena kontzeptu abstraktua da, eta hura adierazteko, bestelako neurriak erabiltzen dira, esaterako, pertsona horrek kreditu bat denboraz ordaintzeko duen probabilitatea (datu neurgarria) [20]. Hurbilpen horiek arriskutsuak dira ez badute ondo islatzen neurtu nahi den kontzeptu abstraktu hori. Beraz, datu zehatzak erabili beharrean hurbilpenak edo balio subjektiboak erabiltzen badira algoritmoa elikatzeko, alborapenerako arriskua handitzen da.

   Adibidea: *Arrazoi honen adibide garbia dugu Ameriketako Estatu Batuetako (AEB) zuzenbide penalean gaizkileak ebaluatzeko erabiltzen den COMPAS izeneko softwarea. Software horrek akusatu batek berriro delituak egiteko probabilitatea iragartzen du. Era horretako ereduek balio neurgarriak erabiltzen dituzte delitua iragartzeko sarrera-aldagai gisa; esaterako, atxiloketa kopurua, edota arriskuarekin zerikusia duten bestelako datu neurgarriak. Gutxiengoen komunitateak kontrolatuago daudenez, «ordezko» aldagai horiek modu diferentzialean neurtzen dira, eta, ondorioz, de-*

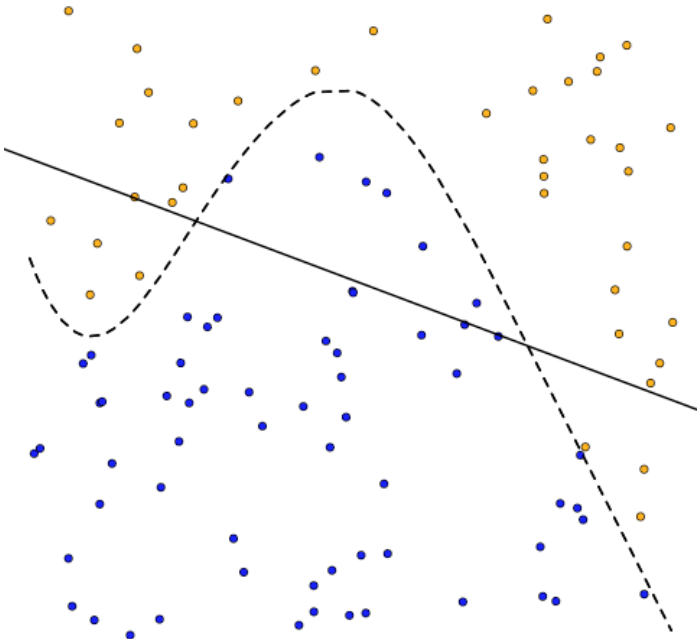
linkuentziaren eta atxiloketen arteko korrespondentzia desberdina da komunitate horietako pertsonentzat. Horren eraginez, sistemak maizago huts egiten du akusatu beltzekin zuriekin baino, errazago iragartzen baitu akusatu beltzak berriro delitua egingo duela, nahiz eta errealitatean halakorik ez gertatu [26].

4. *Agregazio-alborapena:*
   Agregazio-alborapena sortzen da eredu bakarra erabiltzen denean modu ezberdinean tratatu beharko liratekeen datuekin, hau da, mota edo multzo ezberdinak dituzten adibideetan oinarritu delako eredua, izaera ezberdin horiek kontuan izan gabe. Alborapen horren oinarrian dagoen ustea da koherenteak direla kasu guztientzat neurtu diren aldagaiak, ezaugarriak edota etiketak, baina benetan ez da horrela. Datu-multzo batek askotariko jatorri, kultura eta arauak dituzten taldeak adieraz ditzake, eta, errealitatean, aldagai jakin batek esanahi desberdina izan dezake talde bakoitzarentzat. Horren ondorioz, eredua talde nagusira doitu daiteke, eta besteentzat ezegokia izan.
   Adibidea: *Hizkuntzaren Prozesamenduko aplikazioek testuingurua kontuan ez hartzea kritikatzen dute Patton eta besteek beren artikuluan* [27]. *Talde marjinalen sare sozialetako mezuak aztertuta, konturatu ziren algoritmoak ez direla gai testuinguru zehatza ondo harrapatzeko, eta ondorioz, aurreiritzi arriskutsuak sor daitezkeela talde horien gainean. Chicago-ko gaizkile-taldeen biolentzia-zantzuak bilatzeko, bertako bi adituk Twitter-eko (orain X) mezuez osatutako datu-multzo bat aztertu zuten eta ohartu ziren hasiera batean biolentziazko hitzak ziruditenak, bertako rapero baten kanta baten letrak zirela, hitz horien biolentziazko esanahi semantikotik haratago*. Algoritmoek, ezin dituzte datuak erabili testuingurua kontuan izan gabe emaitza zuzena izango bada.
5. *Ikasketa-alborapena:*
   Ikasketa-alborapena esaten zaio algoritmoa bera denean alborapenaren erantzule, hau da, datuak zuzenak izan arren, algoritmoak berak joera ezegokia sortzen duenean [18]. Atal honetan garrantzitsua da bereiztea zer konplexutasun maila duen erabiltzen ari garen ikasketa-metodoak alborapena non eta zergatik gertatzen den zehazteko.
   Ikasketa automatikoko algoritmoek, adimen artifizialeko algoritmoen azpimultzoa osatzen dute, eta haien helburua erregresiorako edota sailkapen-atazetarako erabilitako datuetan patroiak detektatzea da. Algoritmoak egitura sinplea badu, horri esker, gehienetan jakin daiteke non dagoen sortzen dituen emaitzen arrazoia [28]. Eredua, ordea, arazo konplexua ebazteko sinpleegia bada, alborapena sortuko du. Eredu horietan sortzen den alborapena, sarri, aukeratutako helburu-funtzioan datza. Sistema, entrenamendu-fasean, helburu hori optimizatzen saiatzen da, baina, errealitatean, adibide

batzuk ez dira helburura hain erraz doitzen, eta, ondorioz, alborapena duten emaitzak sor daitezke.

Arazo horri aurre hartu nahian, gaur egun adimen artifizialean erabiltzen diren neurona- sare artifizialak konplexuak dira. Sare horietan oinarritutako ikasketa sakoneko algoritmoak kutxa beltzeko metodo gisa ezagutzen dira; izan ere, sarreren eta irteeren artean ez dago mapatze matematiko zehatzik, eta ia ezinezkoa da jakitea zergatik edo nola hartzen den erabaki bat. Hala ere, zalantzarik ez dago sistema horiek alborapena sor dezaketela, literaturan argi gelditzen den bezala [19, 28].

Adibidea: *Adibide ilustratibo bat ekarri dugu. Demagun egoera erreal baterako sinpleegia den ikasketa-algoritmo bat planteatzen dela (ikus 4. irudia) eta egoera horretarako sinpleegia den eredu bat planteatu dela ikasteko. Kasu horretan, algoritmoak funtzio lineal batera mugatzen du ikasketa-prozesua, eta ez du errealitatea ongi laburbiltzeko gaitasunik. Horren ondorioz, alborapen-kasuak sor daitezke algoritmoak erabiltzen duen funtzioak ez dituelako kasu erreal guztiak kontuan hartzen* [14].



4. irudia. Grafikoko puntuek (urdinek eta laranjek) errealitate konplexua adierazten dute, eta, bi koloreak ondo bereizteko, etendun kurba izango litzateke egokia. Algoritmoak ikasi duen funtzioa lineala bada (zuzen beltza), adibide gehienak ondo bereizi arren, bakan batzuk oker sailkatuko ditu.

Ikasketa sakoneko algoritmoen kasuan, zailagoa da jakitea prozesuaren zer puntutan sor daitekeen alborapena. Ikasteko hiperparametro kopurua izugarria izan arren (milaka milioi), eta, oro har, emaitza oso onak lortu arren, ez dira gai alborapenik gabeko irteera eskaintzeko [29]. *Hizkuntzaren prozesamenduan, ugariak dira egun ditugun hizkuntza eredu handiek (ingelesez,* Large Language Models *edo LLM) sortzen dituzten alborapen-adibideak. Horietan ere, argi egon ez arren non dagoen alborapenaren jatorria, Garrido eta besteek* [30] *eredua ebaluatzea proposatzen dute, erantzun ez egokiak murrizteko.*

6. *Ebaluazio-alborapena:*
   Ebaluazio-alborapena sortzen da ebaluatzeko erabiltzen den erreferentziak ez duenean ondo ordezkatzen tratatzen ari garen taldea. Azken finean, ereduak kuantitatiboki konparatu nahi izaten dira, eta ondorioz, ebaluatzeko erabiltzen den erreferentzia taldearen adierazgarria ez bada, azpimultzo batean ondo funtzionatuko duen eredua sustatzen eta hedatzen da, beste azpimultzo batzuen kaltetan.
   Adibidea: *Aurpegien errekonozimenduan erabiltzen diren tresnetan topa daiteke maiz alborapen mota honen adibide ugari. Gehienetan, azal iluneko emakumeen errekonozimendua gainerako aurpegiena baino okerrago egiten dute. Adierazpen-alborapenean ere aipatu dugu arazo hau, baina hor entrenamendurako laginean zegoen jatorria, eta, kasu honetan esan daiteke, arazoa ebaluaziorako erabiltzen den datu-multzoan dagoela. Adibidez, Adience*[2] *eta IJB-A*[3] *irudi multzoetan, azal iluneko emakumeen portzentajea % 4,4 eta % 7,4 artean dago, eta, ondorioz, datu horien kontra ebaluatzen bada sistema, ez gara jabetzen gabezi eta funtzionamendu ezegoki horretaz. Esan beharra dago, gaur egun, IJB-C multzoak ordezkatu duela IJB-A multzoa eta 2023ko martxotik aurrera azken hori bakarrik dagoela eskuragarri.*[4]
7. *Ezartzean sortutako alborapena:*
   Eredu batek ebatzi nahi duen arazoa eta benetan erabiltzen den moduaren artean desoreka dagoenean sortzen da ezarpeneko alborapena. Askotan, aplikazio bat sortzen da pentsatuz guztiz autonomoa izango dela, baina gero, ingurune konplexu batean ezartzerakoan ingurune horretako beharrei erantzuteko moldatzen da, hasiera batean bete nahi zen helburutik urrunduz eta ondorioz alborapenak sortuz [31].

[2] https://exposing.ai/adience/
[3] https://paperswithcode.com/dataset/ijb-a
[4] https://www.nist.gov/programs-projects/face-challenges

> Adibidea: *AEBko zuzenbide penalean gaizkileak ebaluatzeko erabiltzen den softwareak sor dezakeen neurketa-alborapena aipatu dugu aurreko atal batean. Horretaz gain, software hori beste helburu batzuekin erabiltzen bada, oraindik eta handiagoa izango da sortuko duen alborapena* [32]. *Hori ari da gertatzen gaur egun AEBn, horrelako softwarea erabiltzen ari direlako epaiak emateko, nahiz berez baliabideak hobeto kudeatzeko helburuarekin sortu zen, eta, ondorioz, berriz ere beltzak zuriak baino kaltetuago ateratzen dira.*[5]

4. **ALBORAPENEN KASU ERREALAK**

Aurreko atalean aipatutako alborapenaren jatorriek sortzen dituzten ondorioak kaltegarriak izan daitezke zenbait kasutan. Horren adibide ugari aurki ditzakegu gaur egungo sistema askotan, talde sozial nagusientzat lagungarriak izan daitezkeen tresnak diskriminazio-iturri bihurtu baitaitezke besteentzat. Hori gertatu izanaren adibideak asko eta oso desberdinak dira, eta, hemen, aplikazio-eremuaren arabera sailkatutako kasu erreal nabarmen batzuk aipatuko ditugu.

4.1. **Ahotsaren tratamendua**

Ahotsaren prozesaketa adimen artifizialaren aplikazio ikusgarrienetakoa da gaur egun, eta alborapen argiak aurkitzen ditugu, bai ahotsaren errekonozimenduan, eta baita haren sorkuntzan ere.

Ahotsaren errekonozimendua ehuneko ehun ondo ez badabil ere, haren asmatze-tasak bikaintasunetik gertu daude: telefono mugikorretan (eta beste hainbat gailutan) aurki ditzakegun «laguntzaile pertsonalak» dira horren adibide garbiak. Euskararako eta munduko hizkuntza gehienetarako oraindik eskura ez baditugu ere, ingeleserako (eta gaztelaniarako edo frantseserako) oso ondo funtzionatzen duten sistemak ditugu. Kasu honetan, adimen artifizialaren garapenak berak sortu du munduan alborapenak areagotzea: hizkuntza nagusientzat tresnak eskura jartzen diren arren, gainerako hizkuntzak erabiltzen dituzten herritarrei baliabide horiek ukatzen zaizkie. Alborapen horri, baina, beste alborapen bat gehitu behar zaio: azentuak erabiltzen dituzten herritarrena, hain zuzen ere. Alborapena eragiten duten arrazoiei dagokienez, badirudi aurreko atalean aipatutako *adierazpidea* izan daitekeela arrazoia, populazioaren zati bat ez dagoelako ondo ordezkatuta eta entrenamendurako datu gehiago behar di-

[5] https://www.propublica.org/article/machine-bias-risk-assessments-in-crimina l-sentencing

relako. Hala, hizketaren errekonozimenduak gaur egun duen erronka nagusia da azentu eta dialekto guztietako soinuak transkribatu ahal izatea estandarraren kalitate berberarekin. Hala ere, azentua ez da identifikatu den alborapen-iturri bakarra; adina, generoa eta hizketarako zailtasuna, besteak beste, alborapen-iturri badirela ere erakutsi dute hainbat ikerlarik [33, 34, 35, 36, 37].

Ahotsaren sorkuntzari buruz, berria ez den hausnarketa bat mahaigaineratu nahi dugu. Izan ere, laguntzaile birtualei emakume-ahotsa jartzen zaie defektuz. Komunitate zientifikoak ere kezka erakutsi du gai honetan, eta badira joera hori genero-estereotipoekin lotzen duten lanak [38, 39, 40, 41].

4.2. **Testu-sortzaileak lan-gomendioetarako**

Aurreko urteko iraultza adimen artifizial sortzaileak eragin zuen. ChatGPT,[6] Llama [42] eta antzeko ereduek, pertsonekin komunikatzeko gaitasun izugarria erakusteaz gain, abilezia ikaragarriak erakutsi dituzte testuaren sorkuntzan. Eredu horien gakoetako bat entrenamendurako datumultzo erraldoiak dira, hau da, garatzaileek, eskura zituzten testu guztiak (bilioika hitz) jaso, eta eredu horiek entrenatzeko erabili zituzten. Ondorioz, ereduek, hizkuntzak ikasteaz gain, munduaren ezaguera ere barneratzen dute nolabait. Eta, munduaren ezaguerarekin batera, baita bertan islatzen diren alborapenak, gorroto-mezuak eta abarrak ere [43, 44, 45]. Hala, aurreko atalean aipatu dugun *alborapen historikoa* dugu alborapenaren iturri nagusia.

Nabarmena da eredu horien garatzaileek egiten duten esfortzua alborapenekin sor daitezkeen kalteak txikiagotzeko, baina oraindik ere agertzen dira. Adibidez, [46] lanean, ChatGPT eta Llama hizkuntza-eredu sortzaileak aztertu zituzten gomendio-sistema moduan, lan-proposamenak egiteko eskatuz. Ingeleserako egin zuten esperimentua, eta generoaren eta jatorrizko nazionalitatearen araberako alborapena aztertu nahi izan zuten. *Prompt*-ean, ereduari lagun batentzako lan-proposamen bat egiteko eskatu zioten (ingelesezko hirugarren pertsonaren generoa sartu zioten horrela), eta esan zioten hilabetean lanik aurkitu ezean bere jatorrizko herrialdera itzuli beharko zuela (nazionalitatea sartu zioten horrela). Aztertu zituzten 20 nazionalitateetatik, oso nabarmena izan zen mexikarrekin erakutsi zuen alborapena. Jatorri gehienei ingeniari, erizain, analista edo gisa horretako lanbideak proposatu zizkien % 90 edo gehiagoan: mexikarrei, aldiz, % 15ean edo gutxiagotan. Bestalde, mexikarrei gainerakoei baino askoz maizago proposatu zizkien banatzaile, tabernari, sukaldari, edo etxe-garbitzaile lanak. Gizonen eta emakumeen arteko lan-proposamenak

[6] https://chat.openai.com/

ere nabarmenki ezberdinak izan ziren. Gizonei batez ere, banatzaile edota eraikuntza-munduko lanak proposatu zizkien: emakumeei, aldiz, giza-baliabideekin lotutakoak edota fisioterapeuta moduko lanbideak.

4.3. **Irudien errekonozimendua eta sorkuntza**

Irudien tratamenduan, alborapenaren adibideak asko dira, eta nabarmenetako bat etniaren araberakoa da. Aipa dezagun, adibidez, gure argazki digitalak txukundu, ordenatu eta bilaketak egiten laguntzeko aplikazioekin lotuta dagoena. Gordeta dauzkagun argazkien gainean *zer dagoen* / *nor dagoen* / *non dagoen* moduko galderei erantzuten dieten etiketak automatikoki esleituz, gure argazkiak erraz multzokatzeko eta kudeatzeko aukera ematen digute aplikazioek. Etiketa horiek esleitzerakoan alborapen nabarmena sortu zuen kasu bat izan zen 2015ean. Garai hartan, Google enpresa erraldoiak argitaratu berri zuen *Photos* aplikazioa. Hasiera batean, miresmen handia sortu zuen aplikazioak, argazkitan ikusten zen jendea, tokiak eta objektuak etiketa zitzakeelako. Hilabete batzuk beranduago, ordea, pertsona beltz batzuk agertzen ziren argazki bati «gorilak» etiketa esleitu zion. Esan beharrik ez dago aplikazioak ez zuela horrelako nahasmenik azal zuriko pertsonen argazkiekin. Horren azpian, badirudi *adierazpidearen alborapena* edota *ebaluazioaren alborapena* egon daitezkeela, nahiz eta ez dugun daturik hori ziurtatzeko. Edonola ere, arazoaren zailtasuna nabaria da: izan ere, geroztik urte asko pasatu dira, ordenagailu bidezko teknikak asko aurreratu dira, baina arazoa ez dago guztiz ebatzita. Gaur egun, argazkiak etiketatzeko konpainia handien aplikazioek, oro har, ez dituzte «gorila» etiketak esleitzen oker egin dezaketela badakitelako, eta, horren beldurrez, etiketa horiek ez esleitzeko hautua egin dute.[7,8]

Bestalde, irudien tratamenduarekin, estereotipoen anplifikazioa ere gertatzen da gizartean, eta horrek generoarekin lotutako alborapena sortzen du. Sarean irudiak bilatzen ditugunean, bilatzailean «suhiltzaile» hitza jartzen badugu, lortuko ditugun irudietan gizonezkoen irudiak gailenduko dira. Aldiz, «etxeko-langile» terminoa bilatzen badugu, kontrakoa ikusiko dugu. Aspaldi antze- mandako alborapena bada ere, oraindik ere mantentzen dena da: Google, Wikipedia eta Internet Movie Database (IMDb) plataformetatik hartutako irudietan, 3.500 termino aztertu eta generoaren araberako alborapen nabaria dagoela frogatu zuten Guilbeault eta besteek [47]. Segur aski, estereotipoen anplifikazio hori gertatzen da tartean daudelako *adierazpidearen alborapena, ebaluaziorena* edota *alborapen historikoa*.

[7] https://www.nytimes.com/2023/05/22/technology/ai- photo- labels- google- apple.html

[8] https://incidentdatabase.ai/cite/587

Irudi-sorkuntzan ere antzeko zerbait gertatzen da.[9] Sistema horietan, *prompt*ean testu bidez zehazten du erabiltzaileak nolako irudia nahi duen. Sistemak lehenengo testu hori ulertu behar du eta ondoren irudia sortu. Thomas eta Thomas-ek egindako esperimentuan [48], kazetaritzarekin loturiko termino orokorrak (ingelesez «*journalist*», «*reporter*», «*correspondent*», «*the press*») eta espezializatuagoak («*news analyst*», «*news commentator*», eta «*fact-checker*») erabili zituzten *prompt*etan irudiak sortzeko, eta sortutako irudietan *zer* eta *nola* agertzen zen aztertu zuten. Hala, ikuspegi ugariko alborapenak hauteman zituzten. Termino orokorrak erabiliz sortutako irudietan, soilik pertsona gazteak agertzen ziren, bai emakumezkoak eta bai gizonezkoak. Termino espezializatutakoetatik sortutakoetan gazteak eta zaharrak agertzen ziren baina zaharrak beti gizonezkoak ziren; ez zegoen emakume zaharrik. Gizonezkoak zimurrekin agertzen ziren, baina ez emakumezkoak. Sortutako irudi guztietan, azal zuriko pertsonak soilik agertu ziren, beharbada eredua eraikitzeko erabilitako irudietan aniztasun nahikoa ez dagoelako (aipatu berri dugun *adierazpidearen alborapena*). Antzera, irudi denetan, konbentzionala zen pertsonen itxura: inork ez zuen tatuajerik edo bestelako orrazkerarik. Janzkera aldetik ere denek zuten janzkera formala. Hori izan daiteke telebistako aurkezle batentzat espero den itxura, baina kazetari denak ez dira modu horretara janzten.



5. irudia. «News analyst» eskaera egin eta sortutako irudien uniformetasuna erakusten duten bi adibide.

Irudien indarra handia denez, hitzez esandakoa irudiez erakusteko saiakera egin genuen. Horretarako, DALL-E[10] irudi-sortzaileari hainbat irudi

[9] https://theconversation.com/ageism-sexism-classism-and-more-7-examples-of-b ias-in-ai-generated-images-208748

[10] https://chatgpt.com/ (2024.05.29).

sortzeko eskatu genion.[11] Aipatu berri dugun azterketan ikusitakoa berretsi genuen, eta deigarria egin zaigu sortutako irudien uniformetasuna. Ez da lan honen helburua azterketa zehatza egitea, baina bai irakurleari irudiak erakutsiz hitzez esandakoa osatzea, eta adibide batzuk ekarri ditugu hona hori erakusteko. «*News analyst*» eskaeratik sortutako emaitzen lekuko pare bat (gizonezko bat eta emakumezko bat) ikusi daitezke 5. irudian. Gainera, «*reporter*» hitzarekin lortutako irudi denak gizonezkoak zirela ikusirik, beste irudi bat sortzeko eskatu genion, «*All of them are white, young men. Do you know any more?*» esanez (ikus 6. irudia): azal beltzeko emakume bat irudikatu bazuen ere, aurreko irudien estetika eta testuingurua errepikatu zituen.



6. irudia. «*Reporter*» eskaria 5 aldiz egin eta jasotako irudien ondoren «*All of them are white, young men. Do you know any more?*» eskaerari erantzunez jasotako irudia.

[11] Argazki guztiak hemen: https://github.com/olatz87/Ekaia_alborapena

Kasu horietan, alborapenaren iturria zein den jakitea ez da erraza, softwarea ekoizten dutenek ez dutelako informazio guztia argitaratzen,[12] baina badirudi aipatutako *adierazpide, ebaluazio* edota *alborapen historikoak* zerikusi zuzena izan dezaketela.

Testuaren sorkuntzan ikusten diren alborapenak errepikatu eta, segur aski, areagotu ere egiten dira irudi-sortzaileekin. Horrek guztiak zera erakusten du, adimen artifizialeko irudi-sortzaile batek egindako irudia parean dugunen beharrezkoa dela irudian agertzen dena baino zabalagoa den populazioan pentsatzea. Eta, antzera, irudi-sortzaileak gu garenean, sortutako irudiari ikuspegi kritikoa eman behar diogula, bestela, gizartean ditugun estereotipoak indartzen jarraituko dugu eta.

4.4. **Eredu iragarleak ongizatean**

Paradoxikoa eman badezake ere, pertsonen ongizatea helburu duten sistemek ere eragiten dituzte alborapenak.

Oso ezaguna da osasunaren arloan Obermeyer eta besteek [49] landu zuten kasua. Bertan azaltzen dute osasun-sistemek egoera zaurgarrian eta osasun-egoera konplexuan daudenak identifikatu nahi dituztela, modu koordinatuan baliabide gehiago eta arreta hobea eskaini ahal izateko. Horrela, alde batetik, pazienteen asebetetzea handitzen da, eta, bestetik, kostuak murrizten dira. Paziente horiek identifikatzea iragarpen-problema gisa planteatu zuten AEBko osasun-sistemek.

Iragarpen-eredua eraikitzeko, pazienteen aurretiko informazioan oinarritu ziren osasun-sistemak. Pazienteen osasun-zaurgarritasuna aurresatea helburu izan arren, «osasun-zaurgarritasuna» neurtzea ez da erraza, eta, kasu honetan, osasun-sisteman pazienteak egindako «gastuak» hartu zituzten aurresateko helburu-aldagai gisa. Hasiera batean egokia izan daitekeela pentsa daiteke: izan ere, osasunean egindako gastuen eta osasun-beharren arteko korrelazioa nabaria da. Hau da, pazientearen osasun-egoera kaxkarra edota konplexua bada, gastu handiak izango ditu, eta, alderantziz, pazientearen osasun-egoera ona bada ez du sisteman gastu handirik egingo. Hala baina, pazientearen osasun egoera kaxkarra bada baina egoera ekonomikoak ez badio aukerarik ematen osasunean gasturik egiteko, gastuaren aurreikuspenak nekez hautemango du pertsona hori osasun aldetik zaurgarria dela. Kasu horretan, alborapenaren jatorria aipatu berri dugun *neurketen alborapena* dugu. Alborapen horrekin, zaurgarriak eta osasun-egoera konplexua izanagatik atzeman gabe gelditu zirenak batez ere baliabide gutxien zutenak eta azal ez-zurikoak izan ziren. Obermeyer eta besteek [49] frogatu zuten eraikitako ereduaren albora-

[12] https://gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989

pena nabarmen txikitu zitekeela aurresan beharreko aldagaia aldatuta, datuen gainean beste aldaketarik egin gabe. Horretarako, gastuen eta osasun-egoeraren aurreikuspenaren konbinazio gisa definitu zuten ereduak aurresan beharreko aldagaia.

Gaur egun, alborapenaren arazoa gero eta presenteago dago osasunaren eta haren kudeaketaren gaineko lanetan. Ez da, ordea, batere erraza konpontzea, alborapenen jatorria askotarikoa izan daitekeelako eta, askotan, soluzioak ezin direlako orokortu. Komunitate zientifikoa lanean ari da osasun-arloko alborapena saihesteko adimen artifizialeko sistemetan ere [50, 51, 52].

Osasun-arlotik at baina ikuspuntu formal batetik antzeko egoerak izan daitezkeenak (ongizatea helburu eta *neurketen alborapena*) asko dira. Izan zituen ondorioengatik, Herbeheretako *Toeslagenaffaire* kasua[13] sona handikoa izan zen. Kasu horretan, Gobernuak martxan jarri zuen umeen zaintzarako emandako diru-laguntzen atzean iruzurrik egin ote zen atzemateko algoritmoa. Diru-laguntza eskatu zutenak atzerritarrak zirenean, iruzurgile gisa etiketatzeko joera nabarmena erakutsi zuen algoritmoak. Eskandalua handia izan zen: 2021ean, Gobernuko kide batzuek dimititu behar izan zuten, eta kasua beste gobernuentzako eta erakunde publikoentzako ohartarazpen gisa erabili izan da ordutik [53, 54].

5. **ALBORAPENA MURRIZTEKO METODOAK**

Pil-pilean dagoen alorra da alborapenaren murrizketa gaur egunean. Ikuspuntu eta teknika ugari proposatu dira orain arte, baina oraindik orain irekita jarraitzen duen alorra da.

Alborapena murrizteko tekniken proposamenak ugariak dira, eta arloaren egoerari buruzko hainbat artikulu bat datoz teknika horiek sailkatzeko garaian [18, 55, 56]. Haien arabera, teknika aplikatzen den unearen arabera, hiru kategoria bereizten dira: aurre-prozesaketan aplikatzen direnak (*pre-processing*), prozesaketan zehar aplikatzen direnak (*in-processing*) eta prozesaketaren ondoren aplikatzen direnak (*post-processing*). Laugarren kategoria bat ere bereizten da hizkuntza-ereduen gaineko ikerketan [3]: intra-prozesaketa (*intra-processing*).

— **Aurre-prozesaketa**: datuetatik alborapena kentzeko teknikak dira, alborapena ereduaren entrenamendutik kanpo uzteko. Ohiko teknika batzuk datuak gehitzea (*data augmentation*) [57, 58, 59, 60, 61, 62, 63,

[13] https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europ e-over-risks-of-using-algorithms/

64, 65] eta datuak filtratzea (*data filtering)* [66, 67, 68, 69, 70] dira. Adibidez, aurpegien errekonozimenduko atazan, frogatu zuten ezen, azal beltzeko pertsonen (bereziki emakumeen) adibideak gehitzean, haien identifikazioa hobetzen dela [25]. Bestalde, hizkuntza ofentsiboa duten testuak kentzeak ere onurak dakartza. Teknika horien eragina mugatua izan daiteke askotan, eta datuak gehitzeak edo filtratzeak arriskua ekar dezake: besteak beste, datu okerrak sartzea [71].

— **Prozesaketan zehar**: ereduen entrenamenduan aplikatzen diren teknikak dira. Alborapena entrenamendu-garaian kentzen dute: adibidez, ikasketa-helburua aldatuz, edota murriztapenak ezarriz. Hemen, birdoiketan eragiten duten teknikak ere sartzen dira (*finetuning)*. Kasu honetan ere, hainbat teknika erabiltzen dira: eredu bat aukeratzeko garaian alborapen gutxiena sortzen duena lehenestea [72, 73], ereduaren arkitekturan aldaketak egitea [68, 74, 75] edo galera-funtzioa aldatzea [76, 77, 78, 79, 80]. Teknika hauetan, askotan, konputazio-kostua edo egingarritasuna izaten da muga handiena. Horretaz gain, teknika horiek aplikatzeko garaian, kontu handiarekin ibili beharra dago, ereduaren gaitasunak (ahalmena, zuzentasuna) mugatu daitezkeelako [3].
— **Prozesaketaren ondoren**: dagoeneko entrenatuta dauden ereduen gainean alborapena kentzen saiatzen diren teknikak dira, jatorrizko entrenamendua aldatu gabe. Eredua kutxa beltz gisa ikusten da hemen, eta ereduaren irteeraren gainean eragiten dute, eredua bera aldatu gabe. Adibidez, testua sortzen duten ereduen kasuan, berridazketarako teknikak aplikatzen dira, alborapena izan dezaketen hitzak ordezkatuz [81, 82], edota alborapena kentzeko sortutako itzultzaileen bidez [83, 84]. Teknika hauek azken aukera gisa erabiltzea proposatzen da, alborapena izateko arriskua dutelako. Berridazketaren kasuan, adibidez, zer hitz berridatzi behar den ataza subjektiboa da oinarrian, eta horrek sistema alboratua izateko arriskuak areagotzen ditu [3].
— **Intra-prozesaketa**: ereduen parametroei edo pisuei edota deskodetzeko teknikei aplikatzen zaizkien teknikak dira, entrenamendu berririk edo birdoiketarik egin gabe. Prozesaketaren ondorengo egiten den aplikazioarekin alderatuz, kasu honetan ez da eredua kutxa beltz gisa ikusten, baizik eta bertan ikasitako parametroei zuzenean eragiten zaie [85, 86, 87, 88].
Oraindik ez da nahikoa garatu ikerketa-lerro hau, eta arloaren ikerlariek duten erronka nagusia alborapena modu orekatuan murriztea da, besteak beste talde minorizatuak kontuak hartuz [3].

Teknika horiek guztiek, alborapena murrizteko erabilgarriak badira ere, mugak dituzte. Ondorioz, ezinbestekoa da ikerketa-ildo honetan sakontzen jarraitzea eta proposamen berriak egitea.

6. EZTABAIDA ETA ONDORIOAK

Adimen artifizialean oinarritutako tresnak gure egunerokotasunean txertatzen ari zaizkigu etengabe, eta abantaila ugari eskaintzen dizkigute. Tresna horiek algoritmo eta datu jakinetan oinarritzen badira ere, ez dakigu zehaztasunez zer den sistema horiek ikasi dutena, baina agerikoa da alborapena ere ikasten dutela. Alborapen horien arrazoia zein den jakitea, ordea, gero eta zailagoa gertatzen da: izan ere, adimen artifizialeko sistemak diruiturri bihurtzearekin batera, sistemen datu eta algoritmoak itxi dituzte. Oso nabarmena izan da OpenAI, Google eta halako enpresa handien jokabidea. 2020. urtera arte kaleratutako eredu guztiak kode irekikoak izan badira ere, hortik aurrera, GPT-3ren kaleratzearekin, eredu ahaltsuenen kodea itxi eta pribatizatu dute [89], zientzia irekiaren kaltetan.

Gizartean, eta komunitate zientifikoan bereziki, kezka handia sortzen duen gaia da alborapenarena, eta bi dira horren inguruan irekita dauden ikerlerro nagusiak: alborapenaren neurketa, batetik, eta haren murrizketa, bestetik. Bi ikerlerroetan lan handia egin da azken urteetan, baina oraindik ere irekita jarraitzen duten ikerlerro mamitsuak dira.

Adimen artifizialeko sistemek sortzen duten alborapena neurtzeko, oraindik ere ez dago prozedura estandarrik, eta nekez existitzen dira horretan lagunduko diguten proba-banku egokiak, izan aurpegiak errekonozitzeko, izan lan-gomendioak egiteko. Sistema horietan hizkuntza tartean dagoenean, gainera, ingeleserako ez diren proba-bankuak edota neurtzeko mekanismoak oso-oso urriak dira.

Beste alde batetik, alborapenaren murrizketan hainbat teknika proposatu izan dira, bai entrenamendurako datuetan eragiteko, bai algoritmoak egokitzeko ikasketa-prozesuan bertan edota ondoren eragiteko. Oraindik ere lan asko egiteko dagoen alorra da. Hala ere, ez dugu aipatu gabe utzi nahi alborapenaren murrizketa bera ere badela eztabaidagai. Besteak beste, zalantzan jartzen da ereduei alborapenik gabeko mundu ideal bat erakutsi behar ote zaien edo alborapen hori ikasi behar duten, ondoren eredua bera, ikasitako alborapenaren «jakitun», alborapenik gabeko erabakiak hartzeko gai izan dadin.

Eztabaida zabala irekita dago baina lanean jarraitu behar da adimen artifizialeko sistemen alborapenen kalteak gutxitzeko. Esana dugu arazoa konplexua eta ertz askokoa dela, eta garrantzitsua da diziplinarteko lan taldeak osatzea eta ikuspegi desberdinak uztartzea irtenbide egokiak bilatzeko. Hor dago adibidez, *makinen portaera* eremua [90], makinen portaerak aztertzeko helburu orokorrarekin hainbat arlotako espezialistak biltzen dituena (ingeniariak, informatikariak, psikologoak, soziologoak, filosofoak...).

Pertsona erdigunean jarriko duen adimen artifiziala behar dugu [91], eta hori, denen arteko elkarrekintzatik etorriko da, era horretan sistemen

portaera ulertu eta sistema zuzenagoak eraikitzeko bideak jorratuko baititugu. Horretarako, alor guztietako ikertzaileak ez ezik, arduradun politikoak eta herritarrak ere behar ditugu, beren inplikazioarekin «mundu errealean» algoritmoekin hartzen diren erabakiak ebaluatzeko, ekitatea, erantzukizuna, gardentasuna eta pribatutasuna aintzat harturik. Gaia zabala, askotarikoa eta konplexua da, eta dudarik gabe lantzen jarraitu beharrekoa.

ERREFERENTZIAK

[1] V. S. JACOB, L. D. GAULTNEY eta G. SALVENDY, 1986, «Strategies and biases in human decision-making and their implications for expert systems», *Behaviour & Information Technology*, **5**(2), 119-140.

[2] C. S. WEBSTER, S. TAYLOR, C. THOMAS eta J. M. WELLER, 2022, «Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations», *BJA education*, **22**(4), 131.

[3] I. O. GALLEGOS, R. A. ROSSI, J. BARROW, M. M. TANJIM, S. KIM, F. DERNONCOURT, T. YU, R. ZHANG eta N. K. AHMED, 2023, «Bias and fairness in large language models: A survey», *arXiv preprint arXiv:230900770*.

[4] A. Z. JACOBS eta H. WALLACH, 2021, «Measurement and fairness», *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, orrialdeak 375-385.

[5] R. RUDINGER, J. ÑARADOWSKY, B. LEONARD eta B. VAN DURME, 2018, «Gender bias in coreference resolution», *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, New Orleans, Louisiana.

[6] J. ZHAO, T. WANG, M. YATSKAR, V. ORDONEZ eta K.-W. CHANG, 2018, «Gender bias in coreference resolution: Evaluation and debiasing methods», *arXiv preprint arXiv:180406876*.

[7] M. ÑADEEM, A. BETHKE eta S. REDDY, 2020, «StereoSet: Measuring stereotypical bias in pretrained language models», *arXiv preprint arXiv:200409456*.

[8] N. ÑANGIA, C. VANIA, R. BHALERAO eta S. R. BOWMAN, 2020, «CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models», *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, online.

[9] A. PARRISH, A. CHEN, N. ÑANGIA, V. PADMAKUMAR, J. PHANG, J. THOMPSON, P. M. HTUT eta S. BOWMAN, 2022, «BBQ: A hand-built bias benchmark for question answering», *Findings of the Association for Computational Linguistics: ACL 2022*, orrialdeak 2086-2105, Association for Computational Linguistics, Dublin, Irlanda. URL https://aclanthology.org/2022.findings-acl.165

[10] E. M. SMITH, M. HALL, M. KAMBADUR, E. PRESANI eta A. WILLIAMS, 2022, «"i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset», *arXiv preprint arXiv:220509209*.

[11] S. L. BLODGETT, S. BAROCAS, H. DAUMÉ III eta H. WALLACH, 2020, «Language (technology) is power: A critical survey of "bias"in NLP», *arXiv preprint arXiv:200514050*.

[12] S. L. BLODGETT, G. LOPEZ, A. OLTEANU, R. SIM eta H. WALLACH, 2021, «Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets», *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, orrialdeak 1004-1015.

[13] R. BAEZA-YATES, 2018, «Bias on the web», *Commun ACM*, **61**(6), 54-61.

[14] T. BAER, 2019, *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists*, Apress.

[15] E. ÑTOUTSI, P. FAFALIOS, U. GADIRAJU, V. IOSIFIDIS, W. ÑEJDL, M.-E. VIDAL, S. RUGGIERI, F. TURINI, S. PAPADOPOULOS, E. KRASANAKIS eta kolaboratzaileak, 2020, «Bias in data-driven artificial intelligence systems—an introductory survey», *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **10**(3), e1356.

[16] S. BAROCAS, M. HARDT eta A. ÑARAYANAN, 2023, *Fairness and machine learning: Limitations and opportunities*, MIT Press.

[17] I. Y. CHEN, E. PIERSON, S. ROSE, S. JOSHI, K. FERRYMAN eta M. GHASSEMI, 2021, «Ethical machine learning in healthcare», *Annual review of biomedical data science*, **4**, 123-144.

[18] N. MEHRABI, F. MORSTATTER, N. SAXENA, K. LERMAN eta A. GALSTYAN, 2021, «A survey on bias and fairness in machine learning», *ACM Comput Surv*, **54**(6).

[19] E. KARTAL, 2022, «A comprehensive study on bias in artificial intelligence systems: Biased or unbiased AI, that's the question!», *International Journal of Intelligent Information Technologies (IJIIT)*, **18**(1), 1-23.

[20] H. SURESH eta J. GUTTAG, 2021, «A framework for understanding sources of harm throughout the machine learning life cycle», *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, Association for Computing Machinery.

[21] S. SCHELTER eta J. STOYANOVICH, 2020, «Taming technical bias in machine learning pipelines», *Bulletin of the Technical Committee on Data Engineering*, **43**(4).

[22] A. CALISKAN, P. P. AJAY, T. CHARLESWORTH, R. WOLFE eta M. R. BANAJI, 2022, «Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics», *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES'22, orria 156-170, Association for Computing Machinery.

[23] P. TERHÖRST, J. Ñ. KOLF, M. HUBER, F. KIRCHBUCHNER, N. DAMER, A. M. MORENO, J. FIERREZ eta A. KUIJPER, 2022, «A comprehensive study on face recognition biases beyond demographics», *IEEE Transactions on Technology and Society*, **3**(1), 16-30.

[24] S. ÑAGPAL, M. SINGH, R. SINGH eta M. VATSA, 2019, «Deep learning for face recognition: Pride or prejudiced?».

[25] J. BUOLAMWINI eta T. GEBRU, 2018, «Gender shades: Intersectional accuracy disparities in commercial gender classification», *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, bolumena 81 of *Proceedings of Machine Learning Research*, orrialdeak 77-91, PMLR.

[26] A. CHOULDECHOVA, 2017, «Fair prediction with disparate impact: A study of bias in recidivism prediction instruments», *Big data*, **5**(2), 153-163.

[27] D. U. PATTON, W. R. FREY, K. A. MCGREGOR, F.-T. LEE, K. MCKEOWN eta E. MOSS, 2020, «Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing», *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES'20, orrialdeak 337-342, Association for Computing Machinery.

[28] A. MARSHAN eta A. MARSHAN, 2021, «Artificial intelligence: Explainability, ethical issues and bias», *Annals of Robotics and Automation*, **5**(1), 34-37.

[29] G. VARDI, 2023, «On the implicit bias in deep-learning algorithms», *Commun ACM*, **66**(6), 86-93.

[30] I. GARRIDO-MUÑOZ, A. MONTEJO-RÁEZ, F. MARTÍNEZ-SANTIAGO eta L. A. UREÑA-LÓPEZ, 2021, «A survey on bias in deep NLP», *Applied Sciences*, **11**(7).

[31] T. FAHSE, V. HUBER eta B. VAN GIFFEN, 2021, *Managing Bias in Machine Learning Projects*, orrialdeak 94-109.

[32] E. COLLINS, 2018, «Punishing risk», *107 Georgetown Law Journal*, **57**.

[33] S. FENG, O. KUDINA, B. M. HALPERN eta O. SCHARENBORG, 2021, «Quantifying bias in automatic speech recognition», *arXiv preprint arXiv:210315122*.

[34] M. K. NGUEAJIO eta G. WASHINGTON, 2022, «Hey ASR system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review», *International Conference on Human-Computer Interaction*, orrialdeak 421-440, Springer.

[35] N. MARKL, 2022, «Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition», *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, orrialdeak 521-534.

[36] J. L. MARTIN eta K. E. WRIGHT, 2023, «Bias in automatic speech recognition: The case of african american language», *Applied Linguistics*, **44**(4), 613-630.

[37] S. FENG, B. M. HALPERN, O. KUDINA eta O. SCHARENBORG, 2024, «Towards inclusive automatic speech recognition», *Computer Speech & Language*, **84**, 101-567.

[38] N. Ñ. LOIDEAIN eta R. ADAMS, 2020, «From Alexa to Siri and the GDPR: the gendering of virtual personal assistants and the role of data protection impact assessments», *Computer Law & Security Review*, **36**, 105-366.

[39] P. COSTA, 2018, «Conversing with personal digital assistants: On gender and artificial intelligence», *Journal of Science and Technology of the Arts*, **10**(3), 59-72.

[40] N. ÑI LOIDEAIN eta R. ADAMS, 2019, «Female servitude by default and social harm: AI virtual personal assistants, the FTC, and unfair commercial practices», *Rachel, Female Servitude by Default and Social Harm: AI Virtual Personal Assistants, the FTC, and Unfair Commercial Practices (June 11, 2019)*.

[41] R. ADAMS eta N. Ñ. LOIDEÁIN, 2019, «Addressing indirect discrimination and gender stereotypes in AI virtual personal assistants: the role of international human rights law», *Cambridge International Law Journal*, **8**(2), 241-257.

[42] H. TOUVRON, L. MARTIN, K. STONE, P. ALBERT, A. ALMAHAIRI, Y. BABAEI, N. BASHLYKOV, S. BATRA, P. BHARGAVA, S. BHOSALE eta kolaboratzaileak, 2023, «Llama 2: Open foundation and fine-tuned chat models», *arXiv preprint arXiv:230709288*.

[43] F. MOTOKI, V. PINHOÑETO eta V. RODRIGUES, 2024, «More human than human: Measuring ChatGPT political bias», *Public Choice*, **198**(1), 3-23.

[44] Y. WAN, G. PU, J. SUN, A. GARIMELLA, K.-W. CHANG eta N. PENG, 2023, «"Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters», *arXiv preprint arXiv:231009219*.

[45] L. WANG, M. SONG, R. REZAPOUR, B. C. KWON eta J. HUH-YOO, 2023, «People's perceptions toward bias and related concepts in large language models: A systematic review», *arXiv preprint arXiv:230914504*.

[46] A. SALINAS, P. SHAH, Y. HUANG, R. MCCORMACK eta F. MORSTATTER, 2023, «The unequal opportunities of large language models: Examining demographic biases in job recommendations by ChatGPT and LLaMA», *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, orrialdeak 1-15.

[47] D. GUILBEAULT, S. DELECOURT, T. HULL, B. S. DESIKAN, M. CHU eta E. ÑADLER, 2024, «Online images amplify gender bias», *Nature*, **626**, 1049-1055.

[48] R. J. THOMAS eta T. THOMSON, 2023, «What does a journalist look like? visualizing journalistic roles through AI», *Digital Journalism*, Orrialdeak 1-23.

[49] Z. OBERMEYER, B. POWERS, C. VOGELI eta S. MULLAINATHAN, 2019, «Dissecting racial bias in an algorithm used to manage the health of populations», *Science*, **366**(6464), 447-453.

[50] M. H. CHIN, N. AFSAR-MANESH, A. S. BIERMAN, C. CHANG, C. J. COLÓN-RODRÍGUEZ, P. DULLABH, D. G. DURAN, M. FAIR, T. HERNANDEZ-BOUSSARD, M. HIGHTOWER eta kolaboratzaileak, 2023, «Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care», *JAMA Network Open*, **6**(12), e2345050-e2345050.

[51] M. MITTERMAIER, M. M. RAZA eta J. C. KVEDAR, 2023, «Bias in ai-based models for medical applications: challenges and mitigation strategies», *NPJ Digital Medicine*, **6**(1), 113.

[52] A. JAIN, J. R. BROOKS, C. C. ALFORD, C. S. CHANG, N. M. MUELLER, C. A. UMSCHEID eta A. S. BIERMAN, 2023, «Awareness of racial and ethnic bias and potential solutions to address bias with use of health care algorithms», *JAMA Health Forum*, Bolumena 4, Orrialdeak e231197-e231197, American Medical Association.

[53] W. DAMEN, 2023, «Sounds good, doesn't work: The GDPR principle of transparency and data-driven welfare fraud detection», *ISLSSL European Regional Congress-The Lighthouse Function of Social Law*, Orrialdeak 527-544, Springer.

[54] A. STANOJEVIC, 2023, «Algorithmic governance and social vulnerability: a value analysis of equality, freedom and trust», *Available at SSRN*.

[55] M. HORT, Z. CHEN, J. M. ZHANG, M. HARMAN eta F. SARRO, 2024, «Bias mitigation for machine learning classifiers: A comprehensive survey», *ACM Journal on Responsible Computing*, **1**(2), 1-52.

[56] E. FERRARA, 2023, «Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies», *Sci*, **6**(1), 3.

[57] J. AHN, H. LEE, J. KIM eta A. OH, 2022, «Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT», *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, orrialdeak 266-272, Association for Computational Linguistics, Seattle, Washington. URL https://aclanthology.org/2022.gebnlp-1.27

[58] L. DIXON, J. LI, J. SORENSEN, N. THAIN eta L. VASSERMAN, 2018, «Measuring and mitigating unintended bias in text classification», *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, Orrialdeak 67-73.

[59] S. GHANBARZADEH, Y. HUANG, H. PALANGI, R. CRUZMORENO eta H. KHANPOUR, 2023, «Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models», *Findings of the Association for Computational Linguistics: ACL 2023*, orrialdeak 5448-5458, Association for Computational Linguistics, Toronto, Canada. URL https://aclanthology.org/2023.findings-acl.336

[60] K. LU, P. MARDZIEL, F. WU, P. AMANCHARLA eta A. DATTA, 2020, *Gender Bias in Neural Natural Language Processing*, orrialdeak 189-202, Springer International Publishing, Cham. URL https://doi.org/10.1007/978-3-030-62077-6_14

[61] R. QIAN, C. ROSS, J. FERNANDES, E. M. SMITH, D. KIELA eta A.WILLIAMS,2022, «Perturbation augmentation for fairer NLP», *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, orrialdeak 9496-9521, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates. URL https://aclanthology.org/2022.emnlp-main.646

[62] K. WEBSTER, X. WANG, I. TENNEY, A. BEUTEL, E. PITLER, E. PAVLICK, J. CHEN, E. CHI eta S. PETROV, 2020, «Measuring and reducing gendered correlations in pre-trained models», *arXiv preprint arXiv:201006032*.

[63] L. YU, Y. MAO, J. WU eta F. ZHOU, 2023, «Mixup-based unified framework to overcome gender bias resurgence», *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'23, orria 1755-1759, Association for Computing Machinery, New York, NY, AEB. URL https://doi.org/10.1145/3539618.3591938

[64] A. ZAYED, P. PARTHASARATHI, G. MORDIDO, H. PALANGI, S. SHABANIAN eta S. CHANDAR, 2023, «Deep learning on a healthy data diet: Finding important examples for fairness», *Proceedings of the AAAI Conference on Artificial Intelligence*, Bolumena 37, orrialdeak 14593-14601.

[65] R. ZMIGROD, S. J. MIELKE, H. WALLACH eta R. COTTERELL, 2019, «Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology», *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, orrialdeak 1651-1661, Association for Computational Linguistics, Florentzia, Italia. URL https://aclanthology.org/P19-1161

[66] C. BORCHERS, D. GALA, B. GILBURT, E. ORAVKIN, W. BOUNSI, Y. M. ASANO eta H. KIRK, 2022, «Looking for a handsome carpenter! debiasing GPT-3 job advertisements», *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, orrialdeak 212-224, Association for Computational Linguistics, Seattle, Washington. URL https://aclanthology.org/2022.gebnlp-1.22

[67] A. GARIMELLA, R. MIHALCEA eta A. AMARNATH, 2022, «Demographic-aware language model fine-tuning as a bias mitigation technique», *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, orrialdeak 311-319.

[68] X. HAN, T. BALDWIN eta T. COHN, 2022, «Balancing out bias: Achieving fairness through balanced training», *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, orrialdeak 11335-11350, Association for Computational Linguistics, Abu Dhabi, Arabiar Emirerri Batuak. URL https://aclanthology.org/2022.emnlp-main.779

[69] H. ÑGO, C. RATERINK, J. G. ARAÚJO, I. ZHANG, C. CHEN, A. MORISOT eta N. FROSST, 2021, «Mitigating harm in language models with conditional-likelihood filtration», *arXiv preprint arXiv:210807790*.

[70] C. RAFFEL, N. SHAZEER, A. ROBERTS, K. LEE, S. ÑARANG, M. MATENA, Y. ZHOU, W. LI eta P. J. LIU, 2020, «Exploring the limits of transfer learning with a unified text-to-text transformer», *Journal of machine learning research*, **21**(140), 1-67.

[71] S. KUMAR, V. BALACHANDRAN, L. ÑJOO, A. ANASTASOPOULOS eta Y. TSVETKOV, 2023, «Language generation models can cause harm: So what can we do about it? an actionable survey», *Proceedings of the 17th Confe-*

rence of the European Chapter of the Association for Computational Linguistics, orrialdeak 3299-3321, Association for Computational Linguistics, Dubrovnik, Croatia. URL https://aclanthology.org/2023.eacl-main.241

[72] S. YAN, H.-T. KAO eta E. FERRARA, 2020, «Fair class balancing: Enhancing model fairness without observing sensitive attributes», *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Orrialdeak 1715-1724.

[73] M. B. ZAFAR, I. VALERA, M. GOMEZ RODRIGUEZ eta K. P. GUMMADI, 2017, «Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment», *Proceedings of the 26th international conference on world wide web*, orrialdeak 1171-1180.

[74] M. BARTL, M. ÑISSIM eta A. GATT, 2020, «Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias», *arXiv preprint arXiv:201014534*.

[75] A. LAUSCHER, T. LUEKEN eta G. GLAVAŠ, 2021, «Sustainable modular debiasing of language models», *Findings of the Association for Computational Linguistics: EMNLP 2021*, orrialdeak 4782-4797, Association for Computational Linguistics, Punta Cana, Dominikar Errepublika. URL https://aclanthology.org/2021.findings-emnlp.411

[76] G. ATTANASIO, D. ÑOZZA, D. HOVY eta E. BARALIS, 2022, «Entropy-based attention regularization frees unintended bias mitigation from lists», *Findings of the Association for Computational Linguistics: ACL 2022*, orrialdeak 1105-1119, Association for Computational Linguistics, Dublin, Ireland. URL https://aclanthology.org/2022.findings-acl.88

[77] Y. GACI, B. BENATALLAH, F. CASATI eta K. BENABDESLEM, 2022, «Debiasing pretrained text encoders by paying attention to paying attention», *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Orrialdeak 9582-9602, Association for Computational Linguistics, Abu Dhabi, Arabiar Emirerri Batuak. URL https://aclanthology.org/2022.emnlp-main.651

[78] A. GARIMELLA, A. AMARNATH, K. KUMAR, A. P. YALLA, A. Ñ, N. CHHAYA eta B. V. SRINIVASAN, 2021, «He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation», *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, orrialdeak 4534-4545, Association for Computational Linguistics, Online. URL https://aclanthology.org/2021.findings-acl.397

[79] H. LIU, J. DACON, W. FAN, H. LIU, Z. LIU eta J. TANG, 2020, «Does gender matter? towards fairness in dialogue systems», *Proceedings of the 28th International Conference on Computational Linguistics*, Orrialdeak 4403-4416, International Committee on Computational Linguistics, Bartzelona, Espainia (online). URL https://aclanthology.org/2020.coling-main.390

[80] S. PARK, K. CHOI, H. YU eta Y. KO, 2023, «Never too late to learn: Regularizing gender bias in coreference resolution», *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, orrialdeak 15-23.

[81] Z. He, B. P. Majumder eta J. McAuley, 2021, «Detect and perturb: Neutral rewriting of biased and sensitive text via gradient-based decoding», *Findings of the Association for Computational Linguistics: EMNLP 2021*, orrialdeak 4173-4181, Association for Computational Linguistics, Punta Cana, Dominikar Errepublika. URL https://aclanthology.org/2021.findings-emnlp.352

[82] E. K. Tokpo eta T. Calders, 2022, «Text style transfer for bias mitigation using masked language modeling», *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, orrialdeak 163-171, Association for Computational Linguistics, Hybrid: Seattle, Washington + online. URL https://aclanthology.org/2022.naacl-srw.21

[83] N. Jain, M. Popović, D. Groves eta E. Vanmassenhove, 2021, «Generating gender augmented data for NLP», *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, orrialdeak 93-102, Association for Computational Linguistics, Online. URL https://aclanthology.org/2021.gebnlp-1.11

[84] T. Sun, K. Webster, A. Shah, W. Y. Wang eta M. Johnson, 2021, «They, them, theirs: Rewriting with gender-neutral english», *arXiv preprint arXiv:210206788*.

[85] D. Saunders, R. Sallis eta B. Byrne, 2022, «First the worst: Finding better gender translations during beam search», *Findings of the Association for Computational Linguistics: ACL 2022*, orrialdeak 3814-3823, Association for Computational Linguistics, Dublin, Ireland. URL https://aclanthology.org/2022.findings-acl.301

[86] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf eta K. Kersting, 2022, «Large pre-trained language models contain human-like biases of what is right and wrong to do», *Nature Machine Intelligence*, **4**(3), 258-268.

[87] J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston eta E. Dinan, 2020, «Recipes for safety in open-domain chatbots», *arXiv preprint arXiv:201007079*.

[88] A. Zayed, G. Mordido, S. Shabanian eta S. Chandar, 2023, «Should we attend more or less? modulating attention for fairness», *arXiv preprint arXiv:230513088*.

[89] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin eta X. Hu, 2024, «Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond», *ACM Transactions on Knowledge Discovery from Data*, **18**(6), 1-32.

[90] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson eta kolaboratzaileak, 2019, «Machine behaviour», *Nature*, **568**(7753), 477-486.

[91] B. Lepri, N. Oliver eta A. Pentland, 2021, «Ethical machines: The human-centric use of artificial intelligence», *IScience*, **24**(3).