Rivista elettronica di Diritto, Economia, Management

N. 3 - 2022 - Supplemento 1 - pp. 25-30

Discriminazioni algoritmiche?

Gianmarco Gometz[1]

Abstract [IT]: Il funzionamento delle odierne IA basate sul modello del machine learning rende alquanto 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 di un certo trattamento svantaggioso. Se tuttavia i dati di addestramento e gli algoritmi delle IA sono disponibili, v'è talora la possibilità di rieseguire questi ultimi per verificare se avrebbero prodotto gli stessi output qualora i soggetti considerati fossero stati di razza, sesso, religione, orientamento sessuale ecc. diversi, invalidando le decisioni fondate sulla considerazione di elementi che la legge vieta di porre alla base di disparità di trattamento produttive di svantaggi per gli interessati. I costi di questo approccio diventano però estremamente onerosi, e forse insostenibili, quando gli elementi utilizzati dal sistema per elaborare le proprie previsioni non sono ricavati da dataset "statici", bensì da ingenti flussi di dati continuamente aggiornati.

Abstract [EN]: The operation of today's AIs based on the machine learning model makes it somewhat uncomfortable to determine whether or not a certain algorithmic response placed at the basis of choices, decisions and policies productive of legal effects relevant to individuals is censurable as directly or statistically discriminatory, that is, based on the consideration of some characteristic protected by anti-discrimination law as the reason, motive or cause of a certain disadvantageous treatment. If, however, the training data and AI algorithms are available, there is sometimes the possibility of re-running them to see whether they would have produced the same outputs had the subjects considered been of a different race, sex, religion, sexual orientation, etc., invalidating decisions based on the consideration of elements that the law prohibits from being the basis for unequal treatment that is productive of disadvantages for those affected. The costs of this approach, however, become extremely onerous, and perhaps unsustainable, when the elements used by the system to make its predictions are derived not from "static" datasets, but from massive streams of continuously updated data.

Keywords [IT]: discriminazione – algoritmi – profilazione – intelligenza artificiale – discriminazione statistica – discriminazione algoritmica.

[1] Ordinario di Filosofia del diritto presso il Dipartimento di Giurisprudenza dell'Università degli Studi di Cagliari.

Keywords [EN]: discrimination - algorithms - profiling - artificial intelligence - statistical discrimination - algorithmic discrimination.

Alla categoria concettuale delle "discriminazioni algoritmiche" vengono ascritte cose molto diverse: 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. Sorge perfino 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, e ci si può porre addirittura il dubbio che alcune di esse non possano essere propriamente qualificate come "discriminazione". Occorre dunque elucidare un concetto di discriminazione algoritmica sufficientemente determinato da consentirci almeno di individuare gli abusi semantici più lampanti, ciò che tenterò di fare subito.
Se, come mi pare opportuno, ci si accosta alla nozione di discriminazione algoritmica muovendo dal concetto giuridico generale di discriminazione, se ne può evidenziare il connotato eminentemente tecnico-strumentale: è discriminazione algoritmica la prescrizione generale o singolare, il criterio o la pratica che comporta svantaggi relativi per taluni soggetti in quanto portatori delle caratteristiche protette dal diritto di cui parlavo prima, 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" ex art. 22 del GDPR sia ogni altra decisione adottata col supporto di algoritmi che forniscono ai decisori umani stime, descrizioni e previsioni probabilistiche su una certa situazione o caratteristica a cui il diritto ricollega effetti giuridici, direttamente o a seguito della decisione stessa.
Discriminazioni del genere, nella maggior parte dei diritti occidentali, possono venire proscritte a titolo di discriminazione diretta oppure indiretta. La distinzione in oggetto, dibattuta sia sul piano analitico-concettuale che su quello etico-fondazionale, oppone 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 specificamente ad essi e a ciò indirizzate.
Ebbene, considerate nella loro modalità specificamente algoritmica, le discriminazioni indirette non presentano peculiarità degne di nota, giacché esse occorrono indipendentemente dal mezzo impiegato per provocarle, dalla presenza di un intento discriminatorio esplicito o implicito e da qualsivoglia considerazione di un fattore di

protezione a fini discriminatori.
Molto più rilevante in una sede come questa, invece, è il tema dell'accertamento della ricorrenza di una discriminazione algoritmica diretta. Qui occorre infatti elucidare e provare le ragioni del trattamento asseritamente discriminatorio, ossia appurare se una certa decisione o pratica algoritmicamente assistita abbiano svantaggiato dei soggetti a cagione della loro appartenenza a una categoria protetta. Ciò può essere alquanto complicato nel caso di processi decisionali automatizzati mediante l'impiego di reti neurali basate su tecnologie di machine learning.
Come sappiamo, questi sistemi sono particolarmente "abili" nello scoprire automaticamente delle correlazioni tra dati per poi estrapolarne di nuovi, ciò che consente di utilizzarle come formidabili strumenti di profilazione, ossia a fini di analisi e previsione di comportamenti, qualità e disposizioni delle persone fisiche. Tale profilazione, oggi perlopiù applicata nel marketing, potrebbe senz'altro essere impiegata in un prossimo futuro in settori assai più pregnanti sotto il profilo pubblicistico: welfare, sicurezza, sanità, contrasto al crimine ecc.: ove vi sia sufficiente disponibilità dei dati personali degli individui, si potrebbero ad esempio usare le IA per scoprire se un certo soggetto sia più (o meno) bisognoso di particolari prestazioni sociali, o idoneo a ricoprire certi ruoli, 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.
Uno dei problemi che dovremo affrontare, nel momento in cui decideremo se impiegare questi strumenti di profilazione in settori caratterizzati da rilevanti interessi pubblici, proviene per l'appunto dal rischio di discriminazioni algoritmiche dirette.
Allo stato attuale dei progressi tecnologici, infatti, le IA non possono chiarire esplicitamente, se, come e quanto dei fattori di protezione (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 rappresentate nei dati personali accessibili al sistema (residenza, preferenze d'acquisto, dati di geolocalizzazione, "likes", siti visitati, contatti ecc.).
Tale opacità discende 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. Le reti neurali attuali non sono in grado di presentare alcun log circa le ragioni, i motivi, le cause, i fattori o anche solo i singoli passaggi computazionali che le hanno determinate a operare in un certo modo o a produrre un certo risultato, e il log manca perché non c'è alcun logos, almeno ricostruibile al livello di astrazione richiesto in un contesto di controllo e giustificazione. L'attuale intelligenza artificiale è in questo senso un'intelligenza non intelligibile.
Per contrastare le discriminazioni algoritmiche discendenti dall'impiego di reti neurali basate sul paradigma dell'apprendimento automatico abbiamo dunque due stra-

de: 1) proscriverle a titolo di discriminazioni indirette; 2) approntare un quadro normativo rivolto alla verifica a posteriori di eventuali discriminazioni dirette nei responsi delle IA.
La prima strada non presenta criticità peculiari alle discriminazioni algoritmiche: Per accertare l'occorrenza di una discriminazione indiretta non è infatti necessario ispezionare gli algoritmi o i dataset di training 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 ricorra o meno una delle cause di esclusione della discriminazione indiretta previste nel dato diritto positivo; per esempio, sia in ambito europeo che americano non ricorre alcuna discriminazione indiretta tutte le volte in cui una certa pratica 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". Queste formule rimandano ovviamente a qualificazioni altamente discrezionali da parte degli interpreti, ma non si tratta di problemi specifici delle discriminazioni algoritmiche, dunque possiamo tralasciarli in questa sede.
Una seconda direttrice normativa da percorrere per il contrasto alle discriminazioni algoritmiche passa per l'istituzione di controlli ex post sul funzionamento dei sistemi algoritmici di supporto alle decisioni produttive di effetti giuridici che riguardano le persone. Infatti, se pure è vero che i responsi delle odierne IA sono opachi in quanto non corredati né corredabili da un razionale intelligibile, 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 a ulteriori sperimentazioni, può evincersi una spiegazione a posteriori e dunque 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 afferma, 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 essere in generale più trasparenti di quelle umane,

soprattutto sotto il profilo dell'accertamento di eventuali discriminazioni, a patto di adottare i già menzionati accorgimenti relativi alla conservazione dei dati di addestramento, all'accessibilità agli algoritmi e alla loro rieseguibilità in condizioni sperimentali di controllo.
Il limite di questo approccio è che prevede costi che divengono estremamente onerosi, e forse insostenibili, quando gli elementi utilizzati dal sistema per elaborare le proprie previsioni non siano ricavati da dataset "statici", bensì da ingenti flussi di dati continuamente aggiornati, come avviene nel caso delle ricerche sul web o nell'online ad delivery.
Questo è uno dei problemi che dovremo affrontare nel momento in cui dovremo decidere se impiegare questi strumenti di profilazione in settori caratterizzati da una notevole rilevanza degli interessi pubblici in gioco e dalle conseguenti frizioni tra interessi pubblici e diritti individuali, comprese le libertà fondamentali.
Un altro problema discende dalla domanda se le discriminazioni algoritmiche operate dalle odierne reti neurali basate sul paradigma del machine learning possono essere giuridicamente vietate in quanto discriminazioni statistiche, proscritte in buona parte dei diritti occidentali. 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 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: 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.
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. Si noti tuttavia che le stime e previsioni presentate dalle odierne IA 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 convertiti 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 col responso 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, il processo inferenziale che ha condotto alla loro elaborazione e i fattori e gli elementi in esso considerati. Ciò per la semplice ragione che il sistema non compie alcuna inferenza in senso logico-proposizionale, intesa come processo per 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 essere giuridicamente proscritte in quanto discriminazioni statistiche solo a condizione che sia chiaramente accertabile l'impiego d'una caratteristica protetta come fattore rilevante per la produzione del responso a cui vengono ricollegate conseguenze svantaggiose. Una condizione, questa, di ben difficile soddisfazione, alla luce dell'attuale paradigma di funzionamento delle reti neurali utilizzate a fini di profilazione.