TEORIA E STORIA DEL DIRITTO PRIVATO
RIVISTA INTERNAZIONALE ONLINE - PEER REVIEWED JOURNAL
ISSN: 2036-2528

Gianmarco Gometz

Intelligenza artificiale, profilazione e nuove forme di discriminazione

Numero Speciale Anno 2022
Ombre del diritto
(a cura di F. Mancuso e V. Giordano)
Materiali dai seminari del PRIN 2017
'The Dark Side of Law'
www.teoriaestoriadeldirittoprivato.com

Sede della Redazione della rivista:
Prof. Laura Solidoro
Via R. Morghen, 181
80129 Napoli, Italia
Tel. +39 333 4846311

Aut. Tr. Napoli n. 78 del 03.10.2007
Provider Aruba S.p.A
Piazza Garibaldi, 8
52010 Soci AR
Iscr. Cam. Comm. N° 04552920482
P.I 01573850616 – C.F. 04552920482.

Con il patrocinio di:

Ordine degli Avvocati di Salerno

Dipartimento di Scienze Giuridiche
(Scuola di Giurisprudenza)
Università degli Studi di Salerno

I saggi che compongono questo numero speciale di Teoria e Storia del Diritto Privato sono stati sottoposti al giudizio di due Referees con il sistema del 'double blind'.
In Redazione per questo numero speciale: M. Luciano (Univ. Salerno), P. Pasquino (Univ. Salerno).

Intelligenza artificiale, profilazione e nuove forme di discriminazione

SOMMARIO: 1. Intelligenze disumane – 2. Processi decisionali automatizzati produttivi di effetti giuridici: alcuni vincoli – 3. Il concetto di discriminazione algoritmica – 4. Discriminazioni algoritmiche dirette e indirette – 5. Discriminazioni e affidabilità delle stime algoritmiche.

1. *Intelligenze disumane*

In un noto documento del 1955, John McCarthy definisce quello dell'intelligenza artificiale (IA) come il problema di «*making a machine behave in ways that would be called intelligent if a human were so behaving*»[1]. Questa definizione ha tuttora grande successo[2], grazie anche al suo disimpegno

[1] La 'Proposta di Darthmouth' è un documento informale circa un'iniziativa di ricerca sull'intelligenza artificiale presso il Dartmouth College di Hanover, New Hampshire, nell'estate del 1956. Il testo è considerato uno degli atti di nascita del campo di ricerca sull'IA; cfr. J. MCCARTHY, M. L. MINSKY, N. ROCHESTER, C.E. SHANNON, *A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence*, disponibile su http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

[2] Una definizione dello stesso genere è ad esempio quella riportata nell'art. 4 della proposta di Proposta di regolamento del Parlamento europeo e del Consiglio allegata alla Risoluzione del Parlamento europeo del 20 ottobre 2020 recante raccomandazioni alla Commissione concernenti il quadro relativo agli aspetti etici dell'intelligenza artificiale, della robotica e delle tecnologie correlate (2020/2012 [INL]), secondo cui per «intelligenza artificiale», si intende «un sistema basato su software o integrato in dispositivi hardware che mostra un comportamento intelligente, tra l'altro raccogliendo e trattando dati, analizzando e interpretando il proprio ambiente e compiendo azioni, con un certo grado di autonomia, per raggiungere specifici obiettivi». Ancora più ispirata al paradigma dell'IA c.d. 'debole' – ossia, l'IA che si limita alla simulazione/riproduzione di specifiche abilità intellettuali umane – è la definizione

filosofico: lungi dall'affrontare le spinose questioni del significato e dei criteri d'uso del termine 'intelligenza', essa connota l'IA *per relationem*, attraverso il riferimento alla capacità delle macchine di eseguire operazioni che, se fossero svolte da esseri umani, ne rivelerebbero l'intelligenza. In tal modo, è possibile parlare di sistemi, dispositivi, algoritmi intelligenti senza necessariamente affrontare le impervie questioni filosofiche relative alla loro attuale o potenziale (auto)coscienza, soggettività, razionalità e intenzionalità, e soprattutto senza prender posizione circa la configurabilità di una IA c.d. 'forte', connotata da abilità e stati mentali ontologicamente indistinguibili da quelli degli esseri umani intelligenti[3]. Sono infatti senz'altro intelligenti nel senso di McCarthy gli odierni apparati hardware e/o software in grado di riconoscere volti e voci, tradurre discorsi o formularne di nuovi appropriati al contesto[4], giocare a scacchi, a *poker* e a *go*, diagnosticare o prevedere l'insorgenza di malattie, scoprire molecole farmacologicamente efficaci, sintetizzare audio o video perfettamente

dell'art. 3 della Proposta di regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'Unione (COM/2021/206 final), secondo cui per 'sistema di intelligenza artificiale' si intende «un software sviluppato con una o più delle tecniche e degli approcci elencati nell'allegato I, che può, per una determinata serie di obiettivi definiti dall'uomo, generare output quali contenuti, previsioni, raccomandazioni o decisioni che influenzano gli ambienti con cui interagiscono».

[3] L'IA forte si fonda spesso su filosofie funzionaliste della mente secondo cui gli stati mentali di un individuo (ad esempio soffrire o credere che P, dove P è un certo enunciato) vengono definiti esclusivamente sulla base delle loro relazioni reciproche e delle loro relazioni con gli input sensoriali e gli output comportamentali: due stati mentali sono identici se sono identiche le loro relazioni funzionali con gli altri stati mentali e con gli input e gli output del sistema. Secondo il funzionalismo, inoltre, uno stato mentale può essere 'istanziato' da supporti fisici diversi ed è possibile ipotizzare un'identificazione mente/software e sistema nervoso/hardware: come le proprietà formali che definiscono un programma prescindono dalle caratteristiche fisiche dei calcolatori che lo eseguono, così gli stati mentali prescindono dalle caratteristiche fisiche ed anatomiche del sistema nervoso; cfr. M.FRIXIONE, *Logica, significato e intelligenza artificiale*, Milano, 1994, 60-62.

[4] Compresi gli articoli di *humanities* come la filosofia del diritto; cfr. https://gpt3demo.com/apps/essay-writing-by-eduref.

verosimili, guidare veicoli su strada, in aria o in mare, risolvere problemi matematici, suggerire prodotti o servizi conformi ai gusti/interessi di un particolare consumatore, prevedere comportamenti individuali di vario tipo ecc. A corroborare l'antica scelta di parlare di *intelligenza* piuttosto che di *abilità* artificiale contribuisce inoltre la circostanza che le IA odierne funzionano grazie ad algoritmi i quali, piuttosto che eseguire istruzioni predeterminate su cosa fare volta per volta, sono in grado di riconoscere particolari correlazioni tra dati in modo funzionale allo svolgimento di compiti che il sistema 'impara' a svolgere da sé per tentativi ed errori. Questo apprendimento automatico (c.d. *machine learning*) consente alle IA di migliorare autonomamente le proprie capacità operative e raggiungere prestazioni che oramai surclassano quelle dei più abili agenti umani in un numero crescente di campi, con progressi ancora più meravigliosi attesi nel breve e medio termine[5].

Tra gli ambiti d'applicazione dell'IA, particolarmente dibattuto è quello della *profilazione*, ossia l'analisi e la previsione di comportamenti, qualità e disposizioni delle persone fisiche a partire da una serie di dati che le riguardano: geolocalizzazione, contatti, scelte d'acquisto, attività sui *social*, siti web visitati e ogni altra sorta di dati personali, compresi quelli genetici e biometrici.[6] Già allo stadio attuale dello sviluppo

[5] I progressi delle tecnologie dell'IA sono talora accostati a quelli predetti dalla nota *Legge di Moore*, l'osservazione empirica secondo cui la potenza di calcolo dei computer raddoppia all'incirca ogni due anni. In realtà, la varietà delle diverse implementazioni dell'IA e la discontinuità del loro sviluppo, caratterizzato da lunghi periodi di stagnazione e improvvise accelerazioni, non consentono di avanzare previsioni accurate. Nel campo delle traduzioni automatiche, per esempio, dopo decenni di stallo, in circa due anni si è giunti a realizzare sistemi con prestazioni paragonabili a quelle di un traduttore umano non professionale. Non v'è accordo unanime, infine, circa la possibilità e le tempistiche di realizzazione di una c.d. 'Intelligenza Artificiale Generale' (AGI - Artificial General Intelligence) capace di svolgere ogni compito che richieda impegno intellettuale; cfr. H. A. KISSINGER, E. SCHMIDT, D. HUTTENLOCHER, *The Age of AI And Our Human Future*, London, 2021, edizione digitale.

[6] L'art. 4 del Regolamento generale sulla protezione dei dati (GDPR) definisce la profilazione come «qualsiasi forma di trattamento automatizzato di dati personali consistente nell'utilizzo di tali dati personali per valutare determinati aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti riguardanti

tecnologico, un'abbondante disponibilità di questi dati consente alle IA di tracciare dei profili che rivelano assai attendibilmente svariate qualità/disposizioni degli individui (comprese alcune ignote ai loro stessi portatori), ed è ragionevole attendersi che nel prossimo futuro le tecnologie di profilazione 'intelligente' applicate ai *big data* saranno in grado di prevedere la condotta umana più accuratamente di quanto mai siano state in grado di fare la psicologia, la psichiatria e le scienze sociali[7].

La disponibilità di strumenti di tale potenza ha posto una serie di problemi rilevanti per il diritto, soprattutto quando, come accade sempre più spesso, dal loro impiego derivino degli effetti giuridici o giuridicamente rilevanti che incidono significativamente sulla vita delle persone. La potenziale violazione di alcuni diritti fondamentali – privacy, libertà personale, libertà di pensiero, eguaglianza e non discriminazione in primis – ha mosso molti legislatori contemporanei a prevedere vari limiti all'impiego di tecnologie di profilazione, specialmente quando i dati all'uopo utilizzati sono relativi a minori o rivelatori di aspetti particolarmente intimi e privati della persona. Fra tali limiti, vorrei in questa sede trattare quelli derivanti dalle violazioni del *principio di non discriminazione*: produrrò alcuni argomenti a supporto della conclusione secondo cui i timori legati alle c.d. 'discriminazioni algoritmiche' più fondati sono quelli relativi all'eventualità che certe decisioni produttive di effetti giuridici assai incisivi sulla vita degli individui vengano adottate col supporto di sistemi di IA che *sbagliano*, nel senso che non sono in grado di profilare attendibilmente i singoli per via di dati incompleti, obsoleti o *biased*, di errori nella costruzione degli algoritmi o di limitazioni al loro uso ispirate, paradossalmente, dall'intento di evitare effetti discriminatori. Collateralmente, sosterrò che questi timori non riguardano solo le disparità di trattamento giuridicamente proscritte a

il rendimento professionale, la situazione economica, la salute, le preferenze personali, gli interessi, l'affidabilità, il comportamento, l'ubicazione o gli spostamenti di detta persona fisica».

[7] Lo scenario fantascientifico dipinto nel romanzo di D. RIELLI, *Odio*, Milano, 2020, in cui milioni di individui risultano monitorabili e prevedibili in tutti i loro comportamenti grazie a speciali dispositivi indossabili collegati in rete diventa sempre più verosimile col passare del tempo.

titolo di discriminazione, ma riguardano qualsiasi impiego delle tecnologie di IA con funzioni di supporto a decisioni produttive di effetti giuridici o giuridicamente rilevanti.

Prima di addentrarmi nella trattazione, sono opportuni alcuni chiarimenti analitici preliminari. Nelle pagine che seguono aderirò alla tradizione definitoria inaugurata da John McCarthy, e dunque assumerò che un ente denoti IA quando è in grado di effettuare autonomamente[8] delle operazioni che fino a epoche recentissime erano ritenute di esclusivo appannaggio di esseri umani dotati di coscienza, intenzionalità, creatività, ingegno e, per l'appunto, intelligenza. Ciò anche ove tali operazioni producano risultati: *a)* di gran lunga superiori per quantità e qualità (esattezza, attendibilità, rapidità, completezza ecc.) a quelli raggiungibili dagli esseri umani; *b)* siano ottenuti in modalità che sono affatto difformi, anche sotto il profilo logico, dai processi cognitivi e conativi che tipicamente trovano stanza nelle menti umane; *c)* siano ottenuti attraverso metodi che, almeno allo stadio attuale dei progressi tecnologici, non contemplano alcun ricorso a qualcosa che somigli anche lontanamente al senso comune o alla dimensione emotiva cosciente o subcosciente che fanno da sfondo a qualunque pratica umana. Le correlazioni tra dati e le estrapolazioni che sono in grado di operare le IA – almeno allo stadio attuale dei progressi tecnologici – sono infatti 'vuote' o formali e non prevedono alcuna forma di comprensione, concettualizzazione e categorizzazione assimilabili a quelle degli esseri umani dotati di autocoscienza. I processi che presiedono al funzionamento degli odierni sistemi di IA, pertanto, possono essere accostati al pensiero umano solo metaforicamente; essi non danno luogo a *ragionamenti* analizzabili e controllabili nel contesto di giustificazione, ma prevedono la produzione di output a partire dal riconoscimento di particolari pattern nei dati di input secondo un approccio che non è logico-deduttivo, ma meramente statistico. Ciò determina un'importante differenza tra l'intelligenza umana e le attuali IA: gli agenti umani sono

[8] Nel contesto dell'IA, i termini 'autonomo' e 'autonomia' non designano, come in filosofia politica o morale, la potestà di darsi da sé i criteri di condotta, ma soltanto la capacità di funzionare e svolgere i propri compiti senza l'intervento di operatori umani.

perlopiù inconsapevoli dei fattori neurologici, psicologici e sociali che di fatto determinano o influenzano i loro comportamenti, ma sono di norma, grazie al linguaggio, in grado di comprendere e illustrare le ragioni o i motivi che li hanno mossi ad agire come hanno agito (sebbene possano essere più o meno sinceri in questo resoconto, sia col prossimo che con se stessi). Nel caso delle reti neurali funzionanti sulla base dell'odierno paradigma centrato sul *machine learning*, invece, l'imperscrutabilità travalica il livello fisico-causale e si estende a quello logico-giustificativo; le IA attuali non sono in grado di presentare alcun *log* circa le ragioni, i motivi, le cause, i fattori o anche soltanto i singoli passaggi computazionali che le hanno determinate a operare in un certo modo o a produrre un certo risultato; la loro è insomma un'*intelligenza non intelligibile*. Tale limitazione – si badi – non deriva da contingenti scelte di programmazione, bensì da impossibilità matematiche per così dire 'strutturali'; se il *log* manca, è perché non c'è alcun *logos*[9]. Ciò implica che l'unico modo per capire quali output produrranno quegli algoritmi è eseguirli (anche se, come vedremo, l'accesso ai dati usati per 'addestrare' il sistema e le informazioni sugli obiettivi ad esso assegnati possono, a certe condizioni, darci elementi per congetturare *ex post* come mai siano stati prodotti certi risultati e non altri).

Queste differenze fisiche, logiche ed epistemiche, unitamente alla capacità di elaborare enormi quantità di dati e produrre risultati che in certi ambiti risultano proibitivi perfino per le più eccelse menti umane, suggerisce di qualificare questi artefatti non soltanto come intelligenze artificiali nel senso di costruite, sintetiche, non umane, ma come intelligenze *disumane*. Vedremo che questo è un punto rilevante sul piano della valutazione etico-politica e giuridica degli impieghi dell'IA in processi decisionali da cui scaturiscono effetti giuridici in grado di incidere assai significativamente sul godimento dei diritti fondamentali degli individui.

[9] Almeno, un *logos* intelligibile agli esseri umani.

2. *Processi decisionali automatizzati produttivi di effetti giuridici: alcuni vincoli*

Gli impieghi dell'IA in processi decisionali che producono effetti giuridici riguardanti le persone fisiche sono stati oggetto di recenti iniziative legislative mosse da più che giustificate preoccupazioni per la potenziale violazione di alcuni diritti fondamentali degli individui. In Europa, ad esempio, il Regolamento generale sulla protezione dei dati (d'ora in poi GDPR)[10], la coeva Direttiva 680/2016[11] e l'ampio *corpus* del vigente diritto antidiscriminatorio[12] prevedono che i processi decisionali (totalmente o parzialmente) automatizzati, compresi quelli che prevedono l'impiego di sistemi di IA a fini di profilazione, siano consentiti solo a certe condizioni e dietro opportune garanzie. È previsto fra l'altro:

1) Il divieto, salvo che si verifichino particolari condizioni[13], dei processi decisionali automatizzati basati sull'impiego di dati genetici e biometrici intesi a identificare in modo univoco una persona fisica, di dati relativi alla salute, alla vita sessuale o all'orientamento sessuale della persona, nonché di dati personali che rivelino l'origine razziale

[10] 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).

[11] Direttiva (UE) 2016/680 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativa alla protezione delle persone fisiche con riguardo al trattamento dei dati personali da parte delle autorità competenti a fini di prevenzione, indagine, accertamento e perseguimento di reati o esecuzione di sanzioni penali, nonché alla libera circolazione di tali dati e che abroga la decisione quadro 2008/977/GAI del Consiglio.

[12] Una ricognizione delle principali fonti del diritto antidiscriminatorio è in E. CONSIGLIO, *Che cos'è la discriminazione? Un'introduzione teorica al diritto antidiscriminatorio*, Torino, 2020, 6 ss.

[13] Si tratta delle condizioni stabilite dall'art. 9, par. 2, lett. a) o g) del GDPR, ossia se l'interessato ha prestato il proprio consenso esplicito al trattamento di tali dati personali per una o più finalità specifiche (salvo i casi di consenso irrevocabile), e se ricorrono motivi di interesse pubblico rilevante sulla base del diritto dell'Unione o degli Stati membri (interesse che deve essere proporzionato alla finalità perseguita, rispettare l'essenza del diritto alla protezione dei dati e prevedere misure appropriate e specifiche per tutelare i diritti fondamentali e gli interessi dell'interessato).

o etnica, le opinioni politiche, le convinzioni religiose o filosofiche o l'appartenenza sindacale[14];

2) il diritto di chiunque a ricevere informazioni sull'esistenza di processi decisionali automatizzati che lo riguardino e, almeno nei casi di profilazione, a ricevere informazioni significative sulla logica utilizzata nonché sull'importanza e le conseguenze previste di tale trattamento per l'interessato[15];
3) il diritto di chiunque a non essere sottoposto a una decisione che produca effetti giuridici che lo riguardano o che in modo analogo incidono significativamente sulla sua persona, qualora questa decisione sia basata *unicamente* sul trattamento automatizzato, profilazione compresa[16];
4) il divieto di processi decisionali algoritmici o algoritmicamente assistiti produttivi di «effetti discriminatori nei confronti di persone fisiche sulla base della razza o dell'origine etnica, delle opinioni politiche, della religione o delle convinzioni personali, dell'appartenenza sindacale, dello status genetico, dello stato di salute o dell'orientamento sessuale»[17], nonché sulla base di altre caratteristiche protette dal vigente diritto antidiscriminatorio;
5) l'obbligo di far precedere ai trattamenti automatizzati su cui si fondano decisioni produttive di effetti giuridici che riguardano le persone fisiche una *valutazione d'impatto* che dia conto del trattamento

[14] Cfr. artt. 9 e 22 GDPR.

[15] Cfr. artt. 13 e 14 GDPR. La dottrina ha segnalato alcune criticità relative alla possibilità di dare applicazione alle disposizioni che prevedono l'obbligo di fornire di fornire le «informazioni significative sulla logica utilizzata» quando i processi decisionali automatizzati siano assistiti da sistemi di IA di nuova generazione che non operano secondo alcuna 'logica' di natura deterministica e dunque analizzabile, comprensibile e comunicabile; cfr. A. SIMONCINI, S. SUWEIS, *Il cambio di paradigma nell'intelligenza artificiale il suo impatto sul diritto costituzionale*, in *Rivista di filosofia del diritto*, 8.1, 2019, 98 s.; G. CONTISSA, G. LASAGNI, G. SARTOR, *Quando a decidere in materia penale sono (anche) algoritmi e IA: alla ricerca di un rimedio effettivo*, in *Diritto di internet*, 4, 2019, 625-627. Sul tema si tornerà *infra*, § 5.

[16] Cfr. art. 22 GDPR.

[17] Così l'ultima parte del considerando n. 71 del GDPR e *infra* § 3.

e ne valuti la necessità, la proporzionalità e gli eventuali rischi per i diritti e le libertà individuali[18];

6) l'obbligo di adoperare appropriate procedure matematiche o statistiche per la profilazione, con misure tecniche e organizzative adeguate al fine di garantire, in particolare, che siano rettificati i fattori che comportano inesattezze dei dati e sia minimizzato il rischio di errori[19].

Per la verità, le prime tre previsioni sono derogabili in un'ampia serie di situazioni normativamente indicate, comprese alcune assai ricorrenti nella pratica. Le condizioni che escludono il divieto di profilazione basata sui dati delle particolari categorie di cui all'art. 9 GDPR, ad esempio, comprendono il consenso dell'interessato e i motivi di interesse pubblico rilevante sulla base del diritto dell'Unione o degli Stati membri[20]. Il diritto a ricevere informazioni sull'esistenza di processi decisionali automatizzati e sulla logica utilizzata non trova applicazione quando l'interessato già disponga di quelle informazioni, e, ove i dati usati per la profilazione non siano stati ottenuti presso l'interessato, quando comunicare tali informazioni risulta impossibile o implica uno sforzo sproporzionato, nonché quando l'ottenimento o la comunicazione sono espressamente previsti dal diritto dell'Unione o dello Stato membro cui è soggetto il titolare del trattamento, sempreché tale diritto preveda misure appropriate per tutelare gli interessi legittimi dell'interessato[21]. Infine, il diritto a non essere sottoposti a decisioni basate unicamente su un trattamento automatizzato, oltre a essere soggetto a varie eccezioni di portata assai ampia[22], non si estende ai casi,

[18] Cfr. art. 35 GDPR.

[19] Cfr. considerando n. 71 del GDPR e *infra* § 5.

[20] Sempreché siano previste misure appropriate e specifiche per tutelare i diritti fondamentali e gli interessi dell'interessato; cfr. art. 9 GDPR.

[21] Cfr. artt. 13 e 14 GDPR.

[22] Il correlativo divieto non si applica ove la decisione si basi sul consenso esplicito dell'interessato, sia necessaria per l'esecuzione di un contratto con l'interessato, ovvero sia autorizzata dal diritto dell'Unione o del singolo Stato membro. Si tratta di eccezioni di portata pratica amplissima, come sottolineato da A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, in *BioLaw Journal*, 1, 2019, 80.

di gran lunga maggioritari nella prassi, in cui i risultati prodotti dall'IA vengono utilizzati (o presentati) soltanto come ausilio o supporto a decisioni adottate da autorità umane che possono comunque discostarsene, piuttosto che come unica base vincolante per la decisione[23].

Allo stato, insomma, nessuno dei limiti sopraelencati ai punti 1-3 preclude l'apprestamento di un'idonea base giuridica per l'uso di sistemi di IA che assistono le autorità giuridiche nell'adozione di decisioni in settori caratterizzati da un notevole rilievo degli interessi pubblici in gioco (welfare, ordine pubblico, sanità, sicurezza nazionale, contrasto alla criminalità ecc.), eventualmente offrendo ai decisori umani vari supporti informativo-predittivi funzionanti grazie a tecnologie di profilazione[24]. Potrà così ad esempio regolamentarsi il ricorso a strumenti di *evidence-based risk-assessment tools* su modello di quelli che sono stati incentivati dall'amministrazione giudiziaria statunitense[25] e che stanno prendendo piede anche in Europa[26]. È auspicabile, naturalmente, che l'utilizzo di questi sistemi sia subordinato al rispetto delle garanzie e dei limiti stabiliti dalle vigenti discipline in materia di protezione delle persone fisiche con riguardo al trattamento dei dati personali e, più in generale, che sia operato un congruo bilanciamento tra gli interessi

[23] Cfr. 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*, 7.2, 2016, 76 ss.; A. SIMONCINI, S. SUWEIS, *Il cambio*, cit., 99-101.

[24] In Europa, tale base giuridica sarà verosimilmente fondata sulle condizioni stabilite dall'art. 6 GDPR. Nel momento in cui scrivo, molte indicazioni sul punto possono ricavarsi dalla già citata Proposta di regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'Unione (COM/2021/206 final).

[25] L'orientamento maggioritario sviluppato dalle corti statunitensi sull'uso di sistemi algoritmici in ambito penale è favorevole all'impiego di valutazioni automatizzate dei rischi non solo a fini di prevenzione o di esecuzione ma anche in fase di quantificazione della pena, purché la decisione non si fondi esclusivamente su di esse; cfr. G. CONTISSA, G. LASAGNI, G. SARTOR, *Quando a decidere*, cit., 622 ss.

[26] Cfr. https://eurlex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:52015DC0215&from=EN.

pubblici in gioco e i diritti fondamentali degli individui, libertà personale e di pensiero *in primis*[27].

Potenzialmente più stringenti, invece, sono gli ultimi tre vincoli sopra elencati, in particolare quello di cui al punto 4, che è stato chiamato *principio di non discriminazione (per via) algoritmica.* L'ampio *corpus* del diritto antidiscriminatorio vigente a livello internazionale, europeo e nazionale prevede invero dei divieti di discriminazione che si estendono senz'altro alle decisioni algoritmiche/algoritmicamente assistite produttive di effetti giuridici che riguardano le persone o incidono significativamente su di esse. Soprattutto nel caso delle c.d. discriminazioni *dirette*, tali divieti sono assai rigorosi e soggetti soltanto a poche eccezioni tassativamente determinate. Tratterò l'argomento nei paragrafi seguenti.

3. *Il concetto di discriminazione algoritmica*

Quando si parla di *discriminazioni algoritmiche (o per via algoritmica)* si allude a varie forme di discriminazione operate tramite l'impiego di algoritmi, compresi quelli dell'IA. Tra i numerosissimi casi di discriminazione algoritmica denunciati negli ultimi anni possono ricordarsi i seguenti:

1) L'utilizzo del software COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) per valutare, grazie ad un algoritmo predittivo, il rischio di recidiva del reo e la conseguente

[27] Si pensi, giusto per fare un esempio relativo al diritto italiano, al bilanciamento tra le esigenze di sicurezza e le esigenze garantiste poste alla base del divieto del giudice di disporre perizie rivolte ad accertare l'abitualità o professionalità nel reato, la tendenza a delinquere, il carattere e la personalità dell'imputato e in genere le qualità psichiche di costui non dipendenti da cause patologiche (c.d. perizia criminologica, personologica o psicologica: art. 220, comma 2, c.p.p.). Tale divieto può forse estendersi anche all'uso di strumenti predittivi/estimativi come quelli che stiamo trattando in questo articolo, ma non vige nella fase del giudizio circa la ricorrenza delle esigenze cautelari *ex* art. 274 c.p.p. né impedisce gli «accertamenti su altre condizioni e qualità personali dell'imputato» di cui all'art. 299, comma 4-*ter*, c.p.p., funzionali alla revoca di una misura coercitiva o interdittiva.

entità della pena da infliggergli. Secondo uno studio molto citato[28], i cittadini afroamericani avrebbero avuto quasi il doppio delle probabilità rispetto ai bianchi di essere erroneamente classificati da COMPAS come soggetti ad alto rischio di recidiva violenta; di converso, i bianchi sarebbero erroneamente stati classificati come a basso rischio di recidiva assai più spesso degli afroamericani[29]. Successive analisi hanno sollevato obiezioni metodologiche su questo studio, concludendo che le stime del software sono ugualmente ben calibrate sia per gli imputati afroamericani che per quelli bianchi[30]. Tali refutazioni non hanno impedito a COMPAS di diventare eponimo di discriminazioni algoritmiche, almeno nel campo del *law enforcement*.

2) La 'discriminazione di massa' operata secondo alcuni autori dagli algoritmi dei motori di ricerca che, ad esempio, avrebbero associato le persone di colore, e in particolare le donne, a contenuti degradanti per la dignità umana. Il caso forse più noto è dato dalle ricerche sul termine 'gorillas' su Google, che nel 2016 facevano apparire tra i primi risultati svariate immagini di una giovane coppia afroamericana[31].

---

[28] Cfr. J. ANGWIN, J. LARSON, S. MATTU, L. KIRCHNER, *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*, in *ProPublica*, May 23, 2016 disponibile su https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; J. LARSON, S. MATTU, L. KIRCHNER, J. ANGWIN, *How We Analyzed the COMPAS Recidivism Algorithm*, in *ProPublica*, May 23, 2016 disponibile su https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[29] Il *leading case* che ha dato luogo al dibattito su cui insiste lo studio in questione è State of Wisconsin v. Loomis, 881 N.W.W.2d 749 (Wis. 2016). Il caso è noto anche per le sue implicazioni in tema di non esclusività della decisione algoritmica[29] (punto discusso da A. SIMONCINI, S. SUWEIS, *Il cambio*, cit., 99-101. Il caso Loomis è commentato in G. CONTISSA, G. LASAGNI, G. SARTOR, *Quando a decidere*, cit. 622 s. Si veda anche http://www.scotusblog.com/wp-content/uploads/2017/05/16-6387-CVSG-Loomis-AC-Pet.pdf.

[30] Cfr. J. KLEINBERG, S. MULLAINATHAN, M. RAGHAVAN, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, in *8th Innovations in Theoretical Computer Science Conference*, 67, 2017, 2.

[31] Cfr. S.U. NOBLE, *Algorithm of Oppression*, New York, 2018, 38. Un'ampia rassegna di situazioni di questo genere è fin troppo suggestivamente riportata da C. O'NEIL, *Armi*

3) Le *uneven failures* delle IA addestrate sulla base di *dataset* che danno conto solo dei caratteri statisticamente più rappresentati all'interno di una popolazione, con conseguente scadimento prestazionale nel trattamento di casi relativamente atipici, come sono quelli che riguardano gli individui appartenenti a certe minoranze[32].

4) La discriminazione razziale determinata dall'impiego di sistemi di identificazione tramite riconoscimento facciale da parte di varie forze di polizia statunitensi, che farebbe sì che i neri vengano fermati, indagati, arrestati, incarcerati e condannati assai più spesso dei bianchi. Ciò a causa del fatto che: *a)* i neri sono il gruppo storicamente più rappresentato nei database delle foto segnaletiche utilizzate per l'identificazione; *b)* le già accennate *uneven failures* fanno sì che i soggetti di pelle scura vengano riconosciuti meno accuratamente di quelli con pelle chiara, con conseguente maggior probabilità di errori di identificazione che sfociano in controlli ingiustificati e accuse infondate[33].

5) Le discriminazioni risultanti dall'uso di sistemi di IA funzionali alle selezioni per l'accesso a certi impieghi, al credito, a scuole/università e a prestazioni sociali di varia sorta, che in molti casi riproducono e anzi consolidano le posizioni di forza storicamente assunte da determinati

di distruzione matematica. Come i 'big data' aumentano la disuguaglianza e minacciano la democrazia, Milano, 2017.

[32] Un esempio è dato dai sistemi di riconoscimento facciale esaminati da J. BUOLAMWINI, T. GEBRU, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in *Proceedings of Machine Learning Research*, 81, 2018, 77-91, che presentano tassi d'errore dello 0,8% nel riconoscimento dei maschi dalla pelle più chiara e fino al 34,7% nel riconoscimento delle femmine dalla pelle più scura.

[33] Cfr. F. BACCHINI, L. LORUSSO, *Race, Again: How Face Recognition Technology Reinforces Racial Discrimination*, in *Journal of Information, Communication and Ethics in Society*, 17.3, 2019, 321-335. Va aggiunto che i più sofisticati sistemi di riconoscimento facciale basati sull'IA hanno risolto gran parte dei citati problemi di *uneven failures* e hanno esteso il proprio database di riferimento a miliardi di immagini ricavate dai social media, news media e da molte altre fonti aperte, ponendo tuttavia in alcuni casi altri problemi rilevanti per la disciplina del trattamento dei dati personali (cfr. a titolo d'esempio l'ordinanza di ingiunzione del Garante per la protezione dei dati personali nei confronti di *Clearview AI* [una piattaforma di riconoscimento facciale utilizzata da varie forze di polizia per l'identificazione dei sospetti] - 10 febbraio 2022 [9751362]).

gruppi, facendo preferire ad esempio soggetti di sesso maschile e/o di razza bianca a soggetti di sesso femminile e/o non bianchi[34].

6) Le discriminazioni razziali nell'allocazione delle risorse economiche da destinare alle cure sanitarie, calcolate da algoritmi che valutano le condizioni di salute dei pazienti appartenenti a certe minoranze come migliori di quanto non siano realmente poiché prendono in considerazione le spese pregresse come elemento rivelatore dei bisogni sanitari degli individui[35].

La pur rapida rassegna appena operata evidenzia che alla categoria concettuale delle 'discriminazioni algoritmiche' vengono ascritte cose molto diverse: trattamenti svantaggiosi direttamente ricollegati a razza, sesso e altre caratteristiche personali protette dal diritto antidiscriminatorio; situazioni che risultano da *errori* degli (o negli) algoritmi, le cui prestazioni scadono 'selettivamente' con riguardo ai soggetti appartenenti a certe minoranze; disparità di trattamento derivanti da vari *bias* nei dati utilizzati dagli algoritmi per elaborare le proprie stime, che perciò risultano inattendibili; decisioni o policies adottate col supporto di sistemi algoritmici che, pure accurati e attendibili nelle loro previsioni, svantaggiano sistematicamente gli individui appartenenti a taluni gruppi e così perpetuano condizioni di subalternità e svantaggio che sono frutto di inveterate situazioni sociali, eventualmente determinate o favorite da un passato di discriminazioni, ecc.[36]

---

[34] Una trattazione approfondita della materia, corredata da proposte su un uso disciplinato dell'IA in questi ambiti è J. KLEINBERG, J. LUDWIG, S. MULLAINATHAN, C. R SUNSTEIN, *Discrimination in the Age of Algorithms*, in *Journal of Legal Analysis*, 10, 2018.

[35] Cfr. Z. OBERMEYER, B. POWERS, C. VOGELI, S. MULLAINATHAN, *Dissecting racial bias in an algorithm used to manage the health of populations*, in *Science*, 366.6464, 2019, 447-453; R. BENJAMIN, *Assessing risk, automating racism*, in *Science*, 366.6464, 2019, 421 s.

[36] Quando ad esempio i tassi di recidiva previsti dai sistemi del tipo di quelli ricordati al punto 1 sono elaborati considerando dei dati su pregresse condanne che sono *anche* il prodotto di pregiudizi razziali delle autorità rispetto a certi gruppi, allora tali sistemi sopravvalutano i rischi di recidiva dei soggetti che vi appartengono. Il problema si pone anche per gli usi dell'IA nella pianificazione dei controlli di polizia, qualora i sistemi siano addestrati su dati che sovrarappresentano l'incidenza dei crimini in certi gruppi etnici. L'algoritmo indirizzerà infatti le forze di polizia a fermare e controllare più

Queste situazioni chiamano in causa tutte le numerose e talora problematiche 'forme' della discriminazione individuate dalla teoria giuridica degli ultimi due decenni, distinguibili fra l'altro secondo il soggetto che discrimina (personale, impersonale o non identificabile), il destinatario della discriminazione (individuo o gruppo), l'intenzionalità della discriminazione, l'oggetto della discriminazione, gli effetti dell'atto discriminatorio nonché il suo collegamento diretto o indiretto con la caratteristica protetta dalla discriminazione[37]. Sorge tuttavia il sospetto che non tutte le eterogenee istanze di discriminazione algoritmica denunciate in letteratura possano essere ricondotte a un unico concetto giuridico, per quanto generico. Occorre dunque individuare un concetto di discriminazione algoritmica sufficientemente determinato da consentirci almeno di riconoscere gli abusi semantici più lampanti, ciò che tenterò di fare nelle righe che seguono.

Ho osservato sopra che quando si parla di *discriminazioni algoritmiche (o per via algoritmica)* ci si riferisce a discriminazioni operate tramite l'impiego di algoritmi, compresi quelli dell'IA. Questa generica nozione, come pure il correlato *principio di non discriminazione per via algoritmica* di cui si dirà appresso, eredita dal concetto generale di discriminazione elevatissimi e ben noti importi di equivocità[38], sebbene dagli usi linguistici delle autorità giuridiche, dei giuristi e dei teorici del diritto

persone appartenenti a questi gruppi etnici, col risultato che, statisticamente, verranno scoperti più reati commessi da questi soggetti rispetto a quelli commessi da soggetti appartenenti ad altri gruppi. Quando i dati relativi ai nuovi reati scoperti saranno aggiunti al dataset, il tasso di criminalità di quel gruppo etnico verrà ancora più sovrastimato, in una sorta di circolo vizioso di rafforzamento dell'effetto discriminatorio; cfr. K. MILLER, *Total Surveillance, Big Data, and Predictive Crime Technology: Privacy's Perfect Storm*, in *Journal Technology of Law and Policy*, 19, 2014, 105, 127.

[37] Una rassegna critica delle numerose forme della discriminazione è in E. CONSIGLIO, *Che cos'è la discriminazione?*, cit., capp. III e IV.

[38] A constatare l'assenza di una definizione universalmente accettata di 'discriminazione' è W. VANDENHOLE, *Non-Discrimination and Equality in the View of the UN Human Rights Treaty Bodies*, Oxford, 2005, 33. Altri autori rilevano addirittura la polisemia del termine 'discriminazione' e dunque la necessità di individuarne diversi concetti; cfr. P.S. SHIN, *Is There a Unitary Concept of Discrimination?*, in *Philosophical Foundations of Discrimination Law*, ed. by D. Hellman and S. Moreau, Oxford, 2013, 164.

possano ricavarsi alcune indicazioni assai utili all'analisi. Costoro usano infatti il termine 'discriminazione' con riferimento prevalente alle prescrizioni generali o singolari (ad esempio leggi, provvedimenti amministrativi, decisioni giudiziarie), ai criteri (ad esempio classificazioni e distinzioni) e alle pratiche (ad esempio azioni e omissioni) che determinano uno svantaggio relativo[39] subito da taluni soggetti in quanto portatori di certe 'caratteristiche protette'[40] ricomprese in un catalogo non tassativo, aperto e storicamente variabile che nell'Occidente contemporaneo include il sesso, la razza, il colore della pelle o l'origine etnica o sociale, le caratteristiche genetiche, la lingua, la religione, le convinzioni personali, le opinioni politiche o di qualsiasi altra natura, l'appartenenza a una minoranza nazionale, il patrimonio, la nascita, la disabilità, l'età e l'orientamento sessuale[41]. Se, come mi pare opportuno, ci si accosta alla nozione di discriminazione algoritmica muovendo da questo concetto giuridico generale di discriminazione, se ne può evidenziare il connotato eminentemente tecnico-strumentale: è discriminazione algoritmica la prescrizione generale o singolare, la

[39] Questo svantaggio è solitamente ricostruito come una distinzione, esclusione o restrizione nel godimento o nell'esercizio dei diritti umani e delle libertà fondamentali in ogni ambito.

[40] Tali caratteristiche, nel linguaggio specialistico dei cultori del diritto antidiscriminatorio, sono denominate anche 'fattori di protezione' o 'fattori protetti'; i gruppi dei soggetti portatori di tali caratteristiche sono chiamati 'gruppi protetti'. Nel testo mi conformerò a questi usi linguistici.

[41] Questo elenco è tratto dall'art. 21 della Carta dei diritti fondamentali dell'Unione europea, fatta a Nizza il 7 dicembre 2000 ed entrata in vigore il 12 dicembre 2007 (2000/C 364/01). Nel campo delle discriminazioni algoritmiche, si segnala il tentativo di introdurre una definizione di discriminazione che esula da qualsiasi riferimento a un catalogo fisso di caratteristiche protette. L'art. 4 lett. m del Regolamento allegato al Quadro relativo agli aspetti etici dell'intelligenza artificiale, della robotica e delle tecnologie correlate di cui alla Risoluzione del Parlamento europeo del 20 ottobre 2020 definisce infatti la 'discriminazione', come «qualsiasi trattamento differenziato di una persona o di un gruppo di persone per un motivo privo di giustificazione obiettiva o ragionevole e, pertanto, vietato dal diritto dell'Unione». Nonostante il lodevole intento, non può non segnalarsi l'elevatissimo grado di genericità e vaghezza del concetto di 'giustificazione obiettiva o ragionevole', giustificabile forse soltanto alla luce della valenza parenetica, più che direttamente precettiva, della disposizione in oggetto.

qualificazione o la pratica che determina svantaggi relativi per taluni soggetti in quanto portatori delle summenzionate caratteristiche protette dal diritto, *se adottata o attuata (anche) mediante l'impiego di algoritmi, compresi quelli dell'IA*. Uso la formula 'adottata o attuata mediante l'impiego di algoritmi' in un senso sufficientemente ampio da ricomprendere sia i 'processi decisionali automatizzati relativi alle persone fisiche' di cui parla ad es. l'art. 22 del GDPR, sia ogni altra decisione adottata col supporto di algoritmi che forniscono ai decisori umani stime, classificazioni, graduatorie, descrizioni e previsioni su una certa situazione o caratteristica a cui il diritto ricollega effetti giuridici, direttamente o a seguito della decisione stessa. D'ora innanzi qualificherò queste decisioni (pratiche, *policies*, ecc.) informate o coadiuvate da algoritmi come 'algoritmicamente assistite'.

Questa ricostruzione preliminare del concetto di discriminazione algoritmica fa da *pendant* a quello che certa dottrina ha già denominato 'principio di non discriminazione per via algoritmica'[42], che declinerebbe il più generale principio di non discriminazione nel campo dei processi decisionali algoritmicamente assistiti produttivi di effetti giuridici per gli interessati. Nello spazio giuridico europeo, i fondamenti normativi di questo principio sono reperiti in una serie di atti e documenti ufficiali, perlopiù sprovvisti di efficacia vincolante diretta, tra cui spicca il considerando 71 del GDPR nella parte in cui raccomanda che «Al fine di garantire un trattamento corretto e trasparente nel rispetto dell'interessato, tenendo in considerazione le circostanze e il contesto specifici in cui i dati personali sono trattati, è opportuno che il titolare del trattamento utilizzi procedure matematiche o statistiche appropriate per la profilazione, metta in atto misure tecniche e organizzative adeguate al fine di garantire, in particolare, che siano rettificati i fattori che comportano inesattezze dei dati e sia minimizzato il rischio di errori e al fine di garantire la sicurezza dei dati personali secondo una modalità

[42] Tale principio è denominato 'principio di non discriminazione per via algoritmica' da A. SIMONCINI, S. SUWEIS, *Il cambio*, cit., 101 s., e 'principio di non discriminazione algoritmica' da A. SIMONCINI, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, in *BioLaw Journal*, 1, 2019, 84-86.

che tenga conto dei potenziali rischi esistenti per gli interessi e i diritti dell'interessato e che *impedisca tra l'altro effetti discriminatori nei confronti di persone fisiche sulla base della razza o dell'origine etnica, delle opinioni politiche, della religione o delle convinzioni personali, dell'appartenenza sindacale, dello status genetico, dello stato di salute o dell'orientamento sessuale, ovvero che comportano misure aventi tali effetti*» (corsivo mio). A ben vedere, dall'involuta formulazione di questo considerando si evince una prescrizione di portata e forza vincolante alquanto ridotte[43]: ci si limita invero a raccomandare che la sicurezza dei dati personali vada garantita con modalità che impediscono, fra l'altro, effetti discriminatori. Gli altri atti di *soft law* da cui viene ricavato il principio in parola, per parte loro, lungi dallo specificarne ulteriori connotati normativi si limitano a ribadire l'applicabilità del generale principio di non discriminazione anche all'ambito dell'IA e degli algoritmi in genere[44]. Una precisazione tutto sommato ridondante, atteso che non v'è alcun dubbio che l'ampio corpus dell'odierno diritto antidiscriminatorio trovi applicazione indipendentemente dai mezzi contingentemente impiegati per operare le discriminazioni giuridicamente proscritte, soprattutto quando esse incidono sul godimento paritario dei diritti fondamentali degli individui[45].

Il principio di non discriminazione algoritmica può insomma essere ricostruito come la specificazione del divieto generale di discriminazione nel campo delle prescrizioni, valutazioni e pratiche adottate o attuate mediante il ricorso ad algoritmi, compresi quelli dell'IA. Ciò implica che la sua portata prescrittiva vada determinata sulla base della contingente configurazione del principio generale di non discriminazione nel diritto

[43] I considerando dei regolamenti europei, come è noto, sono sprovvisti di efficacia normativa vincolante diretta, e costituiscono elemento integrante la 'motivazione' dell'atto normativo, utile soprattutto a fini interpretativi e applicativi.

[44] Un elenco di tali atti è in G. GIORGINI PIGNATELLO, *Il contrasto alle discriminazioni algoritmiche*, in *Federalismi.it*, 16, 2021, 173 ss.

[45] Sui cui fondamenti normativi per, rispettivamente, il diritto europeo e statunitense si vedano E. Consiglio, *Che cos'è la discriminazione?*, cit., 27-39 e J. KLEINBERG, J. LUDWIG, S. MULLAINATHAN, C. R SUNSTEIN, *Discrimination*, cit., 6-9.

positivo di volta in volta considerato[46]. Tale principio potrà evidentemente contemplare un catalogo più o meno ampio di forme di discriminazione e fattori di protezione, ed essere soggetto a eccezioni più o meno numerose, tassative e specifiche. È su queste basi che occorre verificare se le presunte istanze di discriminazione algoritmica individuate dalla letteratura degli ultimi anni siano giuridicamente qualificabili come "discriminazioni" e dunque siano sanzionabili se si tratta di pratiche, annullabili se si tratta di previsioni normative generali, impugnabili se si tratta di decisioni giudiziali ecc.

4. *Discriminazioni algoritmiche dirette e indirette*

Nella maggior parte dei diritti occidentali, il principio di non discriminazione vieta sia la discriminazione *diretta* sia quella *indiretta*. La distinzione in oggetto, dibattuta sia sul piano analitico-concettuale che su quello etico-fondazionale[47], oppone grossomodo *a)* i trattamenti sfavorevoli di gruppi o persone a causa, in ragione o a motivo di certe loro caratteristiche protette a *b)* le prescrizioni, i criteri e le pratiche che svantaggiano in modo proporzionalmente maggiore gli appartenenti a un gruppo portatore di una caratteristica protetta, pur non essendo

[46] Sul diritto antidiscriminatorio in ambito europeo, si vedano: FUNDAMENTAL RIGHTS AGENCY, *Handbook on European Non-Discrimination Law*, Lussemburgo, 2018; *The European Convention on Human Rights and the Principle of Non Discrimination*, ed. by M. Balboni, Napoli, 2017; O. SIDHU, *The Concept of Equality of Arms in Criminal Proceedings under Article 6 of the European Convention on Human Rights*, Cambridge-Antwerp-Portland, 2017, 63, 76 s., 85 s.; G.P. DOLSO, *Il principio di non discriminazione nella giurisprudenza della Corte europea dei diritti dell'uomo*, Napoli, 2013; S. HAVERKORT-SPEEKENBRINK, *European Non-Discrimination Law - A Comparison of EU Law and the ECHR in the Field of Non-Discrimination and Freedom of Religion in Public Employment with an Emphasis on the Islamic Headscarf Issue*, Oxford, 2012; AGENZIA DELL'UNIONE EUROPEA PER I DIRITTI FONDAMENTALI E CONSIGLIO D'EUROPA, *Manuale di diritto europeo della non discriminazione*, Lussemburgo, 2011; C. FAVILLI, *La non discriminazione nell'Unione Europea*, Bologna, 2011; I. CASTANGIA, G. BIAGIONI, *Il principio di non discriminazione nel diritto dell'Unione europea*, a cura di I. Castangia e G. Biagioni, Napoli, 2011.

[47] Cfr. ad es. G. RUTHERGLEN, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, in *Fordham Law Review*, 74, 2006. Si veda anche *infra*, nota 50.

specificamente ad essi e a ciò indirizzate[48]. Come è noto, le discriminazioni della prima forma sono vietate con particolare rigore sia in ambito europeo che negli Stati Uniti; sarebbe senz'altro illecito in entrambi i contesti, ad esempio, l'uso di un algoritmo funzionale alle selezioni per l'accesso al credito, alla sanità o a scuole/università che fosse programmato per assegnare un minor punteggio ai candidati *in quanto* donne, neri, musulmani, omosessuali ecc. In Europa, tale divieto è soggetto a un'unica clausola di eccezione; la discriminazione diretta è infatti esclusa quando la caratteristica protetta dal diritto antidiscriminatorio sia requisito essenziale, proporzionato e determinante per la selezione che si intende operare, in ragione della natura dell'attività e del contesto e purché l'obiettivo della selezione sia legittimo. Tutte evenienze, queste, riconosciute soltanto in poche ipotesi tassativamente determinate per via normativa e/o giurisprudenziale[49].

[48] Una sintesi delle definizioni legislative e giurisprudenziali di discriminazione indiretta è in E. CONSIGLIO, *Che cosa è la discriminazione?*, cit., p. 95, secondo cui: «Si ha discriminazione indiretta quando una previsione, un criterio o una pratica (un atto, un patto, un comportamento attivo oppure omissivo) apparentemente neutri – che non operano cioè una classificazione sulla base di un fattore protetto (e neppure sulla base di un tratto contiguo o vicino a un fattore protetto, in modo che attraverso questo possa essere individuata la discriminazione) – possono mettere le persone [… portatrici di una qualsiasi] caratteristica protetta, in una situazione di particolare svantaggio rispetto ad altre persone (ovvero trattino in modo proporzionalmente svantaggioso ovvero abbiano un effetto proporzionalmente svantaggioso su un gruppo che possiede una caratteristica protetta), a meno che la disposizione, il criterio o la pratica siano oggettivamente giustificati da una finalità legittima, e i mezzi impiegati per il suo conseguimento siano appropriati e necessari al raggiungimento del fine». L'autrice ricava questa sintesi da una comparazione dell'art. 2, comma 2, lett. b), direttiva sull'eguaglianza razziale; art. 2, comma 2, lett. b), direttiva sulla parità di trattamento in materia di occupazione; art. 2, comma 1, lett. b), direttiva sulla parità di trattamento fra uomini e donne; art. 2, lett. b), direttiva sulla parità di trattamento tra uomini e donne in materia di accesso ai beni e ai servizi.

[49] Ciò accade ad esempio nel campo dei mestieri artistici, per la selezione a ricoprire certi ruoli teatrali o cinematografici particolarmente adatti a soggetti di un certo sesso, di una certa razza o di una certa età. Un'altra eccezione importante riguarda le organizzazioni religiose e quelle c.d. 'di tendenza' (politiche, sindacali, culturali ecc.), che secondo il diritto europeo non discriminano gli individui ancorché li trattino diversamente sulla base della loro religione o delle convinzioni personali quando queste

Meno immediata è invece la qualificazione, e dunque la proscrizione giuridica, di una certa pratica o policy come discriminazione *indiretta*, e ciò sia perché la relativa nozione è di per sé vigorosamente contestata (non senza buone ragioni)[50], sia perché le eccezioni alla configurabilità delle discriminazioni di questa forma, almeno nello spazio giuridico europeo, non sono legate unicamente a poche ipotesi straordinarie e tassative, bensì riconosciute tutte le volte in cui una certa pratica, decisione o *policy*, pur produttiva di effetti svantaggiosi che colpiscono in modo proporzionalmente maggiore gli appartenenti a una categoria protetta, sia oggettivamente giustificata da una finalità legittima perseguita attraverso mezzi appropriati e necessari[51]. Nel diritto

rappresentino un requisito essenziale, legittimo e giustificato per lo svolgimento dell'attività lavorativa, tenuto conto dell'etica dell'organizzazione; per una trattazione di queste e altre eccezioni al divieto di discriminazione diretta si veda E. CONSIGLIO, *Che cos'è la discriminazione?*, cit. 105-109. Sul regime delle eccezioni al divieto di discriminazione diretta (*disparate treatment*) nel diritto statunitense, si veda M.C. HARPER, *Confusion on the Court: Distinguishing Disparate Treatment from Disparate Impact in Young v UPS and EEOC v Abercrombie & Fitch, Inc.*, in *Boston University Law Review*, 96.2, 2016, 543-570.

[50] Contro l'indebita estensione del concetto di discriminazione fuori dal dominio delle discriminazioni dirette si vedano almeno I. M. YOUNG, *Justice and the Politics of Difference,* Princeton, 1990; M. CAVANAGH, *Against Equality of Opportunity*, Oxford, 2002; M. SELMI, *Indirect Discrimination and the Anti-Discrimination Mandate*, in *Philosophical Foundations*, cit., 250-268; B. EIDELSON, *Discrimination and Disrespect*, Oxford, 2015. Alcuni autori sono propensi a riconoscere alla discriminazione indiretta un valore meramente strumentale al contrasto della discriminazione diretta, che sarebbe l'unica ad avere un disvalore specifico; cfr. ad es. H. COLLINS, T. KHAITAN, *Indirect Discrimination Law: Controversies and Critical Questions*, in *Foundations of Indirect Discrimination Law*, ed. by H. Collins, T. Khaitan, Oxford, 2018, 25-27.

[51] Cfr. art. 2, comma 2, lett. b), direttiva 2000/43/CE del Consiglio, del 29 giugno 2000, che attua il principio della parità di trattamento fra le persone indipendentemente dalla razza e dall'origine etnica; art. 2, comma 2, lett. b), direttiva 2000/78/CE del Consiglio, del 27 novembre 2000, che stabilisce un quadro generale per la parità di trattamento in materia di occupazione e di condizioni di lavoro; art. 2, lett. b), direttiva 2004/113/CE, del 13 dicembre 2004, che attua il principio della parità di trattamento tra uomini e donne per quanto riguarda l'accesso a beni e servizi e la loro fornitura; art. 2, comma 1, lett. b), direttiva 2006/54/CE del Parlamento europeo e del Consiglio, del 5 luglio

statunitense, analogamente, le pratiche, decisioni e policies *prima facie* neutre ma inavvertitamente produttive di qualche *disproportionate adverse effect* sui membri di un gruppo protetto (donne, afroamericani ecc.) sfuggono al divieto di discriminazione indiretta ove vengano fornite adeguate giustificazioni della misura in questione[52].

Nell'odierno diritto antidiscriminatorio occidentale, in sintesi, le pratiche, decisioni e policies che, pur apparentemente imparziali, hanno effetti che mettono le persone appartenenti ai gruppi protetti in una situazione di particolare svantaggio, tendono a essere consentite qualora siano assolti gli speciali oneri giustificativi gravanti sui soggetti titolari del potere giuridico di attuarle o disporle.

Quanto al merito di tali giustificazioni, non sfugge al giurista che l'accertamento della ricorrenza di una finalità 'oggettivamente' legittima, dell'"appropriatezza e necessità' dei mezzi impiegati per il suo conseguimento e dell'"adeguatezza' di una giustificazione rimandano a valutazioni altamente discrezionali degli interpreti competenti a operare le necessarie qualificazioni (sebbene, in ambito europeo, dottrina e giurisprudenza abbiano specificato che i mezzi sono 'appropriati' e 'necessari' quando non ve ne siano di altri che evitano qualsiasi svantaggio per i membri del gruppo protetto e quando questo svantaggio è il minimo indispensabile per garantire il raggiungimento della finalità legittima). Va inoltre notato che tali giustificazioni richiedono argomenti più o meno articolati e convincenti a seconda della caratteristica protetta e della materia considerata. Ad esempio, sia in Europa sia negli Stati Uniti gli oneri giustificativi delle disparità di trattamento fondate sulla razza sono generalmente molto più gravosi di quelli delle disparità di trattamento fondate sull'età, in particolare in materia di lavoro e formazione professionale. Analoghe aperture alla discrezionalità degli interpreti risultano infine dalle formulazioni dottrinarie, normative e giurisprudenziali delle eccezioni alla configurabilità di una

2006, riguardante l'attuazione del principio delle pari opportunità e della parità di trattamento fra uomini e donne in materia di occupazione e impiego.

[52] Cfr. J. KLEINBERG, J. LUDWIG, S. MULLAINATHAN, C. R SUNSTEIN, *Discrimination*, cit., 22 ss.

discriminazione diretta, che quando menzionano una 'sproporzione' o una 'particolarità' negli svantaggi raramente chiariscono se si alluda alla maggiore incidenza statistica degli effetti svantaggiosi tra i membri del gruppo protetto oppure alla speciale afflittività degli effetti svantaggiosi per costoro, a cagione di qualche vulnerabilità loro peculiare. Tutte queste indeterminatezze hanno degli ovvi costi sul piano della certezza del diritto, e il problema è accentuato dall'espansione del campo d'applicazione del diritto antidiscriminatorio, che si estende alla protezione di un numero crescente di caratteristiche personali, comprese perfino le opinioni di qualsiasi natura[53]. Si tratta tuttavia di difficoltà che non riguardano soltanto le discriminazioni indirette operate mediante l'impiego di algoritmi, ma tutte le discriminazioni indirette, di qualsiasi sorta, dunque posso interromperne qui l'esame[54].

Considerate nella loro modalità specificamente algoritmica, invece, le discriminazioni indirette non presentano peculiarità degne di nota, giacché esse occorrono indipendentemente dal mezzo impiegato per

[53] Potrebbe sorgere ad esempio il problema di giustificare espressamente, in quanto potenziale discriminazione indiretta, una misura giuridica che, pur adeguata a raggiungere obiettivi di tutela della salute pubblica, produca effetti che svantaggiano 'sproporzionatamente' gli appartenenti a un certo gruppo d'opinione, come accaduto nel caso delle restrizioni correlate al c.d. *Green Pass*, che hanno colpito soprattutto gli appartenenti al movimento *no vax*. Per una trattazione dell'argomento mi permetto di rinviare a G. GOMETZ, *Green pass e discriminazione. Un'analisi*, in *L'ircocervo*, 2, 2021, 156-171.

[54] Aggiungo soltanto che i costi in termini di certezza del diritto, non mancano anche nei casi in cui si intenda verificare la ricorrenza dell'eccezione che giustifica una pratica/policy altrimenti qualificabile come discriminazione diretta, giacché pure le qualificazioni su che cosa costituisca 'requisito essenziale, determinante e proporzionato al conseguimento di un obiettivo legittimo' rimandano a valutazioni discrezionali altamente opinabili e, dunque, contestabili, ancorché in quest'ambito tenda a prevalere un indirizzo secondo cui le deroghe al divieto di discriminazione diretta vadano argomentate e provate in modo particolarmente completo, stringente e persuasivo, e possano essere riconosciute solo in ipotesi che rientrano in un novero tassativo determinato normativamente e/o per via giurisprudenziale. Anche l'accertamento dell'occorrenza di queste eccezioni, peraltro, non presenta criticità peculiari alle discriminazioni algoritmiche, dunque limito a queste poche notazioni la loro trattazione.

realizzarle, dalla presenza di un intento discriminatorio esplicito o implicito e da qualsivoglia considerazione a scopi discriminatori di un fattore di protezione. Ai fini della qualificazione di una discriminazione come indiretta, infatti, non contano le caratteristiche del trattamento che si suppone discriminatorio o la sua eventuale esecuzione per via algoritmica, ma i suoi *effetti* particolarmente svantaggiosi per gli appartenenti ai gruppi portatori di una caratteristica protetta[55]. Per accertare l'occorrenza di una discriminazione di questa forma non è dunque necessario ispezionare gli algoritmi o i loro dataset di addestramento alla ricerca di una qualche prova circa la funzionalizzazione dei dati relativi a caratteristiche protette alla produzione degli *output* del sistema. Sarà invero sufficiente accertare se la decisione o policy algoritmicamente assistita sia produttiva di effetti che nel complesso mettono gli appartenenti a qualche gruppo protetto in una condizione di particolare svantaggio, e valutare se ricorrano o meno le giustificazioni che ricordavo poc'anzi. Anche questi accertamenti e valutazioni non presentano criticità diverse da quelle che riguardano le discriminazioni indirette in genere, dunque esulano dalla presente trattazione. Mi limito a constatare che, una volta appurate queste circostanze, il contrasto alle discriminazioni algoritmiche qualificabili come discriminazioni indirette risulta relativamente più agevole di quello alle discriminazioni algoritmiche *dirette* riconducibili all'impiego di IA funzionanti sulla base dell'odierno paradigma delle reti neurali ad apprendimento automatico. L'accertamento di una discriminazione diretta richiede infatti di elucidare e provare le *ragioni* del trattamento asseritamente discriminatorio[56], ossia di appurare se una certa norma, decisione o pratica algoritmicamente assistita abbiano svantaggiato dei soggetti a cagione della loro appartenenza a una

[55] Cfr. O.M. FISS, *The Fate of an Idea Whose Time Has Come: Antidiscrimination Law in the Second Decade after Brown v. Board of Education*, in *The University of Chicago Law Review*, 41, 1974, 764 s.

[56] Tali ragioni, secondo la teoria giuridica oggi prevalente, possono comunque essere riconosciute 'oggettivamente', senza necessità di affrontare la *probatio diabolica* dell'accertamento dei 'motivi' psicologici soggettivi riguardanti l'intento o il proposito di discriminare.

categoria protetta. Ciò comporta delle particolari difficoltà quando a fornire le ragioni di quel trattamento siano delle IA, eventualmente con l'avallo dei decisori umani. Mi occupo di questo problema nel paragrafo seguente.

5. *Discriminazioni e affidabilità delle stime algoritmiche*

Come osservavo nel primo paragrafo, le reti neurali basate sul machine learning sono particolarmente abili nello scoprire automaticamente delle correlazioni tra dati per poi estrapolarne di nuovi, ciò che ci consente di utilizzarle come formidabili strumenti di analisi e previsione di comportamenti, qualità e disposizioni delle persone fisiche. Tale *profilazione*, oggi perlopiù applicata nel marketing[57], in un prossimo futuro potrebbe essere impiegata in settori assai più pregnanti sotto il profilo pubblicistico: welfare, sicurezza, sanità, *law enforcement* ecc. Ove vi sia sufficiente disponibilità dei dati personali degli individui, e sempre nel rispetto delle vigenti regolazioni del loro trattamento[58], si potrebbero ad esempio usare le IA per scoprire se un certo soggetto sia più (o meno) bisognoso di particolari prestazioni sociali, oppure presenti maggiori probabilità di recidiva dopo esser stato condannato, o di fuga dopo esser stato indagato, o se si stia radicalizzando come terrorista di matrice religiosa, o ancora se sia particolarmente propenso a commettere certi crimini e via dicendo. Il limite di questi strumenti, come si è visto, è che

[57] Ad es. da Amazon, per suggerire ai potenziali acquirenti i prodotti più conformi ai propri gusti e interessi.

[58] Va ricordato che in molti contesti a elevata rilevanza pubblicistica, le odierne regolamentazioni in materia di *privacy* e protezione delle persone con riguardo al trattamento dei dati personali prevedono delle tutele e delle garanzie alquanto ridotte. In Europa, ad esempio, il trattamento dei dati personali non richiede il consenso degli interessati quando sussista una base legittima prevista per legge dal GDPR o dal diritto dell'Unione o degli Stati membri; cfr. considerando 40 e 41 e art. 6 GDPR, specie le lett. d) ed e), in cui si stabilisce che il trattamento è lecito, indipendentemente dal consenso dell'interessato, ove sia necessario per la salvaguardia degli interessi vitali dell'interessato o per l'esecuzione di un compito di interesse pubblico o connesso all'esercizio di pubblici poteri di cui è investito il titolare del trattamento.

possono indicare più o meno esattamente *come, chi e quali* sono tutti quei soggetti, ma non possono dar conto del *perché* lo sono. Quelle IA non possono chiarire esplicitamente, in particolare, se, come e quanto delle caratteristiche protette (ad es. la razza, il sesso, l'orientamento sessuale, l'età, la religione, le opinioni personali ecc.) abbiano influito nelle stime di volta in volta presentate, né quale sia il loro peso relativamente ad altre caratteristiche non protette dal diritto antidiscriminatorio ma parimenti documentate nei dati personali accessibili al sistema (quali ad esempio preferenze d'acquisto, dati di geolocalizzazione, 'likes', 'tag', siti web visitati, contatti ecc.)[59]. Tale opacità discende – lo ricordo ancora – non da limitazioni superabili con idonei accorgimenti adottati in sede di progettazione degli algoritmi, ma da motivi strutturali legati al loro funzionamento come previsori statistico-quantitativi, piuttosto che come operatori razionali capaci di compiere inferenze logico-causali.

Questo richiamo al paradigma statistico di funzionamento delle IA può facilmente ingenerare una confusione. Si potrebbe cioè pensare che la maggior parte delle discriminazioni algoritmiche che sto trattando in questo articolo possano essere ascritte alla categoria delle *discriminazioni statistiche*, e per questa via venire proscritte dai diritti occidentali come discriminazioni dirette o comunque direttamente discendenti da una caratteristica personale dei soggetti discriminati[60]. Una discriminazione statistica, com'è noto, occorre quando un fattore di protezione (razza, sesso, ecc.) viene utilizzato come indicatore statistico di altre caratteristiche o disposizioni ordinariamente non visibili ma ricollegate

[59] Cfr. J. KLEINBERG, J. LUDWIG, S. MULLAINATHAN, Z. OBERMEYER, *Prediction Policy Problems*, in *American Economic Review*, 105.5, 2015, 491.

[60] Le discriminazioni statistiche sono però talora accostate non alle discriminazioni dirette ma a quelle indirette, sulla scorta del rilievo che il trattamento discriminatorio non è ricollegato direttamente alla caratteristica protetta, bensì a una seconda caratteristica probabilisticamente inferita da questa. Anche in questo secondo caso, la discriminazione consiste nel trattamento diseguale operato in ragione di una caratteristica del soggetto discriminato, piuttosto che negli effetti del trattamento stesso sui portatori della caratteristica protetta considerati nel loro complesso, ciò che fa ritenere che le discriminazioni statistiche siano assimilabili più alle discriminazioni dirette che a quelle indirette.

a trattamenti svantaggiosi (quali controlli mirati, misure di sicurezza, esclusione da certe prestazioni sociali ecc.). Se ad esempio in una certa società gli appartenenti a un particolare gruppo etnico versano, in media, in condizioni economiche particolarmente disagiate, magari a causa delle discriminazioni di cui quel gruppo è stato vittima in passato, potrà facilmente spiegarsi il più elevato tasso di criminalità che si registra tra costoro. Se tuttavia questo dato è confermato da successive osservazioni empiriche, allora può razionalmente operarsi quella che Frederick Schauer chiama una *generalizzazione non universale pura*, ossia fondata su una buona base statistica[61]: l'appartenenza di un certo soggetto a un gruppo etnico nel quale si registra un alto tasso di criminalità è un elemento che, singolarmente considerato, determinerà la stima di una maggiore probabilità di quel soggetto di commettere crimini, con possibili conseguenze sfavorevoli che potranno andare da un'intensificazione dei controlli di polizia a giudizi negativi operati nelle sedi in cui si valutino il pericolo di fuga, di recidiva ecc. Ecco allora la discriminazione statistica, consistente nell'aver impiegato la razza come indicatore statistico da cui inferire una maggiore probabilità di commettere crimini e, dunque, come ragione per disporre un trattamento sfavorevole.

Il disvalore di queste discriminazioni è intuitivo: esse implicano trattamenti sfavorevoli estesi anche a individui che, singolarmente considerati, *non* presentano affatto la caratteristica che *giustifica* quegli stessi trattamenti. Eppure, le associazioni tra (ciò che gli esseri umani riconoscono come) le diverse caratteristiche degli individui e le loro attitudini e disposizioni sono proprio il genere di correlazioni che i sistemi di IA basati sul machine learning sono rapidissimi a scoprire automaticamente a partire dai dati di volta in volta disponibili. Questa formidabile capacità associativa potrebbe far ritenere che essi operino delle discriminazioni statistiche censurabili a titolo di discriminazione diretta tutte le volte in cui considerino una caratteristica protetta come *proxy*, ossia indicatore statistico, di un'altra caratteristica a cui vengono ricollegati effetti sfavorevoli (come nel caso poc'anzi portato ad

[61] Cfr. F. SCHAUER, *Di ogni erba un fascio*, Bologna, 2008, 18 ss.

esempio). Ciò – si badi – anche qualora la relazione statistica tra la caratteristica protetta e quella collegata al trattamento sfavorevole sia effettivamente sussistente: molti ordinamenti giuridici vietano infatti la discriminazione statistica anche quando è basata su una generalizzazione statisticamente fondata[62].

La descritta assimilazione tra le discriminazioni algoritmiche e quelle statistiche risente tuttavia di un equivoco. Le discriminazioni statistiche si fondano su argomenti schematizzabili nella forma 'molti X sono Y; Tizio è X; dunque v'è una particolare probabilità che Tizio sia Y', dove Y è una caratteristica o disposizione a cui vengono giuridicamente ricollegati degli svantaggi di qualche tipo. Le stime e previsioni presentate dalle odierne IA, invece, non sono compiute sulla base di inferenze causali del genere di quelle coinvolte nel ragionamento appena schematizzato, ma sono il prodotto un'elaborazione in cui dati personali di qualsiasi sorta vengono tradotti in schemi di attivazione di neuroni artificiali, 'dissolvendosi' in impulsi binari trattati alla rinfusa e ricombinati in processi che rendono impossibile determinare *direttamente* se e quanto un singolo elemento discreto ricavato da un particolare dato personale abbia influito sul responso del sistema. La statistica, qui, non rileva al livello logico-inferenziale della conferma di una delle premesse dell'argomentazione che si conclude con la conclusione da cui discende il trattamento discriminatorio, ma a un livello più profondo, elettronico-quantitativo, imperscrutabile agli esseri umani e per così dire oracolare: possiamo verificare *ex post* l'esattezza, accuratezza, attendibilità e lungimiranza delle stime e previsioni delle IA, quantomeno nel loro complesso, ma non possiamo ricostruire direttamente, specialmente al livello di astrazione richiesto in un contesto di giustificazione/controllo, né il processo inferenziale che ha condotto alla loro elaborazione né i fattori e gli elementi che sono stati effettivamente considerati nel corso di tale processo. Ciò per la semplice ragione che il sistema non compie alcuna *inferenza* in senso logico-proposizionale, intesa come processo per

[62] R.J. ARNESON, *Equality of Opportunity*, in *The Stanford Encyclopedia of Philosophy,* ed. By E.N. Zalta (summer 2015 edition), disponibile su https:// plato.stanford.edu/entries/equal-opportunity/#AntLawDisTreDisImp.

cui si arriva ad affermare una proposizione dotata di significato sulla base di qualche altra proposizione dotata di significato.

Le discriminazioni algoritmiche operate mediante l'impiego delle odierne IA, pertanto, possono giuridicamente contrastarsi in quanto discriminazioni statistiche solo a condizione che si possa comprovare l'impiego di una caratteristica protetta come fattore determinante per la produzione del responso a cui vengono ricollegate conseguenze svantaggiose. Stante il funzionamento delle attuali reti neurali, si tratta di una condizione di ben difficile soddisfazione.

Secondo alcuni studiosi delle discriminazioni algoritmiche, tuttavia, vi sarebbero altri e più forti motivi per vietare l'uso di questi strumenti in processi decisionali da cui scaturiscono conseguenze giuridiche che incidono significativamente sui diritti delle persone. La carenza di una razionalità comprensibile nelle decisioni algoritmiche farebbe infatti venir meno la loro stessa normatività: «la […] forza normativa di una regola (o di un atto giuridico) dipende dalla sua efficacia persuasiva: detto altrimenti, il diritto esprime sempre una ragione per agire in un certo modo e in essa risiede la sua obbligatorietà. Nel momento in cui venisse a mancare questa ratio (dunque la capacità persuasiva), diverrebbe estremamente difficile comprendere la base della sua stessa normatività»[63]. Una variante di questa tesi, meno sospettabile di confusione tra contesto sociologico e contesto di giustificazione[64], potrebbe affermare che tali decisioni sono giuridicamente difettose in quanto carenti sotto il profilo della *motivazione*, intesa come elencazione delle *rationes decidendi* sufficientemente dettagliata da permettere il controllo e l'eventuale annullamento dell'atto da parte di autorità giuridiche di grado superiore.

In realtà, mi pare che a esser prive di ragioni intelligibili non siano le decisioni algoritmicamente assistite complessivamente considerate, ma

[63] A. SIMONCINI, S. SUWEIS, *Il cambio*, cit., 99. La teoria citata dagli autori è compiutamente formulata in B. PASTORE, F. VIOLA, G. ZACCARIA, *Le ragioni del diritto*, Bologna, 2017.

[64] La validità-obbligatorietà della norma in un contesto di giustificazione andrebbe sempre trattata partitamente dalle questioni dell'obbedienza alla norma o della sua effettività nel contesto sociologico.

solo uno dei loro presupposti informativo-fattuali. Come tutte le decisioni produttive di effetti giuridici, anche quelle assistite dall'IA esprimono delle norme (singolari) la cui obbligatorietà discende da una giustificazione fondata sia su presupposti normativi (sostanziali e procedurali), sia su presupposti descrittivo-fattuali[65]. Per controllare in un contesto di giustificazione detti presupposti normativi occorre vagliare la loro validità, rilevanza, coerenza, corretta applicazione e interpretazione ecc., mentre per controllare i presupposti descrittivi occorre vagliare il loro grado di conferma. Ebbene, il supporto offerto dalle IA si dispiega quasi sempre soltanto su quest'ultimo piano descrittivo-fattuale: esse costituiscono strumenti per disporre di giudizi di fatto del tipo «X possiede la proprietà Y», che sono nel complesso più o meno attendibili secondo la loro generale corrispondenza alla realtà che descrivono. Come non occorre conoscere nel dettaglio le esatte cause biochimiche del funzionamento degli antibiotici per giudicarli efficaci per il trattamento di una singola infezione batterica, se la loro efficacia è in generale comprovata dalle opportune osservazioni sperimentali, così non occorre conoscere nel dettaglio il processo attraverso il quale le IA riescono a prevedere la condotta dei singoli individui, se le nostre osservazioni empiriche confermano che ci riescono quasi sempre. Ciò che conta, nella validazione delle premesse descrittive singolari che costituiscono la componente propriamente automatizzata dell'argomentazione che si conclude con la decisione algoritmicamente assistita, non è infatti la loro giustificazione logica o la loro spiegazione eziologica, ma la loro *attendibilità*. Piuttosto, occorrerà definire il grado di attendibilità che reputiamo sufficiente ad autorizzarci all'impiego di questi sistemi, specialmente in ambiti in cui è in gioco il godimento dei diritti fondamentali degli individui; dovremo cioè fissare le soglie e i margini d'errore che siamo disposti a tollerare per poter

[65] Nell'argomentazione pratica, una conclusione prescrittiva, generale o singolare, si fonda sempre sia su premesse normative sia su premesse descrittive, ed è invalidata sia dal rigetto delle prime sia dalla falsificazione delle seconde. Mi sia consentito rinviare a G. GOMETZ, *Le regole tecniche. Una guida refutabile*, Pisa, 2008, 165 ss.

utilizzare le stime e previsioni delle IA al posto di quelle presentate da omologhi previsori umani.

Tornerò su questo punto alla fine di questo paragrafo; per ora mi limito a concludere che una decisione algoritmicamente assistita non è necessariamente priva di giustificazione solo perché non riusciamo a individuare e analizzare i passaggi logici attraverso i quali vengono ricavati i suoi elementi propriamente algoritmici, ossia i *presupposti descrittivi*; ciò che importa, piuttosto, è che questi presupposti siano controllabili e nel complesso corretti, ossia generalmente veri.

È stato d'altronde osservato che un deficit di trasparenza analogo a quello che stiamo discutendo si pone anche per le decisioni e pratiche umane, che spesso discriminano in modo dissimulato, occulto o addirittura inconsapevole: gli esseri umani sono senz'altro in grado di dar conto delle ragioni che li hanno indotti a decidere o agire in un certo modo, e tuttavia essi possono rifiutarsi di fornire tale giustificazione, o rappresentare ragioni diverse da quelle effettivamente considerate, o addirittura ritenere in buona fede di aver deciso/agito sulla base di ragioni ineccepibili sul piano del principio di non discriminazione, anche se di fatto sono stati influenzati da bias impliciti, pregiudizi latenti o altri fattori inconsci[66]. Se le IA sono delle *black box* nel contesto di giustificazione in cui contano le *ragioni* delle proprie conclusioni descrittive, le menti umane sono ancora più oscure nel contesto di spiegazione delle *cause*, ossia dei vari ed eterogenei fattori fisici, neurologici, psicologici, sociali, che laocoonticamente aggrovigliati tra loro contribuiscono a formare ciò che siamo soliti chiamare 'libero arbitrio'[67]. Inoltre, se pure è vero che i responsi delle IA sono opachi in quanto non corredati né corredabili da ragioni intelligibili, la stessa cosa non vale necessariamente per i dati che sono stati usati per addestrare l'algoritmo a svolgere una certa funzione, né del resto per la funzione stessa: già dall'esame di questi elementi, se disponibili, accessibili e aperti

[66] Cfr. J. KLEINBERG, J. LUDWIG, S. MULLAINATHAN, C. R SUNSTEIN, *Discrimination*, cit., 10 ss.

[67] Cfr. almeno D.C. DENNETT, *Elbow Room: The Varieties of Free Will Worth Wanting*, Cambridge, 1984 e ID., *L'evoluzione della libertà*, Milano, 2003.

a ulteriori sperimentazioni, può infatti ricavarsi una spiegazione *a posteriori* e quindi una prova convincente del perché una certa IA ha deciso come ha deciso, in particolare quando si osservino delle disparità nei risultati che facciano sorgere il sospetto di una discriminazione algoritmica in corso. Gli algoritmi delle IA, contrariamente a quanto talora si assume[68], non sono invero 'liberi' di produrre qualsivoglia risultato, ma sono meccanicamente determinati dai processi di *training* attraverso i quali sono stati prodotti, ossia dai dati utilizzati per l'addestramento stesso, dagli *outcomes* da prevedere e dai fattori utilizzati per tale previsione; tutti elementi in buona misura dipendenti da contingenti scelte umane. Se dunque gli algoritmi e i dati di addestramento sono disponibili, v'è la possibilità di rieseguirli per verificare se avrebbero prodotto gli stessi output qualora i soggetti considerati fossero stati di razza, sesso, religione, orientamento sessuale ecc. diversi. In tal modo, è possibile rilevare *ex post* una discriminazione algoritmica diretta, e così invalidare la relativa decisione in quanto fondata sulla considerazione di elementi che la legge vieta di porre alla base di disparità di trattamento produttive di svantaggi per gli interessati. Sulla base di tutti questi rilievi, v'è anzi chi ritiene che le decisioni algoritmiche siano in generale più trasparenti di quelle umane, soprattutto sotto il profilo dell'accertamento di eventuali discriminazioni, a patto di adottare alcuni accorgimenti relativi alla conservazione dei dati di addestramento, all'accessibilità agli algoritmi e alla loro rieseguibilità in condizioni sperimentali di controllo[69].

A tutte queste notazioni va aggiunto che non è affatto detto che la considerazione di dati relativi a qualche fattore di protezione da parte

[68] Cfr. B. YAVAR, *The Artificial Intelligence Black Box and The Failure of Intent and Causation*, in *Harvard Journal of Law and Technology*, 31.2, 2018, 920.

[69] Cfr. J. KLEINBERG, J. LUDWIG, S. MULLAINATHAN, C. R SUNSTEIN, *Discrimination*, cit. Gli autori ammettono tuttavia che i costi di conservazione previsti da questa proposta divengono estremamente onerosi quando gli elementi utilizzati dal sistema per elaborare le proprie previsioni siano ricavati da ingenti flussi di dati continuamente aggiornati, come avviene nel caso delle ricerche sul web, nell'*online ad delivery* e in molti altri campi di utilizzo dei *Big Data* a fini di profilazione; cfr. J. KLEINBERG, J. LUDWIG, S. MULLAINATHAN, C. R SUNSTEIN, *Discrimination*, cit., 32.

degli algoritmi di IA dia necessariamente luogo a stime, previsioni o selezioni discriminatorie, ossia produttive di effetti svantaggiosi per i portatori delle caratteristiche protette. Al contrario, è stato osservato che l'accesso delle IA a questi dati e il loro utilizzo a fini predittivi potrebbero in alcuni casi mitigare gli effetti discriminatori derivanti da una elaborazione resa 'cieca' ai fattori di protezione. Supponiamo ad esempio che un'IA abbia il compito di selezionare gli studenti da ammettere a un'università secondo il rendimento previsto e che, onde evitare discriminazioni di sorta, i suoi amministratori stabiliscano che per operare tale stima possano considerarsi soltanto i voti scolastici e i giudizi riportati nelle lettere di raccomandazione dagli insegnanti delle scuole superiori. Ovviamente, eventuali pregiudizi razziali da parte di questi insegnanti produrrebbero differenze nella qualità media delle lettere degli studenti appartenenti a certi gruppi discriminati, ciò che determinerebbe assai frequentemente la loro penalizzazione nelle graduatorie d'accesso al *college*. Se però l'algoritmo avesse accesso ai dati sulla razza, potrebbe facilmente 'avvedersi' che molti studenti dei gruppi discriminati hanno voti assai migliori di quanto le loro lettere di raccomandazione facciano prevedere; ciò potrebbe indurre il sistema a ridurre autonomamente il peso relativo delle lettere di raccomandazione e correggere i suoi responsi in senso favorevole all'ammissione al *college* di un maggior numero di studenti appartenenti ai gruppi discriminati[70]. Nonostante le migliori intenzioni, insomma, alcuni tentativi di contrastare le discriminazioni delle IA rendendole 'cieche' alle caratteristiche protette possono sortire risultati controproducenti, perché impediscono a quei sistemi di rilevare autonomamente che gli *outcomes* che si intendono stimare o prevedere *non* sono correlati a certi fattori di protezione, ma ad altri elementi inopinati o addirittura inaccessibili alla cognizione dagli esseri umani e ricavabili soltanto a seguito di un'analisi *disumana* (nel senso precisato nel primo paragrafo) di grandi quantità di dati.

[70] Cfr. J. KLEINBERG, J. LUDWIG, S. MULLAINATHAN, A. RAMBACHAN, *Algorithmic Fairness*, in *AEA Papers and Proceedings*, 108, 2018, 22-27. Si veda anche T. GILLIS, J. SPIESS, *Big Data and Discrimination*, in *The University of Chicago Law Review*, 86.2, 2019, 459 ss.

In definitiva, né la difficoltà di accertare l'impiego a fini predittivi o di profilazione di una caratteristica protetta né il suo utilizzo effettivo da parte dell'IA costituiscono ragioni sufficienti per ravvisare una violazione del principio di non discriminazione diretta, il quale, lo ricordo ancora, vieta i trattamenti svantaggiosi operati in ragione, a causa o a motivo di un fattore di protezione e non certo i trattamenti di soggetti che, pur portatori di caratteristiche protette, subiscono svantaggi in ragione, a causa o a motivo di *altre* caratteristiche, che sono incidentalmente proprio quelle che l'IA è chiamata a prevedere: la probabilità di recidiva, il pericolo di fuga, la disposizione a commettere certi crimini, la condizione di particolare bisogno di una certa prestazione sociale ecc. In questa prospettiva, la discriminazione più odiosa è quella che deriva dai deficit prestazionali degli algoritmi che sistematicamente sovrastimano o sottostimano gli *outcomes* di certi individui, dando *erroneamente* alle loro caratteristiche protette una valenza predittiva sproporzionata e dunque risultante in stime inattendibili e inaccurate. Si pensi al cittadino erroneamente indicato dall'algoritmo come più proclive a commettere certi reati semplicemente in quanto nero, o alla cittadina erroneamente ritenuta incapace di svolgere un certo lavoro semplicemente in quanto donna. È in queste pratiche che il disvalore morale della discriminazione, consistente nel dare rilievo a una caratteristica che moralmente è inerte e dunque non dovrebbe attrarre alcuna valutazione negativa[71], raggiunge gli abissi più profondi. Per questo motivo, a me pare che l'accostamento al tema delle decisioni algoritmicamente assistite dall'IA non debba arroccarsi sull'applicazione di una sorta di principio di precauzione teso a vietare qualsiasi impiego di tali tecnologie, ma debba insistere sulla pretesa di puntuali, rigorosi controlli sull'attendibilità e sull'esattezza dei risultati presentati dagli algoritmi eventualmente impiegati a supporto delle decisioni delle autorità giuridiche nonché, correlativamente, sulla veridicità dei dati impiegati come base per operare le stime e le previsioni. Tali controlli dovranno guardare non soltanto *indietro*, ossia alla corretta elaborazione delle stime a partire da dati che documentino veridicamente quanto è già

[71] Così E. CONSIGLIO, *Che cos'è la discriminazione?*, cit., 42.

accaduto nel passato, ma anche e soprattutto *avanti*, confrontando le previsioni e stime con ciò che accade in un tempo successivo alla loro elaborazione, a loro verifica o smentita. Tutte le verifiche, com'è ovvio, andranno operate fin dalla fase di test del sistema, ossia in tempi che precedono il suo impiego come strumento di supporto delle decisioni umane, per poi essere operate successivamente, magari a cadenze periodiche, onde sventare il pericolo di quelle che sono state chiamate '*zombie predictions*', ossia delle previsioni inesatte in quanto basate su dataset che danno conto di situazioni, condizioni e circostanze alquanto differenti da quelle che sussistono al momento della previsione stessa[72].

Questo dei controlli sull'affidabilità delle IA usate a supporto delle decisioni giuridiche è un tema su cui, curiosamente, la letteratura in materia di discriminazioni algoritmiche insiste poco, intenta com'è a denunciare l'insopportabile nequizia di *qualsiasi* trattamento differenziale per via algoritmica di soggetti portatori di caratteristiche protette, indipendentemente dal fatto che tale trattamento sia operato appunto a cagione di quelle caratteristiche o per altri motivi. Anche le vigenti regolamentazioni giuridiche delle decisioni algoritmicamente assistite si occupano dei controlli circa l'affidabilità delle stime dei sistemi quasi solo tangenzialmente. In Europa, ad esempio, il già citato considerando 71 del GDPR, prevede soltanto che «è opportuno che il titolare del trattamento utilizzi procedure matematiche o statistiche appropriate per la profilazione, metta in atto misure tecniche e organizzative adeguate al fine di garantire, in particolare, *che siano rettificati i fattori che comportano inesattezze dei dati e sia minimizzato il rischio di errori*», ma questa raccomandazione non trova riscontro in precetti specificamente rivolti a stabilire dei requisiti o parametri minimi di attendibilità dei risultati delle IA né degli standard per operare i relativi controlli. Si perde così l'occasione di affrontare un problema a mio parere centrale per la questione dell'utilizzabilità dei sistemi di IA a supporto delle decisioni umane produttive di effetti giuridici che incidono significativamente sulle persone. Alludo al margine di tollerabilità degli *errori* nelle stime e

[72] Cfr. J. L. KOEPKE, D. G. ROBINSON, *Danger Ahead: Risk Assessment and The Future of Bail Reform*, in *Washington Law Review*, 93, 2017, 1725-1807.

previsioni avanzate da questi sistemi, materia che a mio avviso tocca il *punctum dolens* della loro ammissibilità sul piano della politica del diritto più ancora che la questione del loro potenziale carattere discriminatorio diretto o indiretto. Nell'argomentazione che si conclude con la decisione algoritmicamente assistita, i responsi conseguenti a una profilazione operata tramite IA vanno invero trattati dai decisori umani come *generalizzazioni non universali*, giacché consistono in stime del tipo «X possiede la proprietà Y», che valgono soltanto probabilisticamente: non tutti gli X di fatto possiederanno la proprietà Y, ma solo una percentuale che dipenderà dalla efficacia predittiva/estimativa del sistema. Quando sono in gioco delle stime o previsioni su caratteristiche personali giuridicamente ricollegate a qualche compressione dei diritti fondamentali degli individui, tuttavia, i costi dell'errore anche in una sola stima/previsione possono essere molto ingenti. Si pensi ai danni conseguenti a una misura di sicurezza detentiva disposta per un imputato erroneamente indicato da un'IA come particolarmente proclive a commettere reati violenti, o si considerino i disagi conseguenti alla denegazione di una certa prestazione sociale per via dell'errata valutazione di un sistema algoritmico circa la condizione di bisogno di un particolare cittadino: sono evenienze che non solo si prestano a provocare conseguenze drammatiche sulla vita degli interessati, ma rischiano di delegittimare l'intero sistema giuridico che le ammetta. Occorre dunque assicurarsi che questi tragici fallimenti siano quantomai rari, possibilmente assai meno frequenti di quelli che derivano dagli errori nelle stime e previsioni degli operatori umani attualmente chiamati a decidere negli stessi ambiti (giudici, funzionari, periti, consulenti tecnici ecc.). Auspicabilmente, inoltre, si coglierà l'occasione per riformare i vigenti regimi di responsabilità per i danni derivanti da chiari errori nei presupposti fattuali delle decisioni pubbliche, compresi i provvedimenti giurisdizionali e amministrativi, prevedendo congrui risarcimenti a favore di chi abbia subito una lesione dei propri diritti soggettivi a causa delle errate stime e previsioni degli algoritmi che forniscono supporto informativo in un qualsivoglia processo decisionale produttivo di effetti giuridici.

Essendo quella dei controlli circa l'affidabilità del supporto informativo offerto dalle IA alle decisioni produttive di importanti effetti giuridici una questione che esula dallo specifico tema delle discriminazioni algoritmiche, non ne proporrò qui alcuna trattazione *ex professo*, limitandomi a menzionarne l'unico aspetto rilevante in questa sede: il problema delle eventuali *uneven failures,* ossia degli scadimenti prestazionali nell'esattezza delle previsioni/stime che riguardano in particolare i soggetti portatori di caratteristiche protette dal diritto antidiscriminatorio, magari perché relativamente atipici e dunque meno rappresentati nei dati di addestramento o di riferimento delle reti neurali. Anche questa evenienza dovrà essere oggetto di specifici e puntuali controlli nel corso delle verifiche su cui ho insistito sul finire di questo articolo. È infatti soprattutto dalla comparazione tra l'affidabilità dei responsi di questi sistemi e quella degli omologhi previsori umani che, in un futuro sempre più prossimo, dipenderanno le nostre scelte circa la loro utilizzabilità a supporto di decisioni da cui discendono conseguenze giuridiche assai rilevanti per le persone e i loro diritti fondamentali.

ABSTRACT

Il funzionamento delle odierne IA basate sul modello del machine learning rende disagevole determinare se un certo responso algoritmico posto alla base di scelte, decisioni e policies produttive di effetti giuridici rilevanti per le persone, sia o no censurabile in quanto direttamente o statisticamente discriminatorio, ossia fondato sulla considerazione di una qualche caratteristica protetta dal diritto antidiscriminatorio come ragione, motivo o causa diretta o indiretta di un certo trattamento svantaggioso. Si sosterrà che detta difficoltà non costituisce di per sé motivo sufficiente per vietare l'impiego di sistemi di IA come strumento di supporto informativo delle decisioni umane, e che i timori legati alle c.d. 'discriminazioni algoritmiche' più fondati sono quelli relativi all'eventualità che certe decisioni produttive di effetti giuridici assai incisivi sulla vita degli individui vengano adottate col supporto di sistemi

di IA che 'sbagliano', nel senso che non sono in grado di profilare attendibilmente i singoli per via di dati incompleti, obsoleti o biased, di errori nella costruzione degli algoritmi, nonché da limitazioni al loro uso ispirate, paradossalmente, dall'intento di evitare effetti discriminatori.

The functioning of today's AI based on the machine learning model makes it difficult to determine whether a certain algorithmic response placed at the basis of choices, decisions and policies that produce legal effects relevant to people, is or is not proscribable as directly or statistically discriminatory, i.e. founded on the consideration of some characteristic protected by the anti-discrimination law as a reason, motive or direct or indirect cause of a certain disadvantageous treatment. It will be argued that this difficulty does not constitute sufficient reason to prohibit the use of AI systems as an information support tool for human decisions, and that the more well-founded fears related to the so-called 'algorithmic discriminations' are those relating to the possibility that certain decisions producing very incisive legal effects on the life of individuals are adopted with the support of AI systems that are 'wrong', in the sense that they are not able to reliably profile individuals by way of incomplete, obsolete or biased data, of errors in the construction of algorithms, as well as limitations to their use inspired, paradoxically, by the intent to avoid discriminatory effects.

PAROLE CHIAVE

discrimination, algorithms, profiling, artificial intelligence, statistical discrimination, algorithmic discrimination

GIANMARCO GOMETZ
Email: gometz@unica.it