UNIVERSITAT RAMON LLULL
Càtedra ETHOS

Ètica algorítmica i perspectiva de gènere: *de l'opacitat a la transparència*

Dr. Francesc Torralba

Dra. Núria Tria

Dra. Mar Rosàs

Sr. Guillem Martí

Setembre, 2021

Càtedra Ethos. Universitat Ramon Llull

Direcció: Dr. Francesc Torralba Rosselló

Coordinació de recerca: Dra. Mar Rosàs Tosàs

c. Claravall, 1-3 | 08022, Barcelona

https://www.url.edu/ca/recerca-i-innovacio/catedres-url/catedra-ethos-url

Contacte: mrosas@rectorat.url.edu

Setembre, 2021

Autors:

Dr. Francesc Torralba Roselló (Director de la Càtedra Ethos de la Universitat Ramon Llull. Professor d'ètica i d'antropologia a la Universitat Ramon Llull)

Dra. Núria Tria Parareda (Investigadora de la Càtedra Ethos de la Universitat Ramon Llull. Professora de bioètica a la Universitat de Vic-Universitat Central de Catalunya)

Dra. Mar Rosàs Tosas (Coordinadora d'investigació de la Càtedra Ethos de la Universitat Ramon Llull. Professora d'ètica i d'antropologia a la Universitat Ramon Llull)

Sr. Guillem Martí Soler (Investigador de la Càtedra Ethos de la Universitat Ramon Llull. Professor d'ètica aplicada a la Universitat Oberta de Catalunya).

Aquesta recerca ha obtingut una subvenció en el concurs competitiu «3a Convocatòria de subvencions per a la recerca en l'àmbit de la transparència, l'accés a la informació pública i el bon govern 2021» de l'Agència de transparència de l'Àrea Metropolitana de Barcelona (AMB), en la sublínia: Contractació de personal i transparència.

Índex

1. Introducció

En els processos de selecció de personal, les organitzacions recorren, cada vegada més, a solucions d'intel·ligència artificial, basades en la programació i la implementació dels anomenats «algoritmes»: sistemes lògics que analitzen automàticament documents com els *curricula vitae* o altres conjunts de dades per identificar en un segon els perfils més pertinents per a una determinada feina.

D'entrada, si ens posem en la perspectiva de les àrees (o empreses) de recursos humans encarregades de dur a terme aquests processos de selecció, això suposa dos avantatges importants tant per a les organitzacions com per a les persones que cerquen feina.

El primer avantatge, de caràcter ètic, és que l'aplicació d'aquesta tecnologia sembla garantir que a tots els candidats se'ls apliquin els mateixos criteris. Així, processos que, fins fa poc, corrien el risc d'estar esbiaixats per elements subjectius, ara es caracteritzen per la seva objectivitat i homogeneïtat.

El segon avantatge és que la tasca de preselecció de candidats i/o d'ofertes de feina, que, fins fa poc, requeria d'una quantitat ingent d'hores de dedicació, és automatitzada, si no totalment, en algunes fases importants. Els algoritmes suposen, com a mínim, un primer filtre. En el cas de les organitzacions, els filtra el nombre de candidats que passaran a un segon estadi del procés de selecció, garantint que hi passen només els candidats més aptes per a la plaça en qüestió sense haver de dedicar temps a mirar els seus currículums. En el cas de les persones que busquen feina, els filtra el nombre d'ofertes que reben, garantint presumptament que només reben aquelles per a les quals estan més preparats.

Ara bé, l'ús d'algoritmes en els processos de selecció pot esdevenir discriminatori. La raó és que els algoritmes mai no són neutres axiològicament, sinó que estan programats segons una sèrie de criteris que algú ha decidit. Aquests criteris poden, doncs, reproduir biaixos humans. A aquesta font "externa" de possibles biaixos (en la mesura que són introduïts en el procés de disseny per part dels responsables de la programació), cal sumar-hi la generació de biaixos o, com a mínim, l'ampliació dels seus efectes, per la pròpia lògica auto-reproductiva d'aquestes tecnologies basades en l'entrenament i l'aprenentatge.

En el cas del gènere, les conseqüències són especialment greus, sobretot si tenim en compte que en els processos de selecció de personal ja solen incidir-hi altres formes de discriminació de gènere. Per exemple, és habitual que les dones hagin hagut de fer molts més mèrits que

els homes per obtenir la mateixa puntuació en l'avaluació d'un currículum[1] o que, en les entrevistes, els avaluadors tendeixin a percebre els homes com a més competents que les dones.

L'objectiu del present estudi no és tractar de dissuadir l´ús de solucions d'Intel·ligència Artificial. Ara bé, sí que és important posar llum en aquells aspectes en què, per exemple, l'automatització dels processos de selecció de personal pot donar lloc a determinades lògiques discriminatòries que atempten contra el marc de la Declaració Universal dels Drets Humans, sobre la qual se sustenten els marcs ètics i legals de les societats democràtiques contemporànies. L'objectiu d'aquest treball és, doncs, reduir l'opacitat que molt sovint envolta l'aplicació social de les tecnologies basades en algoritmes, i contribuir, des de l'ètica, a una reflexió que vetlli perquè els algoritmes no promoguin decisions discriminatòries i biaixos de gènere.

D'acord amb el que acabem de dir, la nostra recerca ha tingut com a objecte d'estudi *els biaixos de gènere que plantegen els algoritmes per a la selecció de personal*. En segon lloc, aquesta recerca ha permès proposar fonamentadament un seguit de *propostes per evitar els possibles efectes discriminatoris de l'ús de tecnologies algorítmiques en els processos de reclutament de treballadores*.

En la primera part d'aquest estudi presentem la identificació i una síntesi dels principals reptes teòrics de l'ètica algorítmica, especialment pel que fa al gènere, així com un recull de casos reals, tant de males pràctiques com de bones pràctiques, que els exemplifiquin i ajudin a la seva comprensió.

En primer lloc, farem una introducció a la noció d'intel·ligència artificial i a l'ètica algorítmica. Un dels principals reptes que trobem a l'abordar aquesta temàtica és la dificultat de comprensió inherent a la matèria de què tracta, a la qual es superposa una certa confusió terminològica fruit de tractaments sensacionalistes i reduccionistes als mitjans de comunicació. Es fa necessària, doncs, una aproximació conceptual que aplani el camí per al posterior abordatge dels reptes ètics.

A continuació, plantejarem com una perspectiva fonamentada en algunes de les principals aportacions teòriques de la teoria feminista, la teoria *queer* i els estudis de gènere ens pot ajudar en la detecció dels biaixos de gènere (junt a d'altres tipus de biaix) en els algoritmes en

[1] En alguns casos, fins a 2,5 vegades els mèrits dels homes (*Libro Blanco. Situación de las mujeres en la ciencia española*, Unidad de Mujeres y Ciencia, pàgina 14).

general. Aquesta teoria ens serà útil tant per detectar i valorar les formes de discriminació dels algoritmes, com a l'hora d'imaginar i construir vies i estratègies per a tractar de prevenir-los o, si és el cas, mitigar els danys reals que, tal i com ens mostra la literatura consultada, està ocasionant a moltes persones l'ús cada cop més generalitzat de la IA en diversos àmbits de la vida (i especialment en el cas particular que ens ocupa).

Finalment ens endinsarem en el context particular del món laboral, i més concretament, en els processos de reclutament i selecció de personal, amb la voluntat i l'objectiu de comprendre què hi passa quan es comencen a implementar processos automatitzats i preses de decisions recolzades (totalment o en part) en IA.

Entrelligats amb l'explicació general aportarem casos reals que ens mostren alguns dels reptes ètics que han hagut d'afrontar empreses i organitzacions que han implementat diverses formes d'automatització dels processos de reclutament de personal. Alguns dels casos han generat molt enrenou als mitjans de comunicació. D'altres, més discrets però no per això amb menys impacte real a les vides de les persones, han estat presentats a congressos, grups de treball i publicacions acadèmiques. Inclourem cada un d'ells al punt exacte en què pugui resultar més aclaridor.

Tota aquesta informació alimentarà la segona part de l'estudi, en la qual proposem un seguit de mesures per tal de detectar/evitar/mitigar els biaixos de gènere i els seus efectes en els processos de reclutament i selecció de personal que es basen o es recolzen en tecnologies algorítmiques.

La segona part del treball parteix de la idea que, des d'una perspectiva ètica, no és lícit assumir que la tendència a l'alça en l'automatització dels processos de reclutament i selecció de personal per mitjà d'enginys basats en IA hagi de suposar inevitablement una opacitat en aquests processos, un efecte de «caixa negra» que n'impedeixi la transparència i la traçabilitat, i que repercuteixi, en conseqüència, en una pèrdua de drets de les persones.

Ara bé, per tal que això no succeeixi, és convenient comptar amb uns principis orientadors i unes mesures, tant generals com concretes, que es puguin adreçar a prevenir, o en el seu defecte, a redreçar o a mitigar, aquests efectes indesitjats de la utilització de software basat en IA.

Si les pràctiques esbiaixades (discriminatòries) de contractació pel sistema tradicional (humà) poden ser evidents, almenys fins a cert punt, i per tant, combatudes de manera també

tradicional amb l'ajuda de sindicats, recorrent a les entitats i organismes reguladors o a la denúncia legal, el perill i el problema inherent als algoritmes rau en què aquests biaixos romanguin «enterrats», passant així desapercebuts, inqüestionats i incontestats.

Calen, doncs, marcs legals, molts d'ells tot just en construcció, que regulin les pràctiques, especialment aquelles que tenen a veure amb el control del sector privat, això és, de la indústria tecnològica.

Tot i així, i malgrat la complexitat palesada, una part important dels reptes que ens ocupen es pot adreçar apropiadament amb uns bons referents ètics que evitin una resposta exclusivament o excessivament tecnològica.

Per tal que aquesta governança ètica de l'entorn digital sigui possible i efectiva cal plantejar mesures a diferents nivells. Això ajudarà a orientar la variabilitat de situacions pràctiques en què es poden trobar les persones responsables dels processos de reclutament i selecció de personal a una organització.

De retruc, l'aplicació dels principis i de les mesures que aquí proposem permetria prevenir el possible «risc reputacional» derivat del fet que una organització, involuntàriament o per negligència, incorri en una discriminació en els seus processos de selecció de personal basats en tecnologies algorítmiques. Aquest és un benefici secundari de les mesures que proposem que, al nostre entendre, no està mancat d'interès.

1.2. Metodologia

Per a l'elaboració d'aquest estudi hem realitzat una revisió bibliogràfica crítica de la literatura acadèmica sobre el tema a les bases de dades: Scopus, Web of Science i Google Acadèmic. Hem considerat els següents criteris d'inclusió: i) pertinència i rellevància del tema, ii) ser un article revisat per parells o una ponència a un congrés internacional, iii) dels darrers 5 anys (entre 2016 i 2021), iv) en idioma anglès, espanyol, francès, italià o portuguès, v) que estigui disponible el text complert. Un cop eliminats els duplicats hem recollit un total de 51 documents (11 de Scopus, 17 de la Web of Science i 25 de Google Acadèmic) que hem passat a analitzar. Volem subratllar que el gruix dels articles, malgrat que l'interval que hem considerat és de cinc anys, correspon principalment als darrers 3 anys. En aquest període s'observa un increment exponencial de la producció acadèmica sobre el tema, tant d'articles com de *proceedings* de congressos. Això ho atribuïm, d'una banda, al fet que es tracta d'una tecnologia relativament nova que tot just s'està introduint a la societat. I de l'altra,

possiblement a l'impacte mediàtic del cas de l'anomenat «algoritme racista» de la companyia Amazon, cas del que parlarem més endavant i que ha servit, malgrat tot, per a posar en el focus d'interès de la ciutadania i de la reflexió acadèmica el tema de l'ètica algorítmica.

Pel que fa a les fonts seleccionades hem considerat important no cenyir la cerca exclusivament als textos més estrictament acadèmics (articles científics amb revisió per parells), per tal de donar cabuda també a les produccions dels diferents *stakeholders*. Un enfocament sistèmic, com justificarem més endavant, ho fa imprescindible. Per aquest motiu incloem també els textos de congressos, on trobem molta més aportació de la pròpia indústria tecnològica. Tot i que la revisió dels continguts dels *proceedings* sol ser una mica més laxa, creiem que aquesta possible pèrdua de rigor queda ben compensada amb un increment en la diversitat de punts de vista que incloem a la mostra, una qualitat important i alineada amb els objectius que ens proposem.

La selecció en les bases de dades l'hem fet emprant els següents termes clau i estratègies de cerca, que presentem a la *figura 1*. La redacció de l'estratègia de cerca la mostrem per mitjà d'una formulació genèrica, en què prescindim de la sintaxi estricta i d'alguns dels operadors booleans a fi de fer-la més llegible i comprensible. Els termes entre parèntesi corresponen als sinònims. Per «termes afins» entenem aquells termes més específics que el terme general de cerca, amb el que estant emparentats, que son rellevants en alguna de les subàrees de la IA, i que hem utilitzat com a formes alternatives a les paraules clau principals..

Artificial intelligence (AI, algorithm*)

Termes afins: algorithmic decision-making, automation computers decisions, decision-making automation, Information Retrieval Systems, Machine Learning, Search Engines, Word Embeddings, Deep Learning, Natural Language Processing, Text Analytics, AI-based interview, Virtual Work Environment, Aumented Intelligence.

AND

Algorithm* bias (bias in AI, fairness bias, algorithmic fairness, bias in digital ratings, discriminat*)

Termes afins: implicit bias, indirect discrimination, implicit association, prejudice, «invisible cage», embodiment, fairness, trust, perceived fair, emigrants and immigrants, ethnic groups, minority groups, stereotyping, racism, social justice, inequity

OR

Gender bias (gender bias, gender inequality, sexism, gender neutral, implicit gender bias)

Termes afins: gender role, gender stereotypes, women's rights, male, female, gender-non-binary, LGTBI*

AND

Recruitment (AI recruitment, automation in HR, hiring algorithm, automated hiring, algorithmic hiring, HR analytics, talent search, talent acquisition, recommendation systems, ranking algorithms)

Termes afins: online recruiting, online job advertisement, job postings, employment, occupations, Personnel Selection, Human Resources, recruiting, selecting, hiring, management, job opportunities, STEM careers, IT workers, platform workers

AND/OR

Mitigate (algorithm auditing)
Termes afins: data protection, evaluation, control, opaque, transparency, impact assesment, accountability, ethics, law, discrimination law.

Figura 1: Paraules clau i estratègia genèrica de cerca

1a PART:

Ética algorítmica i perspectiva de gènere

2. Introducció a la Intel·ligència artificial i l'ètica algorítmica

2.1. Aclariment conceptual

Entenem per ètica algorítmica aquella branca de l'ètica aplicada a la tecnologia que explora i mira de resoldre el conjunt de qüestions que afloren arrel de la programació d'un algoritme.

Un exemple paradigmàtic d'això seria la programació de vehicles autònoms en cas d'un accident inevitable. ¿Com ha de reaccionar el vehicle en qüestió enfront d'un dilema en el qual ha d'escollir x o y? ¿Com arbitrar els riscs que corren els uns i els altres?

Els vehicles autònoms, però, només són un exemple. La major part d'algoritmes són susceptibles de paràmetres i d'opcions per defecte que tenen una immensa transcendència moral.

Els filòsofs estimen les categories, les definicions, en definitiva, els conceptes. Justament un dels camps de l'ètica aplicada consisteix en generar categories útils per comprendre la realitat i, eventualment, transformar-la. La noció d'ètica dels algoritmes ens ajuda a desenvolupar aquesta anàlisi conceptual. L'expressió indica, d'entrada, un àmbit d'aplicació: el conjunt de qüestions morals que es plantegen en la programació.

Es pot situar aquesta àrea dins d'una subàrea de qüestions en ètica de l'IA, que es refereixen a allò que és bo, just o virtuós de fer amb sistema d'IA. L'ètica de la IA és, a la vegada, una branca de l'ètica de la tecnologia o tecnoètica, més precisament, de les tecnologies de la informació.

L'ètica dels algoritmes està, però, molt a prop del que s'anomena ètica de la robòtica o ètica de les màquines. Aquesta branca té com a referent el llibre de Wallash i Allen, Moral Machines. En aquest assaig, editat el 2008, els autors assenten els fonaments d'una qüestió inèdita, fins aleshores, en filosofia pràctica: ¿Com desenvolupar una moralitat computacional?

L'ètica algorítmica es distingeix de la dels robots perquè és més englobant. Un algoritme no és indispensable dins d'un cos robòtic o d'una encarnació concreta per plantejar qüestions morals. Una aplicació de recerca, de recomanació o de traducció pot, de fet, ser avaluada moralment. La qüestió que roman sempre és la mateixa: ¿Com actuar dins del programa i de l'algoritme per conformar-se als estàndards morals?

De fet, l'ètica algorítmica només es distingeix de l'ètica de les màquines per algunes connotacions. Trenca amb la idea unitària que relacionem amb la imatge de la màquina. Les màquines ens semblen, espontàniament, individualitats, artefactes, pacients morals. Amb tot, seria més adequat, des d'un punt de vista ontològic, parlar en plural i emancipar-se d'aquesta concepció unificada de les màquines. Potser caldria veure les màquines dotades d'IA com a entitats especialitzades que romanen cognitivament opaques.

El cor del problema no és l'artefacte, ni la seva forma, sinó la programació que porta instal·lada a dins. Des del punt de vista de la filosofia moral, els robots o les màquines no són altra cosa que embolcalls dels algoritmes.

L'ètica algorítmica estimula els filòsofs a aprofundir en els sistemes interns i a conèixer-los amb precisió i transparència. Dedicar-se a l'ètica algorítmica és concentrar-se en una àrea molt particular del saber, però, també, en una escala especial, molt poc habitual. ¿Com es manifesta aquest canvi d'escala? ¿Què fa diferent l'ètica algorítmica de l'ètica de la IA que l'engloba?

Vegem-ho amb algunes preguntes pràctiques.

- ¿Cal implementar els vehicles autònoms a les ciutats? Ètica de la IA.
- ¿Com programar un vehicle autònom en cas d'un accident inevitable? Ètica algorítmica
- ¿Cal censurar els robots sexuals? Ètica de la IA.
- ¿Aquests robots haurien d'oferir l'opció de simular una resistència? Ètica algorítmica.
- ¿En quines condicions un robot hauria de tenir drets? Ètica de la IA.
- ¿Una aplicació de trobades interpersonals hauria d'automatitzar o reduir certes discriminacions? Ètica algorítmica.

Mentre que l'ètica de la IA sosté que els sistemes d'IA no haurien d'engendrar discriminacions, l'ètica algorítmica s'interroga per la traducció d'aquests principis generals en el codi informàtic, ja que, del que es tracta en ella és, justament, de codificar la moral. Els algoritmes exigeixen que es prenguin decisions terriblement precises i

nítidament clares. Això significa que l'ètica algorítmica no pot quedar-se en el terreny del dubte o de la vacil·lació.

Aquest tipus de treball ens obliga a pensar sobre els models de justícia disponibles i a aplicar-ne un d'ells. La Technology Review va publicar una bella il·lustració del problema a partir del cas COMPAS, un sistema de predicció de la criminalitat criticat per tractar, injustament, les persones negres.

L'ètica algorítmica no està reservada, únicament, a la programació. Desenvolupa una perspectiva descendent. Mentre que la IA afavoreix una perspectiva ascendent, més macro, a prop del costat de l'ètica de la tecnologia, l'ètica algorítmica focalitza l'atenció en el detall, en l'àmbit microcòsmic.

No és estrany que els filòsofs analítics expressin les seves afinitats amb les exigències de transparència i precisió de l'ètica dels algoritmes, mentre que, en canvi, els filòsofs continentals es sentin més a prop dels grans relats de l'ètica de la IA.

Els dos enfocaments són del tot necessaris. Mentre que l'ètica dels algoritmes explora el funcionament intern de les màquines, l'ètica de la IA s'interessa, globalment, per l'ésser humà, per la seva història, pel seu entorn i per les seves relacions, en definitiva, per tot el que la implementació de la Intel·ligència Artificial pot alterar i transformar.

Es fàcil endevinar que la frontera entre aquestes dues àrees d'estudi no és nítida. Si és possible tenir una programació moral satisfactòria per a un robot, tema que estudia l'ètica algorítmica, aleshores, hi ha una bona raó per fabricar-les, tema que pertoca a l'ètica de la IA, però, ¿i si no és així?

Una aplicació que sembli absolutament necessària pot coaccionar la presa de decisions en ètica algorítmica, tot donant a una app antipandèmia que, per exemple, doni la prioritat a la seguretat sobre el respecte de la vida privada. L'anàlisi moral del disseny pertany a l'ètica algorítmica, mentre que la reflexió sobre la tensió entre el dret a la seguretat i a la vida privada correspon a l'ètica de la Intel·ligència Artificial.

La frontera no és gens clara. De fet, **l'ètica de la IA i la dels algoritmes formen, més aviat, dos pols relligats per un contínuum** que no pas dos territoris separats.

Prenem, per exemple, la veu dels assistents personals. ¿Cal que sigui, per defecte, femenina, masculina o neutre? ¿Ha de semblar jove, madura o d'una persona gran? ¿Ha d'imitar una

veu humana o posseir un accent metàl·lic, robòtic, fàcil de distingir? ¿Quin timbre de veu s'ha d'escollir?

Totes aquestes qüestions concerneixen el disseny de l'assistent personal, la manera com interacciona amb els interlocutors humans. Són relativament precises. Depenen, però, d'un conjunt de consideracions molt generals com, per exemple, amb quin tipus de robots volem viure.

Dit d'una altra manera, una decisió que pot semblar circumscrita, únicament, al disseny vocal d'un robot social pressuposa una reflexió global, a escala de l'ètica de la IA. Els programadors coneixen millor els algoritmes, que no pas l'ètica dels algoritmes. Aquesta àrea els és reservada, d'aquí la necessitat del treball interdisciplinari entre filòsofs, enginyers, matemàtics i sociòlegs.

Les màquines dotades d'Intel·ligència Artificial han de prendre decisions. Generalment han de determinar una opció per defecte. És sabut que una gran part d'usuaris no canvia els paràmetres inicials. Convé, doncs, preguntar-se si les veus femenines de Siri o de Google Home no reforcen l'estereotip que vincula la polaritat femenina al servei. Igualment, cal preguntar-se què produirà, en el cervell de la gent, l'associació d'una entitat no biològica a un gènere. Potser permet concloure que no és una bona idea reproduir en els artefactes inanimats la dicotomia masculí/femení.

¿Cal automatitzar certes normes?

Els assistents personals dotats d'Intel·ligència Artificial i altres robots no desembarquen en un món moralment verge ni pur. Operen en un món que està saturat de desafiaments ètics i, per què no dir-ho, farcit d'injustícies. Per ell circulen diverses jerarquies socials, estereotips i tot tipus de biaixos implícits. Una part important de la feina en ètica algorítmica rau en identificar-los i preguntar-se si és legítim automatitzar-los i reproduir-los en les màquines intel·ligents que utilitzem.

Sembla difícil per als programadors sostraure's a aquesta responsabilitat. Igual que l'opció per defecte, l'elecció del menú és ja un arbitratge moralment carregat.

A França, per exemple, els assistents vocals tenen accent francès i al Quebec, accent quebequès. ¿Què és, però, l'accent francès? ¿És el de Tolosa, el de Marsella o el de Poitiers? És el de la capital per oposició al de les províncies. No és un accent popular; és el de la

televisió. Però, ¿Qui no ens assegura que seria una bona idea que a França Alexia respongués amb l'accent quebequès i Siri tingués l'accent d'Haití? ¿Qui sap si això no facilitaria la vida als immigrants quebequesos a França i dels haitians al Quebec? Com a mínim es podria donar l'elecció als usuaris. Es podria oferir aquesta opció entre una desena d'accents. Això és un desafiament del menú.

Cal escollir a l'hora de fer la programació i no és evident que els sistemes d'IA ho facin a partir de les normes i categories en joc.

Una altra àrea específica de l'ètica algorítmica té a veure amb l'autonomia. En molts casos, no es tracta de programar una simple reacció a un senyal, com el fum que activi el sistema d'alarma. Es tracta de dotar el sistema d'una capacitat de decisió tot integrant-li diverses informacions.

La Intel·ligència Artificial del joc de Go s'adapta al seu adversari, pot reaccionar amb cops sorprenents. ¿Què fer amb els comportaments més complexos i moralment pertinents? ¿Com programar una entitat autònoma o parcialment autònoma?

Això implica que no es pot, simplement, implementar una regla de conducta. Cal també enquadrar un conjunt de metaregles. Els vehicles autònoms permeten il·lustrar la qüestió. Cal programar-los, lògicament, perquè s'aturin quan hi ha semàfor en vermell, però el sistema pot ser més intel·ligent que això. Cal que el vehicle passi en vermell si, així, pot evitar un accident. Aquí és on sorgeix la noció d'autonomia aplicada a la Intel·ligència Artificial. Programar moralment un vehicle autònom és donar-lo la capacitat de decidir per si mateix. En certa mesura, és delegar-li, com a subjecte, el procés de presa de decisions.

Aquesta delegació planteja, però, moltes qüestions: Si un robot està programat, ¿en què és autònom? ¿Es pot parlar realment d'autonomia? ¿No és una *contradictio in terminis*?

Aquesta darrera qüestió també pot valer per entitats biològiques com els éssers humans. ¿Estem programats? I, si ho estem, ¿som realment autònoms o bé ens ho sembla? El vell debat entre determinisme i indeterminisme entra, de retruc, en l'escenari.

Un robot de transport no té necessitat de ser lliure en el sentit filosòfic del terme per decidir si ha de respectar o no el semàfor vermell. El que cal és que tingui instal·lades meta-regles i una capacitat de predir el que li passarà.

¿Es pot considerar que programar una entitat autònoma és, més o menys, com educar-la? ¿És aquesta la paraula o hem de guardar el verb educar i només l'hem de fer servir quan ens referim als éssers humans? ¿Es poden educar els animals? ¿I les màquines? ¿És el mateix educar que ensinistrar? El ball semàntic és inevitable quan ens afrontem filosòficament aquest tipus de qüestions.

¿Es pot educar els vehicles autònoms a prendre les corbes amb cura sobre un terra glaçat o evitar que aixafin un animal quan travessi la carretera? ¿Se li pot ensenyar a resoldre dilemes tràgics com el de salvar un infant o un ancià?

Alguns experts consideren que l'ètica algorítmica és, d'alguna manera, una forma d'educació moral artificial. Segons alguns tecnòlegs, educar un vehicle autònom o educar un infant no són tasques tan llunyanes, encara que tampoc no són idèntiques. En ambdós casos, es tracta de preparar-los a desenvolupar-se sols, sense l'ajuda dels enginyers o dels seus pares, perquè quan un estigui sobre la carretera i l'altre confrontat a situacions vitals futures, sàpiguen què cal fer.

El camí de la vida humana està ple de cruïlles i d'esdeveniments imprevistos. A l'infant se li transmeten uns valors, uns criteris, però, és ell, en cada context, el qui ha de valorar què és el més just, el més adequat i sensat de fer. El mateix passa, salvant distàncies, amb el vehicle autònom. Un cop se li han introduït aquests criteris o metaregles generals, l'artefacte circularà sol per la ciutat i haurà de prendre decisions en situacions no previstes pels enginyers a la llum d'aquest conjunt de criteris. Com diu Michel Montaigne, *vaut mieux une tête bien faite qu'une tête bien pleine*.

A l'hora d'educar un infant, però, entenem que cal desenvolupar en ell el pensament crític i el sentit de la justícia. ¿Es pot fer el mateix amb un robot? ¿Pot un robot ser crític respecte el mateix programa que se li ha instal·lat? ¿Té capacitat per prendre distància de si mateix com a actor i esdevenir un espectador de si mateix? En definitiva, ¿la programació moral d'un robot es pot comparar amb l'educació moral d'un infant?

Aquest paral·lelisme té, evidentment, els seus límits. L'infant està subjecte a la feblesa de la seva voluntat (l'acràcia), mentre que el robot, no. A ell se li pot introduir una nova programació moral en pocs segons dins d'un sistema d'IA, mentre que l'educació moral d'un infant requereix molt de temps i molts actors degudament coordinats apuntant a un mateix horitzó.

Els algoritmes es poden controlar perfectament, perquè no fan res més que el que se'ls ha ordenat de fer. Això no passa amb l'infant, ni amb la condició humana en general. Hi ha espai per a la transgressió, per a la disrupció, per a la vulneració del marc normatiu. Forma part de la mateixa naturalesa humana, com diu George Bataille, la seducció per allò prohibit.

El relat arquetípic del Gènesi ho mostra de manera fefaent. Déu crea Adam, però no el programa només per a obeir consignes, les normes que el creador li ha donat. L'ha creat lliure i això inclou la possibilitat de desobeir-lo, de transgredir la norma i de menjar la fruita prohibida.

¿L'aprenentatge automàtic (*machine learning*) es pot relacionar amb l'ensenyament d'aquests criteris ètics? Aquesta tècnica de la IA permet acomplir una tasca, però no seguint unes regles preestablertes, sinó induint-li regles, sovint incomprensibles per a nosaltres, a partir d'exemples o d'assaigs i errors.

Segons alguns experts, és possible posar l'aprenentatge automàtic al servei de l'ètica dels algoritmes. Nick Bostrom, per exemple, considera que el problema del control d'una superintel.ligència no és diferent al d'un infant sordmut que s'emancipa dels seus pares.

¿Com educar-lo? ¿Com orientar-lo vers el bé? ¿Com fer possible que actuï guiat per criteris de justícia?

2.2. El pes dels algoritmes en la construcció de l'actualitat

Nit i dia, estem confrontats als algoritmes. La seva influència sobre les nostres opinions polítiques, el nostre humor o les nostres decisions està més que demostrada.

Lluny de ser neutres, estan farcits dels judicis de valors dels seus programadors i els apliquem, molt sovint, sense tenir-ne consciència. Cal, doncs, qüestionar la seva ètica i trobar solucions reals i aplicables als biaixos que, de manera inconscient, segueixen els seus usuaris.

¿Quin és l'objectiu de Facebook? ¿I el de Twitter?¿Quina és la funció d'una xarxa social? La resposta simplista, però no menys justa, és seleccionar la informació que ens serà presentada al nostre mur, per tal que passem el màxim temps possible en xarxa. Darrera el fil de les actualitats s'hi amaga una selecció de continguts, publicitaris o no, optimitzat per a cada usuari i reforçat per algoritmes.

Gràcies a ells, les xarxes socials determinen què és el que serà més interessant per a cadascú de nosaltres. Sense qüestionar la utilitat d'aquestes pàgines web, el seu funcionament planteja, però, grans i difícils qüestions ètiques.

Segons Christine Balagué, investigadora a *Télécom École de Management* i vicepresidenta del *Conseil national du numérique*, el subjecte de la captació de les dades personals és conegut, però molt menys el tractament de les dades per part dels algoritmes. Encara que els usuaris presten cada cop més atenció al que comparteixen en les xarxes socials, això no vol dir, ni de lluny, que es preguntin com funciona el servei que utilitzen. Aquest coneixement només el té Facebook o Twitter.

Els algoritmes són per tot arreu en les nostres vides, presents en les nostres aplicacions mòbils i serveis web que utilitzem quotidianament. Des del matí fins a la nit, hem de prendre decisions, però no ho fem sobre el buit, sinó confrontats a tota mena de suggeriments, estem exposats a informacions que són tractades per algoritmes.

Netflix, Citymapper, Waze, Google, Uber, TripAdvisor, AirBnb... generen noves vides. Un nombre creixent d'articles d'investigadors de diferents àrees, emfatitzen el poder que els algoritmes tenen sobre els ciutadans, sobre les seves preferències, gustos i eleccions.

El 2015, Robert Epstein, investigador a l'Institut americà de recerca del comportament, mostrava com un motor de recerca podia fer variar els resultats d'una elecció. El seu estudi, fet sobre més de quatre mil persones, li va permetre determinar que la classe social dels candidats en els resultats d'una recerca influïa, com a mínim, un vint per cent els votants inèdits.

Una recerca del 2012 conduïda per Facebook sobre set-cents mil usuaris del seu servei va demostrar que les persones exposades a publicacions d'ètica crítiques amb la creació numèrica posaven majoritàriament un contingut negatiu, mentre que les que estaven exposades a publicacions de caràcter positiu, tenien essencialment continguts positius. Això demostra que els algoritmes són susceptibles de manipular les emocions dels individus sense que en siguin conscients ni n'estiguin informats.

En aquesta opacitat resideix un dels principals problemes ètics dels algoritmes. Sobre un buscador com Google, dos usuaris que fan la mateixa recerca no obtindran el mateix resultat. L'explicació avançada pel servei és la personalització de les respostes a fi d'atendre millor a cadascú, però els mecanismes de selecció dels resultats romanen obscurs.

Entre els paràmetres que es tenen en compte per determinar quines pàgines seran posades en la pàgina, més d'un centenar concerneixen l'usuari que fa la recerca. Sota el pretext del secret industrial, la naturalesa exacta d'aquests paràmetres personals i la manera com són tinguts en compte per part dels algoritmes de Google és, a hores d'ara, desconeguda. És gairebé impossible saber com l'empresa ens categoritza, com determina els nostres centres d'interès o prediu els nostres comportaments.

Un cop s'ha fet aquesta categorització, ¿és possible sortir-ne? ¿Com romandre senyor de la percepció que l'algoritme ha creat de nosaltres? Per causa d'aquesta opacitat, és igualment impossible conèixer el biaix que pot sofrir el tractament de les nostres dades.

Els estudis conduïts per Grazia Cecere, economista de la *Télécom École de Management*, han posat en evidència una discriminació entre homes i dones en els algoritmes d'associació de centres d'interès d'una gran xarxa social. Lluny del mite de la Intel·ligència Artificial malèfica, l'origen d'aquest tipus de biaix cal cercar-lo en la mateixa acció humana. Massa sovint s'oblida, però cal recordar la presència de programadors darrera de cada línia de codi que es dissenya.

Els algoritmes són aquí, abans que res i sobretot, per proposar serveis, el més habitualment de tipus comercial. S'inscriuen, doncs, en el marc d'una estratègia d'empresa. Són creats per respondre a les demandes econòmiques.

Els científics *data* treballen sobre un projecte orientat a l'optimització dels seus algoritmes sense reflexionar, necessàriament, sobre les qüestions ètiques que comporten aquests programes. Els éssers humans tenim percepcions de la societat que integrem de manera més o menys conscient en els marcs lògics que desenvolupem. Projectem en ells el que som i el que pensem sense adonar-nos-en.

El judici de valor d'un algoritme és molt sovint el judici de valor que tenen els seus creadors. Ho mostren els treballs de Grazia Cecere. L'algoritme aprèn el que se li fa aprendre i repeteix mecànicament els estereotips dels seus creadors.

Un exemple emblemàtic d'aquest fenomen el trobem en el diagnòstic per imatge. Un algoritme que classifica una cèl·lula com a malalta o com a sana haurà de ser configurat per fer un arbitratge entre el nombre de falsos positius i el de falsos negatius. Els programadors han de decidir fins a quin punt és tolerable de tenir tests positius en persones sanes per no passar al costat de les persones malaltes que tindrien tests negatius.

Per als metges, és preferible conèixer els falsos positius més que no pas els falsos negatius. Per als científics que desenvolupen l'algoritme, en canvi, és preferible conèixer els falsos negatius que els falsos positius, perquè els coneixements científics són acumulatius. Segons els valors que prioritzen els programadors, privilegiaran una o altra de les professions.

Tal com s'ha dit, l'única manera de lluitar contra l'opacitat dels algoritmes és fent-los més transparents. Des d'octubre del 2016, a França, la Llei de la República numèrica proposada per Axelle Lemaire, imposa la transparència de tots els algoritmes públics. A poc a poc, les empreses se sumen, igualment, a aquesta Llei.

Des del disset de maig del 2017, Twitter proposa als seus usuaris conèixer els centres d'interès que els són associats. Tot i les bones intencions, la transparència amb prou feines és suficient per garantir la dimensió ètica.

La intel·ligibilitat dels codis és sovint negligida. Els algoritmes estan lliurats sota formats que no faciliten la lectura i la comprensió, fins i tot per part dels mateixos professionals. La transparència, doncs, pot acabar essent fictícia.

En el cas de Twitter no s'ha comunicat com s'han atribuït els centres d'interès als usuaris. ¿Quina és la ponderació efectuada pels algoritmes de Twitter entre *Actualitats científiques* i *Afers i finances*, per penjar els continguts en el fil d'actualitat de l'usuari?

Per anar més lluny, cal avaluar el grau de transparència dels algoritmes. Aquest és el sentit de la iniciativa *TransAlgo*. Es tracta d'una plataforma que mesura la transparència tot esguardant les dades que són emprades i les que són produïdes. És la primera plataforma d'aquest gènere a tot Europa. En la mateixa línia, cal subratllar *DataIA*, un institut de convergència sobre les dades iniciat sobre el plató de Saclay per una durada de deu anys.

Aquest programa únic és interdisciplinari i comprèn les recerques sobre els algoritmes en Intel·ligència Artificial, la seva transparència i els desafiaments ètics. Reuneix els equips científics pluridisciplinaris, amb l'objectiu d'estudiar els mecanismes de desenvolupament dels algoritmes.

Les ciències humanes poden aportar molt en l'anàlisi dels valors i de les decisions que s'amaguen darrera l'elaboració dels codis. Cada cop esdevé més necessari descompondre els mètodes algorítmics, fer un *retro-engineering*, mesurar els seus potencials biaixos i discriminacions, fer-los més transparents. Cal fer recerques etnogràfiques sobre els seus

programadors tot submergint-se en les seves intencions i estudiant la fusió sociotècnica dels algoritmes.

A mesura que els serveis numèrics ocupen més importància en les nostres vides, és primordial arribar a definir els riscs que els algoritmes tenen per als usuaris. Partint d'objectes d'estudi de temàtiques tan variades com el medi ambient, la salut, la robòtica o les nano-tecnologies, cal sensibilitzar els dissenyadors de tecnologies en les qüestions ètiques.

2.3. Contra l'opacitat dels algoritmes

La paraula "algoritme" s'està convertint en el vocable de moda i no solament ella, sinó, també, els seus derivats. S'escriu sobre la *cultura algorítmica*, *l'amor als algoritmes, l'ètica algorítmica, el poder dels algoritmes, la societat algorítmica* i, fins i tot, la *governança algorítmica*.

Ho expressen S. Habitable i G. Donem quan diuen que els algoritmes estan transformant les ciències, les indústries i la societat, que alteren les nocions de treball, de propietat, de govern, de vida privada i d'humanitat.

D'una banda, els algoritmes ens faciliten la vida, però de l'altra, ens plantegen molts dubtes i grans preguntes. Els algoritmes vehiculen, tot tipus de pors, però és bo que els mateixos programadors ens ajudin a per extirpar-les del cos social.

Estem completament envoltats per algoritmes. Alguns reaccionen als mercats financers, altres actuen en el terreny de les asseguradores, n'hi ha en el camp dels mitjans de comunicació social, altres treballen en el món de la seguretat; n'hi ha que ens guien en les nostres eleccions de consum i també n'hi ha que piloten les xarxes socials.

L'algoritme ha entrat en la vida de la societat, però ho ha fet, perquè, en certa mesura, se li ha confiat cada cop més les operacions essencials. Decidir l'orientació d'un estudiant, decidir si rebrà o no un ajut, preveure si un detingut té possibilitats de reincidir, anticipar el resultat d'un procés, entre altres, són accions que fan quotidianament els algoritmes, però això no vol dir que ho facin respectant un cert nombre de principis ètics, com el respecte als drets individuals, l'equitat i la no discriminació.

Ara per ara, els algoritmes plantegen problemes molt espinosos, ja que estan embolcallats d'una aura d'objectivitat científica, com si una decisió presa a partir dels seus consells fos

indiscutible perquè, suposadament, és purament, mecànica i està desproveïda de tota mena de prejudicis. Es parteix del fals *supositttum* que són neutres.

Si això és així, és fàcil arribar a la conclusió que és millor sotmetre's al veredicte d'un algoritme més que no pas a un jutge humà, susceptible de prendre decisions que varien en el decurs de la jornada pels efectes de la fatiga i de l'estat d'ànim.

L'algoritme de funcionament matemàtic es presenta, doncs, com una solució per pal·liar la fal·libilitat humana. I, no obstant això, convé un punt de vista ètic i interrogar-se sobre aquesta pretesa objectivitat.

Com diu Dominique Cardon, els algoritmes no són neutres. Reforcen una visió de la societat que els ha estat donada per part dels qui els programen o bé per part dels qui paguen aquests programes. Els artefactes tècnics contenen els principis, els interessos i els valors dels seus dissenyadors. La posada en marxa operativa d'aquests valors passa per decisions tècniques, variables estadístiques i mètodes de càlcul.

En la mesura en què els algoritmes classifiquen, operen, categoritzen o recomanen, entre altres operacions, entren, de ple, en el camp de l'ètica. Per això, és imprescindible reflexionar sobre com ho fan, com arriben a les conclusions que arriben, però això solament és possible si es dissol la seva opacitat. Enfront d'aquesta opacitat, cal reclamar transparència i posar més de manifest que mai la necessitat d'obrir les caixes negres.

La raó d'aquesta petició és evident. **En una societat democràtica, convé que decidim públicament amb quins criteris i amb quins principis desitgem que funcionin els algoritmes** i, fins a quin punt, volem delegar-los les nostres decisions i, a la vegada, quin tipus de control desitgem poder exercir sobre ells.

Els exemples que hem anat veient al llarg d'aquest assaig, mostren que no es pot partir de la neutralitat dels algoritmes. En la mesura en què se'ls implementen criteris de decisió, els algoritmes no són, de cap manera, neutres. Són sistemes automatitzats que contenen, fàcilment, biaixos.

La qüestió dels biaixos ve de lluny. El 1996, B. Friedman i H. Nissenbaum evocaven ja el biaix que podia haver-hi en un sistema informàtic. Segons ells, els biaixos podia revestir tres formes: els preexistents, és a dir, continguts en les actituds o pràctiques i en les institucions que preexisteixen al sistema i que poden marcar de manera explícita i conscient la presa de

decisions o bé de manera implícita i inconscient. En segon lloc, els tècnics que provenen de consideracions tècniques i, finalment, els emergents que apareixen en el context dels usuaris dels sistemes.

Si bé les preguntes ètiques afloren al voltant dels algoritmes procedimentals clàssics, el problema de l'opacitat algorítmica ha tornat al cor de l'actualitat amb l'èxit del *Big data* i de la Intel·ligència Artificial que ens condueixen a concentrar-nos en els algoritmes d'aprenentatge estadístic.

El que es discutia fins fa poc eren els algoritmes de l'aprenentatge supervisat. Com recorda Y. Lecun, el principi bàsic de l'aprenentatge supervisat és sempre el mateix: consisteix a ajustar els paràmetres del sistema per reduir una funció de cost que mesura l'error mitjà entre la sortida real del sistema i la sortida desitjada, calculat sobre un conjunt d'exemples d'aprenentatge. Reduir aquesta funció de cost i entrenar el sistema són una sola i mateixa acció.

La pregunta ètica cal que estigui atenta a la doble dimensió de les dades, d'una banda, i de l'algoritme, de l'altra. Les dades poden, evidentment, contenir biaixos causats per la seva no representativitat o bé perquè tradueixen una realitat que és, en si mateixa, discriminatòria.

Cal recordar la polèmica suscitada pel servei de *Google Photos* que classificava les persones negres com goril·les. En aquest cas, el problema venia del fet que la base de l'entrenament no era representativa de la població real. El sistema havia estat, sense dubte, entrenat a partir de clixés elaborats, majoritàriament, per persones blanques.

Per il·lustrar-ho amb un segon cas, imaginem un algoritme que hagi estat entrenat a partir d'un conjunt de decisions passades. Es evident que reproduirà certs biaixos discriminatoris, per exemple, en relació a les dones.

Aquestes raons fan que es presti atenció en un cert nombre de principis ètics en relació a les dades tractades, com l'absència de biaix en les dades, que aquestes respectin la diversitat de cultures o de grups, que no comportin cap risc de discriminació, que els dissenyadors s'interroguin sobre les variables que són susceptibles de ser socialment discriminades.

La qüestió de les dades és molt rellevant, però cal subratllar, a la vegada, la del tractament i les modalitats d'explotació d'aquestes dades. Els algoritmes treballen sobre aquestes dades a fi de fer-les parlar.

El problema esdevé molt gran quan és difícil de saber, exactament, quins són els criteris que s'han tingut en compte per desenvolupar una classificació i és, justament, aquesta la situació que es dona amb certs algoritmes de la Intel·ligència Artificial. Aleshores és difícil traçar el camí que ha recorregut la màquina per prendre aquella decisió o identificar els criteris que s'han tingut en compte per prendre-la.

La capacitat d'aprenentatge incrementa considerablement la dificultat d'explicació i fa que el dissenyador mateix no estigui en condicions de comprendre el comportament del sistema. Mentre que els algoritmes clàssics tradueixen un model que es presta a explicacions, perquè està fet per analistes, l'aprenentatge calcula un model per ajustament dels seus paràmetres, treballa amb milions de dades.

Aquesta qüestió ens condueix, naturalment, a altres desafiaments ètics respecte els algoritmes, a saber, la nostra capacitat de comprendre'ls a fi de poder-los governar. Volem governar-los i no ser governats per ells.

¿Què estem demanant, doncs, als algoritmes? La resposta és simple. Els demanem que puguin retre comptes, els exigim transparència, lleialtat, així com l'explicació de les decisions preses.

¿De què parlem quan parlem de transparència? ¿Quan és transparent un algoritme? Quan el seu funcionament està clarament explicat i quan les dades que manipula són precises. La transparència d'un algoritme permet verificar les decisions que pren, les eleccions que fa. El principi de transparència dels algoritmes està, així, estretament lligat al de lleialtat i al d'equitat.

Un principi d'aquesta naturalesa és, evidentment, desitjable des d'un punt de vista ètic, però la seva aplicació és problemàtica.

Els algoritmes poden patir una triple opacitat:

i. La primera correspon a una estratègia intencional, la del secret industrial;

ii. La segona resulta del fet que el codi, en si mateix, no és una realitat comprensible per a tots;.

iii. La tercera, la més problemàtica, prové del conflicte entre l'optimització matemàtica en alta dimensió i les exigències semàntiques de l'explicació humana que demana raons.

La primera forma d'opacitat redueix considerablement la petició de transparència. Aquesta transparència necessitaria que es divulgués el codi del sistema algorítmic i això s'oposaria al secret industrial i als drets de la propietat intel·lectual. Fins i tot en el cas que fos revelat, l'algoritme romandria inintel·ligible per a la gran part de ciutadans.

Aquestes dues raons limiten l'intent de posar en marxa una plataforma nacional d'auditoria dels algoritmes, sigui a través d'un cos públic d'experts o bé mitjançant una homologació d'auditories privades.

És una qüestió major la de si realment volem governar els algoritmes i no ser governats per ells, però solament es pot resoldre si disposem del poder de comprendre la lògica general del funcionament de l'algoritme, les criteris que regulen les decisions i que sigui explicable i interpretable. Els ciutadans tenen dret a comprendre les modalitats segons les quals són calculats i classificats.

Com s'ha dit, aquestes exigències són problemàtiques pel que fa als nous algoritmes de la Intel·ligència Artificial. Les iniciatives tècniques es multipliquen per fer-los més intel·ligibles. Això és una bona notícia, però exigeix que la reflexió ètica es recolzi sobre aquesta intensificació de la recerca tècnica.

No hi ha res de fatalitat en el que els algoritmes fan. Tampoc d'atzar. Tot al contrari, estan programats. Poden ser veritablement auditats i se'ls pot exigir que els criteris que els regulen siguin nobles èticament, com el d'enfortir la justícia i l'equitat.

Cal capacitar la ciutadania per a la comprensió dels sistemes informàtics. Dit d'una altra manera, cal reforçar l'autonomia i la reflexió a fi de pal·liar les situacions d'asimetria que poden establir els algoritmes. Cal capacitar la ciutadania perquè prenguin decisions informades i de manera lúcida.

Les exigències ètiques, doncs, són clares. Reforçar la matriu democràtica dels algoritmes i reforçar la llibertat dels usuaris per tal de no cedir a la governança algorítmica.

Més enllà de l'explicació dels algoritmes, alguns estudiosos s'inquieten enfront del que s'ha anomenat la *governança algorítmica.* ¿De què estem parlant?

Amb aquesta expressió es designa un tipus de racionalitat (*a*) normativa o (*b*) política que se sosté sobre la recollida, l'agregació i l'anàlisi automatitzada de dades en quantitats massives a fi de modelitzar, anticipar i afectar, amb anticipació, els possibles comportaments.

La idea subjacent és que es poden governar les persones simplement a partir de la recol·lecció de les dades en estat brut, suposadament neutres i objectives, i que, a través de la seva explotació algorítmica, es pot predir exactament i sense fer hipòtesis els seus comportaments. Cal subratllar el caràcter revolucionari d'aquesta hipòtesi i el canvi que podria comportar a la societat.

El vertiginós desenvolupament de *Big data* i de la IA podria comportar, també, una mutació completa del mètode científic. Amb suficients dades, les xifres parlen per si mateixes. Les hipòtesis es podrien generar automàticament a partir de les dades. S'estima que la lògica inductiva que presideix les tècniques d'aprenentatge estadístic van, igualment, en aquest sentit.

El problema, com subratlla A. Rouvroy, és que es troben sorpreses seguint aquest camí no per millorar el món en el qual vivim, sinó, tot al contrari, per deixar-lo en l'estat amb el qual l'hem trobat tal com hem vist amb la qüestió de les desigualtats que podrien ser consagrades i reforçades, tot neutralitzant la possibilitat de la crítica. La pretensió d'objectivitat és, precisament, l'oblit de l'elecció política. L'obertura de les caixes negres té, doncs, una funció crítica essencial.

No podem oblidar, però, que hi ha una gran diferència entre produir un sistema de recomanació per millorar, per exemple, els suggeriments que es poden fer sobre una plataforma com Netflix i modelitzar el món social amb tota la seva complexitat.

La deontologia científica i l'ètica de la recerca científica han de prevaler. No es poden tractar les dades sense interrogar-se, prèviament, sobre la realitat que representen.

Ens podem preguntar per la posada en marxa de models en els quals la previsió és prioritària a la comprensió o explicació dels fenòmens. Cal mesurar fins a quin punt aquest tipus de models pot comportar un lliberticidi.

2.4. Prejudicis encarnats en els algoritmes

Algunes corporacions vetllen pel desenvolupament de solucions ètiques als problemes que planteja el desenvolupament de màquines intel·ligents i la seva implementació en la societat. Enfront dels sistemes d'Intel·ligència Artificial que engendren o confirmen prejudicis racials o sexistes, volen trobar els mecanismes adients perquè això no es produeixi.

Construïts sobre el model de la *machine learning*, els algoritmes d'Intel·ligència Artificial són capaços d'optimitzar, permanentment, els seus càlculs a mesura que van tractant les dades.

Mitjançant els criteris de funcionament o les dades d'aprenentatge, un algoritme pot, malgrat tot, estar esbiaixat i reflectir les discriminacions que hi ha en la societat.

Hi ha dos tipus de biaixos algorítmics, els que reprodueixen les discriminacions que hi ha en el cos social i els qui els fabriquen, perquè són construïts sobre jocs de dades d'aprenentatge no representatius de la societat,

El 2015, uns investigadors van senyalar el biaix sexista de la plataforma publicitària automatitzada de Google, *AdSense*, en la qual a les dones els proposaven ofertes de treball menys remunerades que les que anaven adreçades als homes amb un nivell semblant de qualificació.

El 2018, uns investigadors van veure que una plataforma de reclutament com la d'Amazon, discriminava les dones, tot i que el sexe no era una dada determinant. Les dones concernides mencionaven menys mots clau en les seves candidatures que els seus homòlegs masculins.

El biaix sexista de l'algoritme reproduïa, mimèticament, un biaix preexistent en la mateixa societat que els enginyers que el van dissenyar no van contemplar críticament.

Verificar que un algoritme no discrimina certs grups de la societat és un desafiament de primera magnitud per als programadors. Els encarreguem que descodifiquin les desigualtats per codificar la igualtat. Per evitar introduir els biaixos d'aprenentatges, però, els científics *data* han de garantir que els jocs de dades siguin suficientment heteròclits. Calen revisions creuades del codi i permetre a altres enginyers que aportin una visió diferent.

Els desafiaments ètics al voltant de la IA consisteixen a evitar de reforçar les injustícies o a crear-ne de noves. Això exigeix, segons Marie Crappe, cap de l'oficina de tecnologia de StaffMe, plataforma d'intermediació per les missions puntuals, vetllar-hi de manera permanent. Aquesta plataforma es dedica a treure totes les dimensions no pertinents per al procés de selecció: el sexe, el lloc de naixement, la nacionalitat... per evitar d'introduir biaixos sexistes.

Per anar més lluny, cal diversificar els perfils dels qui conceben els algoritmes. Com passa amb la política, una assemblea no representativa pot tenir conseqüències sobre la societat.

En aquest sentit, cal proposar a les empreses que s'interroguin, permanentment, sobre els impactes discriminatoris dels algoritmes tot **desenvolupant una ètica *by design*, des de la fase de concepció, enlloc de posar en marxa una avaluació d'impacte sobre la discriminació a posteriori**.

Des de l'empresa creadora de dispositius fins als internautes, cal formar en ètica totes les anelles de la cadena algorítmica. Les persones que no coneixen les capacitats de la Intel·ligència Artificial estan privades d'una part de lliure arbitri. La informació permet posar les bases. Cal recomanar, igualment, a les empreses d'instaurar seminaris de formació i de programes de sensibilització per al conjunt d'actors del món numèric.

El desenvolupament de l'ètica és una ocasió per reforçar la confiança en l'economia numèrica. Algunes empreses han creat un interlocutor privilegiat, el cap de l'oficina de l'ètica digital, encarregat de garantir la coherència global de la política ètica i numèrica de l'empresa, més enllà de la conformitat. És un primer pas.

Lligar la igualtat amb la innovació és una oportunitat extraordinària que no pot passar desapercebuda a les grans corporacions. Complementàriament al desenvolupament legislatiu, a la instauració d'un reglament general sobre la protecció de dades, la introducció de l'ètica en aquests processos és imprescindible per generar confiança en la ciutadania.

2.5. La reproducció dels biaixos en els algoritmes

Les tècniques de la Intel·ligència Artificial tenen un potencial extraordinari. Les oportunitats que s'obren són moltes, però plantegen també un allau de qüestions ètiques.

Aquestes qüestions s'estenen per diferents àmbits com ara la utilització de les dades personals, els impactes ecològics o l'emergència de sistemes d'Intel·ligència Artificial capaços de prendre decisions per si mateixos.

Al llarg de la jornada laboral, ens intercanviem un bon nombre d'emails. Per desplaçar-nos, obrim les aplicacions dels itineraris preferits. Utilitzem serveis numèrics quotidianament. Majoritàriament, aquests serveis són gratuïts, però tenen una altra cara que no es pot negligir.

En la mesura en què els utilitzem, les dades produïdes són compilades i analitzades, especialment per a finalitats publicitàries. El perfil de cadascú de nosaltres és analitzat a partir

dels usos que hem fet de les aplicacions, per proposar-nos la publicitat més adequada, tenint en compte els nostres hàbits i costums observats.

Un adagi resumeix bé aquest mecanisme: "Si és gratuït, és que vostè és el producte". Això es verifica en la gran majoria de gegants de la indústria numèrica i esdevé, sovint, el cor del seu model econòmic. ¿Com, aleshores, puc garantir que les meves dades no seran utilitzades per finalitats no desitjables? Emprant, sobretot, el sentit crític.

Si s'utilitza un servei numèric, cal que es dediqui temps a identificar les dades que són recollides a fi de comprendre com poden ser emprades.

En la mesura en que la màquina ens demana que identifiquem imatges per provar que no som un robot, ja estem participant dins d'algoritmes de reconeixement d'imatges.

Quan acceptem que la nostra activitat sobre una pàgina segueixi endavant, tot consentint els *cookies*, estem permetent als algoritmes que ens proposin publicitat filtrada en funció de les nostres preferències. Algunes pràctiques ens poden semblar acceptables, d'altres, en canvi, no, però la qüestió clau aquí és ser-ne conscient.

A nivell europeu, el *Reglament General sobre la Protecció de dades* (RGPD) va començar a ser efectiu el 2018. Aquest reglament tracta de reforçar el marc de les organitzacions, públiques i privades, que tracten les dades personal.

En alguns països, s'observen certes utilitzacions que poden posar-lo en qüestió. És el cas, per exemple, de la Xina on els aparells de l'Estat vigilen, com en el panòptic de Jeremy Bentham, l'activitat web dels ciutadans, així com els seus desplaçaments i comportaments a partir del reconeixement d'imatges.

La gran majoria de ciutats estan videovigilades. Les dades dels ciutadans i dels sistemes d'Intel·ligència Artificial són emprades per atribuir, a cada ciutadà, una etiqueta en funció de les seves accions, que li permetrà accedir, amb més o menys facilitat, als serveis, com els crèdits o els transports.

Les empreses gegants recullen moltes dades. Aquestes són emprades amb finalitats publicitàries, però no únicament. Amb l'adveniment de la Intel·ligència Artificial, poden automatitzar certes tasques. Així, les decisions són preses no per a éssers humans, sinó per part d'algoritmes.

¿Quin risc de biaix hi ha? Molt alt.

Els algoritmes poden reproduir els biaixos humans. No són mai neutres axiològicament, ja que se sustenten sobre bases de dades d'aprenentatge. Les dades poden ser esbiaixades, per exemple, donant una representació imperfecta del món, com li pot passar a un algoritme de reconeixement visual que solament aprèn a partir de subjectes blancs. Si les dades contenen traces de discriminació, l'algoritme tindrà, a la vegada, un comportament discriminatori.

Prenem un exemple concret, el del reclutament. Les direccions de recursos humans d'avui cada cop empren més solucions d'Intel·ligència Artificial amb aquesta finalitat. Els sistemes lògics analitzen automàticament documents com els *curricula vitae* per retenir els perfils més pertinents per a una determinada funció o rol.

L'any 2015, Amazon, per exemple, va posar en marxa un sistema de reclutament per analitzar les candidatures a les seves ofertes de treball. Amazon havia fet un algoritme expressament esbiaixat per identificar candidatures femenines.

El sistema descartava per si mateix les candidatures de dones a favor de la dels homes. I ho feia així perquè havia estat entrenat a partir de les dades de l'organigrama dels empleats d'Amazon on el vuitanta-cinc per cent dels efectius eren masculins. Un cop aquesta informació va ser revelada, Amazon va decidir abandonar aquest instrument.

Aquestes derives no són sistemàtiques i nombrosos serveis fets amb Intel·ligència Artificial són molt útils. En el cas del reclutament, els serveis proposats són sovint molt pertinents per al buscador de feina. Si una persona consulta, per exemple, les ofertes de treball, els sistemes d'Intel·ligència Artificial són molt adequats per presentar les ofertes que millor corresponen al perfil del buscador.

Cal vetllar, però, perquè tots aquests productes estiguin al servei de tot el món i que no reprodueixin comportaments sexistes o racistes. Cal, doncs, mantenir un esperit crític envers aquestes solucions de la Intel·ligència Artificial i exigir transparència en aquests sistemes.

Hi ha, encara, d'altres problemes que cal pensar a fons. Tots hem vist vídeos surrealistes on Barack Obama insulta Donald Trump o Mark Zuckeberg manifesta manipular els usuaris de Facebook.

Van aparèixer durant el 2018 i, com se sap, són muntatges de vídeo, però el seu realisme és imponent. La tecnologia que ho fa possible es sustenta sobre un tècnica recent

d'Intel·ligència Artificial. Aquests muntatges són, sovint, qualificats de *deep fake*, que es pot traduir com *mentida profunda*, però no sempre es detecten ràpidament, ni hi ha consciència d'aquest fet.

La tecnologia que es fa servir és la de *deep learning* o d'aprenentatge profund. Algú pot objectar dient que el muntatge de fotos ja existia abans de la Intel·ligència Artificial i és veritat, però aquest muntatge està cada cop més a l'abast de més ciutadans i els resultats són, cada cop, més convincents que abans.

Si s'observa la pàgina *This person does not exist* (https://thispersondoesnotexist.com/) ens adonarem de la transcendència de la qüestió. En aquesta pàgina s'apleguen una sèrie de retrats genèrics amb aquesta tècnica d'Intel·ligència Artificial. Es pot fàcilment creure que són cares reals, encara que hagin estat generades artificialment. La *deep fake* afecta, a la vegada, el trucatge de la foto, l'àudio i el vídeo.

El 2019, una aplicació xinesa, Zao, va fer molt soroll en permetre, als seus usuaris, reemplaçar la cara de l'actor d'un clip musical o d'una pel·lícula per un retrat a la seva elecció.

No hi ha dubte que aquestes noves tècniques susciten riscs, especialment, el de la desinformació massiva. ¿Què pot fer l'usuari? ¿Està a mercè de tot? ¿No pot creure's res del que se li presenta com a veritat? ¿Ha de practicar un escepticisme metodològic?

Enfront de tot plegat, és imprescindible que es pregunti per la font de la informació, si és legítima, si el mitjà de comunicació és reconegut. També cal comparar les informacions amb altres pàgines digitals i emprar pàgines de *fact-checking* com les que proposen alguns diaris.

S'ha escrit que les dades són el nou petroli del segle XXI i que la Intel·ligència Artificial carbura aquest nou or negre. ¿Quin impacte té des del punt de vista ecològic tota aquesta indústria? No és menor aquesta pregunta, especialment en un context de crisi mediambiental global i de canvi climàtic.

El món digital com a sector econòmic té un important impacte ecològic. S'estima que va representar al voltant del 3,5 per cent de les emissions mundials de CO2 el 2015. Per fer una petita comparació, el transport aeri emet un 2 per cent i les xarxes de telecomunicacions un zero i mig per cent.

Per a poder rutllar, l'economia digital es basa en tots els nostres dispositius digitals (telèfons intel·ligents, ordinadors, tauletes, etc.), però també en nombroses infraestructures menys

visibles: xarxes fixes i mòbils, xarxes empresarials i centres de dades o centres de dades. Aquests centres allotgen els servidors que fan possible Internet i els seus serveis.

Sovint parlem de *cloud computing* per designar aquests servidors remots. El núvol representa tots aquests dispositius que funcionen remotament en tot moment per fer possible les nostres activitats digitals.

Tan bon punt utilitzem Internet o un dels seus serveis, sol·licitem un servidor al món que processarà la nostra sol·licitud i ens donarà una resposta. Per fer això, el servidor consumeix energia.

¿Com podem limitar el nostre impacte ecològic a escala mundial?

Diàriament podem treballar per reduir la nostra petjada ecològica digital. Hi ha gestos senzills com, per exemple, suprimir correus electrònics antics, cancel·lar la subscripció a butlletins inútils o limitar l'ús de plataformes de transmissió.

El desenvolupament de la Intel·ligència Artificial s'afegeix a aquest impacte ambiental. Requereix grans capacitats d'emmagatzematge i capacitats informàtiques. La raó per la qual els algoritmes quotidians funcionen tan bé és que han estat entrenats durant dies i dies utilitzant servidors extremadament potents que consumeixen energia.

Cal assenyalar, també, que s'estan fent esforços perquè els centres de dades siguin, cada vegada, més responsables amb el medi ambient. El seu consum d'energia està optimitzat, de vegades, gràcies als algoritmes d'Intel·ligència Artificial. La calor que produeixen s'utilitza, de vegades, a les zones urbanes, per escalfar edificis propers o una piscina (que consumeix molta energia).

Tot i que els algoritmes d'Intel·ligència Artificial consumeixen molta energia, hi ha moltes aplicacions que milloren les cadenes de producció, així com el nostre consum de recursos i energia.

2.6. Avaluació ètica dels algoritmes

El mot "algoritme" ha entrat, de ple, en el debat públic sense que el seu objecte hagi estat ben definit. Els algoritmes estan en totes les salses, en les noves tecnologies de la informació i de la comunicació, en el *Cloud Computing*, en els objectes connectats, en la Intel·ligència Artificial, en el *block chain*.

S'han convertit en una font de preguntes inquietants al voltant de la manipulació i de l'explotació de les dades numèriques, al voltant de la confidencialitat i la transparència, de l'interès personal i col·lectiu.

En el camp de la salut, la nova ciència que utilitza els instruments matemàtics per tractar la muntanya de dades del *Big data* és crucial. S'espera una millor comprensió d'esdeveniments com les malalties i una presa de decisió òptima i una activitat predictiva.

Vegem alguns exemples d'utilització d'algoritmes.

L'explotació de les dades del *Big data* pot identificar un perill sanitari, un risc per a la salut i, per tant, el pot prevenir. L'actriu americana, Angelina Jolie, per exemple, va saber, després d'una seqüència del seu genoma, que era portadora d'una mutació en el gen BRCA1 que comportava un risc de més del noranta per cent de desenvolupar un càncer en els propers deu anys. Va decidir fer-se operar per evitar-ho.

En el tractament del càncer, els sistemes experts poden tractar en un temps record al voltant de deu mil milions de dades d'una seqüència d'ADN provenint d'un tumor d'un pacient, una tasca que els metges, per molt qualificats que siguin, no poden fer.

L'instrument *Google Flu Trens* té en compte totes les recerques efectuades a Google i mesura quantes vegades el terme "grip" o "gastroenteritis" o "varicel·la" ha estat cercat per tot el món.

La idea és que els individus tenen tendència a cercar informacions sobre la grip o sobre una altra malaltia quan pensen tenir-ne els símptomes i en la gran majoria de casos, estan, realment, malalts. Així que, quan Google anuncia un augment de recerques sobre una epidèmia gripal, aquesta es desenvolupa, generalment, dues setmanes més tard.

Els algoritmes ens ajuden, igualment, a lluitar contra la propagació d'epidèmies. A Àfrica, per exemple, les dades de geolocalització d'un telèfon mòbil són molt valuoses a fi de poder seguir els moviments de la població a través dels fluxos reals de desplaçament i, així, anticipar el desenvolupament de malalties en un país. Això va permetre, per exemple, predir la propagació del virus de l'Ebola.

Les dades massives que afecten un llarga cohort de pacients tenen el potencial d'accelerar la recerca científica i els protocols experimentals sobre nombroses patologies i tractaments. Per exemple, la xarxa social Facebook ha estat emprada per investigadors per establir una

cartografia dels ciutadans americans amb més risc d'obesitat. Una recerca d'aquest tipus, sense aquests instruments, demanaria molt de temps i costaria una fortuna.

Aquest nou món al voltant dels algoritmes i del *Big data* estudia, de manera permanent, el món real amb l'objectiu de crear predictibilitat. Ara bé, planteja, entre altres, una qüestió major: ¿Com garantir que un algoritme sigui ètic?

S'imposa una reflexió sobre l'avaluació ètica dels algoritmes. L'objectiu està posat en donar sentit, transparència, seguretat i confiança a tots aquests instruments, per tal de concebre'ls, explotar-los i controlar-los millor.

L'acció ètica és, abans que res, una resposta a una situació límit i complexa. Generalment, l'ètica del numèric es tradueix per preguntes sobre el comportament i l'ús dels individus enfront de les noves tecnologies de la informació i de la comunicació i, posteriorment, sobre el comportament, cada cop més autònom, dels instruments tecnològics.

Dins d'aquest marc, l'ètica, en la mesura que és una manera de regular els comportaments basada en el respecte als valors, és essencial per aportar un marc a la utilització dels algoritmes. Cal no oblidar la dimensió temporal i molt sovint irreversible de certes decisions. D'aquí la rellevància de la reflexió ètica.

La responsabilitat dels éssers humans està en el centre de l'ètica. També una empresa ha de saber com introduir l'ètica en les seves accions numèriques.

Les preguntes ètiques han de formar part integrant de la seva missió i, així, construir una reflexió ètica. Cal transcendir l'aproximació interdisciplinària i assolir una veritable transdisciplinarietat, una fusió de disciplines per assolir una veritable ètica del numèric en la qual les qüestions socials i morals s'integrin dins de les noves tecnologies de la informació i de la comunicació.

És essencial establir curoses directrius ètiques específiques en el món numèric, sense perdre de vista la qüestió: ¿Pot allò numèric posar en risc els nostres comportaments ètics?

Cal avaluar allò numèric com un sistema continuat sense voler segmentar els diferents elements que el constitueixen.

L'estudi al voltant del tractament algorítmic es desplega segons tres categories interdependents.

a) Hi ha, d'entrada, *l'ètica de les dades* que inclou la identificació, la construcció, la naturalesa i característiques de les dades tractades per l'algoritme i els intercanvis d'aquestes dades.

b) Després hi ha *l'ètica dels algoritmes* que inclou el funcionament, les operacions i els processos associats a l'algoritme al llarg del cicle de vida de les dades.

c) Finalment, l'*ètica de les pràctiques*, que inclou l'explicació sobre la qualitat de les finalitats i els resultats de l'algoritme.

Cal, doncs, fornir un marc de reflexions i de bones pràctiques ètiques sobre l'elaboració, la posada en marxa i l'ús dels sistemes algorítmics, sobre la intel·ligència artificial i els projectes *Big data* a fi de concebre'ls millor, controlar-los i seguir-los en el si de les empreses.

Un govern rigorós i estricte de sancions al voltant d'allò numèric no és, al nostre entendre, el primer que cal fer. L'aspecte reglamentari no ha de representar el primer recurs, però no es pot oblidar. El que és essencial és examinar la dimensió ètica del tractament d'algoritmes *ex ante*.

El govern de les noves tecnologies de la informació i de la comunicació ha de superar llargament la dimensió rígida, purament tecnològica i normativa per abraçar la dimensió transversal, flexible, dinàmica i evolutiva de l'ètica algorítmica.

Aquesta ètica basada en el principi de l'*ethics by design* és la que s'ha d'imposar en els propers anys i ha de ser la pedra angular de la relació de confiança que cal construir amb la ciutadania. Això contribuirà a què els usuaris estiguin més atents i també les instàncies públiques respecte l'explotació de les *Big data.*

Tot plegat conduirà els ciutadans a preguntar-se fins a quin punt poden entregar la seva vida privada als serveis numèrics.

2.7. La responsabilitat dels algoritmes

La Intel·ligència Artificial i els algoritmes han fet irrupció en la nostra vida quotidiana. S'estenen pertot. Les tasques complexes es troben cada cop més delegades a aplicacions més i més elaborades i autònomes, a mesura que els mètodes de *machine learning* i de *deep learning* es van desenvolupant.

Aquests usos exponencials, associats a impressionants volums de dades, disponibles en quasi tots els sectors, obre un munt de qüestions de naturalesa ètica, jurídica i social.

Estem assistint a un vertiginós desenvolupament d'aplicacions a partir de la Intel·ligència Artificial. El seu ús es troba en tots els sectors: la salut, els transports, l'educació, l'esport, la gestió dels recursos humans, la banca, les companyies d'assegurança, el món militar. Molt sovint, aquestes aplicacions es troben integrades en els robots.

La primera qüestió que brolla en la ment és la de la responsabilitat. Quan es produeix un accident que pot estar causat per un vehicle autònom, apareixen moltes preguntes.

¿Qui és responsable si un cotxe autònom aixafa un vianant o si xoca contra un mur i mata els passatgers? ¿El propietari del vehicle? ¿El dissenyador del programa? ¿El mateix programa? ¿Un algoritme pot ser civilment o penalment responsable?

Per mirar de respondre aquesta bateria de preguntes, cal que ens endinsem en l'univers de la consciència. Partim del supòsit que l'ésser humà està dotat de consciència, que es pot fer responsable dels seus actes, perquè és un ésser capaç d'actes lliures, intencionalment pensats i executats des de la seva voluntat.

¿Poden les màquines o els algoritmes esdevenir conscients? ¿Poden ser responsables? ¿Poden respondre dels seus actes, de les seves decisions? ¿Són realment *seves* o bé són introduïdes per un ésser humà?

Abans d'endinsar-nos en una qüestió d'antropologia filosòfica i ètica fonamental, fixem l'atenció en un altre exemple. Als Estats Units, les aplicacions de selecció automàtica de candidatures a base de l'anàlisi dels *curricula vitae* han estat qualificades de discriminatòries. El sistema de *machine learning* estava nodrit de *curriculae vitae* de candidats blancs i masculins. La màquina aprèn a partir del que l'ésser humà li aporta com a nodriment.

La Intel·ligència Artificial és un conjunt de conceptes i de tecnologies que degudament associats fan emergir un sistema capaç de simular, encara que sigui parcialment, la intel·ligència humana en les àrees del raonament lògic o de l'aprenentatge. Les aplicacions són nombroses, tant en l'ajuda al diagnòstic, com en el manteniment de l'assistència. Proporcionen un millorament dels processos en molts camps com la medicina, les finances, la gestió de recursos humans, la gestió de la relació amb el client.

Per assolir aquests objectius, la Intel·ligència Artificial empra mètodes de resolució de problemes de tipus lògic i/o algorítmic. El desenvolupament de la IA ha estat intensament lligat al desenvolupament extraordinari de la potència de càlcul dels ordinadors (amb la cèlebre llei de Moore que encara no ha estat desmentida) i de volums de dades cada cop més i més grans. És el que anomenem *Big Data.*

Els anys vuitanta marquen l'inici de l'aprenentatge automàtic (*machine learning*). Durant els anys noranta, es van començar a desenvolupar els sistemes d'experts, principalment per les aplicacions a les finances, a la salut o al manteniment. Aquests sistemes oferien una ajuda en el diagnòstic gràcies a l'expertesa d'un home del sector que permetia formalitzar la seva experiència de la seva àrea.

Els sistemes d'experts van compondre una base de coneixement, de regles i un motor d'inferència. Es tractava, aleshores, de poder modelar, formalitzar i conservar l'expertesa dels millors professionals en els seus respectius dominis. Es va fer per no perdre aquests sabers i emprant aquestes aplicacions per formar els experts.

Els sistemes experts no van tenir, malauradament, l'eco que haurien merescut, essencialment per dues raons. En aquella època, els costos que representaven els càlculs i els volums de dades eren prohibitius, però, sobretot, perquè molts experts van renunciar a donar el seu saber, a modelar-lo per transmetre'l.

La via, però, ja estava traçada. A la fi dels anys noranta, el mes de maig de 1997, l'ordinador *Deep Blue* va guanyar Garry Kasparov.

A partir de la primera dècada d'aquest segle, l'impressionant creixement de volums disponibles de dades i el desenvolupament de noves potències i infraestructures de càlcul van permetre a certs ordinadors explorar masses de dades sense precedents. Va néixer un nou mot: *aprenentatge profund* (*deep learning*).

L'aprenentatge profund correspon a l'explotació d'enormes masses de dades estructurades o no, que permet a la màquina crear noves regles en funció de les dades posades a la seva disposició. Avui en dia, l'ordinador dotat d'aplicacions d'aquest nivell és més capaç de resolució, en alguns camps, que qualsevol ésser humà per dotat que sigui.

El maig del 2017, un ordinador va guanyar el millor jugador de Go del món. El coreà Lee Se-Dol, campió del món de Go, va ser vençut per la Intel·ligència Artificial, *AlphaGo* de

Google, des de la primera fase d'aquesta comtessa que es va desenvolupar en cinc partides. Després d'això, Lee Se-Dol va decidir acabar la competició de Go.

Un altre exemple d'aquesta superioritat de la màquina en algunes àrees de la vida humana el trobem en un article de l'*Usine Digitale* del 21 d'agost del 2020. En un concurs organitzat per la DARPA (*Defense Advanced Research Projects Agency*) als Estats Units, un conjunt d'empresaris van confrontar els seus sistemes de pilotatge autònom per determinar quin era el millor. El duel final va consistir en enfrontar un d'aquests sistemes a un pilot de l'exèrcit de l'aire dels Estats Units que no va poder vèncer la màquina.

Per acabar amb aquests exemples, que es multipliquen arreu, cal prestar atenció a un article publicat el 28 d'agost del 2020 a JDN que mostra com les noves professions com els científics *data* poden ser superats per les màquines. La pregunta de l'article és clara i contundent: ¿La *machine learning* automatitzada reemplaçarà el científic *data*?

En l'actualitat, el sector de la Intel·ligència Artificial cerca elaborar una IA capaç de percebre l'entorn, de comprendre una situació i, sobretot, de prendre decisions.

Després d'aquesta digressió, tornem a la suposada responsabilitat del vehicle autònom. La jurisprudència, en aquest camp, és, encara, molt pobra. La responsabilitat en cas de sinistre podria imputar-se a l'usuari de la IA, o al seu propietari, o al seu dissenyador o al seu eventual empleat. Tot és qüestió de contracte. El debat jurídic està lluny d'estar tancat i les situacions molt complexes.

Alguns parlen de dotar la IA de consciència, tot sabent que, també en aquest cas, la consciència i els graus de llibertat que li seran associats estaran programats per éssers humans. Ara per ara, aquest nivell de consciència és difícilment programable, perquè no es codifica el que no es governa. El mateix ésser humà no governa la consciència, ni el subconscient, ni l'inconscient. ¿Què significa, al capdavall, dir que *jo soc conscient que soc un home*? ¿I com programar-ho?

El concepte d'Intel·ligència Artificial a dia d'avui, fa referència a una màquina capaç de reproduir certes intel·ligències que emanen del cervell humà, però això no és la consciència d'un mateix o l'autoconsciència, ni la vida emocional. Alguns investigadors creuen que, un dia no molt llunyà, podran crear una intel·ligència conscient sobre un suport material o totalment immaterial. Si aquest dia arriba, no hi haurà cap límit en la concepció de totes les

intel·ligències artificials. Els únics límits seran els de l'aptitud humana per modelar les idees i els conceptes abstractes i desenvolupar els marcs lògics corresponents.

En la mesura en què una consciència pugui ser modelada sota la forma de fets i de regles, és possible concebre-la i implementar-la dins d'un sistema. Res, però, fa pensar que els robots puguin tenir consciència de la seva pròpia existència de manera autònoma.

Per assolir un cert nivell de consciència, l'arquitectura d'un sistema IA hauria d'incloure, de manera necessària, però no suficient, en el seu nucli una base de regles universals, les regles mínims acceptades pel conjunt dels éssers humans. Al costat d'aquesta ètica universal, si és que fos possible, la IA hauria de disposar, igualment, de referencials reforçant les regles lligades a una religió, cultura o ideologia.

També caldria introduir un referencial particular a una professió, com per exemple, el Jurament hipocràtic, en el cas dels metges. Aquests elements referencials podrien ser diferents i donarien resultat a decisions diferents provocant conseqüències diferents en funció del país, de la cultura o de pertinença ètnica.

Certes tribus d'indis americans, fa tres o quatre segles, tractaven les persones grans d'una manera que avui no podem deixar de reprovar. Quan una dona gran ja no tenia més dents i no podia treballar, li donaven una ració de pa i d'aigua i la deixaven morir a la muntanya. Hi ha països, encara avui, que no deixen conduir un cotxe a les dones. Avui, tant una com altra pràctica, la considerem discriminatòria, però forma part d'aquests marcs culturals.

No cal anar tan lluny, però. Posem per exemple un gran hospital de Londres en el qual només hi ha un llit amb reanimació en el servei de cardiologia. Si el servei d'urgències els truca per portar-los un pacient víctima d'un infart, el cap de guàrdia preguntarà per la seva edat. Prefereixen guardar la plaça per a un malalt jove, més que no emprar-la per a una persona de vuitanta cinc anys o més.

Si es vol bastir una Intel·ligència Artificial que funcioni com el cervell humà i prengui el mateix tipus de decisions, s'hi han d'introduir les mateixes regles, preferències i prejudicis. Al costat d'un referencial universal, s'hi ha d'introduir altres referencials vinculats a religions, cultures, tradicions a partir de les bases de regles conformes a les prescripcions d'aquests universos.

L'usuari d'una Intel·ligència Artificial podria escollir el seu referencial d'ús, conforme als seus desigs i al seu propi sistema axiològic. Només li mancarà assegurar-se després amb una companyia que li proposi les conseqüències de les seves decisions que podrien no sempre ser conformes a la Llei del seu lloc.

La gran qüestió és: ¿Volem explicitar totes aquestes regles que podrien formalitzar-se i ser conegudes per tots? Els dissenyadors podrien, igualment, dotar les intel·ligències artificials de lliure arbitri, permetent que aquestes fessin eleccions personals, dins del marc axiològic referencial d'execució. Aquest lliure arbitri podria funcionar, per exemple, sobre la base de l'atzar o algun altre principi, però necessàriament amb algoritmes programats per éssers humans, tot donant a la màquina certs graus de llibertat.

L'oposició entre predestinació i lliure arbitri ha alimentat, des de fa temps, els debats en aquesta matèria. En un context d'Intel·ligència Artificial, l'un i l'altre podrien ser programats, amb un grau de llibertat més o menys gran deixat a la màquina, tot sabent, que aquesta llibertat i aquesta autonomia de decisió haurà estat programada per l'esser humà.

Aquesta qüestió ens condueix a pensar en la nostra pròpia condició. Té sentit que ens preguntem si nosaltres, com a éssers humans, no som, a fi de comptes, una intel·ligència dotada d'una consciència i d'una porció de lliure arbitri, programada per una força superior, que alguns anomenen Déu i altres, l'evolució o simplement, l'atzar i la necessitat.

3. Intel·ligència artificial i risc de biaix de gènere: podem dir que els algoritmes son sexistes?

Abordar el risc de biaix de gènere en la intel·ligència artificial (IA) és un tema complex, des d'una doble vessant.

D'una banda hem pogut comprovar a les pàgines precedents la dificultat que presenta la pròpia noció d'«intel·ligència artificial» i els riscs i reptes ètics que comporta el seu ús en qualsevol àmbit de l'activitat humana on sigui que s'apliqui. Una de les qüestions clau a tenir en compte és que, malgrat la poca intel·ligibilitat i l'opacitat que presenten aquests recursos de software, les persones solen abraçar-los (la tecnologia en general) amb força entusiasme i de forma totalment acrítica.

A la poca intel·ligibilitat del producte/objecte d'estudi, hem de superposar-hi una segona dificultat: la de conceptualitzar què és un «biaix de gènere», així com ser capaços de copsar-los en la pràctica i veure'n el seu potencial risc discriminatori.

Malgrat que a molts dels texts revisats s'esmenta la teoria feminista i els estudis de gènere com a rellevants per a la comprensió del tema, no hem trobat cap estudi en què se'n faci una revisió a fons aplicada al camp de la IA. Tampoc hem trobat textos que integrin plenament els tres àmbits: ètica algorítmica, biaix de gènere, i processos de reclutament i selecció de personal. La majoria d'ells posa el focus en el biaix de gènere o en la selecció de personal, però no en la intersecció d'ambdós. Hi ha doncs un buit de recerca que és important subratllar. En aquest sentit el nostre estudi és innovador i fa una lectura crítica i integradora de les aportacions, traçant ponts entre els diferents àmbits.

Tot seguit presentarem breument algunes de les construccions teòriques de la teoria feminista, la teoria *queer* i els estudis de gènere, tendint possibles connexions cap a l'ètica algorítmica i el món laboral.

3.1. «Dona no es neix, s'esdevé»

Una de les primeres aportacions històriques a la teoria feminista és la incorporació de la perspectiva constructivista, i amb ella la noció de «construcció social», a l'hora d'analitzar la posició i el rol de les dones en l'estructura social. Simone De Beauvoir, a la seva reconeguda obra «El segon sexe», de 1949, planteja la historicitat i la localitat del que significa «ser dona». «Dona no es neix, s'esdevé» diu la seva clarivident sentència (De Beauvoir, 1981). Avui dia,

aquesta expressió l'entenem com un anti-determinisme biològic. Això és: ser dona, home, o qualsevol altre opció de sexe-gènere no binari que pugui oferir una determinada cultura i entorn social, és quelcom que, malgrat la biologia faci la seva aportació inicial, en gran part s'adquireix, s'aprèn. Aquesta adquisició respon a un procés de socialització i d'enculturació continu, que s'inicia amb el naixement (al si de la família) i perdura al llarg de tota la vida (al cercle social més ampli: escola, treball, societat).

D'aquesta primera aportació de De Beauvoir ja podem extrapolar-ne una premissa per a la nostra anàlisi. Quan pensem els algoritmes i la IA ho farem aplicant la perspectiva constructivista i partirem de la premissa que els algoritmes necessàriament incorporen biaixos, fruit tant de la seva «construcció» com de la seva «socialització». Caldrà que ens preguntem, en conseqüència, quan i com s'adquireixen per tal de poder-hi incidir.

La literatura sobre els biaixos de la IA en general, i els biaixos de gènere en particular, és més que abundant. Una part important de la mateixa ens posa sobre la pista d'on podem localitzar-los i ens en fa una caracterització bo i advertint que els biaixos cal esperar-los i cercar-los en tot el cicle de vida d'una IA (Leavy 2018, West 2020, Nadeem, Abedin & Marjanovic 2020, Sun, Gaut, Tang et al. 2019).

La majoria d'aquests biaixos solen ser designats de forma genèrica com «biaixos implícits» (*«implicit bias»*) i descrits com a inconscients i involuntaris, per oposició al que seria una discriminació explícita i directa envers un grup determinat (De-Arteaga, Romanov, Wallach, Chayes, et al. 2019, Kleinberg & Raghavan 2018). Alguns autors atribueixen la seva gènesi (si més no en part) a la manca d'heterogeneïtat dels equips de creació i producció dels productes basats en IA. Segons aquesta tesi, un producte basat en IA pot ser inconscientment i involuntàriament discriminatori envers les dones si qui el concep i dissenya és un equip de treball format exclusiva o majoritàriament per homes. D'igual manera, pot ser també discriminatori envers alguns grups ètnics i/o culturals si l'esmentat equip està format de forma exclusiva o predominant per persones de fenotip «caucàsic» (Yarger, Payton & Neupane 2020). Altres texts subscriuen la mateixa tesi bo i fent-la extensiva a l'entorn organitzacional més ampli, a la cultura institucional. Així doncs, trobem textos que apunten a la «cultura geek» (*«geek culture»*), entenent aquesta com un conjunt de valors i significats compartits als entorns de les empreses tecnològiques, i més concretament dels programadors informàtics, en què les dones rarament hi tenen cabuda, o no la hi tenen en un pla d'igualtat (Tassabehji, Harding, Lee, & Dominguez-Pery 2021).

3.2. El cas dels «platform-workers»

En la mateixa línia, la cultura del treball a les plataformes («*platform work*», «*gig work*»), una vessant molt important de les noves formes de treball autònom (sovint molt precari) del segle XXI, ocupa també una bona part de la literatura sobre el tema dels biaixos i l'ètica algorítmica. Sense centrar-se específicament en la discriminació de gènere, aquests textos posen de relleu l'alta opacitat dels algoritmes i el seu elevat potencial discriminador, posant de manifest el desemparament dels treballadors a l'hora d'entomar i conduir la seva carrera professional o de, simplement, mantenir el seu lloc de treball (Vyas 2021, Jahanbakhsh, Cranshaw, Counts, Lasecki & Inkpen 2020, Kullmann 2018, Rahman, 2021).

La tesi sobre que els algoritmes reflecteixen la composició dels equips de treball i la cultura institucional més àmplia, tot i ser amplament subscrita, és també rebatuda per d'altres autors que sostindrien que «ser dona» no converteix automàticament una persona en «experta en gènere» (d'igual manera que «ser immigrant» no converteix automàticament una persona en expert en «minories ètniques»). Aquesta postura no s'ha d'entendre com una negació dels possibles biaixos, sinó d'algunes de les causes que hi contribueixen. Repercuteix, en tot cas, en les propostes pal·liatives: mentre que uns autors recomanaran fer els equips de treball en IA més diversos (aplicant quotes, per exemple, la «proporció dels ⅔»… o la «Rooney rule» al context nord-americà), els altres preferiran contractar experts en gènere per tal de sensibilitzar les organitzacions i els equips de treball. En tot cas, les postures no són excloents i el debat resta obert (Köchling & Wehner 2020b).

Si reprenem el fil de les aportacions de De Beauvoir, del seu enfocament constructivista se'n desprèn per a nosaltres una segona premissa important. No n'hi ha prou amb dir: «ser dona és una construcció històrica i local, situada». Si aprofundim una mica més en el seu plantejament en podem extreure conseqüències polítiques. «Ser dona» comporta tenir una posició i uns rols predeterminats en l'estructura social i en les relacions de poder i de dominació d'una societat concreta. A Occident històricament aquesta ha estat, per a les dones, una posició i uns rols subalterns. Els infants son socialitzats des del naixement en la diferenciació sexo-genèrica. En aquest procés, les persones amb atributs biològics femenins aprenen a ser un grup social invisibilitzat, dominat, objectificat, violentat: el «segon sexe», en definitiva. Només amb un constant i elevat grau de constrenyiment, de violència (física i simbòlica) per part del grup dominant (el de les persones amb atributs biològics masculins), ha estat i segueix sent factible que el grup subaltern dugui a terme les tasques reproductives i les productives de menor reconeixement social, amb una aparença de normalitat, fins i tot

de voluntarietat. La perspectiva feminista-constructivista ens ensenya a desnaturalitzar aquestes assumpcions, bo i posant-les en qüestió.

D'aquest enfocament se'n desprèn per a la nostra anàlisi que els biaixos dels algoritmes no només incorporen i reprodueixen els estereotips socials de gènere dels creadors i del seu entorn immediat (també d'altres, com els prejudicis i estereotips amb motiu de la raça, l'edat, les creences religioses o trets culturals diferents als propis), sinó que, amb ells, i això és el realment important, també es reprodueixen les relacions de poder, les relacions de dominació, i les diferents formes, subtils algunes d'elles, d'exclusió (Crawford & Paglen 2021).

3.3. Siri i Alexa: el cas de les assistents personals de veu

Un cas paradigmàtic i prou clar de reproducció d'estereotips socials i, amb ells, de les relacions de dominació que incorporen és el disseny dels Assistents Personals de Veu (*PVA: Personal Voice Assistants*). «Siri» i «Alexa», per citar-ne dos dels més estesos a escala planetària, tenen una caracterització femenina. Podem considerar-ho «casualitat», fruit de l'atzar? No, si tenim en compte que Siri i Alexa son dos enginys basats en IA que realitzen una tasca d'assistent personal que tradicionalment segueix sent una posició laboral subalterna i, per tant, eminentment femenina (Adams & Loideáin, 2019). En aquest cas, les autores fins i tot van un pas més enllà en la seva anàlisi bo i indicant com aquests artefactes amb noms i veus femenines contribueixen a naturalitzar la idea que les dones existeixen per a ser «utilitzades» pels homes.

3.4. Breu apunt sobre la inter-sexualitat i el gènere no binari

Arribats a aquest punt caldria fer un aclariment. En la teoria social sol donar-se per apropiada, en la línia constructivista que venim explicant, la pràctica discursiva d'utilitzar el terme «sexe» per a referir-se als trets biològics, i el terme «gènere», per a la construcció social i cultural de les imatges estereotipades i els rols i comportaments associats a cadascun dels sexes. Ambdues categories solen representar-se sobre la base d'un model binari, per bé que a la categoria «sexe» el binarisme es pressuposa inqüestionable, mentre que a la de «gènere» sovint se li atorga una major flexibilitat, i fins i tots en determinats contexts s'admeten models no binaris. No ens queda dins l'abast d'aquesta breu incursió en la teoria sobre el gènere

qüestionar aquests usos intel·lectuals[2]. En la literatura consultada sobre IA i biaix de gènere no hem trobat cap referència a la discriminació envers el gènere no binari.

Per tant, i en aquesta línia, considerem important revisar el concepte de gènere amb les darreres aportacions de la «teoria queer» i dels estudis LGTBI+. Històricament les persones de sexe i/o gènere no conforme o gènere no binari han estat les "oblidades de les oblidades": encara més invisibilitzades que les dones, relegades a la perifèria o fins i tot excloses completa i permanentment de la societat (Preciado, 2016). En aquest estudi tenim la voluntat de no centrar-nos exclusivament en les pràctiques que discriminen les dones, sinó també a qualsevol persona que s'autoidentifiqui amb un sexe o un gènere «no conforme» o «no binari», així com les que puguin ser discriminades per raó de les seves opcions sexo-afectives. No cal dir que els homes també poden ser discriminats per raó de sexe/gènere en determinats perfils laborals. Per exemple, tradicionalment els treballs relacionats amb el «tenir cura», com ara la majoria de les professions sanitàries o educatives, son considerats femenins, així com els treballs de la llar, les tasques administratives, o aquelles on hi hagi el requisit d'una «bona imatge».

3.5. La performativitat del gènere i la violència simbòlica

Seguint amb el nostre breu recorregut per la teoria sobre el gènere, Judit Butler, en l'estela del treball de De Beauvoir, va fer un pas més enllà en l'anàlisi de com es produeix aquesta incorporació, aquest «esdevenir dona», destacant-ne el paper clau que en aquest procés hi té el llenguatge, i més en concret, la seva qualitat o potència performativa (Butler, 2007). Aquesta consisteix, dit de forma molt sintètica, en la capacitat que té el llenguatge de crear allò que està nombrant. Els científics socials solen referir-s'hi com «l'efecte de la profecia auto-acomplerta» o l'«efecte Pigmalió», i consisteix en què les persones solen aprendre i ajustar-se als rols i a les expectatives socials dels termes (etiquetatges socials) amb què se les designa, siguin aquests categories de tipus socioeconòmic (ric, pobre, classe mitja, sensesostre...), mèdiques (malalt, foll, bipolar, cancerós,...), educatives (fracàs escolar, superdotat, inatent, hiperactiu...)[3], etc. En aquest sentit, una dona ho esdevé quan, des del seu naixement, és nombrada com a tal (amb el seu nom de pila i el sexe-gènere que se li assigna) i es va desgranant al seu voltant, al llarg de la vida, el rosari de pràctiques i

[2] Per a iniciar-se en aquesta temàtica indicarem el text d'A. Fausto-Sterling: «Los cinco sexos» (Fausto-Sterling, 1993) com a referent d'una visió no binària i no exclusivament biològica de la categoria «sexe». I d'A. Bolin «La transversalidad del género» (Bolin, 2003) per a una visió panoràmica i transcultural de la riquesa d'opcions de gènere no binari a diferents contexts socials i moments històrics.

[3] Fem notar el sentit socialment pejoratiu que tenen molts dels qualificatius, que en cap cas subscrivim.

representacions, també els ritus de pas, de la «feminitat». Així doncs, seguint els estudis de Butler, quedem advertits de la capacitat creativa i reproductiva del llenguatge, així com de la seva potència subversiva.

Transposant aquesta idea al món dels algoritmes i els seus biaixos, veurem com molts dels estudis analitzats fan incidència, precisament, en aquest àmbit, bo i desvelant els biaixos de gènere que es produeixen en la traducció o la conversió del llenguatge humà (llenguatge natural, text o àudio), i també les imatges (fotografies o vídeos, per exemple, d'entrevistes professionals), en instruccions informàtiques (per mitjà de processos d'etiquetatge i de codificació), en diferents moments al llarg del cicle de vida d'una IA. Així doncs, una bona part de les anàlisis en la literatura consultada posa el focus d'interès en les àrees tècniques del «Word Embedding» i del «Natural Language Processing», entenent que en el procés i les maneres de classificar, d'etiquetar, de codificar, les dades (en definitiva de subsumir la diversitat del mon a unes quantes categories pre-establertes), es produeix una de les formes més clares d'incorporació de biaixos en els algoritmes (Leavy, Meaney, Wade & Greene 2020, Sun, Gaut, Tang et al. 2019). Aquests biaixos, un cop incorporats a una IA, son reproduïts, i fins i tot amplificats, durant el seu ús. Si no son monitoritzats i corregits, els danys reals sobre persones concretes poden ser de gran importància.

Les diferents aportacions senyalen diferents aspectes d'interès. N'esmentarem uns quants.

Si resseguim els diferents moments del cicle de vida d'una IA trobem que els biaixos s'incorporen en algun (sinó en tots) dels següents processos:

D'una banda, en el moment de la concepció i la implementació, incorporant els biaixos dels creadors i del seu entorn més immediat (equip de treball, cultura institucional) en la pròpia concepció del producte o en la seva lògica interna. Aquí podríem incloure, des d'un exemple senzill com la caracterització femenina d'una assistent virtual, fins el cas d'una IA de suport a una web d'anuncis de treball que, per la seva construcció, fa que algunes demandes –com ara els anuncis de feina de les empreses tecnològiques o de les carreres STEM– impactin menys en les dones; és a dir, fent que les dones tinguin menys possibilitats de rebre l'anunci i optar a la feina (Lambrecht & Tucker 2019, Böhm, Linnyk, Kohl, Weber et al. 2020).

De l'altra, en el moment de l'entrenament/aprenentatge de l'algoritme (Machine Learning, Deep Learning) per mitjà d'una base de dades concreta per a cada cas (*Training Datasets*).

Pel que fa a l'entrenament, aquest pot ser supervisat o no, i en cada cas les repercussions son diferents. Com es pot suposar, l'entrenament no supervisat és més arriscat de cara a la incorporació de biaixos no controlats (Crawford & Paglen 2021, Nadeem, Abedin & Marjanovic 2020).

Pel que fa a les bases de dades amb les que s'ensenya/entrena un algoritme, n'hi ha de pre-establertes: algunes es comercialitzen i d'altres es troben en codi obert. També es pot deixar que un algoritme aprengui directament a la web, acotant-li o no uns determinants llocs (per exemple, Wikipèdia).

Sigui quin sigui el procés, les dades que nodreixen els algoritmes han hagut de ser etiquetades prèviament per una persona (treballador d'una companyia de software o altre) per tal de categoritzar els elements que son rellevants per a la tasca concreta encomanada a la IA. Aquests etiquetatges depenen de la subjectivitat de la persona que els fa i actuen com vectors de transmissió de valors i estereotips socials, de la cosmovisió de la persona, la cultura i la societat a la que pertany. Al biaix fruit dels grups infra-representats a les bases de dades (per exemple, dones, persones afrodescendents, persones amb discapacitat, persones d'una determinada franja d'edat), se l'anomena «biaix de representació» (Gutierrez 2021, Leavy 2018).

Si, a més, tenim en compte que les bases de dades sovint són històriques, és a dir, contenen material del passat, als biaixos prevists cal afegir el que alguns autors denominen un «biaix històric». Això comporta que un algoritme pot estar aprenent, no només del moment present, sinó també del passat. Aquest va ser el cas de «l'algoritme discriminador» de l'empresa Amazon, que va "aprendre" d'una base de dades dels seus propis treballadors que contenia informació dels darrers deu anys.

3.6. El cas de «l'algoritme sexista» d'Amazon

El 2018, el gegant de la logística Amazon va ser notícia als mitjans de comunicació en relació a l'ètica dels algoritmes que utilitzava en la gestió dels seus treballadors. L'empresa va ser acusada d'utilitzar un «algoritme sexista» de contractació que discriminava les dones i afavoria els candidats masculins. Quan es va indagar en el tema es va poder comprovar que el problema estava arrelat en el fet que l'entrenament de l'algoritme s'havia fet sense supervisió i utilitzant la base de dades històrica dels treballadors de la pròpia empresa dels darrers 10

anys. Això havia produït un biaix de gènere en el seu aprenentatge, atès que els treballadors havien estat, al llarg d'aquells anys, majoritàriament homes[4].

Per tant, en relació als moments en què un algoritme pot adquirir biaixos, cal afegir el moment dels seus usos concrets, quan se l'ajusta amb els paràmetres adequats per a la realització d'una tasca productiva concreta, i se'l nodreix amb un material real. En aquest moment ja no estem a l'entorn tècnic i expert dels treballadors de la indústria tecnològica, sinó en els contextos organitzacionals o productius dels diversos agents, públics o privats (departaments de recursos humans, institucions, organitzacions, empreses) que poden comprar un software basat en IA per utilitzar-lo encomanant-li una determinada tasca o procés. Per exemple, el cas prototípic d'una pre-selecció de CV d'un grup de candidats per tal que passin a la fase d'entrevista. En aquestes situacions reals, els algoritmes han de ser ajustats, monitoritzats i avaluats per tal de valorar-ne no només el seu rendiment en termes de cost-efectivitat sinó també la seva precisió (no biaix) amb els paràmetres reals d'una tasca concreta i les característiques de les dades reals. Cal veure si els resultats no només són correctes (apropiats), sinó també si són justos (Fabris, Purpura, Silvello & Susto 2020, Fernández-Martínez & Fernández 2020, Hangartner, Kopp & Siegenthaler 2021, Chen, Ma, Hannák & Wilson 2018).

Si reprenen el nostre fil teòric, un concepte que complementa i ajuda a entendre la noció de «performativitat» del llenguatge, és el de «violència simbòlica». Aquest concepte resulta especialment clarificador a l'hora de copsar el nivell de profunditat en què actuen els mecanismes discriminatoris. Com venim plantejant, la dominació d'un gènere per un altre no és un fenomen natural sinó un fet social naturalitzat. El gènere subaltern no neix amb aquesta condició, sinó que se'l subjuga de forma insidiosa i violenta. Seguint la tesi de P. Bourdieu a «La dominació masculina» (2001), la violència que s'exerceix sobre les dones no és només física, sinó sobretot simbòlica, discursiva. És per mitjà del discurs, de l'univers simbòlic que el dominador i el dominat, l'opressor i l'oprimit, comparteixen quotidianament –una mateixa cosmovisió, uns mateixos valors, ideals, estil de vida– que les classes subalternes assumeixen de forma subtil el seu encàrrec. La pitjor de les discriminacions és, doncs, la que s'infringeixen les dones a sí mateixes: per mitja de les pròpies eleccions i de l'autocensura. Traçant ponts amb el mon de la IA: tal i com ens mostren alguns estudis, son les pròpies dones les que

4 Dustin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 11-10-2018. Recuperat a: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [Consultat 10-6-2021]

sovint no desitgen o desisteixen de les posicions laborals en ambients molt masculins i masculinitzats, com és el cas de les empreses tecnològiques (Tassabehji, Harding, Lee & Dominguez-Pery 2021) i de les carreres tecno-científiques (Lambrecht & Tucker 2019, Böhm, Linnyk, Kohl, Weber et al. 2020), on s'indiquen, a més, la poca freqüència de dones en posicions de lideratge en aquests entorns, i la necessitat que tenen d'orientació i suport. Com hem vist prèviament, i amb això tanquem el cercle, els entorns de treball poc diversos tenen una probabilitat més alta de generar discriminació, tant per mitjà d'algoritmes com directament.

3.7. La intersecció de múltiples discriminacions

Fins aquí hem vist doncs com el sexe/gènere esdevé un marcador social que situa les dones a l'escalafó més baix de l'estructura social, una pràctica que la IA incorpora i reprodueix, quan no magnifica. En la divisió social del treball, el gènere ha estat, i segueix sent en l'actualitat, un factor discriminatori de primer ordre, però no l'únic. Així doncs una bona anàlisi en clau de gènere no pot deixar d'incorporar una de les darreres aportacions, com és la teoria de la interseccionalitat. Kimberlé Crenshaw va difondre la idea a partir de l'estudi d'un seguit de casos de discriminació laboral que van patir unes dones afrodescendents i que ella mateixa, com advocada, va defensar als tribunals i va perdre. El problema real de les seves defensades era que patien múltiples discriminacions alhora, i que aquestes es superposaven multiplicant la magnitud dels danys. La justícia americana en aquell moment no permetia defensar una causa per la via de la discriminació racial i, alhora, de la discriminació de gènere; calia triar una o l'altra. Per aquest motiu l'advocada fracassava reiteradament en la seva defensa. L'analogia que va emprar Crenshaw, i que va fer fortuna en les ciències socials, resulta encara molt didàctica: «una persona, a una cruïlla de carreteres, pot ser atropellada per varis camions a la vegada» (Crenshaw, 1989).

Així doncs, haurem de tenir en compte que els múltiples factors de discriminació, i per tant també els biaixos dels algoritmes –edat, gènere, ètnia, nivell formatiu, discapacitat, religió, opcions sexo-afectives no heterosexuals– es poden superposar, interseccionar, donant lloc a injustícies majors. Segons aquesta teoria, que concorda amb els estudis de la literatura revisada (molts d'ells estudis de tipus experimental, en què es fan tests amb simulacions), una dona afrodescendent tindria moltes més probabilitats de patir una doble discriminació causada per la racialització i el sexisme implícit dels algoritmes, en comparació amb un home caucàsic. Un cas clar que exemplifica aquesta teoria en el mon dels algoritmes el trobem en

la poca eficàcia, malgrat els considerables esforços, dels mecanismes correctors (o sigui, de les mesures anti-biaix), creades per la pròpia indústria de la IA per tal de corregir el seu reconegut problema (Bornstein 2017, Raghavan, Barocas, Kleinberg & Levy 2020, Vasconcelos, Cardonha & Gonçalves 2018). Les mesures anti-biaix no funcionen principalment perquè és molt complex «esborrar» els marcadors de gènere de les dades. No n'hi ha prou amb no posar el nom a un CV, o amb treure els pronoms d'un text, o fins i tot amb no etiquetar el gènere d'un determinat contingut (per exemple, fotogràfic). El gènere deixa una empremta en la vida, la biografia i el currículum de les dones fàcilment detectable per un algoritme avesat a processar infinitud de dades i trobar-hi patrons de co-relació. Algunes de les fonts revisades parlen, en aquest sentit, del «cluster» del gènere (De-Arteaga, Romanov, Wallach, Chayes et al. 2019). Posarem com a exemple un dels casos amb què aquests autors l'il·lustren, en un context de reclutament i selecció laboral. Tot i eliminar els marcadors de gènere més evidents d'un CV, la trajectòria professional de les dones sovint presenta buits temporals o una baixa productivitat (feines a temps parcial o menys qualificades) en l'apartat «experiència professional», fàcilment atribuïbles als anys de procreació i cura dels fills. Una IA adreçada a fer un rànquing de candidats que puntués aquest apartat sense cap mesura compensatòria (des-biaixadora), seria alhora tant objectiva com injusta. Els múltiples factors de discriminació interseccionen, co-relacionen, inexorablement

3.8. La dimensió transformadora

Finalment, però no menys important, no voldríem tancar aquest apartat sense apuntar que de les contribucions dels autors que hem presentat se'n pot extreure una dimensió transformadora, que és important tenir ben present: allò socialment construït pot modificar-se, deconstruir-se. Per tant, també els biaixos de gènere dels algoritmes, per molt ocults i incrustats que puguin estar.

Això, com veurem en el Segon informe, ho podrem fer, com a mínim, en tres moments diferents del cicle de vida dels algoritmes, sense que aquests siguin excloents:

i. Prenent mesures per a què el biaix no es produeixi prèviament a la seva implementació, és a dir, durant la fase de disseny i programació.

ii. Adoptant mesures per a redreçar el biaix si aquest ja s'ha produït, és a dir, amb mesures correctores post-hoc o anti-biaix.

iii. Afrontant una tasca més de fons, obertament política, que condueixi a què la societat, de la que son fruit els algoritmes, sigui globalment més justa i igualitària.

Com es pot preveure és una tasca immensa i en cap cas senzilla. Ens hi anirem endinsant progressivament.

4. La introducció de la IA en l'àmbit del reclutament i la contractació de personal i el problema del biaix de gènere

Com ja hem anticipat, l'automatització total o parcial dels processos de reclutament i contractació de persones va guanyant terreny dia a dia i a escala global (tot i que de forma desigual en els diferents països i sectors productius). Segons Hmoud & Laszlo (2019) hi ha hagut una tendència emergent d'utilitzar tecnologies d'Intel·ligència Artificial (IA) dins de l'entorn empresarial al llarg de les dues darreres dècades. Un informe recent estima que el 98% de les empreses de «Fortune 500» utilitzen algun tipus de software basat en IA per tal de captar treballadors i en el seu procés de contractació (Hmoud & Laszlo 2019).

Per tal d'abordar la qüestió del biaix de gènere en la IA i el seu potencial discriminatori en el cas concret dels processos de reclutament i contractació de personal, així com mesures per tal de prevenir-lo o, en el seu defecte, mitigar-lo, hem de poder conceptualitzar el tema sense perdre'n la complexitat (ni perdre'ns en ella), bo i preservant-ne el màxim de capes de significat. La literatura revisada apunta a una manca de definicions i consens sobre el nucli dur de conceptes clau: *«there is no singular or unified way of interpreting the meaning of discrimination, or how it might feature in hiring practices, nor is there consensus on any computational criteria for how "bias" should be defined, made explicit, or mitigated»* (Sánchez-Monedero, Dencik & Edwards 2020). La complexitat reclama, doncs, una mirada sistèmica.

4.1. L'ecosistema de la IA

Amb aquest propòsit considerarem a la nostra anàlisi que entren en joc tot un seguit d'actors o *stakeholders* que es troben imbricats a diferents àmbits i nivells, en relacions de producció, distribució i consum de la IA. És clar doncs que des d'aquestes posicions diferenciades tindran també diferents interessos i perspectives sobre la viabilitat i els possibles beneficis/perjudicis de la implementació de la IA en el mon laboral, així com de la percepció dels seus possibles biaixos. El gruix de tots ells conforma una globalitat interrelacionada i en equilibri dinàmic que podem entendre metafòricament com un ecosistema. En el nostre estudi parlarem i entendrem «l'ecosistema de la IA» com el marc de referència en el qual prenen sentit les diferents representacions i pràctiques entorn la IA. Vegem tot seguit quins son els actors que configuren aquest ecosistema (l'ordre de presentació és aleatori i no respon a la seva rellevància).

a) D'una banda tenim les persones que cerquen feina i que utilitzen algun dels sistemes basats en IA per tal d'accedir a ofertes de treball. Per exemple, al nostre país portals com «Linkedin.com» o «Infojobs.net».

b) D'altra banda, trobem les empreses i/o organitzacions que cerquen treballadors i que utilitzen algun software basat en IA per a fer una part o la totalitat de la feina de reclutament i selecció, i per a prendre decisions basades en aquests processos.

c) En tercer lloc, considerarem els treballadors de les empreses tecnològiques que, a diferents nivells, intervenen en el disseny i la implementació dels productes de software basats en IA (en el nostre cas software adreçat al sector de mercat dels anomenats «recursos humans»). En aquest apartat podríem incloure també els distribuïdors i venedors d'aquests productes.

d) I finalment, però no menys important, els treballadors freelance la feina dels quals està mediada quotidianament (totalment o en part) per una plataforma web, coneguts en la literatura sobre el tema com a *«platform workers»* o *«gig workers»*. Tot i que el nostre estudi no es centrarà en aquesta modalitat de treball, és un sector laboral molt rellevant i a l'alça, capdavanter en l'automatització de processos basats en IA i sobre el que versen un gruix important dels estudis revisats. No en va el «*platform work*», la *«gig economy»,* o fins i tot la *«gig culture»,* son termes descriptors d'algunes de les noves modalitats de treball precari del s. XXI.

4.2. Algunes representacions i tasques encomanades a la IA

Com dèiem, dins la tendència emergent d'utilitzar tecnologies d'intel·ligència artificial, la perspectiva que prenen els diferents *stakeholders* sobre els recursos d'IA disponibles i la percepció que tenen de els avantatges i inconvenients de l'automatització de determinats processos i funcions, varia notablement segons quins siguin els interessos en joc de cada una de les parts.

Així, per exemple, en l'àmbit de la gestió de treballadors s'està donant un fort impuls a la incorporació de solucions basades en IA, especialment pel que fa als processos de reclutament i selecció de candidats (Laurim, Arpaci, Prommegger & Krcmar, 2021). En aquest context, la IA gaudeix d'una representació molt positiva i es solen considerar i valorar-ne només els seus presumptes beneficis, deixant de banda les visions més crítiques. Tal és

així que un feix important de literatura avala la idea de que la IA «ha vingut per a quedar-se» (Hmoud & Laszlo 2019, Upadhyay & Khandelwal, 2018).

Alguns dels avantatges que es proclamen són: la IA és presumptament més objectiva (fins i tot admetent que pugui tenir biaixos) que els éssers humans. En aquest sentit, les solucions basades en IA se solen publicitar i vendre a les organitzacions com un recurs per a neutralitzar els biaixos de les persones que tradicionalment es dedicaven a fer les tasques de reclutament i selecció, evitant la subjectivitat del reclutador i millorant, en conseqüència, «l'adquisició de talent» per a les empreses. Val a dir que al sector tecnològic (*IT work*) això és quelcom remarcable, atès que un dels seus problemes endèmics és precisament el de trobar persones òptimament qualificades (Laurim, Arpaci, Prommegger & Krcmar, 2021). En aquesta línia, els esmentats autors parlen d'una permanent «guerra pel talent» («*war for talent*»), una imatge prou significativa en què la IA és representada com un bon aliat a l'hora de reclutar els candidats idonis. Des d'aquesta mateixa perspectiva, una altra avantatge seria que el software pot evitar les pràctiques discriminatòries contra grups protegits legalment (dones, minories ètniques, persones amb algun tipus de discapacitat, etc.), bo i promovent uns entorns de treball diversos i inclusius. Com veurem més endavant, aquest punt resulta especialment controvertit, quan no qüestionable, atès que la majoria de softwares que es comercialitzen a l'actualitat no incorporen mesures anti-biaix (Sánchez-Monedero, Dencik & Edwards, 2020).

La IA també sol considerar-se, en aquesta representació acrítica i positiva, força més eficient i eficaç que les persones: més ràpida, més precisa (comet menys errors), no es queixa, no es cansa. Per tant sol defensar-se amb arguments diversos el seu potencial per alliberar els treballadors que es dediquen a reclutar i seleccionar candidats de les tasques repetitives i considerades per ells mateixos com les més tedioses, com ara revisar i fer un primer cribratge de centenars o milers de CV de candidats a una feina.

Finalment, però no menys important, la IA és representada positivament com a «cost-efectiva», un eufemisme per suggerir que, a la pràctica, la inversió en IA pot beneficiar l'empresa o organització estalviant-li el sou d'un o més treballadors.

Pel que fa a les tasques i processos que es solen encomanar als enginys basats en IA, en el context concret que analitzem, tenim per exemple: el reclutament de candidats per mitjà de software de cerca (els anomenats «*Search Engines*»), el filtratge, la verificació i valoració de les seves credencials, i la recomanació dels candidats considerats més idonis o prometedors per a la fase d'entrevista (per mitjà dels anomenats «*Ranking Algorithms*» o «*Recommender engines*»).

Des de la perspectiva del reclutador, aquests sistemes poden ser autònoms o bé híbrids segons el grau d'interacció home-màquina i de supervisió humana dels processos i dels seus resultats. Pel que fa a les dades o inputs, aquestes poden ser estructurades (per exemple per mitjà de formularis que facilita el reclutador i omplen els interessats) o desestructurades (material que aporta el candidat i que no respon a un patró estricte, per exemple, un CV o una carta de motivació). La diversitat de formats, el llenguatge emprat, la incompletud de les dades aportades, poden ser fonts de biaix.

Més enllà d'aquest primer filtratge també està esdevenint una pràctica habitual que els enginys puguin fer i analitzar proves de competències/habilitats als candidats, per mitjà de diferents tipus de tests o videojocs. Aquests es basen en els coneixements de les neurociències i de la psicologia del treball i de les organitzacions, i s'utilitzen per analitzar les reaccions i el comportament dels candidats. A tal fi es genera una mètrica de determinats trets cognitius, emocionals i socials, que al seu torn genera un perfilat i una puntuació dels candidats que finalment serà comparada al perfil dels millors "performers", a partir dels quals s'ha entrenat l'algoritme (Sánchez-Monedero, Dencik & Edwards, 2020).

També és habitual que avui en dia una entrevista de treball es faci per mitjans telemàtics, de forma sincrònica o asincrònica i, per tant, pugui ser gravada i analitzada posteriorment amb l'ajuda d'enginys basats en IA, amb una lògica similar a la que hem vist pels videojocs. Pel que fa a aquest darrer punt presentem tot seguit un interessant estudi de cas que convida a la reflexió.

4.3. L'inquietant cas de l'anàlisi de video-entrevistes

Quines són les preguntes i què s'analitza exactament en aquestes entrevistes gravades i codificades per mitjà de IA, no és una pregunta irrellevant. Alguns autors consideren que aquests processos d'anàlisi són una «caixa negra», un episodi opac del procés de selecció que genera preguntes i dubtes tant en el terreny ètic com en el legal (Köchling, Riazy, Wehner & Simbeck, 2020a; Kim & Heo, 2021; Gutierrez, 2021).

El software adreçat a aquest tipus d'anàlisi és força inquietant ja que algun d'aquests enginys estan preparats per a capturar, codificar i inferir, a partir de dades biomètriques, l'expressió facial i el llenguatge no verbal dels candidats, els seus sentiments, estats d'ànim, trets de personalitat o, fins i tot, l'orientació sexo-afectiva (Crawford & Paglen, 2021).

D'aquests estudis volem ressaltar vàries qüestions que són d'interès general, atès que aquest tipus de pràctiques van a l'alça:

i. Els paranys teòrics implícits en un procés d'inferència com aquest, posen sobre la taula un reguitzell de perills reals per a les persones. Tractarem d'explicar-los. Inferir d'una expressió facial un estat d'ànim o un sentiment (o encara pitjor, un tret estable de personalitat) implica la presumpció que d'un nombre necessàriament limitat d'expressions facials (les IA solen codificar-ne 6 o 7), se'n segueix necessàriament un nombre limitat i precís de sentiments. Això, al seu torn, pressuposa la creença que les persones de tot el mon, de totes les edats i cultures, de tots els gèneres i en tot moment, expressen les emocions d'igual manera (en un nombre limitat de possibilitats i, per tant, sense variacions individuals, ni circumstancials, ni formes híbrides o simplement diferents). D'altra banda, també implica que una expressió facial, capturada un instant per una càmera, després de tota una cadena d'inferències, pot acabar sent l'enunciat objectiu d'un tret psicològic estable. Es tracta doncs d'un model conceptual que, des del nostre punt de vista, és excessivament biologicista, reduccionista, determinista i tancat.

ii. Tot i així, per a nosaltres l'aspecte més inquietant de tots és el fet que una empresa o organització es pugui sentir amb la potestat de poder fer servir aquest tipus d'anàlisi amb motiu d'una entrevista a un lloc de treball i tingui llibertat (o un espai no regulat) per a dur-ho a terme. Aquest punt és essencial doncs comporta la vulneració de drets de les persones en una cruïlla de camins: la intimitat, la igualtat, les dades personals i els drets laborals. La intersecció de tants àmbits i possibles factors discriminatoris fa que sigui difícil la seva aprehensió i, encara més, una possible reclamació.

4.4. Què significa resoldre el problema del biaix de gènere als algoritmes de reclutament i selecció de personal? El cas de «HireVue», «Pymetrics» i «Applied»

Sánchez-Monedero, Dencik & Edwards (2020) analitzen a fons tres softwares basats en IA destinats al reclutament i la selecció de personal, fabricats als Estats Units (HireVue i Pymetrics) i al Regne Unit (Applied), i amplament difosos per tot el món. Els productes van ser triats pels autors per ser els únics del seu àmbit que proporcionen alguna informació pública (a la web i al registre de patents) sobre el seu funcionament (com es dissenyen, validen

i auditen) i sobre les mesures anti-biaix que incorporen. Aquesta informació no és gens habitual i la majoria de productes comercials no la proporcionen.

En l'anàlisi d'aquest software els autors fan una reflexió de gran interès pel que fa a la conceptualització de les mesures desbiaixadores en els algoritmes, de la que volem destacar els següents aspectes.

i) Com ja havíem apuntat prèviament, no existeix a l'ecosistema de la IA una definició unificada de què significa «biaix», i molt menys de com aquest es pot «operacionalitzar» en termes informàtics. Resulta evident que les matemàtiques tenen dificultat a l'hora de capturar el significat de conceptes filosòfics/sociològics com ara «justícia» o «discriminació». Fins i tot una simple aproximació estadística a la noció de biaix com ara «que no es produeixi un desequilibri entre grups (per raó de gènere, ètnia, edat, etc.)» pot ser qüestionable i estar subjecte a molts matisos segons els contexts concrets en què hagi de ser aplicada. En conseqüència, igual o major dificultat presenta la concepció i operacionalització del que és el «des-biaix». Malgrat tot, observant el tipus de mesures des-biaixadores que aplica, del software comercial analitzat pot deduir-se que s'entén el biaix com a sinònim d'inequitat o discriminació. Aquestes responen a una de les següents tres aproximacions:

    a. *La des-classificació*, que implica intentar corregir el biaix mitjançant l'omissió de variables de grup en el model de presa de decisions.

    b. *La paritat*, que comporta intentar corregir el biaix mitjançant l'aplicació de taxes de passatge iguals (o en una proporció determinada) entre grups.

    c. *El calibratge*, que estableix el requisit que els resultats d'un cribratge siguin independents de les variables de grup.

    Del que acabem de dir és destacable que totes les fórmules depenen de definicions clares del que constitueixen els «grups». Exemples de definicions de grup poden ser: el gènere (segons una concepció binària o altra), l'ètnia (descrita per exemple segons les categories oficials del sistema de representació demogràfic propi de cada país), l'interval d'edat, etc. Per tant, és clar que en la definició i delimitació dels trets distintius d'aquests grups tenim un moment important d'infusió de biaixos, inclús quan el propòsit de creació d'aquests grups sigui exclusivament l'aplicació de les presumptes mesures mitigadores.

ii) Les mesures «anti-biaix», de la mateixa manera que el significat que s'atribueix al terme «biaix», reflectiran necessàriament les concepcions socials i legals del context en què han estat creades (en el cas de l'exemple, EEUU i Regne Unit). D'ací se'n deriva que l'exportació comercial d'aquests enginys de software a d'altres contexts, amb altres sistemes axiològics i legals pel que fa a la contractació, genera molts interrogants ja que pot produir un desajust imprevist en els resultats. Per exemple, «l'auditoria de biaix» que incorporen els enginys de HireVue i Pymetrics empren «la regla dels 4/5» a l'hora de valorar l'equitat d'una selecció. Aquesta proporció, que és un ús legal als EEUU que garanteix que cap grup legalment protegit (gènere, edat, creences) no pugui ser discriminat, ens forneix un exemple clar de com el software incorpora els valors i la legislació pròpia del lloc on ha estat creat. A Europa i d'altres contexts on s'exporta aquest software aquesta regla segueix aplicant-se sense que es correspongui amb el seu context legal.

Ara bé, des del nostre punt de vista, la proposta més interessant que fan els autors és considerar que els reptes ètics que presenten els algoritmes, entre ells el de la discriminació de gènere (o altra), poden ser abordats de forma molt més eficaç des de la perspectiva de la protecció de dades que no pas des de la perspectiva de l'equitat i la no discriminació.

Seguirem, doncs, aquests autors en l'anàlisi que fan de la directiva europea de protecció de dades (GDPR EU 2016/679), especialment del controvertit «Article 22», que ha generat força literatura, pel que fa a les conseqüències que se'n poden desprendre a l'hora de construir una ètica algorítmica i reclamar el dret a la transparència dels algoritmes. Donada l'extensió de cada una de les temàtiques que hi conflueixen, ens limitarem a apuntar breument cada una d'elles i indicar on rauen els «punts calents» que mereixen tota l'atenció. Aquests punts, que aquí plantegem en forma de reptes i preguntes, els entomarem novament a l'Informe 2, a l'hora de fonamentar les nostres recomanacions.

1) Dret a la presència d'un ésser humà al procés.

Aquest dret comporta, estrictament, que no es pugui deixar un procés que pot tenir importants repercussions/conseqüències per a les persones, com ara el cas que ens ocupa dels algoritmes de contractació, totalment a mans d'un enginy de software, sense supervisió humana.

Aquest punt desvetlla ràpidament algunes qüestions/reptes des d'una perspectiva ètica, com ara: quina és la intervenció que es pot considerar mínima i suficient per part d'aquest supervisor humà? Per exemple, en un procés de selecció, n'hi ha prou amb «validar» els suggeriments que faci la IA?

D'altra banda, tal com hem apuntat en la introducció a l'ètica dels algoritmes, ens trobem també amb un interrogant sobre la responsabilitat. Qui és, en últim terme, i en quina mesura, el responsable dels possibles biaixos, els dissenyadors del software, el supervisor, la mateixa IA considerada com a subjecte imputable?

D'ací també se'n deriva que si una organització planteja un procés en què no hi haurà supervisió humana (és a dir, en què per exemple la decisió de contractació serà totalment automatitzada) es requereix que l'interessat ho sàpiga i doni un consentiment explícit.

2) Dret a una explicació.

Aquest dret comporta, en el cas dels algoritmes de contractació, que la persona interessada pugui rebre informació clara i pertinent si s'automatitzen parts d'un procés. Aquesta «explicació» ha de tenir la característica de ser comprensible per a una persona no experta en informàtica, i no ser, com apunten irònicament els autors, «una mera regurgitació de codi informàtic».

Per tant, una explicació plausible hauria d'incloure informació significativa sobre el procés de selecció i la seva lògica, com ara informació general dels factors que es tenen en compte per al procés de presa de decisions i sobre el seu pes respectiu en la valoració final. També s'hauria d'indicar si es tracta d'un procés totalment o parcialment automatitzat, inclòs el perfilat dels candidats.

L'objectiu subjacent és clar: que els processos encomanats als algoritmes no siguin opacs, esdevinguin transparents i, per tant, puguin ser qüestionats i reclamats si es dona el cas.

5. Conclusions

Les pràctiques discriminatòries pel que fa al gènere en el reclutament i la contractació de personal són una realitat que afecta, entre d'altres, la capacitat de les persones d'obtenir i mantenir una feina, un aspecte clau tant per obtenir els mitjans de subsistència com per participar en la societat. En l'actualitat, la manera de prendre decisions sobre qui és elegible, per a quin lloc de treball i per què, està canviant ràpidament, com estem veient, amb l'adveniment i l'adopció de sistemes de selecció i contractació automatitzada, és a dir, per mitjà d'enginys basats en IA. Com indicàvem als epígrafs anteriors, aquest procés s'està tornant obscur i és susceptible de produir una vulneració de drets fonamentals consolidats a les societats democràtiques. Aquesta tendència arrela i està abonada, en part, per una representació i percepció social d'aquests enginys com a més eficients i econòmics que els treballadors humans.

Al seu torn, les preocupacions emergents inclouen la manca de transparència i la limitació potencial de l'accés a llocs de treball per a perfils específics, és a dir, les formes de discriminació que, més que directes, sovint són indirectes, latents, implícites.

Aquest punt és essencial atès que comporta la vulneració de drets de les persones en una cruïlla de camins: la intimitat, la igualtat, les dades personals i els drets laborals. La intersecció de tants àmbits i factors discriminatoris possibles fa que sigui realment difícil la seva aprehensió.

Volem doncs emfasitzar que és necessari i urgent monitoritzar i fer transparents aquests processos, especialment els que, com hem vist, impliquen una captura i anàlisi de dades biomètriques i proxèmiques dels candidats, en vistes a protegir degudament la ciutadania de la vulneració de drets fonamentals. Per exemple, el dret a no ser discriminats per raó de gènere, entenent aquest no de forma reduccionista sinó, com hem procurat mostrar al llarg d'aquest informe, amb tota la seva complexitat. Però també molts altres drets que es deriven de l'examen exhaustiu dels algoritmes de contractació des d'una perspectiva ètica, com ara el dret a no ser avaluats en relació a aspectes que no tinguin a veure directament amb el lloc de treball per al que es postula (com vèiem, per exemple, amb la realització gratuïta de «perfilats psicològics» dels candidats). O el dret dels candidats a què es garanteixi la seva privacitat i la seguretat de les seves dades personals durant i un cop acabat el procés de reclutament (qui veurà les entrevistes, per a què s'utilitzaran, com s'analitzaran, què se'n farà a posteriori, etc.).

O fins i tot el dret a rebutjar ser avaluat per un software basat en IA i preferir una avaluació humana, sense que això pugui esdevenir, al seu torn, un nou factor discriminatori.

Tots aquests drets prèviament consolidats als estats democràtics es veuen a l'actualitat tensionats, quan no obertament vulnerats, amb la subrogació d'alguns processos humans als enginys basats en IA. Malgrat que la publicitat de molts d'aquests softwares proclama detectar i mitigar pràctiques discriminatòries contra «grups protegits» i promoure la diversitat i la inclusió a la feina, aquestes afirmacions poques vegades s'examinen de prop i s'avaluen a la llum dels principis ètics, la legislació nacional i les recomanacions internacionals de les diferents temàtiques que intersecccionen.

Al segon informe de recerca, que adjuntem a continuació, recollim els aspectes clau que es desprenen de tot el que acabem de dir i que mereixen, al nostre parer, una major monitorització i intervenció, i fem una proposta de principis generals i mesures específiques que ens servirà de guia per a abordar, en la pràctica, la discriminació de gènere als algoritmes de contractació.

2a PART:

De l'opacitat a la transparència

6. Proposta de principis i mesures

6.1. Proposta de principis ètics generals

En paral·lel al desitjat i obligat compliment de la normativa legal, tenim unes exigències ètiques que arriben allà on la legalitat encara no ho fa o es mostra ambigua. Aquesta demanda l'afrontarem per mitjà d'uns principis generals que en el nostre estudi aplicarem a l'àmbit concret dels processos de reclutament i contractació de treballadors: els principis de *publicitat*, *transparència* i *responsabilitat*.

Per mitjà de l'aplicació raonada i contextualitzada en cada cas d'aquests principis procurarem garantir que els processos de reclutament i selecció de personal de les organitzacions siguin, no només legals, sinó ètics, això és, que tinguin en compte i tinguin cura dels interessos de totes les persones involucrades, sense incórrer en cap discriminació.

La idea principal, que fonamenta les diferents mesures proposades, és que els processos assistits total o parcialment per IA han de ser, de resultes de l'aplicació d'aquestes principis generals: *intel·ligibles*, *traçables* i *auditables* (en vistes a possibles reclamacions per danys ètics o legals), així com *estudiables* (en vistes a la seva millora o, si és el cas, desestimació).

Aquests principis generals els desplegarem tot seguit en alguns dels seus aspectes més concrets, conscients que per a moltes situacions reals caldrà recórrer novament als principis ja que és impossible preveure tota la casuística que es donarà en la pràctica.

6.2. Proposta de mesures generals

Mesura 1

De forma general, els processos de contractació cal que compleixin amb la legislació vigent (estatal i directives comunitàries) en matèria: laboral, d'igualtat, de transparència i de protecció de dades i drets digitals. Malgrat l'autoevidència de l'afirmació, la seva implementació efectiva no és tasca senzilla;, per això, caldrà tenir en compte les restants mesures proposades.

Mesura 2

És recomanable formar equips de treball el més diversos i heterogenis possible en les seves característiques, i paritaris pel que fa al gènere. Animem a aplicar quotes i ràtios, variables segons el context i les exigències tècniques.

Aquesta és una recomanació que se sol rebatre argumentant que una dona, pel fet de ser-ho, no necessàriament és "experta en gènere". Tot i així, creiem que la diversitat/heterogeneïtat en un equip de treball aporta una riquesa (en punts de vista i en experiència personal i professional) a les organitzacions que és positiva en sí mateixa. Com vèiem a l'Informe 1, en base a la bibliografia científica revisada, una de les possibles causes de la presència de biaixos de gènere en els algoritmes, i per tant, una de les possibles maneres d'afrontar-los, és la baixa presència de dones i la manca de continuïtat dels seus llocs de treball en la indústria de la IA.

Mesura 3

Derivada de l'anterior (i de les postures que la rebaten), comptar amb els serveis d'experts en gènere i fer formació als responsables de RRHH és una mesura general de provada eficàcia per tal de sensibilitzar i donar eines i recursos a l'hora d'identificar els biaixos de gènere tant dels algoritmes com dels éssers humans.

Seguint amb els principis generals, però ara endinsant-nos una mica més en les característiques intrínseques de la tecnologia basada en IA:

Mesura 4

No podem incorporar enginys basats en IA i deixar-los sense supervisió humana. La literatura científica aporta evidències que justifiquen que cal monitoritzar els productes d'IA al llarg de tot el seu cicle de vida. Això implica fer-ho en molts moments diferents. La dificultat és clara, donat que no hi haurà una única persona que se'n pugui fer responsable. En aquesta supervisió, caldria fomentar la transversalitat entre diferents agents implicats; això és: no delegar la supervisió exclusivament en mans de les empreses proveïdores dels enginys d'IA, sinó d'equips i processos de supervisió que comptin amb la presència de responsables de les empreses que usen aquests ginys en els seus processos de reclutament, i d'agents claus com els/les representants dels treballadors o els sindicats, i responsables d'*ethical compliance* (si n'hi ha).

Mesura 5

Cal monitoritzar el procés de construcció i entrenament dels algoritmes per tal d'evitar/mitigar els biaixos de gènere: en el concepte, en la construcció, en les bases de dades utilitzades en el seu entrenament. De nou, cal insistir en que els processos de construcció i entrenament dels algoritmes haurien d'estar en alguna mesura participats pels agents que després els faran servir, a fi i efecte d'adaptar-los al context específic de l'empresa, del sector, de la legislació "local", i fins i tot, de polítiques específiques d'igualtat o de discriminació positiva que afectin l'empresa.

Mesura 6

Cal monitoritzar els algoritmes durant la seva implementació pràctica a un context amb una tasca particular, per tal d'evitar/mitigar els biaixos de gènere deguts a: una demanda particular d'una determinada posició (biaix en l'anunci, biaix explícit/implícit/proxy de les característiques seleccionades), biaix en les bases de dades reals d'on s'extreu la informació dels candidats.

Mesura 7

Cal monitoritzar els algoritmes després de la seva implementació: revisió dels outputs, tant de les «recomanacions» («*match*», predicció) com dels «descarts» en relació a les característiques explícites i públiques de la posició, i analitzar tant els «falsos positius» com molt especialment els «falsos negatius». Aquí es reclama una apertura de la informació a l'estudi acadèmic i a l'auditoria tècnico-legal.

Mesura 8

Caldrà que es vagin formalitzant estàndards industrials i agències independents d'auditoria, normalitzant-ne el seu ús. L'auditabilitat dels algoritmes està esdevenint una demanda social de primer ordre. S'espera de les organitzacions la bona pràctica d'auditar els seus enginys d'IA i fer-los transparents.

Mesura 9

Juntament amb la normalització de les auditories internes, cal implementar mecanismes d'informació i reclamació sobre els processos de reclamació basats totalment o parcial en algoritmes. De dues menes: (*a*) Vies clares i eficients d'informació i reclamació per part de l'empresa que usa els ginys d'IA, i (*b*) mecanismes públics de reclamació. .

Mesura 10

No podem incorporar enginys basats en IA i desconèixer si incorporen mesures anti-biaix, i en cas que ho facin, quines són les seves característiques.

Davant el dubte sol·licitarem les característiques tècniques i la seva explicació en llenguatge divulgatiu (comprensible per a persones que no siguin del sector tecnològic) o bé ens abstindrem d'utilitzar el producte i ens decantarem per un altre.

Després d'ingents quantitas de recerca sobre el «biaix implícit» dels algoritmes, i dels demostrats i validats esforços de les mesures anti-biaix o des-biaixadores, no es pot permetre que aquestes mesures no s'implementin de forma obligada i proactiva. No es pot desconèixer si una IA que està realitzant un procés de treball automatizat està discriminant o no: cal assegura-se'n.

6.3. Proposta de mesures específiques

L'operacionalització dels principis enumerats a l'anterior apartat pot implicar, com a mínim, les obligacions de:

Mesura 11

Fer publicitat efectiva, amb la descripció el més acurada possible de:

- Descripció de la posició ofertada
- Descripció de les característiques que efectivament seran avaluades dels candidats
- Descripció detallada del procés de selecció.

Si el procés de selecció es fa, total o parcialment, per mitjà de software basat en IA, que implica que hi ha una presa automatitzada o semi-automatitzada de decisions (és a dir sense supervisió humana, o amb una supervisió humana parcial), aquest fet hauria de ser públic.

Mesura 12

Si el procés de selecció es fa, total o parcialment, per mitjà de software basat en IA, caldria identificar aquest software, com a mínim, amb la següent informació:

- El nom comercial del producte,

- La companyia fabricant,
- Les mesures anti-biaix del software (en llenguatge tècnic i en llenguatge divulgatiu)
- La base de dades («training dataset») amb la que la IA ha estat entrenada.

Mesura 13

Si el procés de selecció es fa, total o parcialment, per mitjà de software basat en IA, caldria identificar les característiques dels candidats que *efectivament* s'han codificat i analitzat i la seva relació amb els requeriments de la posició ofertada.

Mesura 14

Si el procés de selecció es fa, total o parcialment, per mitjà de software basat en IA, caldria identificar les parts del procés de selecció de què s'ocupa la IA i amb quin grau de responsabilitat: total o parcial (per exemple selecció no supervisada o supervisada de CV dels candidats, anàlisi de video-entrevistes, etc.) i sota quina supervisió.

Mesura 15

Si el procés de selecció es fa, total o parcialment, per mitjà de software basat en IA, caldria que, un cop acabat el procés, es fessin públics els resultats (positius i negatius, degudament anonimitzats), per tal que aquests es puguin estudiar i revalidar o, en el seu defecte, reclamar.

Un procés de selecció de personal, especialment en un organisme públic, ha de ser transparent. En el cas de la utilització d'enginys d'IA convé que sigui, a més a més, obert i accessible a l'estudi i la recerca, tant de cara a la millora del propi software, com a la seva desestimació si es detecta que no ofereix resultats ètica i legalment acceptables. Un procés de reclutament i selecció sempre hauria de poder ser reclamat/impugnat si els candidats consideren que hi ha hagut manca de transparència, desprotecció de les seves dades i/o discriminació.

7. Bibliografia

Adams, R., & Loideáin, N. N. (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.

Alexander, V., Blinder, C., & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. Computers in Human Behavior, 89, 279-288.

Ananny, M., & Crawford, K. (2019). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. doi:10.1177/1461444816676645

Ash, S., Gann, D., & Dodgson, M. (2018). The tech industry needs more women. Here's how to make it happen. Retrieved from https://www.weforum.org/agenda/2017/11/the-tech-industry-needs-more-women-heres-how-to-make-it-happen-d1eec473-48cc-4801-bd40-dfba040b0e4a/

BarIlan, J., Keenoy, K., Levene, M., & Yaari, E. (2008). Presentation bias is significant in determining user preference for search results—A user study. Journal of the Association for Information Science and Technology, 6(1), 135–149. doi:10.1002/asi.20941

Beltran, K., Rowland, C., Hashemi, N., Nguyen, A., Harrison, L., Engle, S., & Yuksel, B. F. (2021, May). Reducing Implicit Gender Bias Using a Virtual Workplace Environment. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-7)

Bishop, C. M. (2006). Pattern recognition and machine learning. London, UK: Springer

Bivens, R. (2017). The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. New Media & Society, 19(6), 880–898. doi:10.1177/1461444815621527

Boellstorff, T. (2013). Making big data, in theory. First Monday, 18(10). doi:10.5210/fm.v18i10.4869

Böhm, S., Linnyk, O., Kohl, J., Weber, T., Teetz, I., Bandurka, K., & Kersting, M. (2020). Analysing gender bias in IT job postings: a pre-study based on samples from the German

Job Market. In Proceedings of the 2020 on Computers and People Research Conference (pp. 72-80).

Bolin, A. (2003). La transversalidad de género. Contexto cultural y prácticas de género. Antropología de la sexualidad y diversidad cultural, 231-260.

Bourdieu, P. (2001). La dominación masculina. Barcelona: Anagrama.

Bornstein, S. (2017). Reckless Discrimination. Calif. L. Rev., 105, 1055.

Braman, S. (2009). Change of state: Information, policy, and power. Cambridge, MA: MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities on commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), Conference on Fairness, Accountability, and Transparency (Proceedings of Machine Learning Research) (Vol. 81, pp. 77–91). New York: New York University Press. Retrieved from: http://proceedings.mlr.press/v81/buolamwini18a.html

Butler, J. (2007). El género en disputa: el feminismo y la subversión de la identidad. Paidós.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. doi:10.1126/science.aal4230

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. In Proceedings of the 2018 chi conference on human factors in computing systems (pp. 1-14)

Ciampaglia, G. L. (2018). Fighting fake news: A role for computational social science in the fight against digital misinformation. Journal of Computational Social Science, 1(1), 147–153. doi:10.1007/s42001-017-0005-6

Ciampaglia, G. L., & Menczer, F. (2018, June 20). Misinformation and biases infect social media, both intentionally and accidentally. Retrieved from https://theconversation.com/misinformation-and-biases-infect-social-media-both-intentionally-and-accidentally-97148

Coleman, G. E. (2013). Coding freedom: The ethics and aesthetics of hacking. Princeton, NJ: Princeton University Press.

Collins, R. L. (2011). Content analysis of gender roles in media: Where are we now and where should we go? Sex Roles, 64, 290–298. doi:10.1007/s11199-010-9929-5

Colman, F., van der Tuin, I., O'Donnell, A., & Bühlmann, V. (2018). Ethics of coding: A report on the algorithmic condition. Brussels, Belgium: European Commission.

Couldry, N. (2013). A necessary disenchantment: Myth, agency, and injustice in a digital world. The Sociological Review, 62(4), 880–897. doi:10.1111/1467-954X.12158

Crawford, K. (2013, April 1). The hidden biases in big data. Retrieved from https://hbr.org/2013/04/the-hidden-biases-in-big-data

——— (2016, June 25). Artificial intelligence's white guy problem. The New York Times. Retrieved from https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. AI & SOCIETY, 1-12.

Crenshaw, K. (1989) Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. University of Chicago Legal Forum. Vol. 1989: Iss. 1, Article 8

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved from https://medium.com/datadriveninvestor/amazon-scraps-secret-ai-recruiting-engine-that-showed-biases-against-women-995c505f5c6f

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies, 2015(1), 92–112. Philadelphia, PA. doi:10.1515/popets-2015-0007454

De Beauvoir, S. (1981). El segundo sexo (1949). Buenos Aires: Siglo XX.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., ... & Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes

setting. In proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 120-128).

Diehl, A. B., Stephenson, A. L., Dzubinski, L. M., & Wang, D. C. (2020). Measuring the invisible: Development and multi-industry validation of the Gender Bias Scale for Women Leaders. Human Resource Development Quarterly, 31(3), 249-280

Dijck, J. van. (2014). Datafication, dataism, and dataveillance: Big data between scientific paradigm and ideology. Surveillance & Society, 12, 2197–208. doi:10.24908/ss.v12i2.4776

D'Ignazio, C., & Klein, L. F. (2019). Chapter one: Bring back the bodies. In C. D'Ignazio & L. F. Klein (Eds.), Data feminism (pp. 1–22). Cambridge, MA: MIT Press Open. Retrieved from https://mitpressonpubpub.mitpress.mit.edu/pub/zrlj0jqb

Fabris, A., Purpura, A., Silvello, G., & Susto, G. A. (2020). Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. Information Processing & Management, 57(6), Article: 102377.

Fausto-Sterling, A. (1993). The five sexes: Why male and female are not enough. Sciences, 33, 1-20.

Fernández-Martínez, C., & Fernández, A. (2020). AI and recruiting software: Ethical and legal implications. Paladyn, Journal of Behavioral Robotics, 11(1), 199-216.

Fernández-Martínez, C. & Fernández, A. (2019a). Ontologies and AI in Recruiting. A Rule-Based Approach to Address Ethical and Legal Auditing. In: Proceedings of the International Semantic Web Conference (ISWC).

Fernández-Martínez, C., & Fernández, A. (2019b). AI in Recruiting. Multi-agent Systems Architecture for Ethical and Legal Auditing. In: IJCAI (pp. 6428-6429).

Fernández-Martínez, C., & Fernández, A. (2019c). Ethical and legal implications of ai recruiting software. Ercim News, 116, 22-23

Floridi, L. (2013). The ethics of information. Oxford, UK: Oxford University Press

Fredrickson, B. L., & Roberts, T. A. (1997). Objectification theory: Toward understanding women's lived experiences and mental health risks. Psychology of Women Quarterly, 21(2), 173–206. doi:10.1111/j.1471-6402.1997.tb00108.x

Galarza Fernández, E., Sosa Valcarcel, A., & Castro Martínez, A. (2018). The (un)protection of women in the audiovisual communication law: A legal reality in the Spanish scenario. Congreso Universitario Internacional sobre la Comunicación en la Profesión y en la Universidad de Hoy [International Congress on Today's Professional and University Communication]: Vol. 1. Vivat Academia (pp. 227–230). Madrid, Spain. doi:10.15178/CUICIID.2018

Gallardo, A. (2018, March 20). How we collected nearly 5,000 stories of maternal harm. ProPublica. Retrieved from https://www.propublica.org/article/how-we-collected-nearly-5-000-stories-of-maternal-harm

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In: Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining (pp. 2221-2231)

Girls Who Code. (2018). About us. Retrieved from: https://girlswhocode.com/about-us/

Gitelman, L. (Ed.). (2013). Raw data is an oxymoron. Cambridge, MA: MIT Press

Gorrostieta, C., Lotfian, R., Taylor, K., Brutti, R., & Kane, J. (2019). Gender de-biasing in speech emotion recognition. Proceedings of Interspeech 2019, 823–2827. doi:10.21437/Interspeech.2019-1708International

Gupta, M., Parra, C. M., & Dennehy, D. (2021). Questioning Racial and Gender Bias in AI-based Recommendations: Do Espoused National Cultural Values Matter?. Information Systems Frontiers, 1-17.

Gutierrez, M. (2021). Algorithmic Gender Bias and Audiovisual Data: A Research Agenda. International Journal of Communication, 15, 439-461.

Habler, F., & Henze, N. (2019). Differences between smart speakers and graphical user interfaces for music search considering gender effects. Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia, 30, 1–7. Pisa, Italy. doi:10.1145/3365610.3365627

Hangartner, D., Kopp, D., & Siegenthaler, M. (2021). Monitoring hiring discrimination through online recruitment platforms. Nature, 589(7843), 572-576

Hannon, C. P. (2018). Avoiding bias in robot speech. Interactions, 25(5), 34–37. doi:10.1145/3236671

Hansen, C. H. (1989). Priming sex-role stereotypic event schemas with rock music videos: Effects on impression favorability, trait inferences, and recall of a subsequent male-female interaction. Basic and Applied Social Psychology, 10(4), 371–391. doi:10.1207/s15324834basp1004_6

Helmond, A. (2015). The platformization of the Web: Making Web data platform ready. Social Media + Society, 1–11. doi:10.1177/2056305115603080

Henderson, N. M. (2014, October 8). White men are 31 percent of the American population. They hold 65 percent of all elected offices. The Washington Post. Retrieved from: https://www.washingtonpost.com/gdpr/consent/?next_url=https%3a%2f%2fwww.washingtonpost.com%2fnews%2fthe-fix%2fwp%2f2014%2f10%2f08%2f65-percent-of-all-american-elected-officials-are-white-men%2f

Hess, K. P. (2013). Investigation of nonverbal discrimination against women in simulated initial job interviews. Journal of Applied Social Psychology, 43(3), 544–555. doi:10.1111/j.1559- 1816.2013.01034.x

Hill Collins, P. (1998). It's all in the family: Intersections of gender, race, and nation. Hypatia, 13(3), 62– 82. doi:10.1111/j.1527-2001.1998.tb01370.x

Hirsch, P.B. (2016), The Caliphate of numbers, Journal of Business Strategy, Vol. 37 No. 6, pp. 51-55. DOI: 10.1108/JBS-09-2016-0098.

Hitti, Y., Jang, E., Moreno, I., & Pelletier, C. (2019). Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. Proceedings of the First Workshop on Gender Bias in Natural Language Processing, 8–17. Florence, Italy. doi:10.18653/v1/W19-3802

Hmoud, B., & Laszlo, V. (2019). Will artificial intelligence take over human resources recruitment and selection?. Network Intelligence Studies, 7(13), 21-30.

Jahanbakhsh, F., Cranshaw, J., Counts, S., Lasecki, W. S., & Inkpen, K. (2020). An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality

and Gender. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

Kahana, E. (2018, December 22). Solving algorithmic bias via artificial intelligence computational law applications. Retrieved from https://law.stanford.edu/2018/12/22/artificial-intelligence-computational-law-applications-and-algorithmic-bias/

Kaili, C. (2018). Play it for change. Ljubljana, Slovenia: Mirovni Institut & The Peace Institute

Karsay, K., Matthes, J., Platzer, P., & Plinke, M. (2017). Adopting the objectifying gaze: Exposure to sexually objectifying music videos and subsequent gazing behavior. Media Psychology, 21(1), 27– 49. doi:10.1080/15213269.2017.1378110

Kim, J. Y., & Heo, W. (2021). Artificial intelligence video interviewing for employment: perspectives from applicants, companies, developer and academicians. Information Technology & People. Vol. ahead-of-print No. ahead-of-print. doi: 10.1108/ITP-04-2019-0173.

Kleinberg, J., & Raghavan, M. (2018). Selection problems in the presence of implicit bias. arXiv. Preprint arXiv:1801.03533

Knight, W. (2016, November 23). How to fix Silicon Valley's sexist algorithms: Computers are inheriting gender bias implanted in language data sets—And not everyone thinks we should correct it. Retrieved from: https://www.technologyreview.com/2016/11/23/155858/how-to-fix-silicon-valleys-sexist-algorithms/

Köchling, A., Riazy, S., Wehner, M., & Simbeck, K. (2020a). Highly Accurate, but Still Discriminatory: A Fairness Evaluation of Algorithmic Video Analysis. In Academy of Management Proceedings (Vol. 2020, No. 1, p. 13339). Briarcliff Manor, NY 10510: Academy of Management.

Köchling, A., & Wehner, M. C. (2020b). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 1-54.

Kuchler, H. (2018, March 9). Tech's sexist algorithms and how to fix them. Financial Times. Retrieved from: https://www.ft.com/content/d2a1ab08-f63e-11e7-a4c9-bbdefa4f210b

Kullmann, M. (2018). Platform work, algorithmic decision-making, and EU gender equality law. International Journal of Comparative Labour Law and Industrial Relations, 34(1).

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Management science, 65(7), 2966-2981.

Lanham, MD: Rowman & Littlefield. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 1–21. doi:10.1177/2053951716679679

Laurim, V., Arpaci, S., Prommegger, B., & Krcmar, H. (2021). Computer, Whom Should I Hire? Acceptance Criteria for Artificial Intelligence in the Recruitment Process. In Proceedings of the 54th Hawaii International Conference on System Sciences (p. 5495).

Leavy, S., Meaney, G., Wade, K., & Greene, D. (2020). Mitigating Gender Bias in Machine Learning Data Sets. In International Workshop on Algorithmic Bias in Search and Recommendation (pp. 12-26). Springer, Cham.

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In Proceedings of the 1st international workshop on gender equality in software engineering (pp. 14-16).

León, G. A., Chiou, E. K., & Wilkins, A. (2021). Accountability Increases Resource Sharing: Effects of Accountability on Human and AI System Performance. International Journal of Human–Computer Interaction, 37(5), 434-444.

Leung, W., Zhang, Z., Jibuti, D., Zhao, J., Klein, M., Pierce, C., ... & Zhu, H. (2020). Race, Gender and Beauty: The Effect of Information Provision on Online Hiring Biases. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-11).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. Retrieved from https://arxiv.org/abs/1908.09635

Meyers, M. (2007). African American women and violence: Gender, race, and class in the news. Critical Studies in Media Communication, 21(2), 95–118. doi:10.1080/07393180410001688029

Michelfelder, D. P., Wellner, G., & Wiltse, H. (2017). Designing differently: Toward a methodology for an ethics of feminist technology design. In S. O. Hansson (Ed.), The ethics of technology (pp. 193– 219)

Musik, C., & Zeppelzauer, M. (2018). Computer vision and the digital humanities: Adapting image processing algorithms and ground truth through active learning. VIEW Journal of European Television History and Culture, 7(14), 59–72. doi:10.18146/2213-0969.2018.jethc153

Mylavarapu, S. (2016, May 10). The lack of women in tech is more than a pipeline problem. Retrieved from https://techcrunch.com/2016/05/10/the-lack-of-women-in-tech-is-more-than-a-pipeline-problem/

Nadeem, A., Abedin, B., & Marjanovic, O. (2020). Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies. ACIS 2020 Proceedings.

Niklas, J., & Peña Gangadharan, S. (2018). Between antidiscrimination and data. London, UK: London School of Economics and Political Science

Nussbaum, M. (1995). Objectification. Philosophy and Public Affairs, 24(4), 249–291. doi:10.1111/j.1088- 4963.1995.tb00032.x

Oberst, U., De Quintana, M., Del Cerro, S., & Chamarro, A. (2020). Recruiters prefer expert recommendations over digital hiring algorithm: a choice-based conjoint study in a pre-employment screening scenario. Management Research Review.Vol. 44 No. 4, pp. 625-641. DOI: 10.1108/MRR-06-2020-0356

ObservantVids. (2017, December 5). Ali Rahimi's talk at NIPS (NIPS 2017 Test-of-time award presentation) [Video file]. Retrieved from: https://www.youtube.com/watch?v=Qi1Yry33TQE

Ochmann, J., & Laumer, S. (2019). Fairness as a Determinant of AI Adoption in Recruiting: An Interview-based Study. DIGIT 2019 Proceedings. 16.

Peng, A., Nushi, B., Kıcıman, E., Inkpen, K., Suri, S., & Kamar, E. (2019, October). What you see is what you get? the impact of representation criteria on human bias in hiring. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (Vol. 7, No. 1, pp. 125-134).

Peña, A., Serna, I., Morales, A., & Fierrez, J. (2020). Bias in multimodal AI: Testbed for fair automatic recruitment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 28-29).

Preciado, P. B. (2016). Manifiesto contrasexual. Barcelona: Anagrama.

Pickett, A. (2018, January 26). The dangers of keeping women out of tech. Retrieved from https://www.wired.com/story/dangers-keeping-women-out-of-tech/

Polonski, V. (2016, November 6). Would you let an algorithm choose the next president? Tech Crunch. Retrieved from https://techcrunch.com/2016/11/06/would-you-let-an-algorithm-choose-the-next-u-s-president/

Prey, R. (2017). Nothing personal: Algorithmic individuation on music streaming platforms. Media, Culture & Society, 40(7), 1086–1100. doi:10.1177/0163443717745147

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 469-481).

Rahman, H. A. (2021). The Invisible Cage: Workers' Reactivity to Opaque Algorithmic Evaluations. Administrative Science Quarterly, Article: 00018392211010118.

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to «solve» the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In: Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 458-468).

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., ... & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. arXiv e-prints, arXiv-1906.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751–752. doi:10.1126/science.aat5991

Tallon, T. (2019, September 3). Century of "shrill": How bias in technology has hurt women's voices. The New Yorker. Retrieved from https://www.newyorker.com/culture/cultural-comment/a-century-of-shrill-how-bias-in-technology-has-hurt-womens-voices

Tang, S., Zhang, X., Cryan, J., Metzger, M. J., Zheng, H., & Zhao, B. Y. (2017). Gender bias in the job market: A longitudinal analysis. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1-19.

Tassabehji, R., Harding, N., Lee, H., & Dominguez-Pery, C. (2021). From female computers to male comput♂ rs: Or why there are so few women writing algorithms and developing software. Human Relations, 74(8), 1296-1326.

Tatman, R. (2016, July 12). Google's speech recognition has a gender bias. Making Noise & Hearing Things. Retrieved from https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/

——— (2017). Gender and dialect bias in YouTube's automatic captions. Proceedings of the First Workshop on Ethics in Natural Language Processing, 53–59. Valencia, Spain. doi:10.18653/v1/W17-1606460

Taylor, C. (1994). The politics of recognition. In A. Gutmann (Ed.), Multiculturalism, and the politics of recognition (pp. 25–73). Princeton, NJ: Princeton University Press. Retrieved from http://elplandehiram.org/documentos/JoustingNYC/Politics_of_Recognition.pdf

Taylor, L. (2018, July 10). As technology advances, women are left behind in digital divide. Reuters. Retrieved from https://www.reuters.com/article/us-britain-women-digital/as-technology-advances-women-are-left-behind-in-digital-divide-idUSKBN1K02NT

Trevisan, F. (2013, May 15). Social engines and social science: A revolution in the making. Economic and Social Research Council. doi:10.2139/ssrn.2265348

Tucker, I. (2017, May 28). A white mask worked better: Why algorithms are not color blind. The Guardian. Retrieved from

https://www.theguardian.com/technology/2017/may/28/joy-buolamwini-when-algorithms-are-racist-facial-recognition-bias

Tufekci, Z. (2014). Engineering the public: Internet, surveillance, and computational politics. First Monday, 19(7). doi:10.5210/fm.v19i7.4901

Upadhyay, A. K., & Khandelwal, K. (2018). Applying artificial intelligence: implications for recruitment. Strategic HR Review, (Vol 17, No 5, pp. 255-258).

Vasconcelos, M., Cardonha, C., & Gonçalves, B. (2018). Modeling epistemological principles for bias mitigation in AI systems: an illustration in hiring decisions. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 323-329).

Vyas, N. (2021). Gender inequality-now available on digital platform: an interplay between gender equality and the gig economy in the European Union. European Labour Law Journal, 12(1), 37-51.

Waardenburg, L., Sergeeva, A., & Huysman, M. (2018). Hotspots and blind spots: A case of predictive policing in practice. In U. Schultze, M. Aanestad, M. Mähring, C. Østerlund, & K. Riemer (Eds.), IFIP WG 8.2 Working Conference on the Interaction of Information Systems and the Organization (pp. 96–109). San Francisco, CA: IFIP Advances in Information and Communication Technology. doi:10.1007/978-3-030-04091-8_8

Wachter-Boettcher, S. (2017). Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech. New York, NY: Norton & Company

Wang, E. (2018, May 1). What does it really mean for an algorithm to be biased? The Gradient. Retrieved from https://thegradient.pub/ai-bias/

Waterson, J. (2019, April 12). Daily Star covers up its page 3 girls, signaling end of tabloid tradition. The Guardian. Retrieved from https://www.theguardian.com/media/2019/apr/12/daily-star-covers-up-its-page-3-girls-signalling-end-of-tabloid-tradition

Wellner, G., & Rothman, T. (2019). Feminist AI: Can we expect our AI systems to become feminist? Philosophy & Technology, 33, 191–205. doi:10.1007/s13347-019-00352-z

West, Sarah Myers (2020). Redistribution and Rekognition: A Feminist Critique of Algorithm Fairness.Catalyst: Feminism, Theory, Technoscience, 6(2), 1–24. doi: https://doi.org/10.28968/cftt.v6i2.33043

Wiewiorówski, W. (2020). A preliminary opinion on data protection and scientific research. Brussels, Belgium: European Data Protection Supervisor

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. Journal of Information Policy, 8, 78–115. doi:10.5325/jinfopoli.8.2018.0078

Williams, S. D. (2020). A textual analysis of racial considerations in human resource analytics vendors' marketing. Management Research and Practice, 12(4), 49-63.

Woodland, P. C., Hain, T., Johnson, S. E., Niesler, T. R., Tuerk, A., & Young, S. J. (1998). Experiments in broadcast news transcription. International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE Signal Processing Society. doi:10.1109/ICASSP.1998.675413

Wronkiewicz, M. (2018, June 4). Realistic expectations for applied machine learning. Retrieved from https://medium.com/devseed/realistic-expectations-for-applied-machine-learning-a8171250db28

Wu, L., & Kane, G. C. (2021). Network-Biased Technical Change: How Modern Digital Collaboration Tools Overcome Some Biases but Exacerbate Others. Organization Science, 32(2), 273-292.

Yang, C. S., & Dobbie, W. (2020). Equal protection under algorithms: A new statistical and legal framework. Michigan Law Review, 119(2), 291-395.

Yarger, L., Payton, F. C., & Neupane, B. (2020). Algorithmic equity in the hiring of underrepresented IT job candidates. Online Information Review. Vol. 44, No. 2, pp. 383-395. DOI 10.1108/OIR-10-2018-0334.