LIBER AMICORUM
PER
PASQUALE COSTANZO

ELETTRA STRADELLA

STEREOTIPI E DISCRIMINAZIONI: DALL'INTELLIGENZA UMANA ALL'INTELLIGENZA ARTIFICIALE

30 MARZO 2020

CONSULTA ONLINE
PERIODICO TELEMATICO ISSN 1971-9892

LIBER AMICORUM
PER
PASQUALE COSTANZO

Elettra Stradella

Stereotipi e discriminazioni: dall'intelligenza umana all'intelligenza artificiale

SOMMARIO: 1. Intelligenza Artificiale e stereotipi. – 2. L'apparente neutralità della tecnologia. – 3. La neutralità dei dati: cosa rappresentare e cosa non rappresentare. Dati che esistono e dati ... "che non esistono. – 4. È sufficiente il riferimento al concetto di *bias* per portare nell'intelligenza artificiale equità e giustizia? – 5. Stereotipi e discriminazioni dall'intelligenza naturale (nell'argomentazione giudiziaria) all'intelligenza artificiale

1. *Intelligenza Artificiale e stereotipi*

Prima di intraprendere la riflessione sul (rapido) passaggio dagli stereotipi alle discriminazioni nei sistemi di intelligenza artificiale, cercando di individuarne l'impatto sui principi fondamentali ed in particolare la rilevanza in termini di eguaglianza e di dignità, occorre una premessa.

La trasformazione dello spazio nel quale i comportamenti umani possono trovare collocazione e svolgimento, rappresentata dall'avvento della Rete, è stata osservata attraverso tre possibili lenti: quella ottimista, o utopica, secondo la definizione che qualcuno ne ha offerto[1], quella distopica[2], conservativa, a tratti apocalittica, e una lente più razionale, rappresentata da quei *cyber-realists* che si sono posti e si pongono soprattutto il problema della regolazione[3].

Tutte le riflessioni sulla Rete prendono in fondo le mosse da un'esigenza di tutelare la libertà, di fronte agli sconvolgimenti che "il costituzionalismo sta subendo per effetto del progresso tecnologico di portata globale"[4], sia in una prospettiva che legge la tecnologia come strumento per una piena realizzazione della libertà d'informazione, nella sua declinazione attiva e in quella passiva (libertà di informare o libertà di essere informati), sia nella prospettiva che si potrebbe definire maggiormente oppositiva, o protettiva, che si concentra sul diritto di controllare il trattamento informatizzato dei propri dati personali, quello che è stato definito "*habeas data*", e che ha contribuito a costruire una rinnovata idea di *privacy*, in cui si riassume un insieme di diritti relativi principalmente ad una gestione dei propri dati orientata alla tutela del diritto all'identità personale, e, dunque della dignità, personale e sociale[5].

Questa seconda prospettiva segna anche la riflessione giuridica sull'Intelligenza Artificiale (da ora *AI*), che si concentra ancora di più sull'esigenza di un utilizzo, e finanche di una progettazione,

---

[1] Cfr. L. LESSIG, *Code 2.0*, New York 2006, J.M. BALKIN, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, in *New York University Law Review*, 79, n. 1/2004. In Italia, cfr. da ultimo, M. MONTI, *Introduzione: la disinformazione online e il ruolo degli esperti nell'agorà digitale,* in ID. (a cura di), *Special Issue* di *Federalismi.it*, in corso di pubblicazione, e all'interno del fascicolo cfr. P. PASSAGLIA, *Fake news e fake democracy: una convergenza da scongiurare*, spec. 7-8.

[2] Si vedano le considerazioni espresse in Cfr. A. MORELLI - O. POLLICINO, *Le metafore della rete. Linguaggio figurato, judicial frame e tutela dei diritti fondamentali nel cyberspazio: modelli a confronto*, in *Rivista AIC*, n. 1/2018.

[3] Cfr. D. FREEDMAN, *The Internet of Rules: Critical Approaches to Online Regulation and Governance*, in J. Curran – N. FENTON – D. FREEDMAN (a cura di), *Misunderstanding the Internet*, New York, 2016, e anche J.M. BALKIN, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, in *UC Davis Law Review*, 51, n. 3/2018. Questione della regolazione che si pone naturalmente anche oggi per l'*AI*, da ultimo si veda proprio il contributo in questo *Liber Amicorum* di A. CELOTTO, *Algoritmi e algoretica: quali regole per l'intelligenza artificiale?*, 26 marzo 2020. Per uno stato dell'arte sul dibattito e le proposte di regolazione dell'*AI* sia consentito di rinviare a E. STRADELLA, *La regolazione della Robotica e dell'Intelligenza artificiale: il dibattito, le proposte, le prospettive. Alcuni spunti di riflessione*, in *MediaLaws*, n. 1/2019, 73 ss.

[4] Cfr. P. COSTANZO, *Il fattore tecnologico e le sue conseguenze*, Convegno annuale AIC, Salerno, 23-24 novembre 2012 "Costituzionalismo e globalizzazione", in *www.associazionedeicostituzionalisti.it*, 2012, 2.

[5] Si devono, com'è noto, proprio a Pasquale Costanzo, le prime fondamentali riflessioni sulle trasformazioni del diritto all'informazione, nelle sue diverse sfaccettature, determinate dalla Rete, e sull'impatto della stessa sui diritti fondamentali, cfr. in particolare P. COSTANZO, *Internet (diritto pubblico)*, in *Digesto Quarta Edizione (Discipline pubblicistiche), Appendice*, Torino, UTET, 2000, e ID., *Miti e realtà dell'accesso a internet (una prospettiva costituzionalistica)*, in *Studi in memoria di Paolo Barile*, Passigli Editore, Firenze, 2012.

orientati allo sviluppo umano e alla garanzia del principio personalistico, esigenza "amplificata" dalle caratteristiche di questi sistemi.

Spesso quando si ragiona di applicazioni dell'*AI* in ambito - genericamente - giuridico, sia esso quello normativo o quello processuale, si concentra l'attenzione sul significato dell'algoritmo e sulle caratteristiche dei sistemi di apprendimento automatico.

Ma ciò che non deve sfuggire è che essi, affinché possano funzionare, hanno bisogno di dati "annotati" (*supervised learning*) o per lo meno selezionati e preparati (*unsupervised learning*), dall'essere umano oppure tramite sistemi automatici di produzione della fonte di informazione.

Il tema centrale dunque, non soltanto per i *data scientists* e per tutti i tecnologi che operano nell'ambito delle tecnologie dell'informazione, ma anche per chi, sul versante del diritto e delle politiche pubbliche, intenda investigare la possibilità per le tecnologie di contribuire alla realizzazione di una società più equa e giusta, dove sono garantiti i diritti fondamentali, non è solo o tanto quello dell'algoritmo, ma quello dei dati, e dunque della creazione di condizioni che consentano all'*AI* di utilizzare basi di dati costruite correttamente.

Da questo punto di vista il problema sta per un verso nella genesi dei dati, che può essere umana, tecnologica, ma anche mista, per altro verso nell'organizzazione stessa dei dati, che può e generalmente deve strutturarsi, al fine del loro più efficiente utilizzo, secondo modalità che emulano il ragionamento umano: come altre tecnologie, infatti, anche l'*AI* è *bio-inspired* e *bio-oriented*[6].

Nelle fonti, dunque, si annida il *vulnus* potenzialmente intrinseco ai sistemi che si fondano sull'*AI*, e che sarà oggetto di riflessione, quello dei *bias*. Se nell'approccio tecnologico il *bias* rappresenta un errore di valutazione, un concetto (anche, eventualmente, pre-concetto) che rischia di minare la correttezza e l'affidabilità dei risultati di un'analisi, nella prospettiva giuridica il *bias* rappresenta lo stereotipo pronto a trasformarsi in scelta discriminatoria, e per questo il principale bersaglio del principio di non discriminazione. Non è un caso che tale principio sia tra i più rilevanti che l'*European Commission for the Efficiency of Justice* (CEPEJ) ha recentemente messo in evidenza[7], prendendo le mosse dalla constatazione che proprio certi metodi di processamento dei dati tendono a rivelare discriminazioni che esistono e, potremmo dire, a cristallizzarle[8].

La questione che qui si intende porre è quella della riflessività degli stereotipi e delle discriminazioni, nel passaggio dalla generazione dei dati, spesso frutto dell'intelligenza e del comportamento umano, alla costruzione degli algoritmi. L'errore, come scrive Andrea Simoncini, di "derivare dall'essere il dover essere"[9], diviene particolarmente grave quando l'essere è fatto da una realtà sociale ingiusta, che tende a perpetrare diseguaglianze, e che quindi la cristallizzazione dell'ingiustizia nelle maglie dell'algoritmo rischia di normativizzare.

Gli esempi, i più risalenti nel tempo e i più recenti, ai quali ci si può riferire, confermano questa "ipotesi", ma nello stesso tempo, come si suggerirà in conclusione, possono indicare un mutamento di prospettiva.

Il più "storico" è quello relativo al *Franglen's Admissions Algorithm*[10]. Nel 1970 il dr. Fraglen a Londra inizia a scrivere un algoritmo per selezionare le *applications* degli studenti per l'ammissione

---

[6] Cfr. M.C. CARROZZA *et al.*, Statuto etico e giuridico dell'Intelligenza Artificiale dalla Fondazione Leonardo "Civiltà delle macchine", *AI: profili tecnologici. Automazione e Autonomia: dalla definizione alle possibili applicazioni dell'Intelligenza Artificiale*, in Rivista di BioDiritto, n. 3/2019, da cui emerge anche l'esigenza di un superamento della logica antagonistica dell'*AI*.

[7] Nel recente atto contenente una *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment*, adottata dalla CEPEJ nella 31ma assemblea generale (Strasburgo, 3-4 Dicembre 2018, CEPEJ(2018)14.

[8] E da questo punto di vista l'applicazione dell'*AI* in ambito giudiziario, e penale in particolare, mostra il rischio della riemersione di dottrine deterministiche che, per lo meno in Italia, sono definitivamente superate, anche nel diritto positivo processuale, attraverso l'art. 220, c. 2 del codice di procedura penale che stabilisce espressamente, com'è noto, l'inammissibilità di perizie criminologiche al fine della verifica della punibilità, si veda A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, in Rivista di BioDiritto, n. 1/2019.

[9] *Ibidem*, 86

[10] Recentemente studiato su *IEEE Spectrum*, O. SCHWART, *Untold History of AI: Algorithmic Bias Was Born in the 1980s*, IEEE Spectrum, April 2019.

alla scuola di medicina. All'epoca tre quarti dei candidati ogni anno venivano esclusi soltanto sulla base della loro domanda senza poter accedere all'*interview*. Al fine di rendere meno gravoso lo sforzo di selezionare un numero elevatissimo di domande, Franglen pensa di automatizzare quella fase della procedura, cercando di tradurre all'interno dell'algoritmo i processi che negli anni lui e i suoi colleghi avevano utilizzato per selezionare gli studenti.

Ma invece di produrre risultati migliori e più efficienti, l'algoritmo provoca danni enormi: il semplice fatto di avere un nome non europeo determinava infatti la perdita di 15 punti nella valutazione.

Dal passato al presente più prossimo, basti guardare al progetto *gendershades.org*[11], che ha mostrato come le tecniche di *machine learning* utilizzate per la classificazione di genere da parte di tre compagnie (IBM, Microsoft e Face++) presentano evidenti *bias* etnici e di genere. Ancora una volta il punto non è che il *machine learning* sbaglia, ma *come* sbaglia. Se l'*AI* è in grado di fallire così come lo è l'intelligenza naturale, il problema è che essa è altrettanto in grado di discriminare. In questo caso, tutte le compagnie hanno risultati migliori nel riconoscimento dei maschi che in quello delle femmine, con una differenza fino a circa il 21%. E tutte le compagnie hanno risultati migliori con soggetti di pelle più chiara piuttosto che su soggetti di pelle più scura, con una differenza fino a circa il 20%. E se si analizza la discriminazione intersezionale, si nota come tra il gruppo "maschi chiari" e il gruppo "femmine scure" si arrivi ad un *gap* di accuratezza di circa il 35%.

Conferme provengono anche dagli studi, su cui ci si soffermerà in seguito, che mostrano come gli algoritmi dei motori di ricerca tendano a rafforzare ideologie e sentimenti razzisti[12].

Come anticipato, l'analisi non può che condurre a sottolineare, quale esigenza avvertita peraltro dagli stessi tecnologi, la centralità della qualità dei dati, e la sua preminenza anche rispetto alla scelta della tecnica di *machine learning* o *deep learning* nella progettazione e nell'utilizzo dell'*AI*.

Ma fatto questo, l'approccio alla relazione tra intelligenza umana (naturale) e intelligenza artificiale potrebbe invero aprirsi ad una prospettiva diversa dal consueto richiamo alla necessità di una regolazione, che pur salvaguardando le opportunità non rinunci a porre limiti.

Infatti, constatato come il potenziale discriminatorio dell'*AI* altro non è che l'amplificatore degli stereotipi e delle discriminazioni esistenti[13] nel nostro linguaggio, nel nostro agire, nell'applicazione della nostra intelligenza, vale a dire della nostra capacità di attribuire significati alle cose e alle esperienze che ci circondano, sarebbe forse utile pensare a come utilizzare l'*AI* "contro" quello che non dovremmo essere.

Se sotto il profilo della *data science* questo significherebbe aumentare, e non limitare o controllare, il margine di autoapprendimento, dal punto di vista giuridico sembra aprirsi la possibilità di utilizzare l'*AI* al fine di individuare e rimuovere i *bias* insiti nell'intelligenza umana, e spesso in grado di annidarsi anche nell'intelligenza dei soggetti che, a vario titolo, sono chiamati a produrre il diritto: sia quello politico, sia quello giurisprudenziale.

Prima però di aprire alle potenzialità antidiscriminatorie dell'*AI*, e dunque ad un'applicazione costituzionalmente orientata, occorre approfondire tre aspetti: quello dell'apparente, solo apparente, neutralità della tecnologia; quello della pretesa neutralità dei dati, e dunque della presenza di spazi di invisibilità nei dati; e quello dell'utilità epistemologica, ma soprattutto in termini di politica del diritto, del concetto di *bias*.

---

[11] Cfr. J. BUOLAMWINI – T. GEBRU, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in *Proceedings of Machine Learning Research* 81/2018.

[12] Cfr. S. U. NOBLE, *Algorithms of Oppression. How Search Engines Enforce Racism*, New York University Press, 2018. Sulle criticità insite nelle logiche proprie dei motori di ricerca cfr. già P. COSTANZO, *Motori di ricerca: un altro campo di sfida tra logiche del mercato e tutela dei diritti?*, in *Diritto dell'internet*, 2006.

[13] Come sottolinea M. ONUOHA, *On algorithmic violence: Attempts at fleshing out the concept of algorithmic violence*, in *Github* Feb 22, 2018

2. *L'apparente neutralità della tecnologia*

Esattamente un anno fa, nel marzo 2019, Berlino ha celebrato la giornata della *Gender Pay Equality* offrendo sconti alle donne sui trasporti pubblici. Questi sconti sono stati forniti automaticamente, analizzando i volti delle persone che acquistavano i biglietti, attraverso un sistema di *Automatic Gender Recognition* (AGR), un sistema di *AI* attraverso il quale si tenta di inferire il genere del soggetto di una foto o di un video attraverso l'apprendimento automatico[14].

È evidente che un sistema del genere non può definirsi neutrale. Come ogni tecnologia integra *by design* e *by default* dei presupposti, ed in questo caso integra una specifica e non includente concezione di cosa sia il "genere", rimodellando, e cristallizzando, la definizione del genere, con evidenti conseguenze in particolare per persone *gender non conforming* e *trans*. Infatti, un sistema che utilizza AGR scatta una foto (o realizza un video), ed elabora gli elementi di somiglianza e differenza tra le caratteristiche del viso analizzato e quelle di un modello, classifica quel viso come "maschile" o "femminile", sulla base della struttura ossea, della pelle, della forma del viso. La qualificazione viene quindi inserita in un sistema che registra il genere, ne tiene traccia, e lo utilizza a fini pubblicitari, di comunicazione, di organizzazione degli spazi e controllo dell'accesso agli stessi (si pensi all'accesso ai servizi igienici).

Inferire il genere dalle caratteristiche somatiche è d'altra parte profondamente discriminatorio[15] e tale tecnologia è stata infatti criticata per l'impatto dirompente che produce rispetto al riconoscimento dell'identità di genere, oggi, com'è noto, nominata e tutelata anche nel nostro ordinamento a partire in particolare dalla sent. n. 221/2015 della Corte cost., confermata nel 2017, che superando una connessione biunivoca e necessaria tra morfologia del corpo e identità sessuale, ha consentito il riconoscimento giuridico, attraverso la rettificazione anagrafica, di soggettività in transizione di genere per le quali il sesso anatomico non trova corrispondenza nella dimensione identitaria[16].

Sistemi di questo tipo non soltanto possiedono una significativa carica discriminatoria, ma sono in grado di intervenire nella definizione stessa del genere; una volta integrati all'interno dell'ordinamento giuridico, ovvero nelle strutture di potere economico e sociale, agiscono sull'oggetto della misurazione trasformandolo, ovvero legittimandolo.

Se gli strumenti dell'*AI* in generale sono oggi parte integrante del nostro modo di conoscere e di comprendere la realtà, e "naturalizzandosi" conformano la realtà, attraverso il loro linguaggio, i

---

[14] Questo sistema è integrato nei servizi di riconoscimento facciale venduti da grandi aziende tecnologiche come Amazon e IBM ed è stato utilizzato per la ricerca accademica, ma anche per il controllo degli accessi a strutture pubbliche e private e le profilazioni a fini pubblicitari.

[15] Cfr. O. KEYES, *The Body Instrumental*, in Logic, n. 9/2019

[16] Dal 2015, sia la giurisprudenza costituzionale sia la giurisprudenza di legittimità sono intervenute ad affermare definitivamente quella che era un'interpretazione minoritaria sulla facoltatività degli interventi chirurgici, ponendo al centro dell'argomentazione il diritto all'autodeterminazione e alla salute, nel suo significato più olistico. Su Corte cost. 221/2015, cfr. A. LORENZETTI, *Corte costituzionale e transessualismo: ammesso il cambiamento di sesso senza intervento chirurgico ma spetta al giudice la valutazione*, in *Quaderni costituzionali*, n. 4/2015, 1006 ss.; e si vedano poi sia la sent. n. 180/2017, sia l'ord. 185/2017, su cui ID., *Il cambiamento di sesso secondo la Corte costituzionale: due nuove pronunce (nn. 180 e 185/2017)*, in *Studium Iuris*, 4/2018, 446 ss., ID., *I corpi transessuali di fronte al diritto: alla ricerca di una faticosa rielaborazione degli impliciti normativi in materia di cambiamento di sesso*, in *Rivista di sessuologia*, n. 2/2016, 157 ss. e ID., *Il cambiamento di sesso anagrafico e le sue condizioni: la necessità o meno dell'intervento chirurgico. Brevi riflessioni sulla situazione attuale e sui prossimi sviluppi*, in *GenIUS*, n. 1/2015, 174 ss. Lorenzetti ritiene che le pronunce abbiano comunque confermato la sussistenza in capo al giudice di un potere decisionale, esercitato attraverso il supporto dei sanitari, in grado di limitare ancora lo spazio di autodeterminazione della persona. Da ultimo cfr. G. M. NEGRI, *Il percorso di riconoscimento di genere. I suoi profili applicativi tra norma e prassi*, in E. STRADELLA (a cura di), *Le discriminazioni fondate sull'orientamento sessuale e sull'identità di genere,* Pisa, Pisa University Press, 2019. Il tema è peraltro al centro di un intenso dialogo tra corti, si veda da ultimo L. TRUCCO, *Dialogo tra corti e diritti LGBT (a margine della "opinión consultiva" della Corte IDU "OC-24 de 24 de noviembre de 2017"),* in R. ROMBOLI - A. RUGGERI (a cura), *Corte europea dei diritti dell'uomo e Corte interamericana dei diritti umani: modelli ed esperienze a confronto*, Torino, Giappichelli, 2019, 369 ss.

meccanismi di AGR sono particolarmente significativi per la capacità di modellare il genere misurandolo, esemplificativi del modo di operare proprio del *machine learning*, che conduce verso una conoscenza umana della realtà che si definisce sulla scorta di come lo strumento tecnologico descrive la realtà medesima.

Anche per questa ragione gli algoritmi di riconoscimento facciale sono attualmente considerati come applicazioni dell'*AI* tra le più critiche rispetto alla protezione dei diritti fondamentali[17].

3. *La neutralità dei dati: cosa rappresentare e cosa non rappresentare. Dati che esistono e dati ... "che non esistono"*

I pregiudizi ancora fortemente radicati nei sistemi sociali vengono replicati nei set di dati[18]. Tutto questo accade anche, forse principalmente, per due ordini di ragioni: la prima risiede nelle caratteristiche soggettive di chi è chiamato a progettare i sistemi, poiché la maggior parte dei *team* di ingegneri e *data scientists* non comprende (o ne comprende pochissime) donne e persone appartenenti a minoranze etniche, e perché non esiste pressoché alcuna formazione, nel campo delle scienze c.d. dure, sul significato della *diversity*.

Ma perché la tecnologia, e l'*AI* in particolare, non possono essere neutri? Perché la mancanza di neutralità è propria dei dati: la circostanza stessa che un dato, o, meglio, un set di dati, rappresenti o meno qualcosa, già mette in discussione la presunta neutralità. L'astrattezza della scienza dei dati, che apparentemente trascende dai corpi, per così dire *body blind*, produce così quelle discriminazioni che lo stesso diritto, quando è stato costruito come indifferente ai corpi, alla materia, alle differenze presenti nell'umanità (basti pensare ai danni prodotti da un costituzionalismo che si è fondato sulla sua natura *colour-blindness*), ha generato, e per questa ragione il diritto è chiamato ad investigarla e indirizzarla.

La scienza dei dati, infatti, dipende fortemente dai corpi, benché non si vedano: si basa su di essi come fonti di dati e per prendere decisioni sui dati. Se oggettività o neutralità hanno un significato abbastanza definito dal punto di vista giuridico (pur nelle maglie sempre variabili dell'interpretazione), essi assumono nel rapporto con la scienza dei dati la pericolosa ambiguità di presentare come indipendente da condizionamenti un dato, o un risultato, che invece è frutto di una molteplicità fattori che interagiscono tra loro.

Negare l'oggettività dei dati non significa rinunciare al loro utilizzo o sottovalutare l'importanza che essi rivestono nell'elaborazione di politiche virtuose, ma riconoscere le diseguaglianze che le pratiche di raccolta e trattamento dei dati incorporano.

Se l'eguaglianza parte dai corpi, la potenziale discriminatorietà dei dati si evince dal numero e dalle qualità dei molti corpi non contati, e silenziati. Donne, persone appartenenti a minoranze "razziali", disabili, persone transgenere, *insular minorities* non costruiscono i dati e anche per questo i dati tendono spesso conformarsi al potere, a diventarne specchio per un verso, cassa di risonanza per altro. Com'è stato scritto di recente, sono le persone e i loro corpi che possono dirci quali dati li aiuteranno a migliorare la vita e quali dati li danneggeranno[19]. Il problema della presunta neutralità dei dati è ben evidente in numerose applicazioni dell'*AI*, e in tutte mostra l'impatto dirompente sui diritti costituzionalmente garantiti. Si pensi a quello alla salute, ambito in cui la presenza di *bias* derivanti dai dati rischia di produrre effetti di particolare penalizzazione dei

[17] Cfr. EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS, *Facial Recognition Technology: Fundamental Rights Consideration in the Context of Law*, Report 2020.

[18] Tornando all'esempio degli algoritmi di analisi facciale, scavando nei dati di *benchmarking* per questi algoritmi, Buolamwini mostra che consistevano nel 78% di volti maschili e dell'84% di volti chiari, nettamente in contrasto con una popolazione globale che è per lo più femminile e di pelle non chiara, cfr. J. BUOLAMWINI – T. GEBRU, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, cit.

[19] Cfr. C. D'IGNAZIO – L. F. KLEIN, *Data Feminism*, MIT Press, 2019, *passim.*

soggetti in relazione a caratteristiche personali ad alto grado di vulnerabilità e stereotipizzazione, quali l'identità sessuale o di genere, l'origine etnica, l'appartenenza religiosa.

È il caso degli algoritmi di previsione utilizzati negli Stati Uniti per l'individuazione di pazienti con esigenze sanitarie complesse[20]. In questo caso è stato recentemente mostrato come a un determinato livello di rischio i pazienti neri risultano più malati dei pazienti bianchi e risolvere questa diseguaglianza sociale e assistenziale comporterebbe un aumento dei pazienti neri che ricevono ulteriori cure fino a quasi il 50%. Ma perché l'algoritmo utilizzato per scegliere verso quali pazienti destinare la spesa sanitaria in questo caso produce in maniera così evidente una disparità? Perché automaticamente prevede, e dunque richiede, minori costi per l'assistenza sanitaria per i pazienti neri rispetto ai pazienti bianchi. È probabile che questo derivi, ancora una volta, da una storia di disparità e discriminazioni: la tendenziale minore attenzione alla salute dei pazienti neri, sottoposti ad un'intensità di cura inferiore a fronte di sintomatologie o patologie analoghe o corrispondenti a quelle dei pazienti bianchi, si traduce in un algoritmo che paradossalmente considerandoli meno bisognosi di cure li condanna a peggiori condizioni di salute.

D'altra parte, le questioni relative alla qualità dei dati, e alla presenza di *bias*, paiono assumere sempre maggiore rilevanza, anche nella riflessione dei tecnologi, ragione per cui non sembra irrealistico ritenere che nei prossimi anni sarà possibile progettare sistemi antidiscriminatori *by default* e *by design*, in grado di superare le criticità delineate per lo meno rispetto alla presenza di *bias* all'interno dei sistemi.

Occorre però chiedersi se sia sufficiente il riferimento al concetto di *bias* per tradurre nell'*AI* principi di eguaglianza, ragionevolezza, giustizia sociale.

4. *È sufficiente il riferimento al concetto di* bias *per portare nell'intelligenza artificiale equità e giustizia*?

Il riferimento al concetto di *bias*, oggi piuttosto *mainstream*, non sembra sufficiente a conformare l'*AI* ai principi di equità e giustizia.

Il concetto di "parzialità" individua la fonte della disuguaglianza nel comportamento degli individui o nei risultati di un sistema tecnico (ad esempio un sistema che favorisce i bianchi, o gli uomini, si pensi al caso segnalato in ambito sanitario). Sotto questo modello concettuale, un obiettivo tecnologico potrebbe essere quello di creare un sistema "imparziale", attraverso una progettazione che utilizzi i dati per ottimizzare i suoi parametri, e quindi testi gli eventuali errori che ne risultano. È stato sottolineato che da un punto di vista tecnologico è possibile definire e dunque, potenzialmente, "ottimizzare", un sistema "giusto", intendendolo come un sistema libero da *bias*.

Il limite di un approccio del genere è dato dal fatto che si fonda su un'idea di oggettività che non è propria di questi sistemi. Viene sottolineato infatti quanto sia fuorviante immaginare che i dati, e la tecnologia stessa, siano obiettivi, e dunque limitativo ricondurre i problemi associati a dati e algoritmi "parziali" o "*biased*" esclusivamente alla progettazione e alle caratteristiche dei sistemi "intelligenti"[21].

È stata così messa in evidenza la ristrettezza dell'equità concepita dal punto di vista computazionale, sottolineando come i *data scientists* che mettono a disposizione la loro scienza per riformare i processi decisionali in genere (che si tratti di processi normativi o di decisioni giudiziarie), dovrebbero orientare i loro sforzi verso l'obiettivo della giustizia, piuttosto che farsi ispirare da uno *zeitgeist* di soluzioni tecnologiche[22]. Naturalmente non è immediata la definizione

[20] Cfr. Z. OBERMEYER *et al.*, *Dissecting racial bias in an algorithm used to manage the health of populations*, in *Science*, Vol. 366, 2019, 447 ss.

[21] Si vedano già le prime considerazioni di D. P. BENJAMIN, *Change of Representation and Inductive Bias*, Boston, Klewer Academic Publisher, 1990.

[22] Cfr. B. Green, cit. in C. D'IGNAZIO – L. F. KLEIN, *Data Feminism*, cit.

"tecnologica" di equità. Com'è stato segnalato, un concetto di equità che assuma significato indipendentemente dal contesto, o dalla storia che racconta le esperienze dei gruppi, delle minoranze, delle soggettività discriminate, non sembra in grado di rispondere e di contrastare, non solo semanticamente, ma soprattutto normativamente, le "ingiustizie" sistematiche perpetrate da alcuni gruppi su altri gruppi[23].

Un modello di equità che invece di creare algoritmi "daltonici", porti alla progettazione di algoritmi "giusti", in grado di tenere in considerazione tempo, storia e asimmetrie di potere. D'altra parte, l'evoluzione del principio di eguaglianza in Europa, tra diritti interni e diritto europeo, è stata almeno in parte caratterizzata da una progressiva consapevolezza circa le asimmetrie di potere, basti pensare a come si sviluppa il diritto antidiscriminatorio, alla progressiva astrazione del giudizio di discriminazione dall'elemento della comparazione, con l'elaborazione (prima) giurisprudenziale della figura delle discriminazioni indirette, e all'inserimento delle molestie e molestie sessuali come fattispecie discriminatorie[24]. Tradurre questa consapevolezza nell'*AI* sarebbe quindi come gettare le basi per vere e proprie "azioni positive tecnologiche", cioè, come si dirà in seguito, per un utilizzo delle tecnologie innovative nella prospettiva del superamento delle esistenti, e radicate, disparità, attraverso interventi correttivi comunque orientati alla "rimozione degli ostacoli".

A questi fini, il concetto di *bias* (negli individui, così come negli *data assets*, o negli algoritmi) non è abbastanza forte da poter fungere, per lo meno preso da solo, come risolutivo ancoraggio per i principi di equità e di giustizia sociale.

Non è il *bias* di per sé a generare algoritmi di intelligenza artificiale ingiusti o discriminatori, ma la segregazione tecnologica è il frutto di un complesso di interazioni che rendono necessario pensare l'*AI* e costruire gli algoritmi come orientati *by default* al superamento delle asimmetrie di potere, e dunque, per così dire, *biased* nella misura in cui questo è funzionale a rimuovere una radicata discriminazione, o a neutralizzare uno stereotipo operante.

È stato da tempo chiarito che la discriminazione algoritmica può avvenire in modi involontari o difficili da ricostruire razionalmente, quando i pregiudizi sociali preesistenti si riflettono nei dati attraverso modalità che possono essere poco prevedibili. In questi casi, si dice che i pregiudizi "si intrufolano"[25], "intenzionalmente o per caso"[26], o in maniera sotterranea, che emerge soltanto nel tempo[27].

Significativo in tal senso è il fenomeno dell'amplificazione delle discriminazioni fondate sulla "razza" attraverso algoritmi incostituzionali[28]: esistono decisioni digitali che hanno la capacità di rafforzare relazioni sociali oppressive e di realizzare esse stesse nuovi modelli di profilazione razziale e di genere. La constatazione della presenza di errori di dati guidati algoritmicamente che sono specifici per donne e minoranze razziali, descrivono il fenomeno, che è stato definito di oppressione algoritmica, in base al quale l'*AI* mostra la sua capacità di contribuire alla strutturazione del razzismo e del sessismo: il razzismo sarebbe l'interfaccia del programma

[23] Cfr. S. COSTANZA-CHOCK, *Data and Discrimination*, in *Data Justice Conference*, 21-22 maggio 2018, Cardiff, UK.

[24] Numerosi i possibili richiami sul punto, si veda per tutti E. ELLIS, *EU Anti-discrimination law*, Oxford, 2005; M. Bell, *Anti-discrimination law and the European Union*, Oxford, 2002; M. BARBERA (a cura di), *Il nuovo diritto antidiscriminatorio*, Milano, 2007, A. LORENZETTI, *Il Diritto Antidiscriminatorio europeo: genesi ed evoluzione*, in B. PEZZINI (a cura di), *La costruzione del genere. Norme e regole*, Bergamo, 2012. Da ultimo la relazione presentata in occasione del convegno dell'Associazione Italiana dei Costituzionalisti a Bergamo, "*Eguaglianza e discriminazioni nell'epoca contemporanea*", 15-16 novembre 2019, di A. SCIORTINO, *Eguaglianza di genere nell'UE: categorie giuridiche e tutele.*

[25] Cfr. A. CALISKAN *et al*., *Biased bots: Human prejudices sneak into AI systems*, in *University of Bath*, 13 Aprile 2017

[26] Cfr. S. BAROCAS - A. D. SELBST, *Big data's disparate impact*, in *California Law Review*, vol. 104, 2016, 674

[27] Cfr. B. FRIEDMAN - H. NISSENBAUM, *Bias in computer systems*, in *ACM Transactions on Information Systems (TOIS)*, vol. 14, n. 3/1996, 330 ss.

[28] Secondo l'espressione utilizzata da A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, cit.

applicativo (API *application programming interface*) di Internet[29]. Dunque, i dati e gli algoritmi non si limitano a modellare i risultati distributivi, ma incidono sulla costruzione dei significati, rinforzando alcuni "*frame*" discorsivi rispetto ad altri[30]. Ciò significa che i dati, frammenti personali dell'identità, diventano strumento di collegamento tra i soggetti e quelle entità che cercano di tradurre i dati stessi in risorse finanziarie, attraverso modalità di "estrazione e appropriazione"[31].

A fronte di questo quadro, l'intreccio tra intelligenza naturale e artificiale impone di provare a leggere in un modo diverso il rapporto tra le due, proponendo non soltanto un effettivo orientamento alla giustizia della seconda al fine di non perpetuare stereotipi e discriminazioni secolarmente prodotte dalla prima, ma l'esercizio di una funzione correttiva della seconda sulla prima.

5. *Stereotipi e discriminazioni dall'intelligenza naturale (nell'argomentazione giudiziaria) all'intelligenza artificiale*

L'argomentazione giudiziaria resta uno dei luoghi dove ancora molto significativi sono le pratiche di stereotipizzazione e la costruzione di concettualizzazioni, molto spesso implicite, ma comunque (o forse proprio per questo) assai rilevanti.

Nell'ambito dei reati culturalmente orientati, ad esempio, studi recenti[32] mostrano come "considerazioni di senso comune prendano talvolta il posto di un'adeguata e accurata analisi non solo quando il giudice si approccia ai profili concettuali-definitori della pratica culturale, ma anche quando affronta i profili più operativi, relativi alla considerazione da attribuire concretamente a questo elemento"[33]. Il richiamo al "senso comune", com'è facile immaginare, è frequente soprattutto nei casi in cui si tratta di soggettività diverse, non maggioritarie, gruppi vulnerabili e in vario modo riconducibili a quei *grounds* delle discriminazioni che il diritto costituzionale e il diritto europeo mettono a disposizione per la realizzazione del principio di eguaglianza.

A tale rischio non sembra immune neppure il giudice costituzionale, basti pensare, in tempi recenti, alle argomentazioni della sent. n. 221/2019 della Corte costituzionale, nella quale, con riferimento all'accesso alla fecondazione eterologa da parte di coppie lesbiche si ritiene che non sia "irragionevole [...] che il legislatore si preoccupi di garantirgli [con riferimento al bambino] quelle che, secondo la sua valutazione e alla luce degli apprezzamenti correnti nella comunità sociale, appaiono, in astratto, come le migliori condizioni "di partenza"."[34]. Si considera dunque non irragionevole che il legislatore ritenga preferibile per un bambino avere una madre e un padre e che, per questo motivo, "alla luce degli apprezzamenti correnti della comunità sociale", si escluda il

[29] Cfr. S. U. NOBLE, *Algorithms of Oppression*. *How Search Engines Enforce Racism*, cit.

[30] Cfr. R. BIVENS - A. S. HOQUE, *Programming sex, gender, and sexuality: Infrastructural failures in the "feminist" dating app Bumble*, in *Canadian Journal of Communication*, vol. 43, n. 3/2018, 441 ss.; M. E. SWEENEY, *The intersectional interface*, in S. U. NOBLE - B. M. TYNES (a cura di), *The intersectional internet: Race, sex, class, and culture online*, Peter Lang International Academic Publishers, 2016; M. WILLSON, *Algorithms (and the) everyday*, in *Information, Communication & Society*, vol. 20, n. 1/2017, 137 ss.

[31] Cfr. J. E. COHEN, *The biopolitical public domain: The legal construction of the surveillance economy*, in *Philosophy & Technology*, vol. 31, n. 2/2018, 214, cit. in A. L. HOFFMANN, *Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse*, in *Information, Communication & Society*, vol. 22, n. 7/2019, 900 ss., che ha recentemente ricostruito l'insufficienza degli strumenti del diritto antidiscriminatorio e della riflessione antidiscriminatoria nella rimozione di stereotipi e discriminazioni dai sistemi di *AI*. Su estrazione dei dati e problemi di riservatezza e tutela dell'identità personale nelle decisioni automatizzate cfr. *ex aliis* E. PELLECCHIA, *Privacy, decisioni automatizzate, algoritmi*, in E. TOSI (a cura di), *Privacy Digitale. Riservatezza e protezione dei dati personali tra GDPR e nuovo Codice Privacy*, Milano, Giuffré, 2019, 417 ss., e i riferimenti *ivi* contenuti.

[32] Cfr. P. PANNIA, *Quando la cultura entra nell'aula giudiziaria: uno studio sulle argomentazioni dei giudici italiani*, in *Sociologia del diritto*, n. 3/2016.

[33] *Ibidem*, che cita numerose pronunce individuate attraverso un rigoroso monitoraggio della giurisprudenza.

[34] Corte cost., sent. n 221/2019, *Considerato in diritto*, 13.2.

ricorso alle tecniche di PMA a coppie di donne. Tale passaggio è significativo di come nella costruzione dei ruoli in relazione al genere resti fortissimo l'ancoraggio all'apprezzamento sociale.

Nelle questioni culturali, prima menzionate, spesso proprio i giudici fanno proprie considerazioni che esulano da un riferimento al sistema normativo nazionale, e dunque al diritto applicabile, ma cercano di approdare all'interpretazione di altri sistemi normativi, in realtà non conosciuti, e quindi generalmente letti attraverso la lente dello stereotipo culturale[35]. È il caso del richiamo al patriarcato come elemento integrato nella categoria di "cultura", e quindi concetto ad essa riconducibile, in un processo di duplice stereotipizzazione e discriminazione: da un lato verso chi appartiene al gruppo culturale nel quale si realizza la condotta penalmente rilevante, e quindi verso l'autore del crimine, dall'altro verso la vittima, meno tutelata e protetta. Non è raro trovare nelle argomentazioni di giudici riferimenti ad "ancestrali codici familiari"[36] che potrebbero in qualche modo giustificare alcune condotte violente e criminali, "malintesi diritti maritali"[37], che possono condurre a tensioni emotive, frustrazioni, turbamenti.

Le pratiche giudiziali di stereotipizzazione e conseguente discriminazione assumono un peso particolare nella fenomenologia della violenza di genere, traducendosi anche in una pratica linguistica. Non è proprio soltanto dei giornalisti l'utilizzo di termini che abbandonando la tecnicità e la specificità del linguaggio giuridico trasformano reati in conflittualità, litigi in aggressioni. Questo avviene nel processo e nelle decisioni giudiziali, in cui si mettono in discussione perizie mediche, o si assume la posizione della vittima (che nel femminicidio in particolare non ha evidentemente possibilità di prendere parola)[38]. Nel caso della violenza di genere, così come dei reati culturalmente orientati, emerge tutta la portata stereotipante della giurisprudenza che continua, dunque, a ricostruire, attraverso decisioni e linguaggio, ruoli femminili costruiti intorno a inferiorità, sessualizzazione del corpo, centralità dell'estetica.

È stato correttamente sottolineato che se è vero che le discriminazioni sono nell'intelligenza naturale prima che in quella artificiale, non si può d'altra parte paragonare il pregiudizio del singolo giudice nel singolo giudizio con l'esplosione dell'errore, del *bias*, in tutti i giudizi attraverso l'utilizzo dei sistemi di *AI*[39]. Da questo punto di vista, il potenziale dirompente di una giustizia predittiva algoritmica è stato ampiamente studiato, soprattutto nell'ambito della giustizia penale[40], ben descritto nella Carta etica adottata dalla CEPEJ nel 2018[41], e sono oggetto di approfondimento e di critica sia da parte dei giuristi sia da parte dei tecnologi i principali elementi di tensione rispetto al sistema dei valori costituzionali coinvolti: la ben nota mancanza di trasparenza dal punto di vista tecnico, propria delle decisioni assunte attraverso l'*AI* o comunque sostenute da meccanismi di *AI*[42], e la spiegabilità, nel senso della traduzione della funzione dell'algoritmo in termini comprensibili agli utenti, strumentale all'esercizio del diritto di cui all'art. 22, par. 3 del GDPR, il quale com'è noto stabilisce che i titolari dei dati hanno il diritto di contestare le decisioni basate unicamente su

[35] "In modo approssimativo, o addirittura scorretto" talvolta, riporta Pannia citando tra le altre Cass. Pen. 26.06.2007 n. 34909; Cass. Pen., 18.12.2013 n. 51059

[36] Tribunale di Trento, 19 febbraio 2009, n. 138, in *ibidem.*

[37] *Ibidem*.

[38] Si veda F. FILICE, *La violenza di genere*, Milano, Giuffré, 2019, per un richiamo alla necessità di affrontare la violenza di genere in una prospettiva anche sociologica, volta proprio a mettere a nudo il perdurante radicamento negli uomini e nella società di una visione del genere maschile come dominante e dunque, per così dire, abilitato ad agire sul genere subordinato esercitando coazione e praticando subordinazione.

[39] Cfr. A. MANTELERO, *Report on Artificial Intelligence and Data Protection: Challenges and Possible Remedies*. Strasburgo, 25 gennaio 2019, T-PD(2018)09Rev Consultative Committee of the Convention for the Protection of Individuals with regard to automatic processing of personal data (Convention 108).

[40] In Italia sono fondamentali i contributi di C. CASONATO, *Intelligenza artificiale e diritto costituzionale: prime considerazioni*, in *Diritto pubblico comparato ed europeo*, Speciale 2019, 101 ss. e A. D'ALOIA, *Il diritto verso "il mondo nuovo". Le sfide dell'Intelligenza Artificiale*, in Rivista di BioDiritto, n. 1/2019, oltre a quello già citato di A. SIMONCINI, *L'algoritmo incostituzionale*, cit.

[41] Cfr. *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment*, cit.

[42] Cfr. F. PASQUALE, *The Black Box Society: The Secret Algorithms that Control Money and Information*, Harvard University Press, 2015

processi automatizzati[43]. Ciò non toglie però che i sistemi di *AI* possano svolgere un ruolo anche nei processi, quali assistenti cognitivi dei giudici in particolare per la trattazione di cause "semplici, seriali, ripetitive, interamente documentali ecc." o "nelle procedure alternative di soluzione delle controversie"[44].

Ma soprattutto, quello che forse il diritto potrebbe chiedere in più all'*AI* è di fornire gli strumenti capaci di guidare le decisioni orientandole ai valori del costituzionalismo. Strumenti in grado di rappresentare la realtà senza amplificarne le ingiustizie, ma anzi sfruttando l'infinita possibilità di accuratezza e di completezza che sempre di più caratterizza i dati, e la progressiva espansione della potenza di calcolo, per sradicarle, attraverso algoritmi costruiti come azioni positive.

[43] Su questo, tra i numerosi riferimenti, interessante il contributo di M. ALMADA, *Human Intervention in Automated Decision-Making: Toward the construction of contestable systems*, in *ICAIL 2019*, June 17–21, 2019, Montreal, QC, Canada

[44] Cfr. F. DONATI, *Intelligenza Artificiale e giustizia*, in *Rivista AIC*, n. 1/2020, spec. 431.