GRUPPO di PISA
Dibattito aperto sul Diritto e la Giustizia Costituzionale

CONVEGNO DEL GRUPPO DI PISA
IL DIRITTO COSTITUZIONALE E LE SFIDE DELL'INNOVAZIONE TECNOLOGICA
UNIVERSITÀ DEGLI STUDI DI GENOVA – 18-19 GIUGNO 2021

INTELLIGENZA ARTIFICIALE E DISCRIMINAZIONI[*]

COSTANZA NARDOCCI[**]

SOMMARIO: Introduzione: La discriminazione, la persona e la macchina. – PARTE PRIMA – LA DISCRIMINAZIONE "ARTIFICIALE". – 1. Esiste una *AI-derived discrimination*? – 2. Stesse cause, diversa fenomenologia? – 3. Sulla possibile o auspicabile trasposizione delle categorie del diritto anti-discriminatorio. – 3.1. La discriminazione diretta: quando le tecniche di intelligenza artificiale distinguono ("male"). – 3.2. A cavallo tra discriminazione diretta e indiretta: l'*unconscious disparate treatment.* – 3.3. La discriminazione indiretta: dove risiede la (non) neutralità? – 3.4. Ai confini della discriminazione per associazione: la "*proxy discrimination*". – 4. Discriminazioni strutturali e intelligenza artificiale. – 5. Una prima delimitazione teorica della *AI-derived discrimination*. – 6. Le "vittime": verso nuove identità e appartenenze. – 7. Le "vecchie" minoranze tra genere, razza e etnia. – 8. "Nuove" minoranze e "nuovi" fattori di discriminazione? Il *proxy*. – 9. L'intersezionalità oggettiva e soggettiva nella *AI-derived discrimination*. – PARTE SECONDA – DALLA GENESI ALL'ACCERTAMENTO: IL LEGISLATORE, I GIUDICI. – 10. Il legislatore. 10.1. Sulla regolamentazione della *AI-derived discrimination*: il "se", il "chi"... – 10.2. … il "come". – 11. Il giudice. –11.1. Le criticità di ordine teorico: l'accesso alla giustizia. – 11.2. (*Segue*) l'individuazione del soggetto responsabile – 11.3. (*Segue*) la prova della disparità di trattamento. – 11.4. … come risolverle. – 12. Uno sguardo ai primi casi giudiziari tra Italia e Europa. – 13. La giurisprudenza costituzionale tra eguaglianza, ragionevolezza e automatismi: quale spazio per lo scrutinio sulla "discriminazione algoritmica"? – Conclusioni: L'intelligenza artificiale ha "cambiato" la discriminazione?

[*] Contributo sottoposto a referaggio ai sensi dell'art. 5 del Regolamento della Rivista.

[**] Ricercatrice in Diritto costituzionale, Dipartimento di Diritto pubblico italiano e sovranazionale, Università degli Studi di Milano.

Introduzione: La discriminazione, la persona e la macchina

La discriminazione, quale fenomeno direttamente correlato alla condotta umana, secondo una costruzione relazionale di tipo causale tra la seconda e la prima, è negli ultimi anni esposto a nuove e, in parte, impreviste torsioni[1].

Il collegamento immediato, in termini temporali, e diretto, perché portato fattuale e causale, con l'azione della persona umana si vede, infatti, progressivamente sfumare al cospetto dell'*agere* di un soggetto o fattore di tipo diverso e terzo[2]. L'ingresso nello spazio umano, pubblico[3] e privato, della tecnologia e, con essa, delle tecniche di intelligenza artificiale[4] ha portato con sé conseguenze importanti, tra le altre[5], sulla fenomenologia della discriminazione[6], rendendo problematica la configurazione dei suoi tratti caratterizzanti, prodotto di una convergenza di studi che ha avvicinato nel corso dei decenni le scienze sociali – dalla sociologia, al diritto, alla filosofia – nella ricerca di una sua definizione esaustiva e condivisa[7].

Se sussiste un consenso intorno alla equiparazione tra la discriminazione, nella sua duplice accezione di fenomeno sociale e giuridico, e la differenziazione irragionevole, perché ingiustificata e sproporzionata, tra due fattispecie sovrapponibili, cioè analoghe

---

[1] Riferisce delle criticità che discendono dalla interposizione tra l'uomo e gli eventi della "macchina", A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, in *BioLaw Journal – Rivista di BioDiritto*, 1/2019, 63 ss., e, in particolare, 67 ss.

[2] La letteratura osserva, infatti, in rapporto al nesso causale tra azione umana ed effetto che: «[i]l fatto che oggi la tecnologia (per meglio dire, la potenza cibernetica) non è più soltanto uno 'strumento' per realizzare finalità decise da un soggetto agente umano, ma, è essa stessa a prendere decisioni rilevanti per la libertà e la persona umana», così A. SIMONCINI, S. SUWEIS, *Il cambio di paradigma nell'intelligenza artificiale e il suo impatto sul diritto costituzionale*, in *Rivista di filosofia del diritto*, 2019, 93.

[3] Il riferimento è, come noto, anche al settore della giustizia. In proposito, recente è la notizia secondo cui anche la Corte europea dei diritti dell'uomo sarebbe in procinto di introdurre meccanismi di intelligenza artificiale per la fase preliminare di ricevibilità. Si veda la manifestazione di interesse del Giudice Spano: «[w]e have throughout the last decade been introducing reforms and one of them, certainly, is the use of information technology. We are now in a phase where we are looking at to what extent we can, for example, at the registration phase introduce algorithmic or automated decision making so as to try and reduce the extent to which this classical registration phase has to all be done manually. […] When it is done, we can use the data introduced into the system in a more effective manner. But I do think moving to the future a mass, a bulk case court like ours will slowly start introducing algorithmic tools to facilitate its tasks». Ne dà conto V. FIKFAK, *What Future for Human Rights? Decision-making by algorithm*, in *Strasbourg Observer*, 19 maggio 2021.

[4] Schematicamente, i settori che, ad oggi, fanno maggiormente ricorso alle tecniche di intelligenza artificiale sono: l'amministrazione della giustizia, per quanto attiene alla prevedibilità del rischio di recidiva; la sanità per la diagnostica; le risorse umane per assunzioni e licenziamento; i sistemi educativi per la valutazione e l'assegnazione dei punteggi; la finanza; il settore assicurativo. Per le implicazioni giuridiche che discendono dalla progressiva centralità che stanno assumendo le tecniche di intelligenza artificiale, K.D. ASHLEY, *Artificial Intelligence and Legal Analytics. New Tools for Law Practice in the Digital Age*, Cambridge, 2017, ripreso da A. SANTOSUOSSO, *Intelligenza artificiale e diritto*, Milano, 2020.

In dottrina, si vedano, anche, le relazioni rese nell'ambito del seminario organizzato dall'Associazione "Gruppo di Pisa" in data 26 marzo 2021 dal titolo "Diritto e nuove tecnologie tra comparazione e interdisciplinarità", i cui contributi sono in corso di pubblicazione.

Per un'analisi delle principali implicazioni di carattere etico legate all'espansione delle tecniche di intelligenza artificiale, J. KAPLAN, *Artificial Intelligence. What Everyone Needs to Know*, Oxford, 2016. Interessanti sono anche le conseguenze e i rischi connessi all'impiego delle tecniche di intelligenza artificiale durante la crisi sanitaria legata alla diffusione del Covid-19, su cui M.S. HORIKAWA, *Digitalized Discrimination: COVID-19 and the Impact of Bias in Artificial Intelligence*, in *The Journal of Robotics, Artificial Intelligence & Law*, 2021, 223 ss.

oppure meritevoli di eguale trattamento, emergono criticità a fronte di distinzioni realizzate non più (non solo) per opera dell'azione dalla persona, bensì più o meno intensamente e direttamente influenzate dal funzionamento della macchina[8]. Ci si riferisce a differenziazioni che, nei casi più complessi, costituiscono il prodotto di condotte quasi eterodirette in cui l'azione umana è pressoché estromessa dalla relazione causale azione-evento[9] sulla base di meccanismi dal funzionamento autonomo e, talvolta, dalle dinamiche interne sconosciute e di ardua ricostruzione *ex ante* e *ex post*.

Le difficoltà nel tracciare un collegamento tra la condotta umana e la discriminazione, motivata dalla intermediazione della macchina, si pone così al centro delle problematiche teoriche e applicative, che investono i rapporti tra la discriminazione, il diritto antidiscriminatorio[10] e l'intelligenza artificiale, che ci si propone di approfondire.

Sullo sfondo si staglia il più ampio tema delle relazioni tra la persona e la macchina, cioè tra la prima e le sue azioni, e l'innovazione tecnologica, a cui accede quello del "posto" occupato o che dovrebbe occupare la persona rispetto alla macchina alla luce del principio costituzionale di autodeterminazione letto nel prisma dell'art. 2 Cost. Una macchina che si vorrebbe restasse confinata alla categoria dei "mezzi"[11], ma che sembra invece sempre più tramutarsi in "soggetto agente"[12], rescindendo i legami tra la persona, la condotta e i suoi effetti.

Il rischio che sembra profilarsi è, insomma, che l'abdicazione in favore della macchina, qui intesa in senso volutamente astratto, e della tecnologia che si sta

[5] Niente affatto secondarie e da sfondo al dibattito si inseriscono le implicazioni di carattere etico, su cui insiste A. CELOTTO, *Come regolare gli algoritmi. Il difficile bilanciamento fra scienza, etica e diritto*, in *Analisi Giuridica dell'Economia*, 2019, 47 ss.

[6] L'impatto dell'intelligenza artificiale supera, come noto, la dimensione più circoscritta e propria del fenomeno discriminatorio, tanto da aver indotto autorevole dottrina a coniare l'espressione ormai famosa di realtà «on-life» a enfatizzare come la realtà virtuale si sta oppure si è ormai imposta al fianco di quella materiale e concreta. Così L. FLORIDI, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Milano, 2017.

[7] Si veda, in particolare, in tema il *memorandum* T*he main types and causes of discrimination* della *Sub-Commission on Prevention of Discrimination and Protection of Minorities* delle Nazioni Unite, 7 giugno 1949, UN Doc. E/CN. 4/Sub. 2/40/rev. 1. Per una ricostruzione della nozione di discriminazione, sue origini e cause, si consenta il rinvio a C. NARDOCCI, *Razza e etnia. La discriminazione tra individuo e gruppo nella dimensione costituzionale e sovranazionale*, Napoli, 2016.

[8] Sottolinea come le tecniche di intelligenza artificiale siano ontologicamente preposte a distinguere K. LIPPERT-RASMUSSEN, *Born free and equal? A philosophical inquiry into the nature of discrimination*, Oxford, 2014.

[9] Per una ricostruzione delle novità legate alla intermediazione della tecnologia nella costruzione delle dinamiche relazionali, J. PEARL, D. MACKENZIE, *The book. of why. The new science of cause and effect*, New York, 2018.

[10] La nozione di diritto antidiscriminatorio qui accolta copre quel ramo del diritto che si occupa della definizione dei tratti della discriminazione, della identificazione delle sue forme e tipologie, degli strumenti per la prevenzione ed il contrasto delle sue manifestazioni esterne. In materia, S. FREDMAN, *Discrimiantion Law*, Oxford, 2011.

[11] Richiama e si sofferma sulla distinzione di derivazione aristotelica tra mezzo e soggetto agente nella prospettiva della costruzione teorica delle relazioni tra persona umana e tecnologia, A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, cit., 67.

[12] *Ibidem.*

prepotentemente insediando negli spazi della vita quotidiana, pubblica[13] e privata, lasci indietro le persone[14] e, con esse, i loro diritti[15].

Non fa eccezione, tra questi, il principio di non discriminazione[16].

Che l'intelligenza artificiale distingua è dato intimamente connesso alla sua struttura ontologica, poiché la macchina è chiamata ad operare distinzioni sulla base di una definita e chiusa selezione di dati. Se, quindi, l'intelligenza artificiale per sua natura distingue, si tratterà di interrogarsi non tanto sul *se* le tecniche di intelligenza artificiale distinguono, quanto piuttosto sul *come* lo fanno. La ricerca del *come* costituisce, quindi, obiettivo primario dello studio.

L'indagine sul *come* distingue la macchina impone di accostarne l'analisi alla nozione di discriminazione accolta dal diritto di antidiscriminatorio e di interpretare il funzionamento della "macchina" alla luce dei criteri che qualificano una distinzione quale discriminazione. E, peraltro, che non ogni distinzione costituisca una discriminazione lesiva del principio di eguaglianza è portato pacificamente accolto dal diritto positivo, così come dalla giurisprudenza tanto costituzionale quanto sovranazionale.

La scelta di impostare lo studio della discriminazione prodotta dalla macchina dalla prospettiva dei principi e delle regole del diritto antidiscriminatorio tiene, inoltre, in adeguata considerazione peculiarità ulteriori che insistono, ancora una volta, sul *come* si realizza la distinzione. Il *come*, infatti, non sempre è decifrabile. Talvolta, non è noto il funzionamento interno della macchina; in altre, non è identificabile la causa primigenia della differenziazione: un insieme di criticità che precludono o, perlomeno, rendono particolarmente complesso quello scrutinio sulla ragionevolezza o "bontà" della distinzione tale da renderla esente da conseguenze sanzionatorie.

---

[13] Si pensi allo spazio riservato alle nuove tecnologie nel settore dell'amministrazione della giustizia, su cui svolge riflessioni critiche M. LUCIANI nel suo *La decisione giudiziaria robotica*, in *Rivista AIC*, 3/2018, 872 ss. Uno studio interessante sui settori e sulle modalità di impiego delle tecniche di intelligenza artificiale da parte del governo federale, è stato realizzato negli Stati Uniti da un team di ricercatori della Stanford University e della New York University, *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*, e può essere letto al link: https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.

[14] Sulla centralità che deve riconoscersi alla persona pure al cospetto dell'ingresso della tecnologia nella sfera dell'"umano", si sofferma T. GROPPI, *Alle frontiere dello stato costituzionale: innovazione tecnologica e intelligenza artificiale*, in *Consulta Online*, 3/2020, 675 ss. e, in particolare, 681 - 682.

[15] Per un approfondimento nell'ambito della letteratura nazionale sulle implicazioni derivanti dal ricorso alle tecniche di intelligenza artificiale, U. RUFFOLO (a cura di), *Intelligenza artificiale. Il diritto, i diritti. L'etica*, Milano, 2020 e ai contributi *ivi* contenuti.

[16] Interessante osservare come solo di recente la letteratura abbia cominciato ad occuparsi anche dei rapporti tra funzionamento delle tecniche di intelligenza artificiale e discriminazioni e come, tuttavia, continuino a prevalere studi di scienziati informatici. Offrono una analitica mappatura, corroborata da dati, degli studi ad oggi esistenti in tema di intelligenza artificiale e discriminazione, M. FAVARETTO, E. DE CLERCQ, B. SIMONE ELGER, *Big data and discrimination: perils, promises and solutions. A systematic review*, in *Journal of Big Data*, 2019, 1 e ss. Sull'opportunità di costruire modelli di intelligenza artificiale rispettosi del principio di non discriminazione si veda l'iniziativa del Consiglio d'Europa confluita nell'adozione nel dicembre del 2018 della *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment*. Sui rischi discriminatori legati alle tecniche di intelligenza artificiale, si veda invece lo studio condotto dalla *Federal Anti-Discrimination Agency*, dal titolo *Risks of Discrimination through the Use of Algorithms*, pubblicato in inglese nel 2020, scaricabile al link: https://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/EN/publikationen/Studie_en_Diskriminierungsrisiken_durch_Verwendung_von_Algorithmen.html.

Lo studio dei rapporti tra la discriminazione e l'intelligenza artificiale persegue così uno scopo duplice: verificare se e come l'intelligenza artificiale ha trasformato la nozione giuridica di discriminazione accolta dal diritto anti-discriminatorio di derivazione euro-unitaria, e, in caso affermativo, delineare i tratti della nuova fenomenologia discriminatoria.

Il saggio consta di due parti. La prima è dedicata alle mutazioni della discriminazione legata all'intelligenza artificiale: dal lato *oggettivo*, che investe la tenuta delle nozioni classiche del diritto antidiscriminatorio in termini di adeguatezza delle categorie della discriminazione diretta ed indiretta a ricomprendere le specificità della discriminazione derivante dal ricorso a tecniche di intelligenza artificiale; da quello *soggettivo*, che osserva il fenomeno "dal basso", dalla prospettiva delle vittime di condotte discriminatorie e delle qualità, individuali e collettive, su cui agisce la differenziazione, cioè dei fattori di discriminazione. Nella seconda, ci si occuperà invece degli approcci ed approdi del legislatore continentale e dei giudici, nazionali ed europei, soffermandosi sul ruolo che si vorrebbe o si ritiene dovrebbero assolvere rispettivamente il primo in punto di regolamentazione e di prevenzione e il secondo in fase di contenimento e di repressione del fenomeno.

PARTE PRIMA
LA DISCRIMINAZIONE "ARTIFICIALE"

1. *Esiste una* AI-derived discrimination?

L'interrogativo sulle potenzialità discriminatorie dell'intelligenza artificiale è stato da qualche anno superato e risolto in senso affermativo[17], tanto che ad oggi il tema che si

[17] Il tema si è imposto in modo importante all'attenzione dell'opinione pubblica a partire dal 2014, a seguito della pubblicazione del Report della Casa Bianca statunitense *Big Data: Seizing Opportunities, Preserving Values*, consultabile al link: https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, che in particolare riporta alcuni esempi di effetti discriminatori prodottisi per effetto del ricorso a tecniche di intelligenza artificiale. L'esempio più noto, su cui si avrà modo di tornare perché illustrativa dei caratteri della *AI-derived discrimination*, riguarda le conseguenze prodotte dalla App Street Bump, utilizzata nella città di Boston e realizzata in collaborazione con il *Mayor's Office of New Urban Mechanics*. La App mirava ad utilizzare dati raccolti dagli *smartphones* degli abitanti della città, allo scopo di valutare quali fossero le aree cittadine che avrebbero necessitato interventi di riqualificazione attraverso un'analisi delle condizioni delle strade e dei quartieri. La principale criticità della App derivava dal fatto che, dal momento che le categorie economicamente più povere, avevano meno possibilità di disporre di uno *smartphone* sul quale installare la App in esame, gli interventi a livello comunale si sarebbero concentrati soltanto nei quartieri a più alta intensità abitativa da individui appartenenti alle classi economicamente più agiate. La App, quindi, risultava discriminare alcune categorie di abitanti sulla base delle condizioni economiche. La letteratura ha iniziato ad occuparsi in modo più approfondito delle interferenze tra l'intelligenza artificiale e il fenomeno discriminatorio soprattutto successivamente alla pubblicazione del Report della Casa Bianca del 2014. Una ricerca interessante ha messo in evidenza che, nel novero dei lavori pubblicati dopo il 2014 e sino al 2019, su riviste indicizzate si contano solo 14 studi che si sono occupati del tema in esame dal punto di vista delle sue implicazioni e ricadute giuridiche. Il riferimento è al lavoro di M. FAVARETTO, E. DE CLERCQ, B. SIMONE ELGER, *Big data and discrimination: perils, promises and solutions. A systematic*

impone all'attenzione della dottrina, ma non solo[18], investe non più tanto il *se* la macchina discrimina[19], o meglio, se può discriminare, bensì il *come* lo fa, ai danni di *chi*, in base a *che cosa*.

Quanto al primo aspetto, i dati[20] raccolti sul funzionamento di alcune tecniche di intelligenza artificiale – prevalentemente di *machine learning* – attestano un rischio elevato di discriminazioni, spesso oscure nel loro funzionamento e difficilmente "spiegabili"[21] in termini di rapporto causa / effetto tra le condotte.

Sotto il profilo del *chi* soggiace a trattamenti discriminatori, le evidenze statistiche dimostrano che le ricadute pregiudizievoli derivanti dal ricorso alle tecniche di intelligenza artificiale interessano in modo proporzionalmente maggiore individui appartenenti a minoranze[22], intendendo qui riferirsi non soltanto a gruppi numericamente inferiori rispetto al resto della popolazione, in accordo con la definizione tradizionale di minoranza[23], bensì, e lasciando sullo sfondo il dato quantitativo, a gruppi che occupano

---

review, cit., in particolare 5 e ss. Uno studio ancora più recente dà conto dell'esistenza di una frattura significativa tra gli studi condotti negli Stati Uniti in tema di intelligenza artificiale e discriminazione e lo stato attuale dell'elaborazione teorica sul versante ceuropeo, così S. WACHTER, B. MITTELSTADT, C. RUSSELL, *Why Fairness Cannot Be Automated: Bridging the Gap Between Eu non-Discrimination Law and AI*, in *Computer Law & Security Review*, 2020, 1 e ss. Di recente, anche, K. CRAWFORD, tra cui *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven, 2021.

[18] Si pensi alla proposta in discussione in seno alle istituzioni dell'Unione Europea con cui si vorrebbe introdurre un divieto di utilizzo delle tecniche di IA di riconoscimento facciale a motivo dell'elevato tasso di discriminatorietà che ne connota il funzionamento.

[19] Nell'ambito della dottrina giuspubblicistica, tra gli altri, L. GIACOMELLI, *Big brother is «gendering» you. Il diritto antidiscriminatorio alla prova dell'intelligenza artificiale: quale tutela per il corpo digitale?*, in *BioLaw Journal*, 2019, 269 ss.; P. ZUDDAS, *Intelligenza artificiale e discriminazioni*, in AA.VV., Liber amicorum *per Pasquale Costanzo*, in *Consulta Online*, 2020, 1 ss.

[20] Il riferimento è, anzitutto, alle evidenze relative all'impiego delle tecniche di riconoscimento facciale, s cui J BUOLAMWINI, T. GEBRU nel suo *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, il testo integrale può essere letto al link: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. In questo contesto, si inseriscono anche i dati relativi alle modalità di impiego delle tecniche di IA da parte della polizia statunitense allo scopo di valutare a scopo preventivo le probabilità di recidiva oppure anche, a monte, di commissione di reati. In tema, da ultimo e tra i molti, S. BRAYNE, *Predict and Surveil. Data, Discretion, and the Future of Policing*, Oxford, 2020, e il Rapporto *The leadership conference on civil & human rights et al., predictive policing today: a shared statement of civil rights concerns*, redatto da un gruppo di 17 associazioni attive nel settore della tutela dei diritti civili che nel 2016 mettevano in guardia contro i rischi per le minoranze afro-americane connesse alle ricadute discriminatorie a sfondo razziale delle tecnologie di giusitizia predittiva, consultabile al seguente link: http://civilrightsdocs.info/pdf/final_jo intstatementpredictivepolicing.pdf. Tra i molti contributi in materia, A.D. SELBST, *Disparate impact in big data policing*, in *Georgia Law Review*, 2019, 109 ss. Ancora, alcuni studi hanno messo in luce come i dati raccolti sui *social networks* siano stati impiegati a scopo di profilazione; si veda in proposito M. KOSINSKI, D. STILLWELL, T. GRAEPEL, *Private traits and attributes are predictable from digital records of human behavior*, in *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 2013, 5802 ss. Un altro progetto pilota è *Our Data Bodies* (https://www.odbproject.org), che offre un monitoraggio sulle modalità con cui in alcune città (Detroit, Charlotte, Los Angeles) statunitensi sono raccolti e utilizzati dal potere pubblico dati relativi alle comunità più svantaggiate dei contesti urbani.

[21] La necessità che il funzionamento delle tecniche di IA sia "spiegabile" è sempre più avvertita in letteratura. Ci si riferisce al "right to explanation". In dottrina, tra i molti, M. PALMIRANI, *Big Data e conoscenza*, in *Rivista di filosofia del diritto*, 2020, 73 ss.

[22] Tra i molti, S.U. NOBLE, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, 2018.

[23] Il riferimento è alla nozione proposta da F. CAPOTORTI, nel suo *Study on the persons belonging to ethnic, religious and linguistic minorities*, United Nations, New York, 1979; in tema, anche, la *Sub-Commission on Prevention of Discrimination and Protection of Minorities*, UN.Doc. E/CN.4/641.

una posizione di subordinazione secondo una costruzione gerarchica e stratificata della società su cui tuttora insistono le relazioni di potere entro le società contemporanee[24].

In relazione al *che cosa*, ossia alla qualità individuale motivo della differenziazione irragionevole, si osserva un ampliamento degli elementi su cui poggia il trattamento diversificato non sempre coincidente con i fattori di discriminazione tradizionali e noti alle enumerazioni delle Carte costituzionali e dei trattati di diritto internazionale dei diritti umani.

In simile quadro, si inseriscono le tesi di coloro che ritengono che l'intelligenza artificiale riproduca "vecchie" discriminazioni, già esistenti e note all'elaborazione teorica e dottrinale, e che rifletta società strutturalmente o, nei casi più gravi, istituzionalmente diseguali.

Il tema è se simile caratterizzazione del fenomeno discriminatorio lo mantenga entro "binari" noti oppure se la correlazione tra discriminazione e intelligenza artificiale comporti conseguenze sulla conformazione degli elementi oggettivi della condotta discriminatoria, su quelli soggettivi della identificazione delle vittime e delle qualità individuali e/o collettive, allontanando la definita "*AI-derived discrimination*" dal seminato del diritto antidiscriminatorio.

Lo studio delle modalità con cui l'intelligenza artificiale discrimina è quindi condizione necessaria per valutare l'adeguatezza delle categorie del diritto antidiscriminatorio a fronteggiare il fenomeno, nel senso della sua piana riferibilità alla *AI-derived discrimination*; oppure, al fine di appurare se quest'ultima non sia isolabile, presentando specificità tali da assegnarle una fisionomia autonoma tale da sfuggire alle maglie del sistema vigente.

2. *Stesse cause, diversa fenomenologia?*

«Data are assumed to accurately reflect the social world, but there are significant gaps, with little or no signal coming from particular communities. While massive datasets may feel very abstract, they are intricately linked to physical place and human culture»[25].

Così una delle maggiori studiose delle interazioni tra discriminazione e intelligenza artificiale appuntava l'attenzione sulla neutralità e astrattezza solo apparenti del dato[26]

[24] Per una ricostruzione delle proposte definitorie che hanno investito la nozione di minoranza, si consenta il rinvio a C. NARDOCCI, *Razza e etnia. La discriminazione tra individuo e gruppo nella dimensione costituzionale e sovranazionale*, cit.; nella dottrina costituzionalistica, A. PIZZORUSSO ai cui studi si rinvia diffusamente, tra cui *Le minoranze nel diritto pubblico interno*, Milano, 1967, e *Minoranze e maggioranze*, Torino, 1993..

[25] K. CRAWFORD, *The Hidden Biases in Big Data*, in *Harvard Business Review*, 2013.

[26] Qui il riferimento è al "dato" in senso volutamente generico, a cui si accostano una serie di ulteriori nozioni e classificazioni tra cui quella che più rileva in questa sede è quella di *big data*, definibile quale insieme consistente di dati prodotti dall'uomo ovvero dalle macchine; in tema, si rinvia a R. HENRY, S. VENKATRAMAN, *Big Data Analytics: The next big learning opportunity*, in *Academy of Information and Management Sciences Journal*, 2015, 15 ss. e, in particolare, 17 e ss. Sul concetto di big data, si veda, anche, L. FLORIDI, tra cui, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, cit., 15 ss. Per quanto attiene alla nozione di "dato", si veda, la definizione dell'Enciclopedia Treccani dove il *dato* è descritto come: «la rappresentazione di un'informazione realizzata nell'ambito di un linguaggio formalizzato e non ambiguo, spesso mediante simboli numerici o alfabetici, ma sempre in una forma tale

che, lungi dal riprodurre fedelmente la realtà circostante, ne rifletterebbe soltanto una parte: quella che "vede" o gli è "fatta vedere" dalla persona.

Alla lettura che vorrebbe il dato e la macchina neutrali, dovrebbe cioè sostituirsene una che ne enfatizzi il legame con l'azione soggettiva della persona: più stretto è il legame tra la persona, il dato e la macchina, tanto più il pregiudizio umano si riverbera sul dato e sulla macchina. La dimensione "terrena" del dato è ciò che lo rende, quindi, potenzialmente lacunoso, parziale, non plurale.

Che il dato sia incapace di riprodurre la eterogeneità della società "degli uomini" non dovrebbe sorprendere alla luce della sua natura probabilistica, che dovrebbe guidare la sua *interpretazione* e non la sua *sostituzione* alla persona. Il problema nasce, invece, da una lettura assolutista, che disancora il dato dalla sfera della probabilità per elevarlo a fattore di verità.

Se è vero che il dato nasce dalla persona, ma poi è impiegato quale elemento obiettivo su cui operare differenziazioni presuntivamente imparziali, si rescinde quel collegamento, che esiste tra la persona, il dato e i suoi effetti, che è centrale per la ricostruzione dei meccanismi di funzionamento delle tecniche di intelligenza artificiale, del loro eventuale carattere discriminatorio, dell'accertamento di responsabilità, individuali e collettive.

Le interferenze tra la persona e la macchina non mutano quindi le cause, tutte umane, della nuova discriminazione.

Non solo. A monte della condotta umana, si inserisce un elemento ulteriore, quello culturale che colora e influenza tanto la condotta umana, prima, quanto il dato, poi[27]. Il riferimento alla cultura, che non è esterna ma interna al dato così come lo è rispetto alla persona, avvicina l'elaborazione teorica sulle cause della discriminazione legata all'intelligenza artificiale al ben più arato dibattito su quelle della discriminazione

da poter essere trattata con una metodologia di elaborazione; con un uso più specifico, al plurale, dati (o all'inglese data), l'insieme delle informazioni necessarie a, o prodotte da, una elaborazione automatica, rappresentate in una forma e mediante un supporto conveniente all'elaborazione stessa».

[27] Sulla connessione tra dato informatico e cultura, L. GITELMAN, *"Raw Data" Is an Oxymoron*, Cambridge, 2013.

"classica" [28], interpretata alternativamente quale esito di un conflitto tra gruppi sociali[29] oppure quale prodotto di stereotipi e pregiudizi[30], in particolare, di tipo «collettivo»[31].

Stesse cause, quindi, ma diversa fenomenologia.

Le ragioni. In estrema sintesi, esse risiedono nel funzionamento delle tecniche di intelligenza artificiale che, pure nella loro eterogeneità, presentano alcuni momenti dai quali possono scaturire effetti discriminatori.

La descrizione e la scansione temporale dei passaggi che presiedono alla programmazione delle tecniche di IA diviene allora particolarmente importante, perché in grado di rivelare i momenti in cui si insinua il rischio dell'ingresso di un fattore potenzialmente causa del funzionamento discriminatorio della macchina (c.d. tempo del *bias*).

La letteratura[32]più recente ne individua cinque: la individuazione e selezione delle "*target variables*" e delle "*class labels*"; la raccolta e la selezione dei dati (*data training*); la selezione della caratteristica inserita nel modello ("*feature selection*"); la scelta del "*proxy*"; la discriminazione intenzionale, cioè consapevolmente posta in essere dal programmatore, che discenda da una volontaria e parziale selezione di dati che si riverbera negativamente ai danni di una o più classi protette ("*masking*").

Con riferimento al primo, il rischio di distorsioni discriminatorie risiede anzitutto nella individuazione e nelle relazioni istituite tra la "*target variable*", la caratteristica che il sistema ricerca, e la "*class label*", cioè la categoria che le viene associata. Le tecniche di IA funzionano sulla base di associazioni tra questi due parametri: al ricorrere dell'uno,

---

[28] Nella prospettiva del diritto costituzionale e del diritto internazionale dei diritti umani, le Costituzioni nazionali e i trattati non contengono una definizione di "discriminazione", limitandosi a sancirne il divieto, richiamando elenchi più o meno ampi di elementi in base ai quali non può darsi ingresso a disparità di trattamento. Soccorre, come noto, la giurisprudenza e, sicuramente a partire dagli anni 2000, il diritto antidiscriminatorio dell'Unione Europea che ha occupato un posto di primo piano per quanto attiene alla elaborazione e alla previsione di norme di diritto positivo contenenti alcuni concetti definitori quanto alle tipologie di fattispecie discriminatorie. In dottrina, K. LIPPERT RASMUSSEN (a cura di), *The Routledge Handbook of the Ethics of Discrimination*, New York, 2018. Sulle caratteristiche del diritto antidiscriminatorio forgiato in seno all'Unione Europea, si rinvia, tra i molti, a M. BELL, *Anti-Discrimination and the European Union*, Oxford, 2002; C. MCCRUDDEN, *Anti-Discrimination Law*, Dartmouth, 2004; D. SCHIEK, V. CHEGE, *European Union Non-Discrimination Law. Comparative perspectives on multidimensional equality*, Londra, 2009. Nell'ambito della dottrina italiana, si vedano i contributi pubblicati in M. BARBERA (a cura di), *Il nuovo diritto antidiscriminatorio*, Milano, 2007.

[29] Esponente delle teorie sul conflitto («conflict power theories») è, anzitutto, D.L. HOROWITZ, al cui *Ethnic groups in conflict*, Oakland, 1985, si rinvia diffusamente. In tema, anche, M.N. MARGER, *Race and Ethnic Relations. American and Global Perspectives*, Boston, 2009; e, soprattutto, A.D. SMITH, *The Ethnic Revival*, Cambridge, 1981 e, dello stesso A., anche, *The Ethnic Origins of Nations*, Oxford, 1988.

[30] In tema, si precisa che in letteratura non si riscontrano tesi uniformi quanto alle relazioni tra il pregiudizio e gli stereotipi, da un lato, e la discriminazione, dall'altro. Secondo le elaborazioni dottrinali sviluppatesi in seno agli studi di psicologia, sarebbe il pregiudizio ad essere causa di discriminazione. In dottrina, si vedano, tra i molti, A.W. GORDON, *The Nature of Prejudice*, New York, 1958; G.E. SIMPSON, J.M. YINGER, *Racial and Cultural Minorities*, New York, 1985; J. DUCKITT, *The Social Psychology of Prejudice*, New York, 1992. Viceversa, le tesi diffuse tra gli studiosi di sociologia invertono il rapporto interpretando il pregiudizio quale prodotto di relazioni gerarchiche tra gruppi sociali che consolida relazioni di potere.

[31] Il riferimento è N. BOBBIO, in *Elogio della mitezza e altri scritti morali*, Milano, 2010.

[32] Il riferimento è a S. BAROCAS, A.D. SELBST, *Big data disparate impact*, in *California Law Review*, 2016, 671 ss., ripresi da F.Z. BURGESIUS, *Discrimination, artificial intelligence, and algorithmic decision-making*, Council of Europe Publications, 2018, in particolare, 10 ss. In tema, anche, P. ZUDDAS, *Intelligenza artificiale e discriminazioni*, cit.

la *target variable*, la tecnica di IA vi associa una determinata categoria, la *class label*, operando una scelta.

Il problema delle associazioni istituite dall'algoritmo[33] nasce in presenza di caratteristiche difficilmente inquadrabili entro uno schema binario, che consenta alla macchina di scegliere tra due alternative. Il tema riguarda tutte le ipotesi, in cui si chiede di individuare e selezionare caratteristiche da collegare in via automatica a categorie di difficile definizione. Si pensi alla categoria del "creditore affidabile"[34] o del "buon lavoratore": in questi casi, il processo di selezione delle *target variables* e, quindi, dei parametri predittivi dell'appartenenza alla categoria è complesso e largamente discrezionale, trattandosi di classi di nuova creazione che poggiano su concetti non binari[35].

La circostanza che le scelte siano compiute nella fase di elaborazione del modello evidenzia il rischio che esse celino scelte irragionevoli e discriminatorie, provocando effetti deteriori ai danni delle classi protette[36]. La discriminazione, in definitiva, può derivare dalla individuazione della caratteristica, della categoria, oppure di entrambe, cui si aggiungono criticità connesse a classificazioni fallaci, cioè erronee associazioni tra le caratteristiche selezionate e le classi di appartenenza, che riproducono lo schema falso positivo / falso negativo[37].

Il secondo meccanismo è costituito dal *training* dei dati, che assolve ad una funzione centrale nel funzionamento della macchina: in tanto i dati sono parziali o frutto di pregiudizi, in tanto il modello ne rifletterà la parzialità agendo in modo discriminatorio secondo la regola del «*garbage in, garbage out*»[38].

Le criticità di questa fase possono ricondursi a due situazioni. La prima si verifica in presenza di dati non egualmente rappresentativi delle componenti individuali e collettive della società: la macchina funzionerà a partire da una erronea rappresentazione della realtà, riproducendo schemi che produrranno effetti proporzionalmente deteriori ai danni della categoria sotto oppure sovra-rappresentata. Il riferimento è ai rischi connessi alla marginalizzazione sofferta da alcuni gruppi, che non hanno accesso o non

---

[33] In letteratura e sul concetto di algoritmo, si rinvia a T. GILLESPIE, *Algorithm*, in B. PETERS (a cura di), *Digital Keywords: A Vocabulary of Information Society and Culture*, Princeton, 2016.

[34] Un esempio è offerto dal sistema australiano di calcolo dei pagamenti in eccesso e disporre l'emissione di avvisi di debito, *Robodebt*, cui ha fatto seguito uno scandalo che ha investito l'intera opinione pubblica per gli effetti discriminatori derivanti dall'algoritmo, tanto da essersi posto alla base di una *class action*. Per un approfondimento, *Robodebt: government admits it will be forced to refund $550m under botched scheme*, *The Guardian*, https://www.theguardian.com/australia-news/2020/mar/27/robodebt-government-admits-it-will-be-forced-to-refund-550m-under-botched-scheme.

[35] Per fare un esempio di concetti binari che rendono, viceversa, funzionale e meno pericoloso l'operato della macchina, si veda il concetti di spam, rispetto ai quali è più agevole per il programmatore la selezione dei parametri in base ai quali classificare una mail come spam rispetto ad una che, invece, non lo è. Gli esempi sono di S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit., 677 ss.

[36] Così S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit., 680.

[37] *Ibidem*. Secondo T.B. GILLIS, J.T. SPIESS a complicare ulteriormente il quadro si inseriscono anche le difficoltà legate alla comprensione del ruolo assolto dalle *target variables* nel processo decisionale della macchina, così *Big data and discrimination*, in *The University of Chicago Law Review*, 2019, 475.

[38] In tema, G. RESTA, *Governare l'innovazione tecnologica: decisioni algoritmiche, diritti digitali e principio di uguaglianza*, in *Politica del diritto*, 2019, cit., 214. In tema, anche, A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, cit., 84 ss.

frequentano il *web* e che la dimensione virtuale non conosce, considerandoli inesistenti[39]. Poiché non sempre i dati sono raccolti e selezionati in modo eguale, esistono delle zone grigie dove alcuni gruppi o comunità intere sono, alternativamente, eccessivamente rappresentanti oppure, al contrario, sotto-rappresentati[40]. Il problema si esacerba nel momento in cui il potere pubblico decide di subordinare le proprie scelte a tali dati, traslando una marginalizzazione apparente e solo astratta, poiché virtuale, in una concreta, che si risolve spesso in un aggravamento di condizioni di svantaggio già esistenti. La seconda riguarda, invece, le ipotesi in cui la macchina estende, replicandolo nei confronti di casi "altri", il modello viziato dal pregiudizio originario. Si segnala, peraltro, che può anche accadere che la macchina, oltre al pregiudizio iniziale, ve ne affianchi ulteriori, utilizzando i comportamenti degli utenti quali parametri per l'estrazione dei dati[41]. In definitiva, tutto viene a dipendere dalla qualità dei dati[42]: meno sono accurati, maggiori saranno i rischi che la macchina distingua in modo discriminatorio.

Il terzo meccanismo è la *feature selection*, cioè la selezione delle caratteristiche individuali rilevanti per la costruzione del modello. Anche in questo caso, una delimitazione parziale che ometta di considerare tutti i profili oppure ne accentua taluni a discapito di altri, si traduce nel funzionamento potenzialmente discriminatorio della macchina.

Un ruolo fondamentale nella distinzione discriminatoria è assolto dal *proxy*, l'elemento, cioè, in base al quale la macchina differenzia e che costituisce il quarto meccanismo, secondo lo schema di cui sopra. La scelta del *proxy*, che per ovvie ragioni di rispetto del principio di non discriminazione non dovrebbe ricadere su fattori di discriminazione classici, talvolta ne risulta ugualmente predittivo, ridondando il funzionamento della macchina in una discriminazione ai danni degli appartenenti ad una categoria protetta. Accade, infatti, che tanto più la scelta del *proxy* investe caratteri che accomunano i membri di una categoria protetta, tanto maggiori saranno le probabilità che la macchina, distinguendo in base a quel *proxy*, farà oggetto coloro che appartengono a quel gruppo di un trattamento discriminatorio.

---

[39] Approfondisce questo aspetto, J. LERMAN, *Big Data and Its Exclusions*, in *Stanford Law Review Online*, 2013, che avverte delle criticità che conseguono alla subordinazione di scelte di politica pubblica ed economiche a dati raccolti sulla base di un campione di riferimento che non tiene conto di coloro che vivono ai margini della società, perché individui che non producono dati, perché non possiedono uno *smartphone* oppure non utilizzano assiduamente i *social networks*, non dando ingresso alle proprie preferenze tramite canali che divengono privilegiati perché appannaggio solo di alcuni.

[40] Così, K. CRAWFORD, *Think Again: Big Data*, 2013: http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data.

[41] Si pensi agli annunci pubblicitari di Google (*ads*), che uno studio ha dimostrato accostare nomi propri, più frequentemente impiegati dalla popolazione afro-americana rispetto a quella bianca, a precedenti oppure fedine penali. Lo studio ha rivelato che il nome proprio veniva impiegato dall'algoritmo quale criterio predittivo dell'appartenenza alla razza e collegava al nome proprio, e quindi alla presunta razza di appartenenza, l'altrettanto presunto riferimento ad un precedente penale (ad esempio, un arresto). La ricerca risale al 2013 ed è pubblicata nello studio condotto da L. SWEENEY, *Discrimination in Online Ad Delivery*: *Google ads, black names and white names, racial discrimination, and click advertising*, in *Communicaion of the Acm*, 2013, 44 ss.

[42] Insistono su questo aspetto S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit., 687.

Oltre le specificità di ciascun meccanismo, vi sono due elementi, su cui si tornerà, che li accomunano tutti e che integrano aspetti specifici della *AI-derived discrimination*: da un lato, il fattore oggettivo della imperfezione dell'informazione, cioè del dato, che tradizionalmente si pone alla base della discriminazione statistica cui può ricondursi quella legata alle tecniche di IA[43]; dall'altro, un aspetto di natura soggettiva, il pregiudizio implicito[44], che costituisce uno degli aspetti più ricorrenti e pericolosi della fenomenologia discriminatoria in esame, tanto da aver indotto alcuna dottrina a ricondurre a quest'ultimo, più che a scelte umane consapevoli, la genesi della discriminazione "artificiale"[45].

Quinta e ultima modalità con cui le tecniche di intelligenza artificiale discriminano, la più intuitiva e meno problematica in punto di accertamento, si verifica in presenza della programmazione intenzionalmente discriminatoria del modello: il "*masking*"[46].

Infine, due fenomeni ulteriori possono rendere o aggravare il funzionamento discriminatorio delle tecniche di intelligenza artificiale: l'aggiornamento dei dati, richiesto dal progresso scientifico sopravvenuto, e, aspetto patologico, il loro inquinamento più o meno volontario da parte dell'intervento umano (c.d. *data poisoning*).

3. *Sulla possibile o auspicabile trasposizione delle categorie del diritto anti-discriminatorio*

L'analisi che precede ha dimostrato l'esistenza di peculiarità che paiono separare la discriminazione "artificiale" dalla fisionomia classica del fenomeno discriminatorio.

Si tratta ora di appurare se le categorie del diritto antidiscriminatorio classico possano essere utilmente impiegate anche in relazione alla discriminazione "algoritmica" oppure[47], in quale misura e con quali modalità, quest'ultima devia rispetto alle prime[48].

---

[43] In dottrina, si rinvia a L.J. STRAHILEVITZ, *Privacy versus Antidiscrimination*, in *Chiacago Law Review*, 2008, 363 ss.; K. LIPPERT-RASMUSSEN, *Statistical (And Non-Statistical) Discrimination*, in *The Routledge Handbook of the Ethics of Discrimination*, 2017, 42 ss.

[44] In tema, per un approfondimento sul concetto di *bias* implicito, si rinvia a C. JOLLS, C.R. SUNSTEIN, *The Law of Implicit Bias*, in *California Law Review*, 2006, 969 ss.; in tema, e con specifico riferimento al caso della discriminazione algoritmica, si vedano N. SCHMID, B. STEPHENS, *An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination*, in *ArXiv*, 2019, 130 ss.

[45] Si veda, così, A. CHANDER, *The Racist Algorithm?*, in *Michigan Law Review*, 2017, in particolare, 1207-1208.

[46] In dottrina, su questa forma di discriminazione intenzionale, F.Z. BURGESIUS, *Discrimination, artificial intelligence, and algorithmic decision-making*, 13 ss.; J.A. KROLL e altri, *Accountable algorithms*, in *University of Pennsylvania Law Review*, 2016, 633 ss.

[47] La letteratura recente sembra incline a sostenere l'inadeguatezza degli strumenti del diritto anti-discriminatorio classico a fronteggiare disparità di trattamento legate alle tecniche di intelligenza artificiale. Così, T.B. GILLIS, J.T. SPIESS, *Big data and discrimination*, in *The University of Chicago Law Review*, 2019, 458 ss.; S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit.; C. O'NEIL, *Wapons of math destruction: how big data can increase inequality and threat democracy*, London, 2017.

[48] Una diversa metodologia di analisi è, invece, quella che, emancipandosi dal diritto anti-discriminatorio, guarda alla discriminazione "algoritmica" da tre angolazioni: la prima che investe il i*nput* della decisione, ossia l'elemento sul quale fa perno la scelta di differenziare; la seconda che interessa il processo decisionale che conduce alla disparità di trattamento; la terza che guarda, infine, alla disparità del risultato. Si tratta della proposta su cui insistono T.B. GILLIS, J.T. SPIESS, *Big data and discrimination*, in *The University of Chicago Law Review*, cit.

3.1. *La discriminazione diretta: quando le tecniche di intelligenza artificiale distinguono ("male")*

La prima prospettiva di indagine investe la nozione di discriminazione diretta.

Una lettura che accosti la discriminazione "artificiale" alla teoria della discriminazione diretta (*disparate treatment*) ne mette subito in evidenza alcune deviazioni, che si traducono in altrettante criticità connesse all'impiego della categoria in esame per qualificare e sanzionare come tale la discriminazione algoritmica.

Oltre l'ipotesi del *masking*, di cui si è detto, e dei casi relativi alla costruzione di modelli intenzionalmente preordinati a produrre effetti discriminatori, ad esempio poiché implementati in contesti di razzismo istituzionale[49], la teoria della discriminazione diretta pare male si attagli alla tipologia discriminatoria in esame per alcune ragioni.

La prima. La selezione delle *target variables* e delle *class labels*, nonché le associazioni tra le prime e le seconde che costituiscono come detto uno dei meccanismi in cui si annida il rischio che le tecniche di IA producano un effetto di tipo discriminatorio, difficilmente si fondano esplicitamente su fattori di discriminazione che richiamano in modo diretto classi protette. Più spesso, si tratta di elementi che solo indirettamente risultano predittivi dell'appartenenza individuale alla categoria protetta, come richiede la teoria del *disparate treatment* che, come noto, presuppone che la qualità individuale su cui ricade la distinzione sia *prima facie* riconducibile ad uno dei fattori di discriminazione vietati.

Una secondo elemento di deviazione si scorge con riferimento alla scelta del *proxy* e per ragioni in tutto speculari. L'elemento prescelto può non apparire immediatamente indicativo dell'appartenenza individuale ad un gruppo protetto e, laddove non vi sia una disparità di trattamento che direttamente si fondi sul fattore di discriminazione vietato, quella distinzione non sarà inquadrata nella prospettiva della discriminazione diretta e sanzionata in quanto tale.

Il problema che accomuna queste prime due criticità o deviazioni dal modello classico della discriminazione diretta investe, quindi, il rapporto solo mediato e non diretto che si instaura tra la qualità individuale assunta a *ratio* della distinzione ed uno o più dei fattori che il diritto anti-discriminatorio vieta possano assurgere a motivo di disparità di trattamento.

La discriminazione algoritmica ha così, da un lato, concorso a mettere in luce l'eterogeneità e l'elevato numero di elementi che possono fungere da indici di appartenenza individuale, che ben potrebbe costituire un ausilio nel rendere più pervasiva ed efficace la tutela anti-discriminatoria; dall'altro, però, trattandosi di qualità "altre" rispetto a quelle coperte dai fattori di discriminazione tradizionali, esse non sono interpretate dal diritto anti-discriminatorio quali criteri sintomatici di una discriminazione, sfuggendo pertanto ai sistemi di repressione legislativamente vigenti.

[49] In dottrina, M.R. GOMEZ, *The Next Generation of Disparate Treatment: A Merger of Law and Social Science*, in *The Review of Litigation*, 2013, 553 ss.

Potrebbe così sostenersi che un modo alternativo per stabilire quando una distinzione algoritmica sia ragionevole o meno, per superare le difficoltà appena descritte, sarebbe affermare che una disparità di trattamento non è discriminatoria soltanto qualora non sia possibile stabilire dagli effetti del funzionamento della tecnica di intelligenza artificiale se il soggetto colpito dalla decisione appartenga o meno ad un gruppo protetto[50].

Anche a voler superare le due criticità sopra menzionate ipotizzando un allargamento delle qualità sintomatiche della disparità di trattamento vietata, facendo perno, ad esempio, sul concetto di discriminazione per associazione[51], vi è però una terza criticità che rende difficilmente qualificabile come diretta in senso classico la discriminazione algoritmica. Ci si riferisce al difetto di intenzionalità che spesso accompagna la condotta dell'agente a cui si affiancano le difficoltà connesse alla individuazione dell'atto a cui imputare l'effetto discriminatorio. La prova della intenzionalità richiede, infatti, di stabilire anzitutto *quale* sia la condotta dalla quale dipende l'effetto discriminatorio secondo un nesso di causa / effetto diretto e, quindi, ricercare in quale dei passaggi che presiedono alla programmazione del modello si innesta l'atto censurabile. Entrambi questi passaggi si dimostrano particolarmente complessi nel caso della discriminazione legata al funzionamento delle tecniche di intelligenza artificiale.

Potrebbe, inoltre, accadere, che, a fronte della possibilità di istituire un collegamento causale tra atto ed effetto discriminatorio in relazione ad un elemento ritenuto pianamente predittivo dell'appartenenza ad un gruppo protetto, la intenzionalità dell'agente sia solo apparente. Il caso è quello del *bias* implicito[52], ossia del pregiudizio sconosciuto all'agente, che tuttavia vi conforma la propria condotta in modo del tutto inconsapevole. In questi casi, analogamente, non potrà ritenersi provata la ricorrenza della discriminazione diretta, difettando il *discriminatory intent*.

Si inseriscono in questo quadro anche le ipotesi, contigue, caratterizzate da una frattura più o meno importante tra intenzionalità, consapevolezza e conoscenza, che la dottrina qualifica come *unconscious disparate treatment*.

[50] Così J.A. KROLL e altri, *Accountable algorithms*, cit., che precisano: «[a] different way to define whether a classification is fair is to say that we cannot tell from the outcome whether the subject was a member of a protected group or not. That is, if an individual's outcome does not allow us to predict that individual's attributes any better than we could by guessing them with no information, we can say that outcome was assigned fairly», 690.

[51] Su cui si veda, *infra*.

[52] Il tema è stato approfondito in modo particolare dalla dottrina statunitense. Così C. JOLLS, C.R. SUNSTEIN, *The Law of Implicit Bias*, in *California Law Review*, 2006, 969 ss.

3.2. *A cavallo tra discriminazione diretta e indiretta: l'*unconscious disparate treatment

Come premesso, la discriminazione derivante dal ricorso all'intelligenza artificiale presenta alcune tipicità, che difficilmente la rendono inquadrabile entro la teoria della discriminazione diretta[53].

L'intenzionalità, quale elemento imprescindibile della discriminazione diretta legato alla dimensione soggettiva e che vuole l'agente consapevole del proprio intento discriminatorio, assume contorni peculiari nella discriminazione algoritmica che rende sfumata la distinzione tra la nozione di discriminazione diretta e quella indiretta[54].

Una delle criticità principali del fenomeno discriminatorio legato all'intelligenza artificiale è costituita, infatti, dallo spazio che occupa il pregiudizio implicito nella fase di costruzione del modello, oltre che, eventualmente, in quelle successive connesse all'aggiornamento dei dati. Il funzionamento in senso discriminatorio della macchina può, infatti, derivare dall'influenza che il pregiudizio, di cui il soggetto agente può anche essere inconsapevole[55], esercita nei tanti momenti in cui si snoda il processo di differenziazione operato dalla macchina.

Nella prospettiva di un inquadramento del fenomeno entro le categorie del diritto anti-discriminatorio, la difficoltà principale deriva dalla circostanza che, anche se dal punto di vista oggettivo la disparità di trattamento presenta i caratteri della discriminazione diretta perché si basa espressamente su un fattore di discriminazione oppure su un elemento comunque predittivo dell'appartenenza individuale alla categoria protetta, difetta quello soggettivo della intenzionalità della condotta, cioè dalla volontà di discriminare imputabile all'agente.

Tanto più il pregiudizio sarà implicito, tanto più la disparità di trattamento non potrà inquadrarsi entro la categoria dogmatica della discriminazione diretta. "Tanto più", poichè l'inconsapevolezza rilevante ai fini della disapplicazione della teoria del *disparate treatment* può assumere gradi o livelli diversi, potendosi distinguere il caso della inconsapevolezza, che è anche ignoranza intesa come difetto di conoscenza del *bias* da parte dell'agente, da quello della inconsapevolezza che è, però, conoscenza. Ci si riferisce all'ipotesi dell'impiego di un modello costruito in modo discriminatorio da parte di un soggetto che non ha concorso alla sua programmazione e che, pur tuttavia, ne fa uso nonostante sia a conoscenza degli effetti discriminatori che ne derivano[56].

---

[53] In questo senso, anche, S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit., in particolare 701.

[54] Su cui, *infra*.

[55] Sulla nozione di pregiudizio implicito L.S. RICHARDSON, P.A. GOFF, *Implicit Racial Bias in Public Defender Triage*, in *Yale Law Journal*, 2013, 2626 ss., richiamati da A. CHANDER, *The Racist Algorithm?*, cit., 1028.

[56] Sullo sfondo dei rapporti su cui fa perno il diritto anti-discriminatorio classico nel descrivere e presupporre, quale elemento costitutivo della discriminazione diretta, il concetto di intenzionalità e la sua differenza con quello di conoscenza o conoscibilità, si scorgono strette connessioni con la nozione di colpevolezza, intesa come «valutazione del legame psicologico o, comunque, del rapporto di appartenenza tra 'fatto' e 'autore'; nonché la valutazione delle circostanze di carattere personale che incidono sulle capacità di autodeterminazione del soggetto», cit., 192, così G. FIANDACA, E. MUSCO, *Diritto penale. Parte generale*, Bologna, 2014. Ancora, più diffusamente, quanto alla distinzione tra dolo eventuale e colpa cosciente la cui elaborazione teorica nell'ambito degli studi penalistici assolve rilievo significativo anche

L'*unconscious disparate treament*, quale categoria dogmatica a sé stante, si caratterizza quindi per: l'assenza di intento discriminatorio dell'agente, che non lo vorrebbe ma discrimina; la ricorrenza di un pregiudizio implicito che "infetta"[57] la costruzione del modello rendendone discriminatorio funzionamento ed effetti; la sussistenza di tutti gli altri elementi che qualificano la disparità di trattamento come discriminazione diretta.

Date queste premesse, si profilano due questioni.

La prima riguarda come distinguere, dal punto di vista dell'accertamento della responsabilità individuale, il caso di colui che, affetto da un pregiudizio implicito, vi conforma un modello, da quello di chi invece si "limita" ad utilizzare un sistema di intelligenza artificiale che sa essere viziato da un *bias* non riconducibile però ad una sua azione diretta e che non intende volontariamente discriminare. Seguendo la teoria della discriminazione diretta, la prima condotta andrebbe esente da sanzione, ma anche la seconda poichè, si ritiene, il *disparate treatment* poggia sulla distinzione tra intenzionalità e mera conoscenza[58], nel senso che solo la seconda, non anche la prima, è sufficiente ad integrare gli estremi della discriminazione diretta.

Si potrebbe, quindi, ipotizzare di trasferire lo scrutinio dalla sussistenza dell'elemento soggettivo, legato alla condotta dell'agente, all'effetto discriminatorio. L'attenzione all'effetto è, però, caratteristica tipica della nozione di discriminazione indiretta con la conseguenza che, a voler abbracciare simile impostazione, si uscirebbe dal seminato della teoria del *disparate treatment*. Allo stesso tempo, la fenomenologia discriminatoria in esame non presenta il carattere tipico della discriminazione indiretta, cioè l'apparente neutralità della regola o, in questo caso, del modello, rendendo quindi di difficile inquadramento la discriminazione in esame che si allontana da entrambe le ricostruzioni teoriche del *disparate treatment* e del *disparate impact*. Non a caso, la letteratura si è espressa nel senso che l'*unconscious disparate treatment* costituisca una tipologia di discriminazione autonoma, da collocare a cavallo tra la discriminazione diretta e quella indiretta[59].

Un secondo aspetto, che vorrebbe superare la problematica appena illustrata per restare nel recinto della discriminazione diretta, attiene alla possibilità di attribuire alla macchina più che al soggetto agente l'intento discriminatorio. E, tuttavia, anche tale opzione pare difficilmente spendibile. La discriminazione diretta poggia, infatti, su una ricostruzione che interpreta l'intenzionalità come prodotto per definizione umano e, in quanto tale, non imputabile alla macchina[60]. In quanto alla macchina non può attribuirsi alcuna volontà soggettiva intenzionalmente preordinata a discriminare, cioè a distinguere

nel contesto più specifico della sanzionabilità della condotta violativa del principio di eguaglianza e di non discriminazione, si vedano gli stessi AA., 313 e ss.

[57] Si vuole richiamare il linguaggio, particolarmente evocativo, *infected*, utilizzato da *Justice* R. BADER GINSBURG in alcune delle sue *opinions* in materia di non discriminazione. Tra tutte, Corte Suprema, *Ledbetter v. Goodyear Tyre & Rubber Co., INC.*, 550 US 618 (2007), 29 maggio 2007.

[58] Così S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit., 700, che analizzano il tema dalla prospettiva delle discriminazioni che l'impiego delle tecniche di intelligenza artificiale possono produrre nel contesto lavorativo.

[59] S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit.

[60] *Ibidem.*

"male", in tanto la fattispecie discriminatoria in esame non potrà essere censurata quale discriminazione diretta.

Piuttosto, sarebbe interessante sollecitare le Corti, nazionali e sovranazionali, a riconoscere nel pregiudizio, specie in quello implicito, l'elemento al quale ancorare la genesi della discriminazione e la sua censura[61], soprattutto alla luce di studi[62] che dimostrano la diffusione e il peso ben più importante del *bias* implicito, inconscio, rispetto a quello di cui l'agente è consapevole conformandovi il proprio *agere* sociale.

3.3. *La discriminazione indiretta: dove risiede la (non) neutralità?*

Per fronteggiare disparità di trattamento non intenzionali ovvero almeno apparentemente non intenzionali, il diritto anti-discriminatorio ha introdotto la nozione di discriminazione indiretta[63].

Come il concetto di discriminazione indiretta reagisce al cospetto dell'evoluzione tecnologica è interrogativo particolarmente opportuno a fronte delle descritte difficoltà nel qualificare e sanzionare la seconda quale discriminazione diretta.

Il trasferimento del sindacato dalla volontà / intenzionalità che sorregge l'azione del soggetto agente all'effetto discriminatorio sembrerebbe ad un primo sguardo una strategia funzionale a contenere le criticità che si snodano intorno alla complessa ricostruzione del nesso causale tra condotta ed evento, ma, soprattutto, alla prova della sussistenza dell'intento discriminatorio addebitabile all'agente a motivo dell'intermediazione della macchina.

Il tema è allora se, di fronte ad una fenomenologia discriminatoria che non integra gli estremi della discriminazione diretta, essa possa viceversa ricondursi alla nozione di *disparate impact*.

La risposta, anche in questo caso, si rivela problematica.

---

[61] Qualche spunto è offerto, anzitutto, da un caso, *Texas Department Of Housing And Community Affairs et Al. v. Inclusive Communities Project, Inc., et Al*., deciso dalla Corte Suprema degli Stati Uniti che, nel 2015, ha riconosciuto che il pregiudizio implicito può porsi alla base di una discriminazione, in cui la Corte Suprema rilevava che: «[r]ecognition of disparate-impact liability under the FHA plays an important role in uncovering discriminatory intent: it permits plaintiffs to counteract unconscious prejudices and disguised animus that escape easy classification as disparate treatment». In senso analogo, anche il Comitato europeo dei diritti sociali, CEDS, *International Centre for the Legal Protection of Human Rights (INTERIGHTS) v. Croazia*, decisione nel merito, n. 45/2007, 10 agosto 2009, § 48; e la Corte europea dei diritti dell'uomo *Abdu c. Bulgaria*, n. 26827/08, 11 marzo 2014.

[62] In tema, con riferimento al pregiudizio implicito di tipo razziale, J. KANG, *Trojan Horses of Race*, in *Harvard Law Review*, 2005, 1489 ss., che, a sua volta, richiama C.R. LAWRENCE III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, in *Stanford Law Review*, 1987, 317 ss.

[63] La letteratura più attenta rileva che non vi è unanimità quanto alla *ratio* sottesa alla teoria del *disparate impact*. Se, cioè, essa sia preordinata a fronteggiare unicamente casi in cui l'intento discriminatorio è difficile da dimostrare oppure anche quelli in cui l'intenzionalità è del tutto assente. Il tema è se l'indifferenza nei confronti della dimensione soggettiva per occuparsi dei soli effetti che derivano dalla condotta debba essere ritenuto (implicitamente) sinonimo di assenza, sempre, di intenzionalità, oppure solo espressivo di una volontà legislativa che mira a sanzionare effetti discriminatori quando non si possa conoscere le ragioni che hanno influenzato il soggetto agente. In questo senso, T.B. GILLIS, J.T. SPIESS, *Big data and discrimination*, in *The University of Chicago Law Review*, 2019, 459.

Una prima ragione risiede nella circostanza che la discriminazione derivante dal ricorso all'intelligenza artificiale scaturisce spesso da differenze che poggiano, sebbene non esplicitamente su fattori di discriminazione classici, su elementi tuttavia predittivi dell'appartenenza a categorie protette così da rendere non pienamente integrato il primo elemento costitutivo della discriminazione indiretta che la vuole conseguenza di una norma apparentemente neutra. È vero che, anche nel caso della discriminazione indiretta l'elemento di distinzione, *facially neutral*, risulta ciò nondimeno predittivo dell'appartenenza alla categoria protetta. E, tuttavia, se si accede alla ricostruzione che ritiene sussistere la discriminazione indiretta nelle sole ipotesi in cui l'intento discriminatorio non è difficile da dimostrare, ma è del tutto assente (al netto dei casi di *implicit bias*)[64], la correlazione tra fattore di discriminazione ed elemento impiegato per differenziare, nel senso che il secondo viene utilmente impiegato per escludere chi appartenga alla categoria protetta, determinerà l'impossibilità di qualificare la fattispecie quale discriminazione indiretta.

Un secondo aspetto riguarda, poi, la costruzione del modello.

La discriminazione indiretta si definisce tale sulla base del pregiudizio proporzionalmente peggiore sofferto da una categoria rispetto ad un'altra valutabile in base ai soli effetti della condotta. Nel contesto della discriminazione algoritmica, il problema è che il modello spesso non costituisce una rappresentazione adeguata della realtà, di cui omette di riprodurre fedelmente i rapporti, cioè le proporzioni tra categorie nelle loro effettive consistenze. Ciò comporta che all'interno dei *data-sets* vi siano categorie sotto-rappresentate oppure sovra-rappresentate. In entrambi i casi, la non neutralità che connota i dati su cui si fonda il funzionamento della macchina rende difficile provare la ricorrenza di una discriminazione indiretta in termini di valutazione comparativa degli effetti, perché il paramento di riferimento – cioè il *data-set* – non è in grado di riprodurre la eterogeneità fenomenica esterna alla macchina, cioè quella reale. Dimostrare il *disparate impact* diventa, quindi, difficoltoso non potendo un ipotetico ricorrente utilizzare dati statistici, perché questi ultimi si fonderebbero sulla dimensione "umana" reale e non su quella assunta quale parametro di riferimento dalla macchina sulla base dei dati che le sono stati forniti.

Una terza criticità si scorge, infine, in relazione al *tertium comparationis*[65].

Il funzionamento oscuro della macchina oppure l'impossibilità di conoscere o di avere accesso ai dati può, infatti, impedire l'individuazione del *comparator*, ostacolando il giudizio comparativo ai fini della dimostrazione del *disparate impact*. Una soluzione potrebbe riposare sul ricorso alla figura del *tertium comparationis* ipotetico, che si ritiene, però, poco soddisfacente non consentendo, a motivo della intangibilità della

[64] Si tratta della lettura che si ritiene meglio rispondente alla categoria della discriminazione indiretta, che viceversa, si ridurrebbe ad una sorta di discriminazione diretta per associazione priva di una propria autonomia dogmatica.

[65] Su cui si veda E. LUNDBERG, *Automated decision- making vs indirect discrimination. Solution or aggravation?*, in https://www.diva-portal.org/smash/get/diva2:1331907/FULLTEXT01.pdf.

discriminazione algoritmica[66], di ovviare al problema della fedele rappresentazione della realtà all'interno dei dati di cui si è detto.

3.4. *Ai confini della discriminazione per associazione: la* "proxy discrimination"

L'eliminazione di un elemento di appartenenza individuale ad una categoria protetta, identificabile in base ad un fattore di discriminazione tradizionale, non è sempre criterio sufficiente per assicurare condotte scevre da effetti di tipo discriminatorio. Detto altrimenti, l'esclusione di riferimenti espliciti alla categoria protetta e, dunque, al fattore di discriminazione, non è garanzia di un funzionamento della macchina conforme al principio di eguaglianza.

Si inserisce in questa fase della trattazione la fattispecie della discriminazione per associazione, che costituisce una forma particolare del fenomeno discriminatorio, a cui deve assegnarsi sicuro rilievo in ragione delle caratteristiche della discriminazione algoritmica[67] che fa di frequente perno su elementi di affiliazione individuale solo indirettamente predittivi, cioè associati o associabili, ai fattori di discriminazione tradizionali.

La discriminazione per associazione o *proxy discrimination*[68] assume quindi una posizione di primo piano nella prospettiva dell'indagine.

Il problema nasce, infatti, dalla presenza nei *data-sets* di "redundant encodings", cioè casi in cui l'appartenenza alla categoria protetta risulta codificata in altri dati, che risultano però associati alla medesima categoria protetta[69]. Si osserva inoltre che, a fronte di sistemi di intelligenza artificiale progressivamente più autonomi, la *proxy discrimination* rappresenta la principale sfida al diritto anti-discriminatorio tradizionale che, viceversa, mira a prevenire e contrastare disparità di trattamento fondate su qualità individuali immediatamente predittive dell'appartenenza a gruppi protetti[70].

La centralità da assegnare alla *proxy discrimination* si motiva in ragione dei meccanismi di associazione (*correlations*) istituiti tra i dati forniti alla macchina e la

---

[66] Sulle peculiarità della intangibilità della discriminazione algoritmica, insistono B. D. MITTELSTADT, P. ALLO, M. TADDEO, S. WACHTER, L. FLORIDI, *The ethics of algorithms: Mapping the debate*, in *Big Data & Society*, 2016, 1 ss.

[67] In questo stesso senso, tra gli altri, S. WATCHER, *Affinity Profiling and Discrimination by Association*, in *Berkeley Technology Law Journal*, 2020.

[68] In letteratura, anche per una spiegazione di come si sviluppano le correlazioni tra variabili all'interno dei *data-sets*, B.A. WILLIAMS, C.F. BROOKS, Y. SHMARGAD, *How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications*, in *Journal of Information Policy*, 2018, 78 ss.;, A. DATTA ET AL., *Proxy Discrimination in Data-Driven Systems: Theory and Experiments with Machine Learnt Programs*, 2017, in https://arxiv. org/pdf/1707.08120.pdf.

[69] Così S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit., 691; F.Z. BURGESIUS, *Discrimination, artificial intelligence, and algorithmic decision-making*, cit., 13.

[70] In questo senso, A.E.R. PRINCE, D. SCHWARCZ, *Proxy discrimination in the age of artificial intelligence and big data*, in *Iowa Law Review*, 2020, 1264, che inoltre insistono sull'opportunità, qui condivisa, di mantenere la fattispecie discriminatoria in esame distinta dalla nozione di discriminazione indiretta. In questo stesso senso, anche, J. GRIMMELMANN, D. WESTREICH, *Incomprehensible Discrimination*, in *California Law Review*, 2017, 164 ss. Nell'ambito della letteratura nazionale, P. ZUDDAS, *Intelligenza artificiale e discriminazioni*, cit., 4 ss.

caratteristica che il sistema ricerca (la *target variable*): maggiore è la mole di dati di cui si "nutre" la macchina, maggiori saranno i *proxies* disponibili per identificare caratteristiche predittive dell'appartenenza alle categorie protette[71].

La *proxy discrimination*, in definitiva, rappresenta la fattispecie, che meglio si attaglia a intercettare i tratti ontologici della *AI-derived discrimination*. Proprio perché le tecniche di intelligenza artificiale funzionano sulla base delle associazioni tra dati, è inevitabile che la macchina selezioni quegli elementi che meglio consentono di raggiungere il risultato desiderato poggiando su fattori che possono risultare, direttamente ovvero indirettamente, predittivi di un'affiliazione ad una categoria protetta[72].

Possono, poi, delinearsi due tipologie di *proxy discrimination*: la *proxy discrimination* intenzionale (o causale), cioè diretta, e quella non-intenzionale, ossia indiretta.

Con *proxy discrimination* intenazionale[73], si suole riferirsi a disparità di trattamento che poggiano su un elemento che, ancorché non riconducibile ad un fattore di discriminazione tradizionale, risulta ciò nonostante predittivo dell'affiliazione dell'individuo alla categoria protetta. Si definisce diretta, perché la disparità di trattamento risulta direttamente e causalmente ancorata ad un elemento sintomatico dell'appartenenza individuale al gruppo protetto e perché l'agente è consapevole della istituenda associazione tra elemento di distinzione e fattore di discriminazione e se ne serve volutamente allo scopo di discriminare[74]. «The proxy – si dice – is a mean to an end»[75]: è lo strumento, cioè, per identificare una classe di persone contraddistinte da un elemento che non rientra nel novero dei fattori di discriminazione tradizionali. Ciò rende la *proxy discrimination* diretta una tipologia particolarmente insidiosa di discriminazione, perché ricorre in tutte le ipotesi in cui la categoria che si vuole colpire non è immediatamente individuabile perché "coperta" da un elemento di affiliazione individuale diverso e fuorviante[76].

La *proxy discrimination* diretta ricorre qualora siano rintracciabili nel *data-set* dei nessi causali tra la caratteristica protetta (e vietata) e uno o più fattori direttamente predittivi della prima. Si tratta di eventualità che, per inciso, sussiste e si ritiene permanga anche nell'ambito di *data-sets* che non contengono riferimenti espliciti a fattori di

---

[71] A.E.R. PRINCE, D. SCHWARCZ, *Proxy discrimination in the age of artificial intelligence and big data*, cit., 1275, che, a loro volta, richiamano S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit., 695.

[72] Così A.E.R. PRINCE, D. SCHWARCZ, *Proxy discrimination in the age of artificial intelligence and big data*, cit. Gli AA. rilevano che: «AIs use training data to discover on their own what characteristics can be used to predict the target variable. Although this process completely ignores causation, it results in AIs inevitably 'seeking out' proxies for directly predictive characteristics when data on these characteristics is not made available to the AI due to legal prohibitions», cit. 1264 e 175 e ss.

[73] In tema L. ALEXANDER, K. COLE, *Discrimination by Proxy*, in *Constitutional Commentary*, 1997, 453 ss.

[74] Ci si riferisce al fenomeno del *redlining*, forma specifica di *proxy discrimination* che veniva impiegata dagli istituti finanziari statunitensi per evitare di dover servire zone ad alta densità abitativa di afro-americani. Non potendo assumere la razza ad elemento in base al quale discriminare, gli istituti finanziari utilizzavano come *proxy* le aree geografiche, discriminando la minoranza afro-americana pur non facendo esplicitamente riferimento al fattore di discriminazione vietato.

[75] D. HELLMANN, *Two types of discrimination. The familiar and the forgotten*, in *California Law Review*, 1998, 318.

[76] *Ibidem.*

discriminazione vietati, come il genere oppure la razza. Le tecniche di IA sono, infatti, in grado di realizzare, come detto, associazioni suscettibili di produrre effetti discriminatori ai danni di categorie protette anche sulla base dei collegamenti casuali istituiti tra i dati estenti, con la conseguenza che la mera eliminazione di riferimenti espressi ai fattori di discriminazione tradizionali può non essere sufficiente ad assicurare il funzionamento non discriminatorio della macchina[77].

La dottrina ha, inoltre, ulteriormente precisato la tipologia discriminatoria in esame, distinguendo tra la *proxy discrimination* diretta di tipo causale (*causal proxy discrimination*) e la *proxy discrimination* "opaca" (*opaque proxy discrimination*)[78], ricorrendo la prima quando sia rintracciabile un collegamento di tipo causale e diretto tra il *proxy* e la caratteristica protetta[79]; la seconda, laddove l'elemento predittivo dell'appartenenza alla categoria protetta non sia quantificabile oppure risulti di difficile individuazione all'interno del *data-set*[80] .

In entrambi i casi, però, ciò che rileva è la sussistenza di una relazione causale di tipo diretto tra la variabile utilizzata dalla macchina e l'appartenenza alla categoria protetta: in un caso, essa è spiegabile; nell'altro, si appalesa, invece, sfumata perché l'elemento del *data-set* che "aggancia" la categoria protetta non è sempre noto ma se ne rivela tuttavia direttamente predittivo.

Fattispecie discriminatoria ancora diversa è la *proxy discrimination* indiretta, in cui il *proxy* e il fattore di discriminazione non risultano direttamente correlati, nel senso che il primo non è immediatamente – o, almeno, non apparentemente – predittivo del secondo.

Può, cioè, accadere che la macchina istituisca un collegamento tra una variabile e un dato di cui dispone e che l'associazione che ne risulta, nonostante formulata in termini apparentemente neutri, finisca però con l'identificare una categoria protetta sulla base di un elemento, il fattore di discriminazione, non presente nel *data-set*.

---

[77] Sulla inutilità di strategie di regolamentazione che prevedano l'eliminazione dei dati sensibili dai *data-sets* allo scopo di assicurarne il funzionamento non discriminatorio, la dottrina è prevalentemente coesa. Tra i molti, P.T. KIM, *Data-Driven Discrimination at Work*, in *William & Mary Law Review*, 2017, 857 ss.; A.E.R. PRINCE, D. SCHWARCZ, *Proxy discrimination in the age of artificial intelligence and big data*, cit.; S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit.

[78] Così A.E.R. PRINCE, D. SCHWARCZ, *Proxy discrimination in the age of artificial intelligence and big data*, cit., 1277 ss.; J. GERARDS, *Algorithmic discrimination in Europe: Challenges and Opportunities for EU equality law*, consultabile al link: https://www.europeanfutures.ed.ac.uk/algorithmic-discrimination-in-europe-challenges-and-opportunities-for-eu-equality-law/.

[79] Un esempio classico di *proxy* direttamente e causalmente predittivo dell'appartenenza ad una categoria protetta e, dunque, di un fattore di discriminazione vietato è il codice postale per la sua associazione con la razza e l'etnia.

[80] La proposta ricostruttiva è di A.E.R. PRINCE, D. SCHWARCZ, *Proxy discrimination in the age of artificial intelligence and big data*, cit., 1278, che, come esempio di *proxy discrimination* diretta di tipo "opaco" richiamano le discriminazioni realizzate ai danni delle donne da parte delle compagnie assicurative. Gli Autori illustrano l'esempio come di seguito: «[s]ex (Y) is predictive of auto insurance claims (Z) in part because young girls tend to drive more safely than young boys.75 Of course, it is possible to obtain more direct information about care levels (A). But such data is not widely available, as driver 'care' is difficult to quantify. For this reason, simply banning the use of sex-based discrimination will predictably lead to proxy discrimination by AIs because sex is directly predictive of care levels in ways that are not mediated through any alternative, presently quantifiable, variables». In tema, anche, T.B. GILLIS, J.T. SPIESS, *Big data and discrimination*, cit., 462.

Quale esempio[81] può portarsi il caso di un algoritmo impiegato per selezionare possibili candidati a ricoprire una posizione lavorativa per la quale l'altezza è requisito necessario e, tuttavia, la macchina non dispone di dati sull'altezza dei candidati e delle candidate, in ragione della appurata correlazione tra altezza e sesso[82]. Il sesso non potrà essere impiegato quale *proxy* e non viene inserito nel *data-set* trattandosi di fattore di discriminazione vietato oltreche dato sensibile. Potrebbe allora accadere che la macchina, al fine di raggiungere il proprio obiettivo (la selezione del/lla candidato/a migliore), utilizzi un *proxy* diverso, come le preferenze di serie tv. In questo caso, il ricorso ad un elemento apparentemente neutrale e non direttamente predittivo del fattore di discriminazione vietato (il sesso) può tuttavia produrre un effetto proporzionalmente deteriore ai danni delle donne, che saranno escluse dalla selezione.

Il riferimento è al fenomeno del *omitted variable bias*, in forza del quale l'algoritmo, ancorché non faccia riferimento a dati sensibili esclusi e indisponibili nel *data-set*, è comunque in grado di pervenire ad un risultato discriminatorio in ragione delle associazioni istituite a partire da altri dati che sono indirettamente predittivi dell'appartenenza a categorie protette[83] anche solo in modo apparente.

Trattandosi di una forma di discriminazione che si avvicina alla discriminazione statistica[84], l'incidenza della *proxy discrimination* indiretta potrebbe contenersi laddove venissero forniti alla macchina dati più corretti volti a renderne più sofisticato il funzionamento[85].

Oltre le classificazioni, resta da rilevare che la discriminazione "artificiale" si atteggia spesso quale prodotto di condotte plurali che si sommano e si intrecciano a formare una spirale in continuo movimento, con una discriminazione risultante all'esito dal funzionamento della macchina che costituisce il prodotto dell'azione correlata di più *proxy discriminations* che agiscono simultaneamente[86].

---

[81] L'esempio è tratto da A.E.R. PRINCE, D. SCHWARCZ, *Proxy discrimination in the age of artificial intelligence and big data*, cit., 1280.

[82] Qualora il *data-set* disponesse di dati sull'altezza dei candidati e delle candidate, quest'ultima costituirebbe un *proxy* per il sesso e si verserebbe in un'ipotesi di *proxy discrimination* di tipo diretto e causale.

[83] Per un approfondimento anche in relazione alla nozione di discriminazione statistica, si rinvia a B.A. WILLIAMS, C.F. BROOKS, Y. SHMARGAD, *How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications*, cit., 90 e ss.

[84] Gli studi della teoria economica distinguono tra la discriminazione di tipo statistico e la *taste-based discrimination*. Per un approfondimento, tra i molti, G. BECKER, *The Economics of Discrimination*, Chicago, 1971; E.S. PHELPS, *The Statistical Theory of Racism and Sexism*, in *The American Economic Review*, 1972, 659 ss.; J. GURYAN, K.K.CHARLES, *Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots*, in *The Economic Journal*, 2013, 471 ss.; T. CALDERS, I. ŽLIOBAITĖ, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, in *Discrimination and Privacy in the Information Society*, 2013, 43 ss.

[85] A.E.R. PRINCE, D. SCHWARCZ, *Proxy discrimination in the age of artificial intelligence and big data*, 1280-1281.

[86] *Ibidem*, cit., 1282, secondo cui «a suspect classifier (age) may proxy for a facially neutral category (years since graduation) which proxies for some unquantifiable data (comfort with learning new technology), which predicts a desired outcome. Alternatively, an AI may proxy for one suspect classifier, which proxies for another suspect classifier, which proxies for a facially neutral characteristic that is casually linked to the target variable».

4. *Discriminazioni strutturali e intelligenza artificiale*

Tipologie discriminatorie, che poggiano sulle relazioni tra fenomeno discriminatorio, società e potere pubblico, sono poi la discriminazione istituzionale e quella strutturale.

La ragione per cui tali fenomeni mantengono una importanza che ne giustifica la trattazione in questa sede si deve al legame inscindibile di cui si è detto tra la persona e la macchina, che rendere la seconda riflesso di diseguaglianze che fondano le dinamiche relazionali inter-gruppi.

È la persona, immersa in una realtà non scevra di pregiudizi, a fornire i dati alla macchina e, pure a fronte di tecniche che palesano un livello di progressiva autonomia dalla decisione o programmazione originaria, esse rimangono pur sempre il prodotto di un'azione umana, non sempre, non necessariamente, imparziale. Si tratta di tema che risulta bene esemplificato dal fenomeno già indagato del *masking* e, più efficacemente, dall'ipotesi più problematica del *bias* implicito. Ancora, si possono richiamare quegli studi che dimostrano come le tecniche di intelligenza artificiale acuiscano i livelli di povertà e le diseguaglianze sociali[87], anziché contenerle.

Le contiguità tra la discriminazione "artificiale" e quelle istituzionale e strutturale potrà essere apprezzata in modo particolare se si considerano le vittime della *AI-derived discrimination*. Come si avrà modo di apprezzare, l'intelligenza artificiale non è soltanto foriera del rischio di generare discriminazioni "nuove", bensì, e soprattutto, di esacerbarne di esistenti rendendone però sfumata la caratterizzazione discriminatoria e pertanto più complesso l'accertamento e la repressione. Si versa, così, nell'ipotesi del c.d. «algoritmo strutturalmente incostituzionale»[88], *reo* dell'«errore di derivare dall'essere (in questo caso dall'essere della realtà sociale, spesso ingiusta, parziale o distorta) il dover essere»[89].

5. *Una prima delimitazione teorica della* AI-derived discrimination

In un articolo pubblicato sul *Financial Times*, «What separates humans from AI? It's doubt»[90], l'autore si occupa di tracciare una linea di confine tra il pensiero e l'*agere*

[87] Sul rapporto tra intelligenza artificiale e povertà, si rinvia diffusamente a M. MADDEN, M. GILMAN, K. LEVY, A. MARWICK, *Privacy, Poverty and Big Data: A Matrix of Vulnerabilities for Poor American*s, in *Washington University Law Review*, 2017, 54 ss.; M. GILMAN, *Poverty Lawgorithms. A Poverty Lawyer's Guide To Fighting Automated Decision-Making Harms On Low-Income Communities*, consultabile al link: https://datasociety.net/wp-content/uploads/2020/09/Poverty-Lawgorithms-20200915.pdf.

[88] A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, cit., 86.

[89] *Ibidem.*

[90] Il riferimento è a S.M. FLEMING, *What separates humans from AI? It's doubt*, *Financial Times*, 26 aprile 2021. Si accostano a questa linea argomentativa altri autori che si sono soffermati sulla eterogeneità comportamentale e cognitiva tra la persona e la macchina. Così A. ROUVROY, *The end(s) of critique: Data-behaviourism vs. due-process*, in M. HILDEBRANDT, K. DE VRIES (a cura di), *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, Milton Park and New York, 2013, 143 ss.

umano e quello della macchina. Si sostiene, in particolare, che ciò che distingue il primo dalla seconda sarebbe la meta-cognizione, cioè la capacità, propria della persona ma non della macchina, di sapere ciò che non si conosce, di avere consapevolezza della eventualità di poter incorrere in un errore e di riconoscerne uno come tale. Ancora, che di fronte a qualcosa che non conosce, la macchina, a differenza della persona, non si rende conto che *non* conosce e opera una scelta sulla base di qualcosa che *non* conosce e che produce una conseguenza decisiva: la soluzione prescelta non si esaurisce nell'effetto limitato al caso singolo, ma rinforza e influenza il funzionamento successivo, perché dalla scelta sbagliata la macchina trae dati ulteriori, sebbene fallaci, che guideranno le sue scelte e capacità di discernimento future.

La meta-cognizione consente, quindi, di distinguere "bene" e "male", il "bene" dal "male" e, in definitiva, di non decidere sulla base di dati mancanti oppure parziali. La macchina, al contrario, decide sempre e *in ogni caso*, traendo indicazioni che diventano vincolanti. Il dato che manca, perché sconosciuto, non esclude la decisione della macchina, ma la fonda e costituisce il presupposto logico di decisioni future.

La differenza tra la persona e la macchina non poteva che riflettersi su un fenomeno così intrinsecamente legato all'"umano" come la discriminazione.

L'analisi che precede conferma questa differenza tra la persona e la macchina e, calata nella prospettiva del fenomeno discriminatorio, consente di stabilire alcuni punti fermi.

Il primo è costituito dalla tipicità della discriminazione legata all'intelligenza artificiale. A prescindere dalla tipologia di tecnica, lo scollamento tra la condotta e l'evento, dovuto alla interposizione della macchina, è sufficiente a incidere sulla fenomenologia discriminatoria in esame.

Se allora la discriminazione applicata all'intelligenza artificiale è *species* diversa, ad essa – secondo apporto – difficilmente si attagliano le categorie classiche del diritto anti-discriminatorio incardinate sulle teorie del *disparate treatment* e del *disparate impact*[91].

Piuttosto – terzo esito –, più opportuno è isolare la fattispecie in esame, su cui concorda la letteratura, che parla di *AI-derived discrimination* oppure di discriminazione algoritmica ad enfatizzare il reciso legame con la discriminazione "umana". L'assenza del legame, in termini di *agency*, tra condotta umana e discriminazione, identifica, inoltre, una caratteristica aggiuntiva della tipologia discriminatoria in esame, cioè il suo essere non più intuitiva[92] e, in ultima analisi, prevedibile, dal legislatore, e sanzionabile, dal giudice[93].

---

[91] Guardando al sistema di diritto anti-discriminatorio dell'Unione Europea, in questo senso M. LEES, *The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union*, in *Security Dialogue*, 2014, 494 ss.. Così, anche, il CAHAI, l'*Ad hoc Committee on Artificial Intelligence*, nel suo *Feasibility study on a legal framework on AI design, development and application based on CoE standards*, 17 dicembre 2020.

[92] Insiste su questa peculiarità della discriminazione algoritmica S. WACHTER, *Affinity Profiling and Discrimination by Association in Online Behavioural Advertising*, in *Berkeley Technology Law Journal*, 2020, 1 ss.

[93] Su cui, *infra*, Parte Seconda.

Quarto aspetto centrale è l'inquadramento della discriminazione "artificiale" entro la figura della "*proxy discrimination*".

Altro aspetto peculiare, il quinto, riguarda il *bias*[94]. Ci si riferisce al "luogo" del *bias* cioè alla sua individuazione in seno al complesso e frazionato funzionamento della macchina, allo scopo di verificare se e in che misura la discriminazione costituisca il prodotto (solo) della condotta umana, della macchina oppure di entrambi; al "tempo" del *bias*, cioè la sua collocazione temporale in rapporto alle fasi che preludono alla costruzione del modello e al suo funzionamento; alla sua qualificazione in rapporto al concetto di intenzionalità della condotta. Qui, ritornano le contaminazioni del diritto anti-discriminazione classico, perché è sulla sussistenza o meno della intenzionalità che poggia, anche per la *AI-derived discrimination*, la possibilità di distinguerne una versione diretta, secondo cui il *bias* deve essere conosciuto dall'agente e intenzionalmente preordinato a produrre un effetto discriminatorio, da una indiretta, in cui, all'opposto, il *bias* non è intenzionale o, almeno, non lo è apparentemente, ma l'effetto dell'azione è discriminatorio.

Ultimo portato dall'analisi dei profili oggettivi riguarda l'attitudine della discriminazione algoritmica a risolversi nella simultanea e cumulatività di plurime discriminazioni. Si realizza, cioè, uno "spostamento" della nozione di discriminazione intersezionale dal lato soggettivo, quello, dei fattori di discriminazione, a quello oggettivo della condotta e della sua attitudine a generare una pluralità di *proxy discriminations* che si intersecano. Non una condotta che poggia su due o più fattori, come nella discriminazione multipla e intersezionale; non una pluralità di condotte che, nel loro insieme, qualificano la discriminazione, come nella molestia nella discriminazione su base salariale, bensì una condotta discriminatoria iniziale su cui se ne innestano di successive con caratteristiche proprie tra di loro eterogenee.

6. *Le "vittime": verso nuove identità e appartenenze*

La prospettiva soggettiva segue quella oggettiva e similmente alla prima attesta la specificità della discriminazione "artificiale".

Quanto alle vittime, essa si allontana solo in parte da quella classica, accostando a gruppi tradizionalmente offesi da condotte discriminatorie aggregazioni nuove generate dal *proxy*.

Trattando della "nuova" discriminazione algoritmica, si è più volte fatto riferimento al *proxy*, al suo atteggiarsi da fattore di discriminazione, alla incisività solo apparentemente sfumata e quasi residuale dei tradizionali elementi di divisione tra esseri umani. Il fatto che il diritto anti-discriminatorio proibisca il riferimento espresso ai fattori di discriminazione tradizionali ha prodotto così un mutamento nella selezione delle

[94] Per una categorizzazione del pregiudizio (*bias)*, S. QUINTARELLI, F. COREA, F. FOSSA, A. LOREGGIA, S. SAPIENZA, *AI: profili etici. Una prospettiva etica sull'intelligenza artificiale: principi, diritti e raccomandazioni*, in *BioLaw Journal*, 2019, 218 e ss., e, anche, H. SURESH, J.V. GUTTAG, *A Framework for Understanding Unintended Consequences of Machine Learning*, Cambridge, 2020.

qualità individuali e nelle modalità entro cui si definisce l'appartenenza identitaria del singolo al gruppo. Essa non si fonda più, almeno esplicitamente, su elementi come la razza, l'etnia, il sesso oppure il genere, bensì su criteri che vanno da fattori quantificabili e misurabili, il quartiere di residenza, il codice postale, le preferenze di siti internet, film, serie televisive, ad altri dai contorni discrezionali, l'affidabilità del creditore[95], il "buon" lavoratore, l'automobilista a basso rischio di sinistri stradali.

Il richiamo ad elementi che, come i fattori di discriminazione classici, concorrono a qualificare l'identità del singolo e a determinarne l'appartenenza collettiva, contribuisce così alla emersione di gruppi "nuovi"[96], che condividono però con i "vecchi" la posizione subordinata all'interno della società.

La discriminazione algoritmica fa, quindi, nascere minoranze "nuove"[97], che poggiano su criteri di appartenenza individuale e collettiva diversi, che costruiscono nuove forme identitarie[98]. Allo stesso tempo, se è vero che la discriminazione algoritmica ha portato ad una ridefinizione del concetto di affiliazione individuale, di gruppo, di identità[99], è altrettanto vero che le sue vittime sono tutt'affatto nuove; o meglio, lo sono solo in parte.

Potrebbe dirsi, cioè, che a fattori e gruppi nuovi non corrispondono necessariamente vittime nuove. E, ancora, che si assiste ad una complessa intersezione tra forme di appartenenza classica e tipologie "nuove", che non rimangono separate ma si sovrappongono creando inedite relazioni e intrecci tra fattori di discriminazione tradizionali e criteri di differenziazione nuovi.

Lo studio dell'impatto soggettivo delle torsioni discriminatorie nell'impiego delle tecniche di intelligenza artificiale mostra pertanto una convergenza fra le vittime della discriminazione "umana" e quelle della discriminazione algoritmica, che però non sorprende fino in fondo in considerazione dell'ineliminabile ruolo assolto dalla persona nella progettazione della macchina. Quello che cambia, quindi, più che le vittime sono i

---

[95] In prospettiva comparata, si richiama l'esperienza statunitense ove l'utilizzo di algoritmi cui subordinare la concessione oppure il diniego del prestito da parte di un istituto di credito, si veda quanto riportato da *The Atlantic*, K. WADDELL, *How Algorithms Can Bring Down Minorities' Credit Scores (Analyzing people's social connections may lead to a new way of discriminating against them)*, 2 dicembre 2016, consultabile al link: https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/.

[96] Sulla eterogeneità, o meglio, sul superamento dei *protected groups* per effetto della *AI-derived discrimination*, convergono, tra gli altri, B. MITTELSTADT, *From Individual to Group Privacy in Big Data Analytics*, in *Philosophy and Technology*, 2017, 475 ss.; S. WACHTER, *Affinity Profiling and Discrimination by Association in Online Behavioural Advertising*, in *Berkeley Technology Law Journal*, 2020.

[97] Le minoranze "nuove" a cui si fa riferimento, in questa sede, sono gruppi minoritari distinti rispetto a quelle che la dottrina tradizionalmente qualifica come "nuove minoranze", ossia gruppi sociali che si caratterizzano per elementi che in parte si allontanano dalla definizione classica di minoranza proposta da F. CAPOTORTI, per il fatto di essere composte da individui che non sono cittadini dello Stato nel quale risiedono.

[98] Sul ruolo del *proxy*, *infra*.

[99] Così K. DE VRIES, *Identity, profiling algorithms and a world of ambient intelligence*, in *Ethics Inf Technol*. 2010, 71 ss., che si sofferma su una ricostruzione del concetto di identità che, inquadrata nella prospettiva della tecnologia, assume una importanza decisiva, consentendo di distinguere «a device used to decide who is in and who is out; who is us and who is them; who is likely to be a good customer and who is not; who is allowed to pass the border and who is not», 76.

criteri in base ai quali si delineano le categorie e che giustificano una disamina del fenomeno, che incrocia vecchie e nuove minoranze.

La trasformazione soggettiva non si arresta, infine, alla enucleazione delle vittime e dei rispettivi tratti identitari, riflettendosi sulla identificazione del *tertium comparationis*, anch'esso esposto ad una ridefinizione teorica ed applicativa.

7. *Le "vecchie" minoranze tra genere, razza e etnia*

Prodotto della porzione "umana" della discriminazione "artificiale" è la sovrapponibilità delle "vecchie" minoranze alle vittime della nuova fenomenologia discriminatoria.

Le relazioni che le nuove tecnologie intrattengono con il genere[100] e con il fattore etnico-razziale hanno velocemente conquistato la scena negli anni in cui l'intelligenza artificiale ha invaso i più vasti settori della vita quotidiana.

Con riferimento alle prime, che l'intelligenza artificiale non rispecchi egualmente entrambi i sessi e, anzi, si traduca in una discriminazione strutturale ai danni delle donne è portato ormai pacifico, supportato da analisi statistiche[101] e da progetti accademici[102]

[100] Nell'ambito della dottrina costituzionalistica nazionale, si vedano, diffusamente, M. D'AMICO, *Una parità ambigua. Costituzione e diritti delle donne*, cit., 313 ss., che parla di «intelligenza artificiale 'contro' le donne»; l'A. pone l'accento, tra gli altri aspetti, sul tema del «monopolio maschile», che contraddistingue il settore dell'informatica, composto in modo prevalente da uomini, legandolo agli effetti discriminatori prodotti dagli algoritmi che deriverebbero dalla difettosa *diversity* dal lato di chi «costruisce gli algoritmi», in particolare 314 e ss.; della stessa A., sul tema degli effetti che l'assenza di donne produce non solo piano della non discriminazione, ma, più in generale, sulla qualità e sul buon funzionamento di organi e istituzioni, si rinvia, tra i molti, *a Il difficile cammino della democrazia paritaria*, Torino, 2011; sempre nell'ambito della letteratura nazionale, si rinvia a E. STRADELLA, *Stereotipi e discriminazioni: dall'intelligenza umana all'intelligenza artificiale*, in AA.VV., Liber amicorum *per Pasquale Costanzo*, in *Consulta Online*, 2020, 1 ss. In una prospettiva sovranazionale, spunti sono offerti dai contributi resi in occasione della conferenza *Data Justice Conference*, svolta a Cardiff nel 2018, con particolare riferimento alle relazioni rese nell'ambito del panel *Data and Discrimination*, le cui registrazioni possono essere consultate al link: https://cardiff.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=d132281d-8bbc-4980-8013-a8e8007c788d. In tema, si vedano, anche, C. D'IGNAZIO, L.F. KLEIN, *Data Feminism*, Cambridge, 2019; S. DILLON, C. COLLETT, *AI and Gender: Four Proposals for Future Research*, 2019.

[101] In questo senso, i dati del *World Economic Forum* del 2018 (https://reports.weforum.org/global-gender-gap-report-2018/assessing-gender-gaps-in-artificial-intelligence/?doing_wp_cron=1621003660.5886778831481933593750) e del 2021, dell'UNESCO nel 2020 (https://unesdoc.unesco.org/in/documentViewer.xhtml?v=2.1.196&id=p::usmarcdef_0000374174&file=/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_ab07646d-c784-4a4e-96a1-3be7855b6f76%3F_%3D374174eng.pdf&locale=en&multi=true&ark=/ark:/48223/pf0000374174/PDF/374174eng.pdf#AI%20Gender_pages.indd%3A.11061%3A142)

[102] Il riferimento è al già citato progetto *Gender Shades* sul carattere discriminatorio ai danni delle donne afro-americane di alcuni sistemi di riconoscimento facciale, su cui, anche, J BUOLAMWINI, *Hearing on: Artificial Intelligence: Societal and Ethical Implications*, Washington, DC, United States House Committee on Science, Space and Technology, 2019.

che convergono nell'attestare la natura non *gender-neutral* dei modelli[103], tra gli altri, di *machine learning* e *deep learning*[104].

Oltre ai dati e al *gender divide*[105] nelle discipline c.d. *stem*, le relazioni tra genere e intelligenza artificiale sollecitano ulteriori spunti.

Il primo attiene alla caratterizzazione non nuova della vittima, nel senso che a fronte di una fenomenologia discriminatoria "nuova" la vittima non ne ripete la novità, iscrivendosi le donne tra le categorie maggiormente esposte a subire limitazioni a sfondo discriminatorio nel godimento dei propri diritti.

Vi è poi un secondo elemento che, invece, separa la discriminazione "artificiale" da quella tradizionale e che deriva dal "come" si realizza la distinzione. Il sesso non compare, infatti, (quasi) mai nei modelli così come elementi predittivi dell'appartenenza all'uno o all'altro sesso. Il tema è, quindi, come la *proxy discrimination* arriva comunque a incidere negativamente sulla categoria protetta. Una domanda a cui si potrebbe rispondere richiamando la inadeguatezza degli istituti classici del diritto antidiscriminatorio, ma anche l'insufficienza di strumenti di contrasto che poggino sulla sola eliminazione o non selezione di *proxys* che richiamano, direttamente o indirettamente, il sesso. Piuttosto, l'effetto discriminatorio rivela l'influenza del pregiudizio implicito ma strutturale, che influenza la fase "umana" della programmazione della macchina e che si salda con la composizione prevalentemente mono-genere degli esperti di intelligenza artificiale[106].

---

[103] Alle ricerche di organismi sovranazionali, si affiancano alcuni casi emblematici che hanno avuto come protagonisti multinazionali quali Uber, con riferimento a sistemi di riconoscimento facciale, a LinkedIn, in relazione a sistemi di decisione automatizzata che pregiudicavano in modo proporzionalmente maggiore le donne in punto di visualizzazione delle offerte lavorative. Richiama questi esempi, inserendoli nel più ampio dibattito sulla portata discriminatoria in base al genere delle tecniche di intelligenza artificiale, M. D'AMICO, *Una parità ambigua. Costituzione e diritti delle donne*, cit., 315 e ss. In tema, anche, i contributi di G. DE MINICO e A. PAPA nell'ambito del convegno *La pandemia: nuove asimmetrie o uguaglianze di genere?*, organizzato da Astrid, in data 17 maggio 2021.

[104] Si ricordano il già citato progetto *Gender Shades* sul carattere discriminatorio ai danni delle donne, in particolare afro-americane di alcuni sistemi di riconoscimento facciale (su cui, anche, J BUOLAMWINI, *Hearing on: Artificial Intelligence: Societal and Ethical Implications*, Washington, DC, United States House Committee on Science, Space and Technology, 2019); la vicenda di Amazon in tema di reclutamento (su cui J. DUSTIN, *Amazon scraps secret AI recruiting tool that showed bias against women*, Reuters, 11 ottobre 2018, che per primo diede la notizia del funzionamento discriminatorio dell'algoritmo impiegato, e, anche, J. LAURET, *Amazon's sexist AI recruiting tool: how did it go so wrong?*, in *becominghuman.ai*, 16 agosto 2019); dei sistemi di assistenza vocale (su cui R. ADAMS, N.N. LOIDEAIN, *Addressing Indirect Discrimination and Gender Stereotypes in AI Virtual Personal Assistants: The Role of International Human Rights Law*, in *Annual Cambridge International Law Conference New Technologies: New Challenges for Democracy and International Law*, 2019, 1 ss.).

[105] Il *World Economic Forum* nel 2018 attesta che solo il 22% dei professionisti che si occupano di intelligenza artificiale a livello mondiale sono donne contro il 78% di uomini, con un gap che si avvicina al 72%.

[106] In tema, M. BRUSSEVICH, E. DABLA-NORRIS, S. KHALID, *Is Technology Widening the Gender Gap? Automation and the Future of Female Employment*, in *IMF Working Papers, Working Paper No. 19/91. Washington, DC, International Monetary Fund*, 2019. L'argomento è diffusamente approfondito anche dalla Commissione dell'Unione Europea in un recente rapporto, che può leggersi al link: https://ec.europa.eu/info/sites/default/files/aid_development_cooperation_fundamental_rights/mlp_summary_report_nl_november_2020_en.pdf.

Terzo elemento, su cui non ci si sofferma, investe, infine, i rapporti tra linguaggio non paritario e intelligenza artificiale, nella misura in cui si ritiene che il primo costituisca una delle cause del funzionamento non *gender-neutral* della seconda.

Altrettanto note storicamente sono le relazioni tra la discriminazione e il fattore etnico-razziale[107], così che non sorprende che il fattore etnico-razziale mantenga una posizione di primo piano, dimostrando le evidenze statistiche una imponente incidenza delle tecniche di intelligenza artificiale ai danni di minoranze etnico-razziali[108]. Si segnalano i purtroppo celebri sistemi di sorveglianza, di giustizia predittiva[109], sperimentati in particolare dall'esperienza statunitense, alle diseguaglianze nell'accesso al servizio sanitario[110], al mercato del lavoro tramite annunci pubblicitari[111], ai servizi

[107] La scelta di legare unitariamente i due fattori si motiva alla luce dell'opzione linguistica e concettuale prescelta dal diritto dell'Unione Europea, che poggia sulle criticità che tuttora circondano la nozione di razza così come sulla sua non agevole demarcazione semantica rispetto al contiguo concetto di etnia. Il riferimento è alla Direttiva 2000/43/CE che espressamente sceglie di unire i due fattori. Per un approfondimento sull'utilizzo della nozione di razza anche alla luce del dibattito sviluppatosi circa l'opportunità di conservarne il riferimento nel testo dell'art. 3, comma 1, Cost., si consenta il rinvio a C. NARDOCCI, *Dall'invenzione della razza alle leggi della vergogna: lo sguardo del diritto costituzionale*, in *Italian Review of Legal History*, 2019, 482 ss. Sul significato della nozione di razza, si rinvia diffusamente a L.L. CAVALLI SFORZA, *Storia e geografia dei geni umani*, Torino, 1997. Interessante in questo quadro anche il documento *The Race Question* dell'UNESCO del 1949, che può essere letto al link: http://unesdoc.unesco.org/images/0012/001282/128291eo.pdf.

[108] In tema, C. INTACHOMPHOO, O.D. GUNDERSEN, *Artificial Intelligence and Race: a Systematic Review*, in *Legal Information Management*, 2020, 74 ss.

[109] In letteratura, M. HAMILTON, *The biased algorithm evidence of disparate impact of Hispanics*, in *American Criminal Law Review*, 2019, 1553 ss., che richiama i dati pubblicati dallo studio realizzato da T. GEST, *Civil rights advocates say risk assessment may "worsen racial disparities" in bail decisions*, in *The Crime Report*, 31 luglio 2018; R.M. O'DONNELL, *Challenging Racist Predictive Policing Algorithms under the Equal Protection Clause*, in *New York University Law Review*, 2019, 545 ss., che riporta in apertura del suo lavoro il caso dell'algoritmo, noto come *Strategic Subject List* (SSL), quale strumento di polizia predittiva. Per un approfondimento sul caso citato dall'A., si veda J. GORNER, *Chicago Police Use 'Heat List' as Strategy to Prevent Violence*, 21 agosto 2013, consultabile al link: https://www.chicagotribune.com/news/ct-xpm-2013-08-21-ct-met-heat-list-20130821-story.html.

[110] Lo riporta H. LEIDFOR, *Millions of black people affected by racial bias in health-care algorithms*, in *Nature*, 2019, che richiama lo studio di Z. OBERMEYER, B. POWERS, C. VOGELI, S. MULLAINATHAN, *Dissecting racial bias in an algorithm used to manage the health of populations*, in *Science*, 2019, 447 ss.; in tema, anche, R. HART, *If you're not a white male, artificial intelligence's use in healthcare could be dangerous*, in *Quartz*, 10 luglio 2017, https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-bedangerous. Interessanti i dati riportati da Z. OBERMEYER et al., *Dissecting racial bias in an algorithm used to manage the health of populations*, in *Science*, 2019, 447 ss.

[111] Si rinvia allo studio di D.J. DALENBERG, *Preventing discrimination in the automated targeting of job Advertisements*, in *Computer Law & Security Review*, 2018, 615 ss., che analizza la questione dell'accesso al mercato del lavoro tramite annunci online.

abitativi[112], oppure, ancora, ai sistemi di riconoscimento facciale[113] diffusi anche nell'ambito delle politiche migratorie[114].

In senso analogo a quanto già rilevato, anche le discriminazioni algoritmiche a sfondo etnico-razziale sono da ricondurre a cause che agiscono "in entrata": la scarsa, se non assente, eterogeneità etnico-razziale, in termini di sovra-rappresentanza dei caucasici, che presiede alle fasi di ideazione, costruzione, sviluppo e messa in opera della macchina[115]; il pregiudizio implicito.

Più complesse le ipotesi in cui il fattore etnico-razziale è inciso "in uscita" per effetto della sua associazione, diretta oppure indiretta, con l'elemento, predittivo, cioè con il *proxy*. Tratto caratterizzante e che sfavorisce le minoranze etnico-razziali in queste forme di diseguaglianze "in uscita" è, infatti, lo stretto legame che la razza e l'etnia intrattengono con una vasta serie di fattori di discriminazione classici e "nuovi", con un aumento esponenziale delle probabilità di subire effetti pregiudizievoli per effetto del combinato operare di questa eterogenea moltitudine di elementi di difficile prevenzione e contenimento. Si pensi a sistemi algoritmici che poggiano sull'area geografica di residenza e che uniscono fattore etnico-razziale e condizione economica o grado di povertà; ai meccanismi che guardano alle statistiche di commissione di reati, di recidiva, sino alla controversa inclinazione alla delinquenza; alla presunta inaffidabilità alla restituzione di prestiti; a sistemi di ricerca e selezione dei candidati che utilizzano quale paramtro il nome proprio.

8. *"Nuove" minoranze e "nuovi" fattori di discriminazione? Il* proxy

Accanto alle minoranze vecchie e ai fattori classici, vi é però qualcosa di diverso: il *proxy*.

Il *proxy* non è soltanto elemento costitutivo e tipico della discriminazione algoritmica, assolvendo ad un ruolo centrale anche sotto il profilo soggettivo.

Se il *proxy* identifica l'elemento in base al quale la macchina distingue, ci si chiede se esso possa o meno considerarsi un sinonimo dei fattori di discriminazione classici

[112] In tema, soccorrono gli studi realizzati sul sistema statunitense, tra cui D.K. LEVY a altri, *U.S. Dep't of Hous. & Urban Dev., Discrimination in the Rental Housing Market Against People Who Are Deaf and People Who Use Wheel- chairs, National Study Findings*, 2015.

[113] In dottrina, N. FURL, P. PHILLIPS, A.J. O'TOOLE, *Face recognition algorithms and the other- race effect: Computational mechanisms for a developmental contact hypothesis*, in *Cognitive Science*, 2002, 797 ss. Ancora, dati utili sono offerti dal già più volte richiamato progetto "Gender Shades". Nell'ambito della dottrina nazionale, M. D'AMICO, *Una parità ambigua. Costituzione e diritti delle donne*, cit. e, superando la prospettiva di genere, anche, E. CURRAO, *Il riconoscimento facciale e i diritti fondamentali: quale equilibrio?*, in *Diritto Penale e Uomo*, 2021, 1 ss.

[114] Interessante il caso canadese, in cui i tribunali amministrativi affidavano ad un algoritmo il compito di concedere oppure negare il permesso di asilo, su cui il report già citato P. MOLNAR, L. GILL, *Bots At The Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System.*

[115] Si vedano i dati forniti dallo studio di che attestano, a titolo di esempio, che la percentuale di afro-americani assunti si attesta al 2, 5% a Google e al 4% presso Facebook e Microsoft. Si rinvia a S.M. WEST, M. WHITTAKER, K. CRAWFORD, *Discriminating Systems. Gender, Race, and Power In Ai*, 2019, AI Now Institute, https://ainowinstitute.org/discriminatingsystems.pdf.

oppure, all'opposto, se si traduca in qualcosa di diverso, cioè un elemento che *può* assumere tale attitudine per effetto della sua associazione ed interazione con i *suspect grounds of discrimination*.

Il quesito ha una sua pregnanza se analizzato dalla prospettiva, duplice, della regolamentazione e dell'accertamento di una disparità di trattamento sanzionabile alla luce delle norme di diritto positivo vigenti. Se, infatti, si considera il *proxy* quale fattore di discriminazione dovrà procedersi ad una attenta delimitazione degli elementi inquadrabili entro la nozione in esame al fine di assoggettare ed estendere il trattamento riservato ai fattori di discriminazione tradizionali anche al *proxy*. Il *proxy* verrebbe, quindi, a condividere il carattere di elemento che non può essere mai esplicitamente incluso quale fattore di distinzione, perché protetto e sospetto, con il rischio, però, forse, di imbrigliare eccessivamente il funzionamento delle tecniche di intelligenza artificiale per gli ineliminabili e indefiniti collegamenti tra i tratti identificativi dell'essere umano[116].

Viceversa, a voler sposare la tesi della eterogeneità tra fattore di discriminazione e *proxy*, occorrerà stabilire in che misura il secondo vi si distanzia alla luce della sua riscontrata e potenziale attitudine ad assimilarsi al primo sotto il profilo dell'effetto discriminatorio ai danni della vittima.

Accanto al tema della qualificazione del *proxy*, si affianca, poi, la ricaduta applicativa del suo impiego in relazione alla delimitazione delle appartenenze individuali, cioè alla identificazione dei gruppi e delle componenti strutturali di ciascuno secondo la classificazione imposta dal *proxy*.

A prescindere, quindi, dalla sua natura quale fattore di discriminazione "nuovo" oppure qualità associabile ai fattori tradizionali, il *proxy* traccia confini nuovi tra categorie e gruppi, incide sulle affiliazioni individuali e, non da ultimo, sulla nozione di minoranza[117]. In sintesi, l'intelligenza artificiale si riverbera, comportandone il superamento, sul concetto classico di gruppo sociale e di minoranza[118].

A differenza, però, dei fattori di discriminazione tradizionali che riposano su qualità di divisione tra esseri umani note e prevalentemente esteriorizzabili e misurabili, il *proxy* si appalesa mutevole, talvolta sconosciuto al programmatore e alla macchina che ne fa uso.

Accanto alla prospettiva "dall'alto", cioè di chi costruisce l'algoritmo, vi è poi quella "dal basso" di chi subisce le conseguenze derivanti dalla propria associazione

[116] Si osserva che: «[t]he notion of a protected class remains a fundamental legal concept, but as individuals increasingly face technologically mediated discrimination based on their positions within networks, it may be incomplete. In the most visible examples of networked discrimination, it is easy to see inequities along the lines of race and class because these are often proxies for networked position. As a result, we see outcomes that disproportionately affect already marginalized people», così D. BOYD, K. LEVY, A. MARWICK, *The Networked Nature of Algorithmic Discrimination*, in *Open Technology Institute*, New America, Data & Discrimination, 2014, 53 ss.

[117] Così K. DE VRIES, *Identity, profiling algorithms and a world of ambient intelligence*, in *Ethics Informatic Technology*, 2010, 71 ss.

[118] «This typology of 'group' or 'crowd' differs from the traditional understanding of groups, since the people involved in the "group" might not be aware of (1) their membership to that group, (2) the reasons behind their association with that group and, most importantly, (3) the consequences of being part of that group», così M. FAVARETTO, E. DE CLERCQ, B. SIMONE ELGER, *Big data and discrimination: perils, promises and solutions. A systematic review*, in *Journal of Big Data*, cit., 18.

inconsapevole ad un gruppo rispetto al quale l'involontarietà non significa solo omessa conoscenza dell'essere ascritti a quel gruppo, ma anche soggezione alle conseguenze[119] che ne derivano sul piano delle posizioni giuridiche soggettive.

Questo comporta che l'effetto della *proxy discrimination* sarà del tutto inconsapevole per il singolo, poiché l'affiliazione non è volontaria, nè conosciuta all'individuo, prima, ed alla vittima, poi.

Ci si riferisce ad un ulteriore aspetto tratto peculiare della discriminazione algoritmica, che incide sul diritto fondamentale all'autodeterminazione individuale, comprensivo, anzitutto, del diritto del singolo di decidere se fare o non fare parte di un gruppo secondo una costruzione della dimensione negativa e positiva del diritto individuale. Un diritto, che la Costituzione sancisce al suo articolo 2 e che il diritto internazionale dei diritti umani ha tradotto, nel quadro della sua elaborazione in tema di diritti del gruppo, nel diritto di *self-identification*, come testimonia l'art. 3 della Convenzione Quadro sulla protezione delle minoranze nazionali del Consiglio d'Europa[120].

Ma l'involontaria e sconosciuta affiliazione generata dal *proxy* produce anche altri effetti, che includono la impossibilità per il singolo di sottrarsi alle conseguenze negative derivanti dalla appartenenza tramite l'"uscita", cioè esercitando il diritto di *exit*. Ancora, la capacità aggregatrice del *proxy* determina[121] uno spostamento da una dimensione puramente individuale della discriminazione ad una collettiva, poiché i dati raccolti possono riverberare effetti pregiudizievoli per coloro che sono inseriti nel *data-set*[122], per i sotto-rappresentati, cioè non inclusi nel *data-set*[123], per la generalità degli individui in ragione dell'impiego delle tecniche di intelligenza artificiale in settori a diffusione globale[124].

A voler chiudere il cerchio, la *AI-derived discrimination*, oltre a stimolare un ripensamento degli elementi oggettivo-fattuali della discriminazione come atto, ne

---

[119] Su questo aspetto, A. MANTELERO, *Personal data for decisional purposes in the age of analytics: from an individual to a collective dimension of data protection*, in *Computer Law and Security Review*, 2016, 238 ss., ripreso da M. FAVARETTO, E. DE CLERCQ, B. SIMONE ELGER, *Big data and discrimination: perils, promises and solutions. A systematic review*, in *Journal of Big Data*, cit., 18.

[120] In letteratura, H.J. HEINTZE, *Article 3*, in M. WELLER (a cura di), *Oxford Commentaries on International Law. The Rights of Minorities. A Commentary on the European Framework Convention for the Protection of National Minorities*, Oxford, 2005, 124 ss. Quale esempio delle implicazioni cui può dare luogo l'applicazione di tale diritto nel contesto della Convenzione europea dei diritti dell'uomo, si veda Corte EDU, *Tasev c. North Macedonia*, [Prima Sezione], n. 9825/13, 16 maggio 2019, su cui si veda il commento di K. HENRARD, *Tasev v North- Macedonia: (blurry) dimensions and boundaries of the right to free self-identification*, in *StrasbourgObserver*, 2019; si consenta, inoltre, il rinvio a C. NARDOCCI, *Esiste un diritto individuale alla scelta della propria etnia? A margine di Corte europea dei diritti dell'uomo, Tasev c. North Macedonia*, in *Forum di Quaderni costituzionali*, 2019, 1 ss.

[121] M. FAVARETTO, E. DE CLERCQ, B. SIMONE ELGER, *Big data and discrimination: perils, promises and solutions. A systematic review*, in *Journal of Big Data*, cit., 11.

[122] P. MACDONNELL, T*he European Union's proposed equality and data protection rules: an existential problem for insurers?*, in *Economic Affairs*, 2015, 225 ss.

[123] L.P. FRANCIS, J.G. FRANCIS, *Data reuse and the problem of group identity*, in *Studies in Law, Politics and Society*, 2017, 143 ss.

[124] Ne danno conto H. KENNEDY, G. MOSS, *Known or knowing publics? Social media data mining and the question of public agency*, in *Big Data Society*, 2015, consultabile al link: https://doi.org/10.1177/2053951715611145.

trasforma anche la dimensione soggettiva enfatizzandone quella collettiva sino a favorire una diversa costruzione delle nozioni di minoranza e appartenenza, e delle relazioni *intra* ed *inter*-comunitarie.

9. *L'intersezionalità oggettiva e soggettiva nella* AI-derived discrimination

Vi è un ulteriore tema classico del diritto-antidiscriminatorio che si apprezza nella prospettiva soggettiva e che mantiene sicura pregnanza nel contesto della discriminazione "artificiale". Il riferimento è alle teorie sulla intersezionalità che, applicate all'intelligenza artificiale, presentano caratteri fenomenologici autonomi.

Tradizionalmente, l'intersezionalità presuppone l'interazione simultanea e l'altrettanto contestuale effetto, deteriore, ai danni della vittima ovvero del gruppo[125].

La simultaneità che si realizza dal lato causale e da quello degli effetti poggia, però, su fattori tradizionali noti, dei quali l'individuo possiede una propria consapevolezza dal punto di vista della sua, anche solo potenziale, affiliazione collettiva.

L'ingresso della discriminazione legata alle tecniche di intelligenza artificiale, viceversa, non soltanto incide sulla trasformazione del concetto di gruppo e di appartenenza, ma concorre a creare doppie minoranze o, se si preferisce, dei *sub-groups* che derivano non dall'intersezione tra fattori di discriminazione classici, ma tra questi e il *proxy*.

L'intreccio tra fattori di discriminazione classici e il *proxy* oppure una moltitudine di *proxys* innesca conseguenze in tutto assimilabili a quelle raffigurate dalla Crenshaw[126] nel noto esempio dell'incrocio tra strade: qui, però, le strade non sono solo rappresentate dai fattori discriminazione classici, accostandosene di ulteriori che palesano il ruolo di protagonista assoluto del *proxy*.

L'approccio intersezionale interessa, però, anche la dimensione oggettiva. Il *proxy*, cioè non crea solo nuove categorie, ma anche nuove tipologie di discriminazione. Si pensi a discriminazioni che poggiano su *proxy*, che valorizzano le abitudini di vita del singolo, così come il salario o la posizione lavorativa, come nella *price discrimination*[127].

---

[125] In questo senso, se ne apprezza la differenza rispetto alla pur contigua nozione di discriminazione multipla, che non condivide con la discriminazione intersezionale l'effetto simultaneo e deteriore dell'intreccio tra fattori ai danni della vittima oppure del gruppo. Sulla distinzione tra le due categorie concettuali, si rinvia all'*Handbook of Non-Discrimination, Fundamental Rights Agency* dell'Unione Europea, 2018, 64 e ss. Il testo integrale è consultabile al link: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-handbook-non-discrimination-law-2018_it.pdf. In tema, anche, il documento della Commissione europea, *Tackling Multiple Discrimination. Practices, policies and laws*, 2007.

[126] Tra i numerosi scritti, K. CRENSHAW, *Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color*, in *Stanford Law Review*, 1991, 1241 ss.; della stessa A., *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics*, in *University of Chicago Legal Form*, 1989, 57 ss. Nella dottrina costituzionalista, M. D'AMICO, *Una parità ambigua. Costituzione e diritti delle donne*, cit.

[127] Su cui A. GAUTIER, A. ITTO, P. VAN CLEYNENBREUGEL, *AI algorithms, price discrimination and collusion: a technological, economic and legal perspective*, in *European Journal of Law and Economics*, 2020, 405 ss.

L'incedere sulla scena del *proxy* non ha quindi solo ridefinito i contorni della caratterizzazione oggettiva della discriminazione "artificiale", ma ha ridisegnato il concetto di identità individuale, prima, e di appartenenza al gruppo, poi, rischiando di accentuare le divisioni e le possibili fratture tra esseri umani e gruppi di fronte all'opposto tentativo delle Carte costituzionali e del diritto internazionale dei diritti umani di promuovere un concetto di umanità che tenga sì conto delle differenze, ma secondo una prospettiva che ne salvaguardi coesione e unità.

PARTE SECONDA
DALLA GENESI ALL'ACCERTAMENTO: IL LEGISLATORE, I GIUDICI

10. *Il legislatore*

10.1. *Sulla regolamentazione della* AI-derived discrimination: *il "se", il "chi"…*

Approdando alla dimensione istituzionale, primo tema su cui soffermarsi attiene alla regolamentazione delle tecniche di intelligenza artificiale ed alle sue implicazioni sul principio di non discriminazione.

Oltre la ricognizione dello stato dell'arte in punto di positivizzazione di norme preposte a contrastare la discriminazione algoritmica, è utile porre l'attenzione su alcune questioni teoriche.

In estrema sintesi, queste riguardano: la scelta del *se* regolamentare il funzionamento dell'intelligenza artificiale e della discriminazione che ne discende; l'individuazione del soggetto istituzionale deputato ad assolvere a tale compito, se cioè sia da preferirsi una normazione di livello nazionale oppure uno spostamento del baricentro decisionale verso il diritto dell'Unione Europea, seguendo la linea già tracciata in materia di diritto anti-discriminatorio, se non addirittura in favore di organizzazioni internazionali; il *come*, ossia la utilizzabilità degli istituti del diritto anti-discriminatorio classico alla luce della ricostruita natura della(e) *AI-derived discrimination(s).*

Sul primo aspetto, è opportuno muovere dal difetto di norme di diritto positivo, di livello nazionale e sovranazionale, che si occupino della discriminazione "artificiale".

L'opportunità di incardinare i nuovi sviluppi della tecnologia entro schemi normativi predefiniti, così come quello, ad esso contiguo, dei limiti che incontra il legislatore al cospetto della scienza e dell'innovazione tecnologica – con cui la giurisprudenza costituzionale si è ampiamente confrontata[128] – è, infatti, al centro di

[128] Il riferimento è alla nota giurisprudenza costituzionale inaugurata dalle decisioni n. 282 del 2002 e n. 338 del 2003, poi ripresa e sviluppata dal filone giurisprudenziale in tema di fecondazione medicalmente assistita a partire dalle decisioni n. 151 del 2009 e n. 162 del 2014. Valga soltanto precisare che la riferibilità della giurisprudenza costituzionale richiamata al fenomeno che qui si indaga dovrà, però, tenere conto delle ineliminabili differenze, poggiando quelle decisioni su materie in cui il legislatore si confrontava con scoperte scientifiche, che investivano la scienza medica, ma soggette al controllo umano. La Corte indugiava, infatti, da un lato, sui limiti che incontra l'attività normativa al cospetto dell'arte medica, ma, vi accostava allo stesso tempo, il riconoscimento di uno spazio di autonomia del medico, cioè dell'uomo,

posizioni non sempre condivise sulla doverosità di disciplinare la materia, così come sulla individuazione della fonte del diritto e del soggetto istituzionale competente a farsene carico.

In questa sede, non ci si chiede se l'intelligenza artificiale complessivamente considerata debba essere oggetto di interventi di regolamentazione[129], bensì se la "nuova" discriminazione algoritmica debba essere destinataria di una normativa *ad hoc*, aderendo alle tesi che ne valorizzano l'eterogeneità rispetto a quella classica.

Sul *se* regolamentare la discriminazione "artificiale", si registrano approcci dicotomici tra le due sponde dell'oceano Atlantico, sia a livello nazionale che sovranazionale.

Il *favor* verso logiche di *non-regulation* appare prescelto, in particolare, dagli Stati Uniti d'America, sebbene si registrino sollecitazioni di segno opposto alla luce delle provate violazioni di diritti fondamentali, anzitutto, di eguaglianza e riservatezza. Analoga impostazione si rinviene nell'ambito delle Nazioni Unite[130], sebbene anche in questo caso si assiste ad un avvicinamento all'approccio del Consiglio d'Europa, da qualche anno impegnato nella selezione di strumenti di *soft law* in grado di orientare gli Stati contraenti nel contrasto, tra gli altri, della discriminazione "artificiale".

Quanto all'individuazione del soggetto deputato a disciplinare la discriminazione algoritmica, le risposte preliminari del continente europeo paiono inclini ad assegnare un ruolo di primo piano al diritto sovranazionale. Il riferimento è al Consiglio d'Europa e al all'*Ad hoc Committee on Artificial Intelligence* (CAHAI), che, nel dicembre 2020, ha adottato il documento *Feasibility study on a legal framework on AI design, development and application based on CoE standards*[131], al fine di dare seguito alle criticità emerse in letteratura sulla utilizzabilità dello strumentario vigente in materia di anti-discriminazione

---

chiamato a scegliere, sulla base di proprie conoscenze, il trattamento terapeutico più adeguato al caso concreto. Nel caso, invece, della tecnologia legata alla intelligenza artificiale, l'esautoramento della persona dal funzionamento della macchina, potrebbe costituire un elemento decisivo per ragionare sulla utilità di quella giurisprudenza che poggia sul riconoscimento di limiti ad interventi "dall'alto" del legislatore, ma che allo stesso tempo valorizza competenze propriamente umane. Questa giurisprudenza è, però, particolarmente utile nel segnalare, che deve essere sempre e comunque riconosciuto uno spazio di autonomia all'uomo nei suoi rapporti con la scienza e con la tecnologia.

[129] Sul tema sono diverse le voci che in letteratura nonché a livello delle organizzazioni internazionali e della stessa Unione Europea si sono espresse per la valorizzazione del ruolo sempre più attivo dell'uomo nel funzionamento delle tecniche di intelligenza artificiale. Sulla opportunità del *keeping humans in the loop*, si vedano F.M. ZANZOTTO, *Viewpoint: Human-in-the-loop Articial Intelligence*, in *Journal of Articial Intelligence Research*, 2019, 243 ss.; C. CATH, L. FLORIDI, *The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights*, in *Science and Engineering Ethics*, 2017, 449 ss. In materia, anche il report del Consiglio d'Europa, *Algorithms and Human Rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*, Council of Europe Publications, 2018. Si interroga sui limiti di un intervento umano sulla disciplina delle ricadute discriminatorie delle tecniche di intelligenza artificiale, P. ZUDDAS, *Intelligenza artificiale e discriminazioni*, cit., 14 ss.

[130] Interessante sottolineare come le tecniche di intelligenza artificiale siano utilizzato anche dalle agenzie che fanno capo alle Nazioni Unite soprattutto, ma non solo, nella selezione del personale. Un esempio interessante è offerto dal *software* utilizzato dall'Agenzia delle Nazioni Unite per i rifugiati, UNHCR, *Jetson tool*, per prevedere e stabilire il numero di rifugiati che arriveranno nei campi profughi della Somalia.

[131] Il testo integrale dello studio può essere letto al link: https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da.

e di approfondire le specificità della *AI-derived discrimination*, inquadrata nell'ambito della categoria dogmatica della *proxy discrimination* più che dalla prospettiva delle categorie classiche della discriminazione diretta ed indiretta[132].

Accanto all'attività del Comitato, che discute dell'adozione del primo trattato di diritto internazionale dei diritti umani in tema di intelligenza artificiale[133], si inserisce tutta una serie di provvedimenti di *soft law* approvati negli ultimi anni dal Consiglio d'Europa e che vanno dalla Carta etica europea sull'utilizzo dell'intelligenza artificiale nei sistemi giudiziari e nei sistemi ad esso connessi del 2018[134] sino alle più recenti linee guida in materia di riconoscimento facciale del gennaio 2021[135], che rappresenta uno dei terreni su cui più acceso è il dibattito, così come la proposta di vietare in radice l'impiego delle tecniche di IA in ragione dei risvolti discriminatori.

Più timidi, si sono, invece, rivelati gli interventi del legislatore dell'Unione Europea che, nonostante il GDPR[136], non si occupa esplicitamente della discriminazione "artificiale". La parola "discriminazione" ricorre, non a caso, solo una sola volta nel testo del Regolamento[137] e ad analoga sorte soggiacciono il plurale, "discriminazioni"[138], e il riferimento espresso a trattamenti "discriminatori"[139].

Più ampio e in accordo con le scelte degli estensori del Regolamento è, invece, lo spazio dedicato alla "profilazione"[140], che compare ben 22 volte e che costituisce, però,

---

[132] Il testo appare in linea con gli aspetti più salienti della fattispecie discriminatoria in esame, così come riassuntivamente descritti all'esito della Parte Prima di questo studio; si vedano, in particolare, il § 28 e ss.

[133] Il riferimento è al POLITICO AI Summit, 31 maggio 2021.

[134] La Carta è stata adottata il 3 e 4 dicembre 2018. Il testo integrale può essere letto al seguente link: https://rm.coe.int/carta-etica-europea-sull-utilizzo-dell-intelligenza-artificiale-nei-si/1680993348.

[135] Le linee guida sono state adottate il 28 gennaio 2021. Si rinvia al link: https://rm.coe.int/guidelines-on-facial-recognition/1680a134f3.

[136] Regolamento (Ue) 2016/679 del Parlamento Europeo e del Consiglio del 27 aprile 2016, *relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati)*. Il regolamento UE in tema di protezione delle persone fisiche con riguardo al trattamento dei dati personali costituisce il più avanzato impianto di norme che disciplinano la materia considerata.

[137] Così il punto n. 85 del *Considerando* dove si legge che: «[u]na violazione dei dati personali può, se non affrontata in modo adeguato e tempestivo, provocare danni fisici, materiali o immateriali alle persone fisiche, ad esempio perdita del controllo dei dati personali che li riguardano o limitazione dei loro diritti, discriminazione, furto o usurpazione d'identità, perdite finanziarie, decifratura non autorizzata della pseudonimizzazione, pregiudizio alla reputazione, perdita di riservatezza dei dati personali protetti da segreto professionale o qualsiasi altro danno economico o sociale significativo alla persona fisica interessata».

[138] Si veda il punto n. 75 del *Considerando*.

[139] Punto n. 71 del *Considerando*.

[140] Alla profilazione è dedicato l'art. 22 del GDPR, *Processo decisionale automatizzato relativo alle persone fisiche, compresa la profilazione*, che riconosce all'individuo il diritto di opporsi a procedimenti di decisione automatizzata, fatta eccezione per tre ipotesi, cioè quando la profilazione: «sia necessaria per la conclusione o l'esecuzione di un contratto tra l'interessato e un titolare del trattamento; sia autorizzata dal diritto dell'Unione o dello Stato membro cui è soggetto il titolare del trattamento […]; si basi sul consenso esplicito dell'interessato». Sulle criticità della norma, si condividono le perplessità di P. ZUDDAS, *Intelligenza artificiale e discriminazioni*, cit., in particolare, 16 e ss., quanto al tema del consenso, di A. SIMONCINI, S. SUWEIS, *Il cambio di paradigma nell'intelligenza artificiale e il suo impatto sul diritto costituzionale*, cit. La profilazione è definita una «forma di trattamento automatizzato dei dati personali che valuta aspetti personali concernenti una persona fisica, in particolare al fine di analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze o gli interessi personali, l'affidabilità o il comportamento, l'ubicazione o gli spostamenti dell'interessato[140], ove ciò

solo una delle forme della discriminazione "artificiale", tanto da aver indotto la letteratura a riferirsi alla non discriminazione come «principio mancante»[141] nell'impianto del GDPR.

Il GDPR, quindi, non fornisce indicazioni né sul "se", sul "chi", né tantomeno sul "come" assoggettare ad un *corpus* di norme la discriminazione derivante dall'impiego delle tecniche di intelligenza artificiale. Il legislatore dell'Unione Europea si astiene, cioè, dal disciplinare il fenomeno della discriminazione " artificiale", così come non si occupa di istituire un raccordo coerente tra la nuova fenomenologia e le direttive di seconda generazione.

Maggiore sensibilità nei confronti delle implicazioni discriminatorie dell'intelligenza artificiale[142] emerge, invece, nella Risoluzione del Parlamento europeo del 20 gennaio 2021 *sull'intelligenza artificiale: questioni relative all'interpretazione e applicazione del diritto internazionale nella misura in cui l'UE è interessata relativamente agli impieghi civili e militari e all'autorità dello Stato al di fuori dell'ambito della giustizia penale*, che dimostra una più spiccata propensione nei confronti del fenomeno discriminatorio[143], trattando non solo sul dato esperienziale delle ricadute lesive dell'eguaglianza derivanti dal ricorso a processi di decisione automatizzata, ma enfatizzando la centralità che riveste la verifica del "come" «le tecnologie di IA ad alto rischio giungano a una decisione». Nella stessa direzione si inseriscono ulteriori snodi della Risoluzione, tra cui l'invito agli Stati membri ad assoggettare ad un controllo adeguato, umano, le decisioni delle pubbliche amministrazioni che poggiano su decisioni automatizzate in ragione del rischio di ricadute pregiudizievoli sui diritti individuali, tra cui la non-discriminazione. Importante anche il monito verso forme di regolamentazione che rispondano al principio di trasparenza, elevato a precondizione per uno scrutinio effettivo a tutela del diritto al giudice[144].

Degna di nota è, infine, la recente proposta di regolamento del 21 aprile 2021, *Proposal for a Regulation laying down harmonised rules on artificial intelligence*[145], che

produca effetti giuridici che la riguardano o incida in modo analogo significativamente sulla sua persona», così l'art. 4 del GDPR, *Definizioni*, che riprende il punto n. 71 del *Considerando*.

[141] Così A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, cit., 84.

[142] Prima della Risoluzione, l'Unione Europea si era già mossa con specifico riferimento al fenomeno della intelligenza artificiale. Si richiamano, in questa sede, il Libro Bianco sull'Intelligenza, del 19 febbraio 2020, e gli Orientamenti etici per un'IA affidabile del Gruppo di esperti ad alto livello sull'intelligenza artificiale, consultabile al link: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[143] Nella letteratura costituzionalistica, insiste sulla necessità che i parametri con cui valutare il fenomeno dell'intelligenza artificiale siano i principi cardine su cui si fonda, anzitutto, la Costituzione, ma anche la Carta dei diritti fondamentali dell'Unione europea, così come, a volersi limitare alla dimensione sovranazionale continentale, alla Convenzione europea dei diritti dell'uomo B. CARAVITA, *Principi costituzionali e intelligenza artificiale*, cit., 461 ss.

[144] Cfr. punto n. 52. Sul tema della trasparenza, si veda anche il successivo punto n. 62.

[145] Il testo integrale può essere letto al link: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. Tra gli aspetti cruciali della proposta, può ricordarsi l'enfasi posta sul concetto di *human oversight*, a cui è dedicato l'art. 14, e, ancora costruzione di un meccanismo di *governance* incardinato su un complesso reticolo di relazioni tra nuovi corpi intermedi

palesa una maggiore sensibilità verso la discriminazione algoritmica, sebbene l'approccio non appaia preordinato ad una sua tipizzazione e alla enucleazione di meccanismi alternativi a quelli vigenti.

L'impostazione, che traspare dal *memorandum* esplicativo, non isola la discriminazione "artificiale" da quella classica, ma si limita ad affermare che la proposta *completa* le misure già apprestate dal diritto anti-discriminatorio euro-unitario, chiarendo che l'obiettivo della proposta è minimizzare i rischi legati alla discriminazione algoritmica[146] in relazione a tutte le fasi della ideazione e della messa in funzione delle tecniche di intelligenza artificiale; una scelta, quindi, importante e che valorizza la centralità dei meccanismi di realizzazione della *AI-derived discrimination*[147] .

La proposta di regolamento è, poi, meritoria per aver delineato le aree maggiormente a rischio di generare effetti discriminatori[148], evidenziando, contestualmente, quali sono le tecniche di IA potenzialmente più lesive per i diritti fondamentali.

In definitiva, la proposta di regolamento sembra "prendere sul serio" la discriminazione algoritmica.

Desta, invece, qualche perplessità l'assenza di riferimenti espliciti al *corpus* normativo in materia di non-discriminazione, così come alle categorie classiche del diritto anti-discriminatorio. La proposta, si dice, *completa* lo strumentario esistente, ma non chiarisc fino in fondo come le regole delineate integrano e andranno ad innestarsi su quelle delle direttive degli anni 2000 nonché sulle rispettive normative attuative degli Stati membri.

Occorrerà, quindi, verificare, in caso di definitiva approvazione ed entrata in vigore della proposta di regolamento ed alla luce dei casi concreti, come reagiranno le norme e come risponderà il diritto dell'Unione ad una discriminazione che rimane "nuova" per tutte le ragioni già descritte.

10.2. … *e il "come"*

Il "come" disciplinare la discriminazione "artificiale" interroga sulla effettiva utilizzabilità delle norme di diritto positivo vigenti anche per il contrasto della prima e su

istituiti dalla proposta tra Unione e Stati membri, su cui si vedano le norme di cui al Titolo VI della proposta di regolamento.

[146] La proposta suggerisce una graduazione delle tecniche di IA in dipendenza del rispettivo grado di rischio, assoggettandole a regole differenziate. Si vedano i Titoli II e III, rispettivamente dedicati alle *Prohibited Artificial Intelligence Practices* e ai *High-Risk AI Systems.*

[147] Cfr. punto n. 1.2., 4.

[148] Schematicamente, la proposta di regolamento fa riferimento a: «AI systems providing social scoring of natural persons for general purpose by public authorities or on their behalf» (17); sistemi di intelligenza artificiale che operano nei settori della sanità e della sicurezza (28); dell'istruzione (35); del lavoro (36); dell'accesso e del godimento di servizi pubblici e privati (37); della giustizia, specie penale, con particolare riferimento alla sorveglianza, alla fase esecutiva delle sanzioni (38). In analogo contesto, si colloca la recente notizia, riportata dalla stampa, del divieto imposto dall'Unione Europa opposto all'utilizzo delle tecniche di riconoscimento facciale nei luoghi pubblici.

quali strumenti legislativi potrebbero essere impiegati allo scopo di intervenire sul funzionamento delle tecniche di intelligenza artificiale, in quali fasi, con quali limiti.

La prima questione si ritiene di più agevole soluzione.

Le teorie della discriminazione diretta ed indiretta non si attagliano a quella "meccanica", prodotto di decisioni automatizzate, per la descritta eterogeneità tra fattispecie che si risolve, a monte e per la già menzionata interposizione della macchina, nell'ardua collocazione delle nuove forme di discriminazione entro le tipologie classiche, tanto che la dottrina converge nel senso della inadeguatezza del diritto anti-discriminatorio di derivazione euro-unitaria a rispondere alle specificità della *AI-derived discrimination*.

Se si guarda, ad esempio, alla discriminazione diretta, il legislatore dovrebbe soffermarsi sulla diversa o alternativa costruzione del requisito della intenzionalità che scolorisce al cospetto di condotte, parzialmente ovvero integralmente, imputabili alla macchina, decidendo dove risieda la volontarietà, centrale per la definizione come diretta della disparità di trattamento per le norme vigenti e di così complessa ricostruzione dall'angolo prospettico della discriminazione algoritmica. Ancora, sarebbe doveroso ragionare dei confini tra intenzionalità e conoscibilità, valutando (sarebbe possibile?) uno spostamento della responsabilità individuale in capo al programmatore, magari affetto da un pregiudizio implicito, ma anche all'utilizzatore finale dell'algoritmo, cioè il datore di lavoro[149], l'istituto di credito, l'autorità giudiziaria, anche in presenza di una mera conoscenza del carattere discriminatorio della macchina.

Un altro esempio di scollamento tra il sistema di diritto anti-discriminatorio tradizionale e la "discriminazione artificiale" si coglie in relazione al *proxy*. La predeterminazione di quali siano gli elementi che fungono da *proxy* nella *AI-derived discrimination* da parte del legislatore, secondo elenchi accostabili a quelli di cui all'art. 3, comma 1, Cost., ovvero di altre Carte costituzionali nazionali o trattati sovranazionali, oltre il rapporto da ricostruire con questi ultimi, pare opzione difficilmente percorribile: gli aspetti predittivi di appartenenza individuale sono potenzialmente ben superiori in numero ai fattori di discriminazione classici; la loro selezione pare affetta da un forse troppo elevato tasso di discrezionalità in difetto di criteri che possano orientarne in modo univoco e rispettoso dei principi di legalità e di certezza del diritto la identificazione e la successiva positivizzazione; rinvenire un consenso su quali essi siano appare da ultimo particolarmente complesso, pur rimanendo fondamentale nel contesto delle sempre più strette relazioni tra livelli di tutela.

Il secondo quesito, invece, è più complesso e, per alcuni profili, supera il recinto dell'analisi che si propone. Si ritiene, però, interessante offrire qualche spunto.

Il primo riguarda la scelta in favore del necessario coinvolgimento della persona nel governo dell'intelligenza artificiale. La *ratio* di un'eventuale traduzione legislativa della disciplina della discriminazione algoritmica dovrebbe, quindi, muovere da un approccio all'intelligenza artificiale che non la renda sostitutiva, bensì collaterale e di

[149] Approfondisce le relazioni tra intelligenza artificiale e diritto del lavoro S. MAINARDI, *Intelligenze artificiali e diritto del lavoro*, in U. RUFFOLO (a cura di), *Intelligenza artificiale. Il diritto, i diritti. L'etica*, cit., 363 ss.

supporto all'azione umana. Su questo, si precisa, inoltre, che l'ausilio della tecnica non dovrebbe comunque mai risolversi nella delegazione alla macchina di funzioni, o meglio, di abilità solo umane, di cui la citata meta-cognizione rappresenta una delle capacità più significative.

L'avallo di un'impostazione incardinata sul protagonismo della persona implica, quale secondo aspetto, che dovrebbe verificarsi la possibilità che l'intervento di regolamentazione possa svilupparsi lungo tutte le scansioni logico-temporali[150] che presiedono alla costruzione dell'algoritmo[151], incluse le fasi dell'aggiornamento e dell'inquinamento dei dati.

Un terzo spunto riguarda, invece, gli effetti delle condotte derivanti dalla discriminazione algoritmica. Appare utile enfatizzare che la distinzione meno netta tra le teorie della discriminazione diretta e indiretta non si risolve solo nella loro inutilizzabilità in punto di disciplina della discriminazione algoritmica. Piuttosto, la dedotta difficoltà di agganciare quest'ultima all'una oppure all'altra categoria potrebbe essere superata laddove il legislatore scegliesse di disinteressarsi della prospettiva definitoria per "limitarsi" a qualificare come discriminatoria ogni condotta suscettibile di produrre effetti forieri di una differenziazione irragionevole, ripiegando cioè sui soli criteri del *disparate impact*.

Questa opzione avrebbe il pregio di focalizzare l'attenzione sulla dimensione oggettiva, agevolando la reazione dell'ordinamento soprattutto di fronte alla prova della disparità di trattamento che sarebbe soddisfatta dalla sola dimostrazione di una distinzione di trattamento non sorretta da alcuna giustificazione oggettiva e ragionevole anche sulla base di evidenze statistiche, senza dover ricercare evidenze dell'intento discriminatorio. Sotto altro versante, tuttavia, l'oscuramento della dimensione soggettiva, legata cioè alla intenzionalità dell'agente, potrebbe comportare il rischio di favorire la costruzione di norme rigide che poggiano su presunzioni e automatismi lesivi del principio costituzionale di ragionevolezza.

Norme siffatte, inoltre, male si concilierebbero con l'impostazione che esige la valorizzazione della persona umana nel processo di funzionamento delle tecniche di intelligenza artificiale ed esporrebbero a tensione il principio di autodeterminazione individuale, che dovrebbe viceversa rimanere centrale laddove ci si proponga di studiare strategie di accertamento della responsabilità del singolo, appunto.

[150] Su cui si rinvia, *supra*, Parte Prima.

[151] In tema, P. ZUDDAS, *Intelligenza artificiale e discriminazioni*, cit., riferisce invece di due dimensioni di cui il legislatore dovrebbe tenere conto. L'A. osserva, in proposito, che: «[i] possibili effetti discriminatori prodotti dalle decisioni algoritmiche possono essere mitigati anche attraverso una disciplina che operi non soltanto, per così dire, 'a monte' (nei termini finora illustrati) rispetto a tali decisioni, ma anche 'a valle', richiedendo un intervento umano di controllo che consenta di individuare e correggere le eventuali distorsioni generate dall'intelligenza artificiale», 15.

11. *Il giudice*

11.1. *Le criticità di ordine teorico: l'accesso alla giustizia*

Appurata l'inesistenza di norme che integrino la disciplina in materia di diritto anti-discriminatorio vigente oppure, come sembrerebbe preferibile laddove si opti per la regolamentazione del fenomeno, che isolino la discriminazione "artificiale" valorizzandone la tipicità sia dal lato oggettivo che soggettivo, si tratta ora di soffermarsi sulle relazioni tra il giudice e la nuova fenomenologia discriminatoria intrecciando alcuni profili teorici all'analisi della giurisprudenza, invero scarna, formatasi sul continente europeo.

Il ridotto numero di casi evoca un problema antico ma comune al fenomeno discriminatorio sin dai suoi albori.

L'accesso al giudice da parte della vittima rappresenta, infatti, una criticità persistente dei sistemi di diritto anti-discriminatorio e, nel caso della discriminazione "artificiale", soffre di due elementi aggiuntivi. Da un lato, l'accesso alla giustizia tende ad acuirsi al cospetto di fenomenologie discriminatorie sconosciute oppure poco indagate e normate; dall'altro, la circostanza che le tecniche di intelligenza artificiale discriminano in modo proporzionalmente maggiore individui appartenenti a categorie già discriminate, ne accentua la condizione di vulnerabilità.

A questo, e sempre nella prospettiva soggettiva, si aggiunge l'ulteriore aspetto della carenza di consapevolezza, da parte della vittima, della propria affiliazione ad un gruppo inciso dalle tecniche di intelligenza artificiale, che pregiudicherà la stessa percezione dell'essere vittima di trattamenti discriminatori ostando all'esercizio del diritto al giudice.

Non si riscontra, però, soltanto un problema di accesso alla giustizia.

Ad esso si affiancano la effettiva accertabilità, prima, e sanzionabilità, poi, della condotta discriminatoria, che largamente dipendono da come sono scritte le norme. Criticità destinate ad acuirsi per la discriminazione legata all'impiego delle tecniche di intelligenza artificiale che, in difetto di norme specifiche, vede aggravarsi non soltanto la fase di accesso alla giustizia, ma anche le successive di accertamento e di sanzione della condotta discriminatoria[152].

---

[152] Illustrano esemplificativamente le difficoltà che si insinuano nella fase di accertamento giurisdizionale dell'effetto discriminatorio delle tecniche di intelligenza artificiale alcuni casi giurisprudenziali statunitensi e canadesi relativi all'impiego dei *evidence-based risk assesment tools*, diffusi nell'esperienza d'oltreoceano almeno dal 2004, quando la Conferenza giudiziaria degli Stati Uniti, deputata ad adottare le linee di indirizzo del sistema giudiziario federale, ha optato per l'adozione su larga scala di sistemi automatizzati ai fini della concessione di provvedimenti di *parole*, nonché nelle fasi che precedono l'avvio del procedimento giurisdizionale. Ne dà conto A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, cit., 72 che ricorda, a titolo di esempio, che nel 2017 lo Stato del Massachussets ha addirittura adottato una normativa che impone il ricorso a titolo obbligatorio agli strumenti automatizzati di *assessements-risk* nelle fasi di *pretrial*. Offre un approfondimento, aggiornato però al mese di giugno 2018, degli strumenti automatizzati adottati dalle Corti federali statunitensi il rapporto *Administrative Office of the United States Courts Probation and Pretrial Services Office. An Overview of the Federal Post Conviction Risk Assessment*, che può essere letto integralmente al link: https://www.uscourts.gov/sites/default/files/overview_of_the_post_conviction_risk_assessment_0.pdf.

Quanto ai casi, il riferimento è alla decisione del 2018 del New York State Trial Court che ha stabilito che la richiesta e l'ottenimento dei dati impiegati dai sistemi algoritmici di polizia predittiva utilizzati dal

11.2. (Segue) *l'individuazione del soggetto responsabile*

Se si volge lo sguardo alla fase preposta all'accertamento giurisdizionale del carattere discriminatorio della tecnica di intelligenza artificiale, si osserva come questa sconta alcuni problemi che investono, anzitutto, l'individuazione del soggetto, pubblico o privato, responsabile della condotta discriminatoria[153].

Un primo aspetto attiene alla scelta tra chi ha costruito l'algoritmo e chi lo ha impiegato nell'esercizio delle proprie funzioni. Ci si chiede – come già incidentalmente nella prima parte del lavoro, laddove si discuteva dei confini tra conoscibilità e intenzionalità in relazione alla categoria della discriminazione diretta – se responsabile e, se sì, a quali condizioni, debba oppure possa esserlo: chi ha programmato l'algoritmo; chi ha utilizzato l'algoritmo; oppure, ancora, se possano esservi delle commistioni, cioè ipotesi di corresponsabilità addebitabili ad entrambi i soggetti, siano essi pubblici o privati.

Il tema della individuazione del soggetto responsabile riflette molte delle criticità che caratterizzano il funzionamento della tecnica di intelligenza artificiale nei suoi

---

New York City Police Department sono legittimi nell'ambito di un procedimento giurisdizionale che aveva ad oggetto la verifica del carattere discriminatorio a sfondo etnico-razziale di alcune misure restrittive adottate dalla polizia sulla base di un algoritmo. La pronuncia, caso *Brennan Center for Justice v. New York City Police Departmnet*, n. 160541/2016, può essere letta al link: http://www.nycourts.gov/reporter/pdfs/2017/2017_32716.pdf. Ancora, può richiamarsi la famosa vicenda decisa nel 2013 dalla Corte Suprema dello Stato del Wisconsin nel caso *Loomis* relativa all'impiego del software COMPAS, il cui testo può essere letto al link: https://www.giurisprudenzapenale.com/wp-content/uploads/2019/04/Supreme-Court-of-Wisconsin.pdf. Per l'analisi del funzionamento di COMPAS si veda: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Per un commento critico alla sentenza si veda *Harvard Law Review*, 2017, 1530 ss. In tema, si ricorda che il dibattito sulle criticità legate ai sistemi automatizzati preposti alla valutazione del rischio si è intensificato a seguito della pubblicazione, nel 2016, dell'articolo di J. ANGWIN, J. LARSON, S. MATTU, *Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks*, in *ProPublica*, il 23 maggio 2016; nell'ambito della letteratura nazionale, si veda S. CARRER, *Se l'amicus curiae è un algoritmo: il chiacchierato caso Loomis alla Corte Suprema del Wisconsin*, in *Giurisprudenza Penale Web*, 2019, 1 ss. Sulle difficoltà di accertamento della discriminatorietà degli algoritmi si rinvia, in particolare, a R.M. O'DONNELL, *Challenging Racist Predictive Policing Algorithms under the Equal Protection Clause*, cit., 574 ss. Da ultimo, si segnala, anche, la pronuncia resa dalla Corte Suprema sul caso *Texas Department of Housing & Community Affairs v. Inclusive Communities Project, Inc*., in tema di assegnazione di alloggi in base ad una tecnica di intelligenza artificiale. Per un commento, si rinvia a L. RODRIGEZ, *All data is not credit data: closing the gap between the fair housing act and algorithmic decisionmaking in the lending industry*, in *Columbia Law Review*, 2020, 1843 ss.

Sul versante canadese, si veda il caso *Ewert c. Canada*, 13 giugno 2018, SCC 30, [2018] 2 S.C.R. 165, primo caso di discriminazione algoritmica deciso da una Corte Suprema, che riguardava un uomo, detenuto in un centro di detenzione federale ed appartenente alla comunità indigena dei *Métis*, che lamentava il carattere discriminatorio degli strumenti informatici utilizzati dal *Correctional Service of Canada* (CSC) allo scopo di stabilire il livello di pericolosità sociale dei detenuti. Richiama la pronuncia L. GIACOMELLI, *Big brother is «gendering» you. Il diritto antidiscriminatorio alla prova dell'intelligenza artificiale: quale tutela per il corpo digitale?*, cit., 284 ss.

[153] Non ci si occuperà, invece, delle criticità che riguardano il giudice nell'esercizio delle sue funzioni con riferimento ai principi di autonomia e di indipendenza potenzialmente incisi dalle tecniche di intelligenza artificiale. Per questi aspetti, si vedano N. ZANON, F. BIONDI, *Il sistema costituzionale della magistratura*, Bologna, 2019.

rapporti con l'azione umana. Tanto più la macchina si distanzia dal controllo e dalla supervisione della persona, tanto più quest'ultima difficilmente potrà essere considerata responsabile di una condotta discriminatoria che derivi dalla condotta della prima. La questione concerne, quindi, sia l'identificazione del soggetto a cui addebitare responsabilità da accertare in sede giurisdizionale, ma anche e prima, la ricerca dell'intervento umano cui addebitare la discriminazione.

L'ipotesi più semplice è quella del *masking*, che richiede la dimostrazione della intenzionalità, cioè del *bias* non implicito, del programmatore, che potrà emergere ad esempio dalla selezione dei dati, nonché dal collegamento tra l'intento discriminatorio e, di regola, il suo effetto. Qualora alla *voluntas* discriminatoria del programmatore, che costruisce la macchina *per* discriminare, si affianchi anche quella di colui che la impiega, si tratterà di stabilire se la responsabilità censurabile in sede giurisdizionale sia solo del primo ovvero anche del secondo. Si verserà nella seconda ipotesi qualora il secondo utilizzi l'algoritmo anch'egli *per* discriminare; più complesso, invece, il caso in cui quest'ultimo sia a conoscenza dei rischi potenzialmente discriminatori della tecnica di intelligenza artificiale ma scelga di utilizzarlo ugualmente pur non volendo discriminare[154].

Oltre il *masking*, la casistica diviene più intricata perché da un lato, la discriminazione che deriva dalle tecniche di intelligenza artificiale si allontana dalle categorie della discriminazione diretta e indiretta[155]. Dall'altro, la ricerca del soggetto a cui addebitare la responsabilità presuppone la capacità di individuare la causa del funzionamento discriminatorio della macchina, spesso ignota e prodotto di numerosi passaggi e di una serie altrettanto numerosa di concause, umane ed artificiali insieme.

Le stesse difficoltà che si riscontrano nel comprendere come discrimina la macchina si riflettono, quindi, anche nella fase preposta alla individuazione del soggetto da ritenersi responsabile della discriminazione.

Naturalmente, come insegna l'esperienza statunitense, laddove si tratti di tecniche di intelligenza artificiale impiegate dal pubblico ministero oppure dal giudice nel quadro di meccanismi di polizia oppure di giustizia predittiva, l'azione giudiziaria è stata promossa nei confronti dello Stato e non dell'azienda che aveva costruito l'algoritmo. In tutti questi casi, e a prescindere dalla tipologia di discriminazione in esame, potrebbe sostenersi la tesi secondo cui la responsabilità risiede laddove l'algoritmo diviene realtà, ossia produce effetti concreti collegati al suo utilizzo discriminatorio[156]. Questa è la tesi che presuppone un addebito di responsabilità in capo al soggetto pubblico a fronte dell'esistenza di un collegamento stretto tra l'azienda privata committente e il beneficiario, ad eccezione dell'ipotesi sopra descritta del *masking* che non reciderebbe la responsabilità della prima.

---

[154] È il caso di cui riferiscono S. BAROCAS, A.D. SELBST, *Big data disparate impact*, cit.

[155] Ad esempio, per la discriminazione diretta, la prova della intenzionalità dell'agente e l'individuazione di una condotta che sia esplicitamente fondata su uno dei fattori di discriminazione classici; per quella indiretta, la prova dell'effetto ma non invece della dimensione soggettiva, cioè della intenzionalità del soggetto agente e, al contempo, l'apparente neutralità della regola da cui si sostiene promani la discriminazione.

[156] In questo senso, R.M. O'DONNELL, *Challenging Racist Predictive Policing Algorithms under the Equal Protection Clause*, cit., 577.

In ogni caso, l'insieme delle problematicità evidenziate potrebbe trovare una risposta adeguata se si decidesse di garantire piena effettività al diritto individuale alla spiegazione, cioè a consentire al singolo e alla collettività di essere informati e di conoscere come funziona la tecnica di intelligenza artificiale[157]. Il *right to an explanation*[158] potrebbe, così, ovviare ad alcuni ostacoli che si frappongono alla ricerca del *bias* e, in definitiva, sopperire al *vulnus* esistente in punto di individuazione e di verifica della sussistenza di responsabilità, individuali e collettive, legate all'intelligenza artificiale tramite il disvelamento dei suoi meccanismi di azione.

11.3. (Segue) *e la prova della disparità di trattamento*

Si affiancano alle questioni descritte quelle che interessano la prova del carattere discriminatorio della tecnica di intelligenza artificiale.

I problemi, svariati, variano in dipendenza dello schema teorico – *disparate treatment* oppure *disparate impact* – a cui si associa la doglianza lamentata davanti al giudice, impiegando gli strumenti del diritto anti-discriminatorio classico.

Quanto alla discriminazione diretta, la questione principale concerne la dimostrazione dell'intento discriminatorio, poichè raramente le tecniche di intelligenza artificiale distinguono in base a fattori di discriminazione tradizionali. Altrettanto difficile è appurare che la macchina sia stata intenzionalmente programmata per discriminare. Le difficoltà legate alla prova del *discriminatory intent* seguono quelle dell'individuazione del soggetto agente e sono ulteriormente complicate dal pregiudizio implicito, che costituisce allo stesso tempo tratto caratterizzante della fattispecie discriminatoria e causa della sua non inquadrabilità dogmatica e non sanzionabilità come discriminazione diretta.

Un altro aspetto attiene alla qualità individuale in base al quale la tecnica di intelligenza artificiale distingue, cioè al *proxy*. Si tratterà di individuare quest'ultimo dando prova dell'esistenza della sua correlazione diretta con un elemento identificativo di una categoria protetta.

Non fosse per le conclusioni affatto diverse, il ragionamento muta allorché si ritenga che la fattispecie presenti i caratteri della discriminazione indiretta. In questo caso, si tratterà accertare che l'effetto e non anche l'intento sia discriminatorio. Lo spostamento dalla dimensione soggettiva, la prova della intenzionalità dell'agente, a quella oggettiva, la dimostrazione dell'effetto discriminatorio, sembrerebbe favorire strategie processuali

[157] Per comprendere che cosa si intenda per "spiegazione" con riferimento ai sistemi di decisione automatizzata, S. WACHTER, B. MITTELSTADT, L. FLORIDI, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, in *International Data Privacy Law*, 2017.

[158] Interessa in materia ricordare che il GDPR non contempla il diritto ad una spiegazione nemmeno quando fa riferimento alla profilazione. In letteratura, si vedano, diffusamente, S. WACHTER, B. MITTELSTADT, L. FLORIDI, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, cit., che riferiscono piuttosto dell'esistenza di un «limited right to be informed". In tema, anche, M. BRKAN, *Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond*, in *International Journal of Law and Information*, 2019, 91 ss.

che leggano la discriminazione algoritmica sempre come discriminazione indiretta, così da ovviare alle delineate difficoltà probatorie.

E, tuttavia, la discriminazione che deriva dalle tecniche di intelligenza artificiale difetta di quella anche solo apparente neutralità che invece connota, come visto, la nozione classica di discriminazione indiretta. Il *proxy* agisce spesso quale fattore di discriminazione vero e proprio poiché fa perno su qualità individuali predittive dell'appartenenza a categorie protette sì da rendere recessiva anche la distinzione tra fattore di discriminazione ed elemento puramente predittivo dell'appartenenza alla categoria protetta.

Tanto nella ipotesi in cui si faccia ricorso alla categoria della discriminazione diretta, tanto a quella indiretta, le difficoltà di accertamento si irrobustiscono non solo per effetto del ricorrente difetto di accesso ai dati ed al meccanismo di funzionamento dell'algoritmo, ma soprattutto perché la rigida bipartizione discriminazione diretta / indiretta male si concilia con i tratti peculiari di quella "artificiale".

Le criticità legate alla fase di accertamento confermano, quindi, l'impossibilità di assoggettare la discriminazione algoritmica all'una ovvero all'altra categoria, dimostrandosi piuttosto preferibile valorizzare le specificità della *proxy discrimination*, il ruolo del *proxy* e del pregiudizio implicito, non costituendo l'avallo piano dello schema della discriminazione indiretta idoneo a sopperire alle osservate difficoltà di accertamento[159].

11.4. … *come risolverle*

Oltre le gli aspetti di ordine teorico, interessano le soluzioni che si stanno affacciando sulla scena globale e che si distinguono in due tipologie di intervento: del diritto e della tecnica.

Alla prima fanno capo le elaborazioni teoriche, che ricercano principi generali – il *right to know*, la trasparenza[160], l'accesso ai dati – a cui assoggettare le aziende che programmano tecniche di intelligenza artificiale, così come i loro utilizzatori finali.

---

[159] Così R.M. O'DONNELL, *Challenging Racist Predictive Policing Algorithms under the Equal Protection Clause*, cit.

[160] Si tratta di un tema particolarmente importante quando le tecniche di intelligenza artificiale sono impiegate nel settore pubblico e da enti pubblici e istituzioni. L'esperienza comparata offre alcuni esempi virtuosi. Si segnala l'esperienza delle città di Amsterdam (https://algoritmeregister.amsterdam.nl/en/more-information/) e di Helsinki (https://ai.hel.fi/en/get-to-know-ai-register/) che hanno reso pubblico l'elenco degli strumenti di intelligenza artificiale impiegati e, in senso analogo, può richiamarsi anche il caso della città di New York. Interessante è la presentazione della piattaforma sviluppata dalla città di Helsinki dove si legge: «AI Register is a window into the artificial intelligence systems used by the City of Helsinki. Through the register, you can get acquainted with the quick overviews of the city's artificial intelligence systems or examine their more detailed information based on your own interests. You can also give feedback and *thus participate in building human-centred AI in Helsink* [corsivo aggiunto]". Con riferimento al caso statunitense, merita uno sguardo il rapporto redatto dal *Algorithms Management and Policy Officer (AMPO)* nel 2020, allo scopo di mappare le istituzioni che fanno uso di tecniche di intelligenza artificiale, catalogarle, rendere trasparente l'obiettivo perseguito dalla macchina ed il suo funzionamento. Il testo è consultabile al link: https://www1.nyc.gov/assets/ampo/downloads/pdf/AMPO-CY-2020-Agency-Compliance-Reporting.pdf.

Sull'altro *cotè* si muove, invece, la tecnica allo scopo di costruire strumenti di supervisione, di c.d. *algorithmic auditing*[161], che dovrebbero colmare l'«*accountability gap*»[162] delle tecniche di intelligenza artificiale, testandone il buon funzionamento in termini di *fairness* e di assenza di *bias*. In proposito, ci si limita a rilevare le criticità che discendono dalla scelta di demandare ad un secondo sistema, anch'esso artificiale, il compito di supervisionare il funzionamento di tecniche di intelligenza artificiale. Il controllo sarebbe, infatti, devoluto ancora una volta ad un sistema che vede solo una compartecipazione umana. Ancora una volta e anche in questo caso, si renderebbe cioè necessario verificare in che misura si costruisce il rapporto tra la persona e questa modalità di controllo del funzionamento dell'algoritmo e interrogarsi, a monte, sulla stessa possibilità non solo di controllare gli algoritmi ma, ed è questo il punto cruciale, che tale compito possa realmente essere assolto da una persona.

Si torna, così, al tema dei limiti della persona al cospetto della tecnologia e dell'esigenza di tracciare una linea di equilibrio tra i due, che non si risolva nell'esautoramento della prima dall'*agere* della seconda.

12. *Uno sguardo ai primi casi giudiziari tra Italia ed Europa*

Quanto ai casi, si assiste, anzitutto, ad accertamenti che reagiscono alla discriminazione algoritmica tentando di ricondurla entro i binari del diritto anti-discriminatorio classico[163].

In questa prospettiva, si inserisce la decisione del Tribunale di Bologna sul caso "Deliveroo"[164] che riguardava il meccanismo automatico di accesso, prenotazione e

[161] In tema, interessante richiamare che l'esigenza di sottoporre a controllo le tecniche di intelligenza artificiale, tramite meccanismi di *auditing*, è stata di recente sollecitata negli Stati Uniti da alcuni componenti del Congresso che hanno scritto una lettera ad alcuni Chief Officers di aziende *leader* nell'impiego delle tecniche di intelligenza artificiale (YouTube, Alphabet Inc., Google). Il testo integrale della lettera, datata 4 giugno 2021, è consultabile al link: https://clarke.house.gov/clarke-schakowsky-and-colleagues-pen-a-letter-to-google-demanding-an-audit-on-racial-equity/.

[162] Si rinvia allo studio di C. WILSON et al., *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, 2021, https://evijit.github.io/docs/pymetrics_audit_FAccT.pdf.

[163] Nelle more della presenta ricerca, sono sopraggiunte due ordinanze del Garante per la protezione dei dati personali del 10 giugno 2021 e del 22 luglio 2021 che si hanno dato applicazione all'art. 22 del GDPR in tema di profilazione.

[164] Tribunale di Bologna, ordinanza del 31 dicembre 2020. A commento, D. TESTA, *La discriminazione degli algoritmi: il caso Deliveroo, Trib. Bologna, 31 dicembre 2020*, in *IusinItinere.it*, 26 gennaio 2021. Un altro caso da segnalare, sebbene non direttamente incentrato sulla dimensione discriminatoria, è offerto dalla sentenza del Tribunale Amministrativo del Lazio, Sez. III *bis* n. 3769 del 2017 in relazione all'impiego di un *software*, incaricato di stabilire la sede di assegnazione di ciascun/a docente. A commento di questa seconda pronuncia, I. FORGIONE, *Il caso dell'accesso al software MIUR per l'assegnazione dei docenti - T.A.R. Lazio Sez. III bis, 14 febbraio 2017, n. 3769*, in *Giornale di diritto amministrativo*, 2018, 647 ss.; L. VIOLA, *L'intelligenza artificiale nel procedimento e nel processo, amministrativo: lo stato dell'arte*, in *Foro Amministrativo*, 2018, 1598 ss. Sullo stesso caso, anche la successiva pronuncia del Consiglio di Stato, Sez. VI, n. 2270 del 2019. In senso analogo e sempre in tema di conoscibilità e trasparenza dell'algoritmo, si vedano, sempre del Tribunale Amministrativo del Lazio, le sentenze n. 9224, n. 9225, n. 9226, n. 9227, n. 9228, n. 9229 e n. 9230 del 2018. Per un commento in prospettiva comparata con l'esperienza tedesca, si rinvia a N. FIANO, *La robotizzazione delle decisioni amministrative e della decisione giudiziale. Problematiche di diritto costituzionale in chiave comparata con la Germania*, in corso di pubblicazione, relazione svolta in occasione del seminario, organizzato dall'associazione "Gruppo di Pisa",

cancellazione delle sessioni di lavoro da parte dei *riders* impiegati dalla società. Il *software* non consentiva, tra le altre cose, la cancellazione della prenotazione indipendentemente dalla ragione addotta dal lavoratore, che si vedeva pregiudicata la scelta del turno successivo anche qualora quest'ultima fosse stata motivata da ragioni connesse all'esercizio di diritti di rilievo costituzionale, come, in quel caso, il diritto di sciopero.

Nella sua decisione, il giudice ha censurato la "cecità" della macchina, ritenendo integrati gli estremi della discriminazione indiretta, escludendo qualsiasi intenzionalità e conoscibilità da parte di Deliveroo del funzionamento discriminatorio del *software*. Ipotesi, invero, poco probabile, tanto che meglio avrebbe potuto delineare i contorni della discriminazione censurata la citata fattispecie del *masking* oppure della *proxy discrimination*.

Interessa, però, rilevare, accanto alla qualificazione della condotta come indirettamente discriminatoria, la scelta del giudice di identificare il soggetto responsabile con l'utilizzatore finale della macchina, escludendo quindi il programmatore, condividendo opzioni già peraltro avallate dal giudice amministrativo[165].

Accanto alla scelta di interpretare unitariamente programma e utilizzatore, vi è un altro aspetto che merita di essere sottolineato. Si è detto che il Tribunale di Bologna qualifica la fattispecie come discriminazione indiretta. Il giudice fa perno però non solo sulla formulazione solo apparentemente neutra della regola e sull'effetto discriminatorio (*disparate*) della condotta, ma vi affianca l'elemento della conoscibilità, cioè ritiene che Deliveroo Italia fosse a conoscenza del funzionamento della piattaforma in senso lesivo dell'eguaglianza dei lavoratori. Non arriva a dire che parte convenuta "voleva" discriminare, non era questo il caso, e tuttavia chiama in causa un elemento soggettivo riferibile alla *voluntas* dell'agente. Ci dice, cioè, che Deliveroo sapeva che la piattaforma discriminava, ma che se ne è disinteressata. Questa commistione tra elementi oggettivi e soggettivi è interessante perché sfuma la distinzione tra discriminazione diretta e indiretta confermando la tesi della tipicità della discriminazione "artificiale"[166].

Diritto e nuove tecnologie tra comparazione e interdisciplinarità, 26 marzo 2021. Sul tema del ruolo assolto dall'algoritmo nell'ambito della giustizia amministrativa, C. NAPOLI, *Algoritmi, intelligenza artificiale e formazione della volontà pubblica: la decisione amministrativa e quella giudiziaria*, in *Rivista AIC*, 2020, 318 ss.; N. MUCIACCIA, *Algoritmi e procedimento decisionale: alcuni recenti arresti della giustizia amministrativa*, in *Federalismi.it*, 2018, 1 ss.

[165] Nonostante si tratti di due casi differenti, è però interessante richiamare quanto affermato dal Tribunale Amministrativo del Lazio nel caso sopra citato sulla mobilità degli insegnanti, nella parte in cui indugia sulle relazioni tra *software* e utilizzatore e, in particolare, si sofferma sulla natura del primo. Afferma, così, il giudice amministrativo: «è con il software che si concretizza la volontà finale dell'amministrazione procedente che costituisce, modifica o estingue le situazioni giuridiche individuali anche se lo stesso non produce effetti in via diretta all'esterno. Il software finisce per identificarsi e concretizzare lo stesso procedimento». La identificazione suggerita tra la macchina e l'uomo potrebbe essere utilmente valorizzata nella prospettiva del superamento delle criticità in punto di accertamento della responsabilità della condotta, nel quadro di questo studio, di tipo discriminatorio. Allo stesso tempo, però, questa lettura non tiene, forse, in adeguato conto che l'azione umana si inserisce anche a monte, cioè nella fase di programmazione del programma, cosa che sconsiglierebbe una piana identificabilità tra macchina e utilizzatore finale.

[166] Si precisa, tuttavia, che il diritto anti-discriminatorio a livello europeo sta iniziando a conoscere e a scrutinare ipotesi di discriminazione indirette intenzionali, attestando, cioè la ricorrenza di una discriminazione indiretta anche in presenza di un intento discriminatorio. Si tratta di un'interpretazione che

In altri casi, e volgendo lo sguardo al panorama europeo, più che sindacare la disparità di trattamento e soffermarsi sulla tipologia di discriminazione, i giudici hanno sanzionato la non conoscibilità del funzionamento del sistema da parte dell'utilizzatore finale.

Così, nel caso deciso dalla Corte distrettuale dell'Aja[167] nel marzo del 2020, con riferimento ad una tecnica di intelligenza artificiale, "SiRI" (*Systeem Risicoindicatie*), utilizzata dal governo dei Paesi Bassi per identificare alcune forme di frode ai danni della pubblica amministrazione, ritenuta dalla Corte contraria all'art. 8 CEDU. In questo caso, pur non incentrandosi su violazioni esplicite del principio di non discriminazione, il giudice ha indugiato sulle potenzialità discriminatorie di sistemi preposti a stabilire attitudini comportamentali individuali, aggravate dalla difettosa conoscenza dei dati e del funzionamento della macchina.

In senso analogo, la Corte d'Appello dell'Inghilterra e del Galles[168] in un caso che riguardava un *software* di riconoscimento facciale utilizzato dalle forze di polizia del sud del Galles per ragioni di *law enforcement* e di sorveglianza pubblica, denominato "*AFR Locate*", in cui la Corte inglese ha condannato le forze di polizia, poiché ignoravano le modalità di funzionamento del sistema. Il giudice di secondo grado ha rilevato, in particolare, che le forze di polizia non avevano prodotto in giudizio prove sufficienti a dimostrare che il *software* fosse scevro da pregiudizi a sfondo etnico-razziale o sessuale, precisando che detta carenza probatoria è da imputare all'omesso accesso al *data-set*. L'impossibilità di avere accesso ai dati costituisce, quindi, l'elemento su cui poggia lo scrutinio sull'accertamento della responsabilità dell'utilizzatore. Potrebbe dirsi, cioè, che, ai fini della responsabilità dell'utilizzatore di una tecnica di intelligenza artificiale, può essere sufficiente dimostrare che quest'ultimo non fosse a conoscenza dei dati forniti alla macchina. Pur non soffermandosi sulla tipologia di discriminazione, la decisione si segnala, perché costituisce uno dei primi esempi di sindacato sul carattere discriminatorio di un sistema di riconoscimento facciale e perché la Corte inglese pone l'accento su un elemento che potrebbe costituire una prima linea guida per future applicazioni giurisprudenziali: la doverosa conoscenza e conoscibilità del meccanismo di funzionamento della macchina e dei dati. Basta questa omissione, perché sussista la responsabilità dell'utilizzatore a prescindere dal suo coinvolgimento nella fase di ideazione e costruzione del sistema.

Ultima pronuncia, significativa perché affronta direttamente il tema della discriminatorietà di una tecnica di intelligenza artificiale, è stata definita dal *National*

---

allarga le maglie della nozione classica di discriminazione indiretta, che nasce nella costruzione dogmatica come effetto accidentale di una condotta, viceversa, non volutamente discriminatoria. In dottrina, sul dibattito attuale, H. COSSETTE-LEFEBVRE, *Direct and Indirect Discrimination*, in *Public Affairs Quarterly*, 2020, 340 ss.

[167] Corte distrettuale dell'Aja, n. C/09/550982 /HAZA 18-388, 5 febbraio 2020. Il caso è stato promosso da alcune associazioni della società civile, inclusa la *Dutch Section of the International Commission of Jurists (NJCM)* e due privati cittadini. Il testo della pronuncia può essere letto al link: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878.

[168] Corte d'Appello dell'Inghilterra e del Galles, *R (Bridges) -v- CC South Wales & ors*, 11 agosto 2020. La decisione può essere letta al link: https://www.judiciary.uk/wp-content/uploads/2020/08/R-Bridges-v-CC-South-Wales-ors-Judgment.pdf.

Non-discrimination and Equality Tribunal finlandese[169], che ha accertato la ricorrenza di una discriminazione diretta e multipla ai danni di un uomo che si era visto negare un prestito da un istituto di credito. La sentenza ha il pregio di aver posto, per la prima volta, l'accento sulle specificità della disparità di trattamento, che viene accostata alla discriminazione statistica ritenuta dal giudice più aderente ai tratti di quella "artificiale".

13. *La giurisprudenza costituzionale tra eguaglianza, ragionevolezza e automatismi: quale spazio per lo scrutinio sulla "discriminazione algoritmica"?*

Assente, almeno sinora, dalla scena è la Corte costituzionale. Non sono state sollevate questioni di costituzionalità che investano norme di legge che impongono il ricorso a tecniche di intelligenza artificiale ed altrettanto scarna è la giurisprudenza delle Corti Supreme di altri ordinamenti, come dimostra l'esperienza comparata.

Oltre le ragioni di siffatta scarsa attivazione delle Corti costituzionali, nazionali e sovranazionali[170], è però opportuno ricordare i binari lungo cui si è mossa la giurisprudenza costituzionale quando è stata chiamata a pronunciarsi in relazione a norme di legge costruite in modo rigido, che stabilivano regole da applicarsi in via automatica al ricorrere di predeterminate condizioni.

Ci si riferisce alla nota giurisprudenza costituzionale in materia di automatismi legislativi[171] che, negli ultimi decenni, ha visto il Giudice costituzionale censurare in più di una occasione, perché irragionevoli[172], norme rigide che impedivano al giudice la

[169] Il riferimento è alla decisione del *National Non-discrimination and Equality Tribunal of Finland*, n. 216/2017, 21 marzo 2018. Il testo della pronuncia può essere letto al link: https://www.yvtltk.fi/material/attachments/ytaltk/tapausselosteet/45LI2c6dD/YVTltk-tapausseloste-_21.3.2018-luotto-moniperusteinen_syrjinta-S-en_2.pdf.

[170] Non si conoscono ad oggi casi decisi dalla Corte europea dei diritti dell'uomo che si occupino di intelligenza artificiale e delle sue ricadute sul principio di non discriminazione. Viceversa, la Corte di Strasburgo si è trovata ad occuparsi più di frequente di casi che riguardavano l'accesso ad internet e l'intreccio tra gestione della rete e diritto di libera manifestazione del pensiero *ex* art. 10 CEDU.

[171] Il riferimento è, anzitutto, alla giurisprudenza costituzionale che, in materia di adozione, si è opposta alla previsione di una differenza di età fissa tra adottato e adottanti (Cfr. Corte cost. sent. n. 303 del 1966; n. 140 del 1990; n. 148 del 1992; n. 283 del 1999); alle sentenze in cui la Corte costituzionale ha riscritto la norma nel senso di consentirne un'applicazione rispondente al superiore interesse del minore (Cfr. Corte cost. sent. n. 31 del 2012; n. 7 del 2013). La Corte ha poi, in diverse occasioni, censurato la norma oggetto per irragionevolezza a motivo della sua applicazione automatica in materia penale. Si richiamano, le decisioni con cui la Corte costituzionale ha censurato la presunzione assoluta di adeguatezza della misura della custodia cautelare in carcere in relazione ad alcune categorie di delitti (Cfr. Corte cost. sent. n. 265 del 2010; n. 164, n. 231, n. 331 del 2011; n. 110 del 2012; n. 57, n. 213, n. 232 del 2013). Ancora, può richiamarsi, la decisione costituzionale n. 151 del 2009 che, in tema di fecondazione medicalmente assistita di tipo omologo, ha censurato l'applicazione rigida della norma della legge n. 40 del 2004, *Norme in materia di procreazione medicalmente assistita*, che limitavano a tre il numero massimo di embrioni fecondabili e destinati ad un unico e contemporaneo impianto nell'utero della donna. In dottrina, S. LEONE, *Automatismi legislativi, presunzioni assolute e bilanciamento*, in *Rivista Gruppo di Pisa*, cit.; E. CRIVELLI, *Gli automatismi legislativi nella giurisprudenza della Corte costituzionale*, in D. BUTTURINI, M. NICOLINI (a cura di), *Tipologie ed effetti temporali delle decisioni di incostituzionalità*, Napoli, 2014, 85 ss.; L. PACE, *Gli automatismi legislativi nella giurisprudenza costituzionale*, in *Rivista Gruppo di Pisa*, 2014, 1 ss.

[172] Sul principio costituzionale di eguaglianza e ragionevolezza, per tutti, A. CERRI, *Ragionevolezza delle leggi*, in *Enciclopedia Giuridica*, Roma, 2005, XXV, 1 ss.; L. PALADIN, *Il principio costituzionale*

valutazione del caso concreto. Si tratta di una giurisprudenza, che se letta nel prisma di questo studio, evidenzia la sanzione di meccanismi automatici, ove il ruolo di decisore finale della persona cede il passo alla regola astratta e presuntivamente giusta, valorizzando, all'opposto, l'abilità *umana* di piegare la regola al cospetto della realtà *umana*.

Il meccanismo che sorregge tali norme, automatico e insuscettibile di rettificazione al verificarsi di circostanze diverse da quelle legislativamente predefinite, ricorda molto da vicino quello che soggiace al funzionamento delle tecniche di intelligenza artificiale. Si è osservato che le tecniche di intelligenza artificiale sono ontologicamente preposte a distinguere e che "a fare la differenza" è la modalità, il "come", è operata la scelta. La capacità di una macchina di inchinarsi al cospetto della eterogeneità della realtà esterna potrebbe anche rendere recessivo il problema del "come" differenziano le tecniche di intelligenza artificiale, essendo una caratteristica intrinsecamente umana la capacità di intravedere dettagli e di adattare la regola alle peculiarità del caso.

L'accostamento tra l'automatismo legislativo, sanzionabile poiché incostituzionale, e il funzionamento di alcune tecniche di intelligenza artificiale appare, quindi, utile almeno allo scopo di interrogarsi sulla conformità a Costituzione di strumenti che sostituiscono la persona e che compiono scelte, in modo rigido ed automatico, che se facesse la prima sarebbero esposte a censure di costituzionalità. In questo stesso senso, si sostiene che la «concretezza»[173] e il «pluralismo»[174] costituiscono i pilastri dello Stato costituzionale che si oppone, a norme di diritto inadeguate a riflettere i mutamenti esterni, esautorando la persona dal processo. Così come la giurisprudenza costituzionale esclude l'applicazione di norme rigide, che prevedono che al ricorrere di una data condizione vi debba seguire sempre e comunque la conseguenza prestabilita, perché l'unica corretta[175], si espongono ad analoghe criticità quei meccanismi di intelligenza artificiale che parimenti non consentono l'adattabilità della macchina al caso concreto.

In entrambe le ipotesi, volendo guardarla dal lato del ruolo assolto dalla persona, potrebbe sostenersi che la giurisprudenza costituzionale è poco incline a relegare quest'ultima nelle retrovie, preferendo che le scelte si realizzino per via automatica; e ciò, soprattutto a fronte della trasformazione dello Stato costituzionale e, con esso, del principio di eguaglianza che è sempre più tutela delle differenze[176].

Ancora, che difficilmente la tutela dei diritti fondamentali può essere rimessa in via esclusiva a meccanismi automatici, siano essi il frutto di norme oppure di tecniche di intelligenza artificiale, che escludono la componente umana e il suo diritto di

---

d'eguaglianza, Milano, 1965 e, dello stesso A., anche *Ragionevolezza (principio di)*, in *Enciclopedia del Diritto*, Milano, 1997, 899 ss.; A. MORRONE, *Il custode della ragionevolezza*, Milano, 2001.

[173] T. GROPPI, *Alle frontiere dello stato costituzionale: innovazione tecnologica e intelligenza artificiale*, cit., 682, che precisa in proposito che «le decisioni che riguardano le persone umane, al plurale, debbono essere assunte 'in concreto', tenendo conto della irripetibile specificità di ciascuno».

[174] *Ibidem.*

[175] Per uno studio della giurisprudenza costituzionale in tema di ragionevolezza, F. MODUGNO, *La ragionevolezza nella giustizia costituzionale*, Napoli, 2007; M. CARTABIA, *I principi di ragionevolezza e proporzionalità nella giurisprudenza costituzionale italiana*, 2013, consultabile al link: https://www.cortecostituzionale.it/documenti/convegni_seminari/RI_Cartabia_Roma2013.pdf.

[176] In questo senso, G. ZAGREBELSKY, V. MARCENÒ, *Giustizia costituzionale*, Bologna, 2013, 213.

autodeterminarsi. Non a caso, la Costituzione si regge sul principio personalista e così come pone lo Stato al servizio della persona e non viceversa, l'ingresso nello spazio pubblico – e, quindi, costituzionale – della dimensione artificiale non può che portare con sè profili di dubbia conformità al disegno costituzionale, di cui la irragionevolezza lesiva dell'art. 3, comma 1, Cost., è solo uno degli aspetti da cui affrontare il problema.

Conclusioni: L'intelligenza artificiale ha "cambiato" la discriminazione?

L'intelligenza artificiale ha, quindi, cambiato la discriminazione?

Sì e no.

Sì, perché la discriminazione non è più fenomeno solo umano, ma nasce dalla macchina oppure, più spesso, dall'azione congiunta, umana e artificiale, che si uniscono soprattutto laddove la tecnologia non si sostituisce alla persona ma si limita a supportarne l'azione.

Il mutamento, o meglio, lo spostamento dalla dimensione umana a quella artificiale della macchina, si coglie guardando alle origini del "nuovo" fenomeno discriminatorio, mentre gli effetti della condotta continuano a riverberarsi solo sulla persona, mantenendo salda la componente umana. Quasi a dire, che ciò che cambia nella discriminazione "artificiale" interessa il "*come*" conosce realizzazione la differenziazione irragionevole, più che i suoi effetti, giustificando quel "sì e no", qui soprattutto il "no", con cui si aprono queste riflessioni conclusive.

La conquista dello spazio umano da parte della macchina ha, però, eroso le categorie del diritto anti-discriminatorio. Ha sfumato la distinzione tra discriminazione diretta e indiretta, rendendo la seconda incapace di sopperire alle carenze della prima; ha dato ingresso ad un elemento di divisione nuovo, il *proxy*, su cui innesta la nuova tipologia discriminatoria della *proxy discrimination*. Sempre dal lato oggettivo, si è smarrita l'unicità della condotta, risolvendosi, più spesso, in una molteplicità di atti che si sommano, si intersecano, si costruiscono l'uno sull'altro sino a produrre l'effetto discriminatorio.

L'intermediazione della macchina nello spazio lasciato scoperto tra l'azione discriminatoria e i suoi effetti si apprezza anche sul versante soggettivo. La discriminazione "artificiale" ha travolto le nozioni di appartenenza e di identità individuale, ridisegnando i contorni di categorie e gruppi; ha riletto i fattori di discriminazione classici, sostituiti o accostati dal *proxy*.

La discriminazione "artificiale" non ha, però, solo sfidato le categorie del diritto anti-discriminatorio.

La costruzione di meccanismi automatici di affiliazione ha prodotto ripercussioni anche sui diritti costituzionali: sull'eguaglianza e la non discriminazione, sul diritto

individuale di scegliere se fare o non fare parte[177] di una formazione sociale[178]; sul principio sovranazionale di auto-identificazione.

L'intelligenza artificiale non ha, invece, mutato le cause, umane, della discriminazione, che si colorano tuttavia di un tasso di inconsapevolezza e involontarietà dimostrato dalla centralità del *bias* implicito.

Alla tipicità della discriminazione "artificiale" fa allora da contraltare la scarsa pregnanza del diritto antidiscriminatorio, imponendosi la ricerca di un inquadramento teorico nuovo di diritto positivo e lo studio di strategie processuali che sappiano intercettare la discriminazione anche quando prodotto di meccanismi artificiali.

Poiché la causalità è spesso difficile da dimostrare, si potrebbe preferire, alla ricerca della intenzionalità che sostiene l'azione dell'agente, quella dei soli effetti della condotta. Fintantoché questi discriminano e in virtù della opacità della discriminazione "artificiale", poco dovrebbe importare l'individuazione del chi – se esiste un "chi" – voleva intenzionalmente discriminare. Questa ipotesi potrebbe colmare l'inapplicabilità della teoria della discriminazione diretta, ma non reagirebbe in modo adeguato alla causa primigenia della discriminazione "artificiale".

Oltre il controverso impiego della tecnologia di *algorithm auditing*, la letteratura discute dell'impiego di azioni positive per correggere algoritmi discriminatori, secondo approcci più o meno sensibili alla cruciale e ineliminabile dimensione intersezionale[179].

L'intelligenza artificiale ha, quindi, cambiato la discriminazione? Sì, ma solo in parte.

Vi è, infatti, un elemento che rimane immutato, la sua causa primigenia, ancora tutta umana, della discriminazione. Se così è, un'opzione che guardi ai soli effetti, disinteressandosi del dato soggettivo, condurrebbe a soluzioni parziali, incapaci di incidere sul pregiudizio, individuale e collettivo, che continua ad affliggere la persona nelle sue relazioni *inter*-individuali con il mondo esterno.

Si tratterà, allora, "solo" di attendere e di verificare se a prevalere sarà la persona (e i suoi diritti) oppure la macchina.

---

[177] Sull'esigenza che la formazione sociale, e l'appartenenza individuale al gruppo non si risolva in una lesione dei diritti dell'appartenente, A. BARBERA, *Art. 2*, in G. BRANCA (a cura di), *Commentario della Costituzione*, Zanichelli, Bologna, 1975, 50 ss., cit. 113 e ss.; F. PIZZOLATO, *Formazioni e ... deformazioni sociali*, in *Quaderni costituzionali*, 2005, 137 ss. Si consenta, infine, il rinvio anche a C. NARDOCCI, *Razza e etnia. La discriminazione tra individuo e gruppo nella dimensione costituzionale e sovranazionale*, cit.

[178] La dottrina costituzionalistica si è diffusamente occupata del ruolo dei corpi intermedi nel loro rapporto con gli individui che ne siano parte. Per tutti, C. MORTATI, *La persona, lo Stato e le comunità intermedie*, Torino, 1971; P. RESCIGNO, *Persona e Comunità*, Padova, 1966; G. LOMBARDI, *Potere privato e diritti fondamentali*, Torino, 1970.

[179] Così J.R. BENT, *Is algorithmic affirmative action legal*?, in *The Georgetown Law Journal*, 2020, 803 ss.; A. XIANG, *Reconciling legal and technical approaches to algorithmic bias*, in *Tennessee Law Review*, 2021, 3 ss.