The Lab's Quarterly

2018 / a. XX / n. 4 (ottobre-dicembre)

CONTATTI
thelabs@sp.unipi.it

I saggi della rivista sono sottoposti a un processo di double blind peer-review.
La rivista adotta i criteri del processo di referaggio approvati dal Coordinamento delle Riviste di Sociologia (CRIS): cris.unipg.it
I componenti del Comitato scientifico sono revisori permanenti della rivista.
Le informazioni per i collaboratori sono disponibili sul sito della rivista:
https://thelabs.sp.unipi.it

ISSN 1724-451X

"The Lab's Quarterly" è una rivista di Scienze Sociali fondata nel 1999 e riconosciuta come rivista scientifica dall'ANVUR per l'Area 14 delle Scienze politiche e Sociali. L'obiettivo della rivista è quello di contribuire al dibattito sociologico nazionale ed internazionale, analizzando i mutamenti della società contemporanea, a partire da un'idea di sociologia aperta, pubblica e democratica. In tal senso, la rivista intende favorire il dialogo con i molteplici campi disciplinari riconducibili alle scienze sociali, promuovendo proposte e special issues, provenienti anche da giovani studiosi, che riguardino riflessioni epistemologiche sullo statuto conoscitivo delle scienze sociali, sulle metodologie di ricerca sociale più avanzate e incoraggiando la pubblicazione di ricerche teoriche sulle trasformazioni sociali contemporanee.

The Lab's Quarterly

2018 / a. XX / n. 4 (ottobre-dicembre)

Gli algoritmi come costruzione sociale

A cura di
Antonio Martella, Enrico Campo e Luca Ciccarese

LIBRI IN DISCUSSIONE

DISCRIMINAZIONE E ALGORITMI
Incontri e scontri tra diverse idee di fairness

di *Mattia Galeotti*[*]

Abstract

The growing spread of Machine Learning algorithms in our society, was accompanied in the last years by various cases of discrimination that have been bred by automatized procedures, in particular with respect to race, gender, religious affiliation or sexual orientation. For this reason the subject of algorithmic fairness became central, but finding a solution appears to be a difficult problem. In this work we intend to show different mathematical models that are needed to define various notions of fairness, to investigate the conditions of compatibility and incompatibility between the models, and the relations and frictions with the material conditions. A feature of today governmental apparatuses that are based on automatized processes, is the reduction of political and social problems to the optimization of some functions; this is also the case when we talk about discrimination, that in this approach can be solved by measuring and deleting the bias encoded in algorithms. In this work we try instead to read from a political and sociological point of view the mechanisms of numerical optimization, and we investigate the relations between the subject of discrimination as it has historically appeared, and the multiple statistical formalizations developed with the purpose of governing this problem.

Keywords

Bias; fairness; statistics; movements; algorithms

[*] MATTIA GALEOTTI è post-doc presso Università degli Studi di Trento.
Email: galeotti.mattia.work@gmail.com

1. INTRODUCTION

Sempre più processi decisionali funzionano oggi attraverso la costruzione di modelli predittivi matematici che, basandosi su database storici, cercano di automatizzare i meccanismi di selezione; esempi molto conosciuti sono la pubblicità in rete personalizzata sull'utente, i punteggi individuali ai clienti di istituzioni bancarie per decidere se concedere o meno un prestito, lo screening automatico dei curricula dei candidati per un posto di lavoro, i punteggi che nel sistema giudiziario americano determinano il rischio di recidività per individui condannati, e molti altri. Questi strumenti di *Machine Learning* stanno dunque divenendo essenziali in sempre più funzioni della vita collettiva, e parallelamente alla loro diffusione si afferma come centrale il problema delle discriminazioni che possono accompagnare questi processi o direttamente essere veicolate tramite essi. In particolare, come vedremo meglio nella seconda sezione, specifici casi di studio hanno riscontrato la riproduzione di discriminazioni rispetto alla razza[1], al genere, all'appartenenza religiosa o all'orientamento sessuale, veicolati attraverso processi valutativi di questo tipo, ponendo la questione della discriminazione strutturale, oggi centrale nel dibattito politico, anche nello specifico contesto delle decisioni automatizzate. Più in generale il problema della discriminazione algoritmica si inserisce nel dibattito sul rapporto tra saperi tecnoscientifici e sistemi sociali, un terreno molto fertile per la sociologia della scienza e gli *Science and Technology Studies*.

Ogni processo di apprendimento, valutazione e/o selezione basato su dati statistici, funziona attraverso una formalizzazione matematica, la definizione di specifici strumenti di calcolo e quindi un calcolo effettivo tramite questi strumenti. Sebbene nella narrazione e nel marketing dell'Intelligenza Artificiale si tenda a identificare gli algoritmi con il semplice momento del calcolo, la formalizzazione e la definizione degli strumenti sono essenziali per organizzare e dare senso ai dati su cui si vuol far operare un processo algoritmico, e quindi al processo stesso. Riprendendo Mazzotti (2015) ci soffermiamo sull'aspetto di mutua costituzione tra scienza e società, un aspetto che è stato studiato soprattutto dal punto di vista dei sistemi tecnologici, e molto meno per quanto riguarda le procedure formali e deduttive. Nelle scelte assiomatiche proprie di ogni modello matematico sono codificate delle

[1] In tutto il testo utilizzeremo il termine "razza" riferendoci ad un costrutto storico che ha carattere performante al livello sociale, dunque come elemento reale ma non biologico di auto-riconoscimento degli individui, e potenzialmente di discriminazione. Per questo uso del termine vedere ad esempio Curcio, Mellino (2012).

ipotesi sociali e delle visioni del mondo, e la scelta di un modello per una specifica task deriva dall'incontro tra queste ipotesi, gli obiettivi contingenti e specifici interessi di gruppo; in più, queste cornici di senso vengono riprodotte dall'algoritmo stesso, tramite la sua azione di valutazione e selezione.

Ogni popolazione è caratterizzata da alcuni attributi sensibili come la razza, il genere, l'appartenenza religiosa o l'orientamento sessuale, cioè attributi rispetto ai quali potrebbe attuarsi una discriminazione, la *fairness* di un algoritmo è la proprietà di non discriminare rispetto a questi attributi; seguendo Dwork, Hardt, Pitassi, Reingold, Zemel, (2012) e Friedler, Scheidegger, Venkatasubramanian (2016) possiamo individuare due filoni negli approcci alla formalizzazione del concetto di fairness: da una parte la fairness come descrizione corretta del contesto reale, in cui ogni individuo riceve una "giusta" valutazione senza discriminazioni; dall'altra la fairness come parità statistica di valutazione tra i differenti gruppi sociali, con le stesse percentuali di successo per ogni valore dell'attributo sensibile.

La linea di studio che ci interessa approfondire è quella delle relazioni tra i modelli di fairness e le scelte politiche e organizzative connesse con i processi di selezione. Come abbiamo detto i modelli matematici codificano delle visioni del mondo, in particolare le diverse definizioni di fairness codificano idee distinte di discriminazione, tra queste diverse impostazioni esistono linee di incompatibilità e di frizione, ed ogni modello formale deve confrontarsi con i vincoli matematici e quelli del contesto storico. L'analisi delle incompatibilità e dei limiti di ogni formalizzazione permette di mettere in luce in che maniera l'algoritmo opera, ed in quali dinamiche strutturali viene ad inserirsi. Un esempio chiarificatore è descritto nel recente articolo di Rachel Courtland (2018): l'autrice sottolinea che è matematicamente impossibile attuare nel sistema giudiziario delle forme di fairness individuale e contemporaneamente di parità statistica tra gruppi, se la probabilità di arresto rimane più elevata per alcuni gruppi razziali; per questo se l'algoritmo viene utilizzato nel contesto giudiziario senza interferire col lavoro di polizia, allora dovrà scegliere una tra le due forme di fairness incompatibili.

Per valutare il tipo di azione degli algoritmi sul contesto sociale in cui agiscono, risulta particolarmente utile il concetto di "*governance by the numbers*" per com'è utilizzato in Katz (2017): attraverso la creazione di graduatorie e soglie di inclusione, le modellizzazioni statistiche partecipano ad una forma di disciplina della vita sociale. In questo senso gli algoritmi divengono dei veri e propri dispositivi di

governo, secondo l'utilizzo introdotto da Foucault di questo termine[2]. La "governance by the numbers" è quindi da intendersi come messa in pratica di una certa governamentalità, diffusa e legittimata in particolare con una narrazione secondo cui gli algoritmi sono processi in grado di descrivere oggettivamente alcuni comportamenti sociali: gli algoritmi diventano quindi portatori di una "*vision from nowhere*". Quest'ultima nozione, come abbiamo già detto, dissimula completamente la contingenza storica e sociale in cui il dispositivo è stato concepito ed è messo in azione.

In quest'ottica la ricerca di un algoritmo privo di *bias* (di razza, di genere, etc.) risulta epistemologicamente poco interessante, la nozione stessa di *bias* appare inadeguata, in quanto suggerisce l'esistenza di un processo oggettivo a cui i diversi algoritmi si avvicinano per approssimazione. Piuttosto la nostra trattazione suggerisce che l'analisi matematica dei modelli può permettere di capire quali forme di discriminazione sono proprie ad ogni algoritmo, aprendo ad una descrizione degli algoritmi come fattori operanti in un determinato periodo storico ed all'interno di specifiche ipotesi di normazione dello spazio sociale.

Nella sezione 2 mostreremo alcuni noti e documentati casi di discriminazione riscontrati in vari campi d'utilizzo di algoritmi decisionali. Nelle sezioni 3 e 4 introdurremo gli strumenti matematici necessari alla nostra trattazione, ed approfondiremo alcuni quadri assiomatici entro i quali è possibile approcciare il problema della *fairness*, esplorando le compatibilità ed incompatibilità tra questi differenti approcci. Infine, nella sezione 5 mostreremo in che modo gli algoritmi organizzano costruzioni di senso e divengono strumenti di *governance*.

2. IL PROBLEMA DELLA FAIRNESS NELL'UTILIZZO SOCIALE DEGLI ALGORITMI

La diffusione di algoritmi decisionali basati sul Machine Learning in ogni ambito della vita collettiva, si è accompagnata negli ultimi anni a un sempre maggior numero di casi in cui quegli stessi algoritmi hanno dimostrato di veicolare o riprodurre discriminazioni basate sulla razza, il genere, l'appartenenza religiosa, l'orientamento sessuale e altre caratteristiche; il tema della giustizia algoritmica si è dunque affermato come centrale. L'analisi del problema ha immediatamente rivelato la necessità di approfondire il contesto di applicazione, i dati di apprendimento e le funzioni di questi processi automatizzati, e di considerare gli algoritmi come strumenti dentro contesti storici contin-

[2] Per il concetto di dispositivo in Foucault si veda Agamben (2006).

genti, non nettamente separabili dalle condizioni materiali in cui operano e dai soggetti sociali che li utilizzano. Seguendo il punto di vista degli "Science and Technology Studies", non soltanto gli algoritmi fanno parte di un'infrastruttura tecnoscientifica che opera dentro il contesto sociale, ma è vero anche l'inverso, i dispositivi scientifici e tecnici, tra cui gli algoritmi, sono costituiti da processi di natura sociale.

Nel 2017 l'articolo della rivista online Quartz (Sonnad, 2017) osservava uno strano comportamento di *Google Translate*: nelle traduzioni dal turco all'inglese, la parola turca "o", pronome corrispondente alla terza persona singolare e di genere neutro, veniva tradotto nell'inglese "he" oppure "she" in base alle parole che la accompagnavano, rivelando una chiara discriminazione di genere. Parole come "soldier", "doctor" oppure "hardworking" portavano ad una traduzione maschile, mentre "teacher", "nurse" e "lazy" portavano ad una traduzione femminile. Facile immaginare in questo contesto che sia direttamente l'insieme di dati a cui Google ha accesso a contenere una discriminazione di genere, perché quei dati corrispondono al linguaggio utilizzato dagli utenti in rete. Allo stesso tempo la pervasività di questo strumento fa temere per un effetto di rinforzo della discriminazione linguistica.

Ancora più problematica è la questione degli algoritmi che aiutano nella selezione dei candidati per un posto di lavoro fornendo un punteggio sulla base dei curricula forniti. Nel recente articolo di Chen, Ma, Hannak, Wilson (2018), è stato studiato l'impatto del genere in questo tipo di selezioni in venti città degli Stati Uniti, ottenendo una vasta gamma di esempi di discriminazione. Chiaramente dispositivi di questo tipo non soltanto svantaggiano ingiustamente alcuni individui, ma più in generale rischiano di riprodurre condizioni di svantaggio sistemico, come un minor tasso di impiego o salari più bassi per uno specifico gruppo.

Il caso forse più conosciuto è quello di uno strumento denominato COMPAS, per *Correctional Offender Management Profiling for Alternative Sanctions*, sempre più utilizzato nelle corti giudiziarie degli Stati Uniti con l'obiettivo di determinare il rischio che individui condannati per un crimine divengano recidivi. Nel 2016 il sito giornalistico ProPublica pubblicava l'articolo di Angwin, Larson, Mattu, Kirchner (2016), dimostrando che il COMPAS era chiaramente discriminante verso gli individui neri. In particolare, l'inchiesta metteva in luce che questo strumento ha in alcuni casi tassi di accuratezza molto bassi: un alto tasso di falsi positivi (cioè individui falsamente indicati come a rischio di recidività) nel caso di individui neri, ed un maggior

tasso di falsi negativi (individui falsamente indicati come non a rischio) tra individui bianchi rispetto ai non-bianchi. Il COMPAS ed altri strumenti di polizia preventiva stanno subendo numerose critiche, in particolare dai *community group* degli Stati Uniti che si organizzano contro le discriminazioni della polizia, ma la loro diffusione sembra al momento più rapida degli strumenti in grado di regolarli[3].

Le istituzioni locali, nazionali ed internazionali sono da alcuni anni alle prese con la costruzione di strumenti legislativi e prassi amministrative in grado di affrontare questi temi.[4] Negli ultimi anni si è comunque osservata una rinnovata spinta per "responsabilizzare" gli algoritmi e renderli più trasparenti. Come riportato nell'articolo di Courtland (2018), il consiglio cittadino di New York ha messo in piedi una *task force* per incentivare la condivisione pubblica degli algoritmi, ed indagare i loro funzionamenti discriminanti; recentemente il governo francese si è impegnato a rendere *open* tutti gli algoritmi utilizzati nella burocrazia di selezione statale, ed il governo inglese ha invitato alla responsabilizzazione e trasparenza dei *data* nel settore pubblico.

L'indagine della discriminazione rimane comunque un campo difficile da regolamentare, proprio perché i confini tra l'algoritmo e gli altri comportamenti sociali sono sfumati ed imprecisi. Nel seguito approfondiremo in che modo gli algoritmi entrano in relazione con i fenomeni discriminatori e li interpretano, apprendono e/o riproducono, a partire da specifici modelli matematici.

3. SCHEMATIZZAZIONE DELL'APPRENDIMENTO E PROBLEMA DELLA AWARENESS

Esiste una vasta letteratura scientifica in cui i processi di Machine Learning e di selezione tramite algoritmi sono descritti come raccolta e sintesi di alcuni dati di per sé già presenti nel contesto "naturale", in questa visione il *bias* è semplicemente la misura di un errore commesso nel processo di apprendimento, una differenza numerica tra il risultato

[3] Nei nostri esempi non abbiamo trattato, e non approfondiremo nel seguito, il carattere performante e discriminante presente anche nella definizione degli spazi di possibilità di ogni attributo, cioè l'organizzazione della raccolta di dati secondo griglie che prevedono specifiche risposte possibili per ogni attributo. Chiaramente anche questi aspetti contribuiscono ad influenzare il contesto sociale a cui l'algoritmo viene applicato, e quindi costituiscono uno degli elementi della governamentalità tramite algoritmi, sebbene nella nostra analisi ci soffermeremo su altri aspetti.

[4] Per una panoramica del rapporto tra strumenti matematici e vari settori di applicazione, tra cui le istituzioni legislative, sono referenze importanti gli studi Žliobaitė (2015), Romei, Ruggieri (2014) e Barocas, Selbst (2016).

dell'algoritmo ed il dato "vero". Dal nostro punto di vista invece le procedure di apprendimento e misura statistica partecipano alla costruzione di senso ed alla visione del mondo in cui gli stessi fenomeni di discriminazione vengono rilevati. La nostra analisi approfondirà dunque in che modo gli specifici quadri analitici e statistici contribuiscono a queste costruzioni di senso.

In questa sezione introdurremo alcuni elementi di teoria statistica centrali nella nostra trattazione. Successivamente descriveremo una prima formalizzazione della fairness di un algoritmo tramite la nozione di *unawareness*, spiegando il motivo della inadeguatezza di questo concetto rispetto ai contesti concreti di applicazione dei processi algoritmici. Concentreremo la nostra analisi sui quadri assiomatici necessari per definire diverse nozioni di fairness e discriminazione, anche se non entreremo nel dettaglio dei diversi strumenti di calcolo (regressioni lineari, reti neurali, *deep learning*, etc.), i modelli assiomatici da noi trattati permetteranno un'analisi approfondita del rapporto tra algoritmi e discriminazione.

Vediamo di seguito alcuni concetti statistici fondamentali. Un attributo sarà una variabile aleatoria discreta. Come esempio pensiamo ad un attributo X corrispondente alla razza di un individuo di una certa popolazione. Alla variabile sarà associato un insieme $V=\{v_1,v_2,...\}$ di possibili valori ed una distribuzione di probabilità tale che $p(X=v_i)$ sia un numero reale compreso tra 0 ed 1 per ogni v_i nell'insieme V, ed inoltre

$$\sum_{v_i \in V} p(X = v_i) = 1.$$

Quest'ultima condizione corrisponde alla certezza che l'attributo assuma un valore nell'insieme V. Nel nostro esempio, V è l'insieme delle razze presenti nella popolazione di riferimento e per ogni razza v_i, $p(X=v_i)$ è la probabilità che un individuo scelto casualmente sia di razza v_i, cioè equivalentemente $p(X=v_i)$ è la frazione di popolazione di razza v_i. Quando non c'è rischio di confusione, faremo un piccolo abuso di notazione indicando con p(X) la distribuzione di probabilità dell'attributo X.

Dati due attributi X_1, X_2, indichiamo con $X_1|X_2$ la variabile aleatoria X_1 condizionata alla variabile X_2: resta invariato l'insieme dei possibili valori di X_1, ma le distribuzioni $p(X_1|X_2)$ e $p(X_1)$ possono essere differenti, in particolare $p(X_1|X_2)$ dipende dal valore assunto dall'attributo X_2. Ad esempio, se consideriamo la popolazione di una

città, indichiamo con X_1 il solito attributo razza, e con X_2 l'attributo corrispondente al quartiere di residenza di un individuo della popolazione, allora $p(X_1=v_i|X_2=q)$ indica la probabilità che X_1 assuma valore v_i sapendo che X_2 è il quartiere q. Chiaramente questo valore potrebbe variare molto marcatamente, per ogni razza v_i, al variare del quartiere X_2. Nel caso in cui $p(X_1|X_2)=p(X_1)$ diciamo che i due attributi sono indipendenti.

Un *dataset* è il dato di:

- un insieme I che indicizza gli individui di una popolazione. Nel nostro caso si tratterà semplicemente della numerazione progressiva degli individui;
- una serie di attributi $X_1,X_2,...$ per ogni individuo.

Possiamo immaginare ad esempio un dataset con quattro attributi X_1, X_2, X_3, X_4 corrispondenti a razza, quartiere di residenza, età e reddito. In questo caso $X_1^{(i)}$, $X_2^{(i)}$, $X_3^{(i)}$, $X_4^{(i)}$ saranno rispettivamente la razza, il quartiere, l'età ed il reddito dell'i-esimo individuo, dove i è un indice nell'insieme I. Nelle nostre analisi avremo uno o più attributi la cui previsione è l'obiettivo del dispositivo algoritmico e li chiameremo attributi di classificazione, denotandoli usualmente con la lettera Y. Un attributo non di classificazione è detto attributo descrittivo. Ad esempio, nel caso precedente Y potrebbe essere la probabilità di vincere un concorso di ammissione universitario, o di commettere un crimine.

Introduciamo un ultimo elemento di notazione: come già detto la nozione di attributo di classificazione indica gli attributi la cui previsione è l'obiettivo dell'algoritmo; indichiamo con $\hat{Y}$ il predittore di un tale attributo di classificazione Y, cioè il risultato della predizione algoritmica, in quanto distinto dal "vero" valore Y. L'accuratezza, cioè la proprietà del predittore di essere statisticamente prossimo al vero valore, è una proprietà centrale di un processo algoritmico. Ovviamente la correlazione tra i due valori (quello stimato e quello vero), ed il senso di questa distinzione, dipende fortemente dal modello utilizzato, come vedremo meglio nel seguito.

Il funzionamento di un algoritmo di apprendimento e previsione può essere schematizzato come segue. A partire da un dataset, l'algoritmo stabilisce una correlazione tra gli attributi descrittivi e l'attributo di classificazione. L'algoritmo viene quindi utilizzato su individui (della medesima popolazione) di cui sono conosciuti gli attributi descrittivi, per prevedere l'attributo di classificazione. L'accuratezza del risultato può essere misurata, come vedremo, in molti modi. Questa schematizzazione generale non distingue gli specifici tipi di apprendimento, che non saranno oggetto della nostra trattazione.

Tra gli attributi descrittivi del dataset saranno presenti degli attributi indicati come sensibili (ad esempio razza, genere, appartenenza religiosa, orientamento sessuale, etc.), la nozione di discriminazione e quella di fairness saranno introdotte a partire dalla correlazione tra questi attributi e l'attributo classificatore. È importante sottolineare, come già detto nell'introduzione, che lo studio della discriminazione si concentra sia sul dataset di apprendimento che sul risultato dell'algoritmo, perché per dare senso al procedimento valutativo, e quindi anche alla discriminazione da esso veicolata, è necessario studiare la relazione che viene a crearsi tra questi due oggetti.

Un'idea intuitiva (non formalizzata matematicamente) di discriminazione ad opera di un processo algoritmico si fonda sul concetto di informazione: si ha discriminazione quando l'esito della valutazione di un individuo fornisce informazione sull'attributo sensibile (e viceversa). A partire da questo punto di vista nasce la proposta di oscuramento degli attributi sensibili come rimedio alle pratiche discriminatorie, ad esempio attraverso la creazione di database che non prevedono la registrazione di questi attributi: l'ignoranza del decisore dovrebbe cancellare anche il rischio di discriminazione. La questione della consapevolezza (*awareness*) degli attributi sensibili si può ritrovare anche in ambito legislativo e istituzionale, in particolare il comportamento rispetto agli attributi razziali e religiosi costituisce una grande differenza tra il sistema anglosassone e quello francese: nel primo la razza e la religione sono attributi registrati nei processi di censimento della popolazione, mentre le leggi della Repubblica Francese si muovono in senso diametralmente opposto.

Questa idea di eliminazione della discriminazione attraverso la unawareness è oggi però ritenuta totalmente inadeguata per comprendere i fenomeni discriminatori codificati nei dispositivi algoritmici, ed un semplice esempio permette di capirne il motivo: supponiamo che un attributo protetto X_1 sia correlato con un attributo non protetto (o più di uno) X_2, cioè che la distribuzione di probabilità $p(X_1|X_2)$ dipenda in maniera non trascurabile dal valore del secondo attributo X_2; la cancellazione dell'attributo sensibile allora non corrisponde alla cancellazione totale dell'informazione, qui utilizzata in senso intuitivo, su quello stesso attributo[5].

[5] Una nozione di informazione che corrisponde a questo utilizzo intuitivo del termine si può ritrovare nella Teoria dell'Informazione di Shannon. Dati due attributi X_1, X_2 si definisce una grandezza positiva $H_{X1|X2}$ denominata entropia condizionale di X_1 rispetto a X_2: se V_{X1} e V_{X2} sono gli insiemi dei valori assunti da X_1 e X_2 rispettivamente, e denotiamo $p(x_1|x_2)$ la probabilità $p(X_1=x_1|X_2=x_2)$ per tutti i valori x_1, x_2 in V_{X1},V_{X2} rispettivamente, allora l'entropia condizionale è definita come

Un caso classico è quello di un processo di selezione, pensiamo alla valutazione necessaria per la concessione di un prestito bancario, in cui uno degli attributi descrittivi è il quartiere di residenza: in molte situazioni l'attributo razziale è fortemente dipendente dal quartiere di residenza, pertanto cancellare il primo non elimina la possibilità che individui di una specifica razza siano avvantaggiati o svantaggiati. Il processo di selezione dipenderà ancora dall'attributo sensibile, in maniera potenzialmente discriminatoria. In aggiunta, la cancellazione dell'attributo dal dataset non permetterà nessuna forma di controllo della discriminazione. Questo fenomeno prende il nome di *redlining* (o codificazione ridondante). Le origini del termine sono da ricercare nella pratica di alcuni istituti bancari ed assicurativi americani di negare servizi finanziari a specifici quartieri; in particolare negli anni del *New Deal* venne istituita una vera e propria mappatura dei territori cittadini in cui alcune zone a prevalenza afroamericana e di basso reddito erano contrassegnate col colore rosso. Lo studio di questo fenomeno ha permesso di evidenziarne non soltanto il carattere chiaramente discriminatorio, ma anche gli effetti performanti nel rafforzare le condizioni di svantaggio delle zone marcate in rosso. Risulta chiara la natura governamentale, in senso foucaultiano, di una tale mappatura: l'organizzazione della popolazione non è ottenuta tramite l'obbedienza, ma attraverso l'utilizzo di alcuni saperi sulla popolazione, in particolare grazie alla definizione di una "natura" del corpo sociale iscritta nella cartografia (Vagnarelli, 2017) sul concetto di governamentalità in Foucault). Come esposto nello stesso articolo di Vagnarelli, i dispositivi governamentali funzionano in modo tale per cui «Non vi sarà più la natura da un lato e il sovrano dall'altro ma la "natura" della popolazione farà il suo ingresso all'interno delle tecniche di potere» (Vagnarelli, 2016, 150). Questa descrizione del redlining si applicherà anche ai dispositivi algoritmici che affronteremo più avanti.

Nel seguito rispetteremo sempre il principio per cui gli attributi sensibili, potenzialmente soggetti a discriminazione, vengono registrati nel dataset di apprendimento. Sottolineiamo che questa scelta, largamente accettata nella letteratura scientifica, è comprensibile solo a partire da una nozione intuitiva di discriminazione che mette in evidenza l'inadeguatezza della "fairness come unawareness".

$$H_{X_1|X_2} := \sum_{x_2 \in V_{X_2}} p(X_2 = x_2) \cdot \left(\sum_{x_1 \in V_{X_1}} p(x_1 \mid x_2) \log(p(x_1 \mid x_2))\right).$$

Questa grandezza viene considerata una misura dell'informazione portata dall'attributo X_2 sull'attributo X_1: quanto più $H_{X1|X2}$ è prossima allo zero, tanto più la conoscenza di X_2 dà informazione su X_1. In particolare, $H_{X1|X2}$ vale zero quando la conoscenza del valore di X_2 determina esattamente il valore di X_1, ed è massima quando X_1 e X_2 sono indipendenti.

Osserviamo insomma con questo primo esempio che la definizione di *fairness* entra necessariamente in relazione con il significato storicizzato della discriminazione.

In più vedremo in seguito che a partire da altre nozioni di fairness, il principio della *unawareness* può essere anche completamente ribaltato. Con questo intendiamo dire che dentro altre modellizzazioni, un algoritmo fair può potenzialmente applicare delle forme di "azione positiva", o *affirmative action*, andando cioè a compensare direttamente alcune categorie svantaggiate.

4. DUE DIVERSE IDEE DI FAIRNESS IN UN PARADIGMA DI AWARE DATA MINING

In questa sezione analizzeremo due impostazioni del problema della fairness all'interno di un paradigma di *aware data mining*: con questa locuzione intendiamo che il processo di costruzione del dataset di apprendimento, la raccolta dei dati anche detta "data mining", registra gli attributi sensibili. Le due definizioni che proponiamo non sono da intendersi come contrapposte o divergenti, ci interessa però mostrare le diverse implicazioni sociali codificate dai due approcci e provare a leggere alla luce di queste il rapporto tra gli algoritmi e altri dispositivi sociali di valutazione, controllo e selezione.

Con l'obiettivo di semplificare la nostra trattazione, considereremo un attributo sensibile X di tipo binario: X ha come valori possibili 1 e 0, e 1 corrisponderà alla classe svantaggiata, ad esempio se X è il genere di un individuo scelto tra due possibili, il valore 1 dovrebbe corrispondere al genere potenzialmente discriminato. Anche l'attributo di classificazione che cercheremo di prevedere, indicato con $\hat{Y}$, sarà di tipo binario: in generale indicheremo con $\hat{Y}=1$ l'esito positivo nel processo di selezione, e con $\hat{Y}=0$ l'esito negativo; nel caso di un processo di valutazione per la concessione di un prestito bancario, $\hat{Y}=1$ corrisponderà alla scelta di concedere il prestito, mentre $\hat{Y}=0$ a quella di non concederlo. La scelta della notazione $\hat{Y}$ è dovuta al fatto che in molti casi pratici questo attributo è considerato un predittore (vedi sezione precedente) per il "vero" attributo Y: nel caso del prestito bancario si può ad esempio considerare che esista un attributo Y corrispondente alla capacità di solvenza dell'individuo, che l'algoritmo cerca di predire[6].

[6] Ovviamente la possibilità di ben definire l'attributo Y varia per ogni specifico procedimento, nel caso di un prestito bancario è in realtà molto difficile individuare una definizione per la "vera" capacità di ripagare un prestito quando questo prestito non è stato

Denominiamo accuratezza la precisione con la quale il predittore $\hat{Y}$ predice correttamente il valore Y. L'accuratezza si misura d'abitudine considerando la probabilità che $\hat{Y}$ sia corretto quando Y=1 e quando Y=0, cioè tramite i due valori[7]

$$p(\hat{Y} = 1 \,|\, Y = 1), p(\hat{Y} = 0 \,|\, Y = 0).$$

Il primo approccio che prendiamo in considerazione definisce come non-discriminanti delle procedure di classificazione per le quali l'accuratezza è indipendente dall'attributo protetto X. I due lavori che utilizzeremo come punto di riferimento sono Dwork *et al.* (2012) e Hardt, Price, Srebro (2016); in particolare nel secondo abbiamo due definizioni formali di predittori non discriminanti:

- un predittore $\hat{Y}$ rispetta l'*uguaglianza delle opportunità* se per un individuo con Y=1, la probabilità che venga classificato correttamente è indipendente dall'attributo protetto X, cioè

$$p(\hat{Y} = 1 \,|\, X = 1, Y = 1) = p(\hat{Y} = 1 \,|\, X = 0, Y = 1)$$

- un predittore $\hat{Y}$ rispetta l'*uguaglianza delle probabilità* se oltre alla condizione precedente, la stessa cosa vale per gli individui con Y=0, cioè

$$p(\hat{Y} = 0 \,|\, X = 1, Y = 0) = p(\hat{Y} = 0 \,|\, X = 0, Y = 0).$$

Le condizioni di uguaglianza delle opportunità ed uguaglianza delle probabilità attestano che l'accuratezza nella descrizione, cioè la probabilità di uguaglianza tra $\hat{Y}$ e Y, sia indipendente dal fattore sensibile X. La dipendenza tra X ed il valore "vero" Y, non viene considerata in questo senso un fattore discriminante. Secondo questo approccio dunque a partire dal contesto su cui sono raccolti i dati, l'algoritmo apprende la correlazione tra gli attributi descrittivi (tra cui X) e Y, e la fairness corrisponde alla precisione nella predizione di Y

concesso; diversamente nel caso di un algoritmo che cerca di prevedere l'interessamento di un individuo per un particolare prodotto venduto in un supermercato, è sempre ben definito un attributo che indichi se quel prodotto è stato acquistato o meno. I diversi quadri epistemologici della nozione di predittore non sono comunque oggetto della nostra trattazione.

[7] Ricordiamo che la probabilità condizionale $p(\hat{Y}|Y)$ è definita come la probabilità che $\hat{Y}$ assuma uno specifico valore quando è noto il valore di Y.

(tramite il predittore $\hat{Y}$).

Nei casi pratici, una simile nozione di fairness permetterà di individuare tra diversi algoritmi applicati su uno stesso dataset, quello meno discriminante; allo stesso tempo questi strumenti d'analisi non permettono di fare una valutazione della discriminazione intrinseca al dataset.

Prima di introdurre il secondo approccio consideriamo un'altra misura della discriminazione denominata differenza media

$$d_1 := p(\hat{Y} = 1 \,|\, X = 0) - p(\hat{Y} = 1 \,|\, X = 1).$$

Questa grandezza[8] esprime la differenza tra la probabilità di avere un esito positivo condizionato al valore sensibile X nella classe 0, e la stessa probabilità con valore sensibile X nella classe 1. La differenza media è una misura della discriminazione che non è condizionata a nessun altro attributo oltre che quello protetto, e la condizione $d_1=0$ descrive una parità statistica, rispetto all'algoritmo, tra i vari gruppi caratterizzati da uno stesso valore dell'attributo sensibile. Un'altra misura[9] definita in modo analogo è l'*impact ratio*

$$I_1 = \frac{p(\hat{Y} = 1 \,|\, X = 0)}{p(\hat{Y} = 1 \,|\, X = 1)}.$$

Il secondo approccio che introduciamo utilizza queste misure "di gruppo" per quantificare la fairness dell'algoritmo: un algoritmo è ritenuto non-discriminante se la differenza media d_1 è prossima allo zero (o l'impact ratio I_1 è prossimo all'unità). Nei casi pratici queste misure della discriminazione possono essere utilizzate per valutare il risultato di un algoritmo ma anche il dataset di apprendimento. In Pedreschi, Ruggieri, Turini (2008; 2009) il rapporto tra la fairness del dataset e quella del risultato dell'algoritmo è indagata approfonditamente, in questo senso il dataset iniziale diventa una componente dell'algoritmo stesso, poiché è possibile valutare se il processo di apprendimento ha corretto la discriminazione presente nei dati iniziali.

---

[8] Ovviamente è possibile definire la stessa differenza per l'esito negativo, ma la grandezza così ottenuta

$$d_0 := p(\hat{Y} = 0 \,|\, X = 0) - p(\hat{Y} = 0 \,|\, X = 1)$$

è semplicemente l'inverso di d_1, cioè $d_0=-d_1$

[9] Per una presentazione generale di queste ed altre misure della discriminazione utilizzate in letteratura, vedere lo studio Žliobaitė (2015).

In Kamiran, Calders (2009; 2010) vengono studiate alcune tecniche per aumentare la fairness di gruppo di un algoritmo, in particolare modificando il dataset di apprendimento o i risultati dell'algoritmo in modo da favorire il gruppo svantaggiato, attuando in questo modo una forma di "azione positiva". Senza sviluppare una descrizione di queste tecniche, risulta importante sottolineare che la riduzione della discriminazione di gruppo potrebbe entrare in conflitto con l'altra nozione di fairness che abbiamo dato sopra. Gran parte degli articoli che abbiamo citato si occupano infatti anche di individuare le condizioni necessarie per avere un algoritmo accurato e fair sotto entrambi i punti di vista, esistono però dei casi di incompatibilità, in particolare ci soffermeremo qui sull'esempio riportato dall'articolo di Courtland (2018). Consideriamo una popolazione di individui arrestati ed un algoritmo che valuta la possibilità che ogni individuo divenga recidivo subendo un secondo arresto, sia Y=1 se l'individuo diviene recidivo e Y=0 in caso contrario, e sia $\hat{Y}$ il predittore di Y. Se vogliamo un predittore che rispetti l'ipotesi di uguaglianza delle probabilità descritta sopra, ma allo stesso tempo esiste una discriminazione sistemica in cui gli individui di attributo protetto X=1 vengono arrestati con maggior frequenza, allora è impossibile eliminare la discriminazione secondo una metrica di gruppo come l'impact ratio: gli individui di classe X=1 verranno considerati più probabilmente a rischio di recidività degli altri, ottenendo con più frequenza una classificazione $\hat{Y}$=1. L'incompatibilità tra i due punti di vista si manifesta proprio dal rapporto con una condizione di discriminazione sistemica, cioè la maggior frequenza assoluta di arresti per alcune categorie di individui. Quantificando con più precisione, se denominiamo d la differenza tra la probabilità di arresto per individui con X=1 e per individui con X=0, allora esiste una dipendenza lineare tra l'impact ratio I_1 e la differenza d.

Un'interessante prospettiva sui due approcci di cui sopra, viene portata dall'articolo di Friedler, *et al*. (2016). L'idea di fairness come uguaglianza delle opportunità o uguaglianza delle probabilità, viene descritta come un tentativo di cancellare le discriminazioni a partire dalla supposizione che i metodi di raccolta dati (test, esami, analisi a campione) producano una descrizione accurata della realtà, una prospettiva indicata con l'acronimo WYSIWYG, per *What You See Is What You Get*. I metodi basati sul WYSIWYG partono dunque dall'idea che esista una classificazione "giusta" degli individui, ed il compito di un algoritmo fair è quello di ricostruire questa classificazione eliminando le discriminazioni individuali basate sugli attributi protetti.

L'idea di fairness come parità statistica tra gruppi si basa, secondo

Friedler *et al.* (2016), sulla supposizione che in un contesto non discriminatorio, i gruppi di individui caratterizzati da uno stesso valore degli attributi sensibili, otterrebbero percentuali simili di risultati positivi, una prospettiva indicata con l'acronimo WAE, per *We're All Equal*. I metodi basati sul WAE si fondano dunque sull'idea che l'unica classificazione "giusta" è quella in cui si ha una parità statistica dei risultati tra i diversi gruppi ed il compito di un algoritmo fair è quello di operare una selezione che rispetti il più possibile questo valore di parità. Il diverso focus tra discriminazione individuale e discriminazione di gruppo è evidenziato anche in Dwork *et al*. (2012), a partire da un punto di vista WYSIWYG: l'idea di parità statistica tra gruppi viene criticata a partire dalle discriminazioni individuali che potrebbe produrre.

La prospettiva dei due approcci WYSIWYG e WAE non riesce a cogliere la complessità e varietà delle proposte per la costruzione di algoritmi fair, in particolare come abbiamo già detto in diverse situazioni vengono utilizzati entrambi i metodi, laddove esistono condizioni di compatibilità, come mostrato sempre in Dwork *et al*. (2012) e in Friedler *et al.* (2016), ma anche in Kleinberg, Ludwig, Mullainathan, Rambachan (2018) e Corbett-Davies, Pierson, Feller, Goel, Huq (2017). In quest'ultimo lavoro in particolare la fairness viene propriamente analizzata in quanto *constrained optimization*: si lavora cioè in un paradigma WYSIWYG inserendo però alcuni vincoli sulla disparità statistica tra gruppi.

La classificazione di questi due approcci è comunque utile ad evidenziare le differenti visioni del mondo implicite nelle varie definizioni di fairness, cioè come queste visioni vengano codificate dagli specifici strumenti matematici utilizzati. In questo senso ci interesseremo in particolare agli elementi di incompatibilità tra i diversi modelli, in modo da evidenziare gli elementi sostanziali di queste costruzioni di senso:

Anche se abbiamo introdotto questi due assiomi come differenti visioni del mondo o sistemi di credenze, questi possono anche essere una scelta strategica. Qualsiasi sia la motivazione (che è in fin dei conti matematicamente irrilevante), la scelta degli assiomi è fondamentale per un processo di decisione. Gli assiomi scelti determinano il significato di fairness (Friedler *et al.*, 2016, 9) [10].

Possiamo collegare queste analisi con lo studio del concetto di "struttura" sviluppato in (Crammond, Carey, 2016). Gli autori si soffermano sulla relazione tra il concetto di disuguaglianza strutturale e

[10] Traduzione dell'autore.

quello di *habitus* in Bourdieu, nel caso particolare delle disuguaglianze sanitarie, rilevando l'esistenza di alcuni discorsi socialmente pervasivi (com'è il caso della nozione di "rischio" associata con l'abitudine di fumare) che strutturano l'habitus individuale senza determinare univocamente i comportamenti.

Nel nostro studio della fairness rileviamo la presenza di diversi discorsi con differenti livelli di pervasività. La fairness, o complementarmente la discriminazione, è un concetto non aggirabile all'interno dello studio di algoritmi; questo primo elemento può venir sviluppato secondo differenti metodi (come quelli WYSIWYG e WAE che abbiamo descritto sopra) che vanno quindi a specificarne il significato; ognuno di questi approcci genera conseguentemente un proprio campo di pervasività che abbiamo caratterizzato con alcune assunzioni fondative: ad esempio WYSIWYG suppone l'esistenza a priori di una classificazione individuale corretta, mentre WAE assume la condizione di parità statistica tra gruppi. Una prospettiva di questo tipo permette di approfondire in che modo il dispositivo algoritmico influisce sull'habitus, e quindi anche di rilevare gli aspetti di incompatibilità tra varie e compresenti nozioni pervasive.

5. LA DISCRIMINAZIONE ALL'INTERNO DELLA "GOVERNANCE BY THE NUMBERS"

Yarden Katz (2017) ipotizza che la centralità assunta nell'ultima decade dal termine Intelligenza Artificiale (in ambito accademico, industriale, politico, sui media, etc.) sia spiegabile in particolare attraverso il tentativo di affermare alcune agende politiche di stampo neoliberista. Nella sua trattazione Katz utilizza il concetto di "*governance by the numbers*" ripreso da (Supiot, 2012): questa locuzione sta ad indicare l'introduzione di metriche quantitative in numerosi ambiti sociali, dal lavoro, alla selezione scolastica, al sistema giuridico, fino alle valutazioni di rischio per la concessione di prestiti bancari o la stipula di contratti assicurativi; attraverso queste metriche è così possibile istituire forme di governance che funzionano sulla base di graduatorie e soglie di inclusione. Da questo punto di vista la cultura dell'Intelligenza Artificiale, e la connessa proliferazione di algoritmi valutativi e di selezione, hanno la funzione di rendere efficaci e conosciute queste metriche quantitative, oltre che di disciplinare la popolazione al loro utilizzo.

Una caratteristica fondamentale individuata da Katz nel processo di diffusione di procedure algoritmiche, è la creazione di un punto di vista

esterno ("*vision from nowhere*" nell'articolo) rispetto allo specifico contesto sociale in cui l'algoritmo introduce la sua metrica; da questo punto d'osservazione sarebbe possibile una descrizione oggettiva del terreno sociale su cui l'algoritmo agisce, ed è proprio questa pretesa di oggettività che fornisce al processo di selezione una naturalizzazione e dunque una legittimità sociale e politica. Riconosciamo in questo meccanismo il funzionamento di un dispositivo di governo secondo l'utilizzo del concetto che ne fa Foucault. La "natura" della popolazione diviene una tecnica di potere, e questo meccanismo diviene produttivo della popolazione stessa, cioè «Il termine dispositivo nomina ciò in cui e attraverso cui si realizza una pura attività di governo senza alcun fondamento nell'essere. Per questo i dispositivi devono sempre implicare un processo di soggettivazione, devono, cioè, produrre il loro soggetto» (Agamben, 2006, 16).

In realtà la "vision from nowhere" corrisponde ad un processo di invisibilizzazione delle condizioni contingenti in cui l'algoritmo è stato concepito, dei contesti materiali, storici, sociali a cui deve attenersi, degli interessi di gruppo che hanno agito per la sua diffusione, del sistema assiomatico che gli permette di funzionare e della costruzione di senso codificata in questo sistema assiomatico. L'algoritmo è legittimato in quanto oggettivo, ma è allora necessario riconcettualizzare questo termine, come suggerito in Mazzotti (2015), secondo termini durkehimiani: «per un concetto matematico essere oggettivo significa essere istituzionalizzato, inserito in una rete di concetti e pratiche supportate dagli interessi collettivi di un gruppo» (Mazzotti, 2015, 467).

La marginalizzazione del formalismo matematico necessario per la definizione di una procedura valutativa è strettamente necessaria alla costruzione fittizia dell'esternità della "vision from nowhere", perché vengono nascosti proprio i vincoli metodologici ed assiomatici a cui l'algoritmo è interno.

Nel momento in cui il tema della fairness viene inserito tra le caratteristiche formalizzate di un algoritmo, anche il problema della discriminazione rientra potenzialmente all'interno di quel paradigma oggettivista che legittima la governance statistica, e l'eliminazione della discriminazione si trasforma da problema politico a problema di ottimizzazione numerica. Attraverso i punti che abbiamo sollevato nelle sezioni precedenti, vogliamo invece mostrare che la costruzione matematica di un'idea di fairness avviene sempre all'interno di uno specifico modello assiomatico calato in un insieme di condizioni materiali a cui deve adattarsi, e ogni paradigma universalista nasconde inevitabilmente la contingenza di questi vincoli. Mettendo in luce le

frizioni tra le diverse modellizzazioni matematiche della discriminazione, e quelle con l'attuale organizzazione sociale, è possibile evidenziare le forme di discriminazione strutturale soggiacenti ai sistemi di governance.

Ha particolare utilità riprendere le considerazioni di Longo sul rapporto tra matematica e costruzione di conoscenza, e più precisamente tra geometria ed organizzazione dello spazio, per cercare di fare un'analogia con il rapporto tra modelli statistici ed interpretazione sociologica della discriminazione: «ogni discorso sull' "esistenza" di una struttura matematica [...] è fuori luogo» (2014, 6); così come le strutture materiali della visione non spiegano di per sé i concetti geometrici di linea e bordo, allo stesso modo le condizioni storiche della discriminazione non spiegano di per sé i diversi modelli di fairness algoritmica; piuttosto le condizioni di possibilità storiche permettono la definizione nella «comunità simbolica» (Ibidem) di modelli matematici di discriminazione in grado di organizzare lo spazio sociale tramite specifici strumenti, che nel contesto attuale assumono nella maggioranza dei casi la forma di dispositivi governamentali.

L'esempio riportato da Courtland (2018) che abbiamo descritto nella sezione 4 risulta particolarmente chiarificatore di questo rapporto con le condizioni materiali e dei rapporti di reciproca (in)compatibilità e frizione tra modelli matematici distinti. Se un dispositivo algoritmico viene concepito per assistere il processo giudiziario, la retorica che legittima la "governance by the numbers" descriverà il problema dell'eliminazione della discriminazione come l'ottimizzazione di una specifica funzione. La trattazione che abbiamo fatto ci permette però di osservare che separando il processo giudiziario dalle prassi concrete di controllo e repressione, si deve accettare come vincolo strutturale la discriminazione negli arresti, e questo pone una condizione di incompatibilità tra due modelli di fairness. L'applicazione del modello di valutazione si fonda quindi su una serie di scelte politiche multilivello (separazione tra giudiziario e amministrazione delle pratiche di controllo, applicazione dell'algoritmo al campo giudiziario, scelta di un modello di fairness tra i due modelli incompatibili) che vengono negate se il problema della discriminazione si pone meramente come problema di ottimizzazione.

Non si possono inoltre trascurare gli effetti di riproduzione della discriminazione inerenti a questa dinamica: come detto nella sezione precedente, se l'algoritmo giudiziario è modellizzato per approssimarsi il più possibile all'uguaglianza delle probabilità (una nozione di fairness che viene preferita in ambito giudiziario), allora la discriminazione di

gruppo riprodotta dall'algoritmo sarà direttamente correlata alla discriminazione di gruppo attuata dalla polizia, e l'algoritmo avrà come effetto quello di propagare le discriminazioni strutturali. Più in generale questo fenomeno si inserisce all'interno di tutti quegli effetti di performatività propri ad ogni procedura formalizzata. Nonostante non sia l'oggetto di questa trattazione, ci preme ricordare che la stessa definizione degli insiemi di valori possibili per ogni attributo codifica una normazione dello spazio sociale: quando all'attributo "genere"‘ vengono assegnati due valori possibili (o comunque un insieme di valori prestabiliti), si escludono dalla trattazione tutte le soggettività che si identificano in maniera non binaria, e questo oltre ad avere degli effetti di invisibilizzazione immediati, produce forme di discriminazione nella misura in cui la raccolta di dati ha concrete conseguenze politiche, sociali, lavorative, etc.

Come abbiamo provato a spiegare fino a qui, è necessario considerare il modello assiomatico alla base di un algoritmo come facente parte dell'algoritmo stesso per dare una cornice di senso alla sua azione. Nella retorica oggi più affermata, i processi automatizzati sono invece dei sistemi di calcolo più o meno interpretabili, il cui obiettivo è ridotto alla descrizione tramite il calcolo di un ambiente sociale autonomamente provvisto di senso. Più precisamente, le metriche che permettono la misurazione della realtà vengono sempre più spesso naturalizzate, nascondendo l'atto del dispiegamento di queste metriche come creativo di informazione. Ad essere oscurato è anche il fatto che la stessa cultura degli algoritmi favorisce in molti casi il dispiegamento di queste metriche, incanalando forme di disciplinamento collettivo e qualificandosi quindi anch'essa come dispositivo governamentale.

Cerchiamo di osservare in che modo gli specifici quadri assiomatici aderiscono al contesto materiale nel caso dei due modelli di fairness che abbiamo visto nella sezione 4:

- la nozione di fairness individuale formalizzata con l'uguaglianza delle possibilità, e l'uguaglianza delle probabilità, si fonda su un'idea di ranking individuale "giusto": la metrica specifica di questo ranking dipende dalle condizioni del dataset e dall'algoritmo utilizzato, e una condizione del tipo descritto (l'indipendenza tra $\hat{Y}$ e l'attributo sensibile X, condizionata alla conoscenza del vero valore Y) permette l'esistenza di una classificazione veritiera basata sull'attributo Y. Nel caso descritto in Courtland (2018), sono le pratiche di controllo della polizia a determinare l'attributo Y, e quindi appare chiaramente come una condizione frutto di contingenze potenzialmente discriminanti sia incorporata nella metrica;

- nel caso della parità statistica, invece, la metrica utilizzata è quella della differenza media (o dell'impact ratio) tra le probabilità di successo per i differenti gruppi (ogni gruppo è caratterizzato dal valore dell'attributo sensibile). In questo caso le condizioni materiali in cui l'algoritmo si inserisce sono descritte dal dataset, e il processo di valutazione cerca di approssimare una condizione di parità statistica. Nel caso di Courtland (2018) un meccanismo di parità statistica in campo giudiziario può essere approssimato anche in presenza di dataset con discriminazione (prodotta dalle pratiche di controllo e repressione) a costo di rinunciare alle forme di fairness come uguaglianza delle probabilità. In questo senso l'algoritmo ha la funzione di correggere in sede giudiziaria la discriminazione del dataset creato dalle prassi della polizia.

Da entrambi questi esempi si ricava che il tema della fairness, se posto come semplice ottimizzazione di una funzione di discriminazione, rientra completamente dentro il paradigma del dispositivo di "governance by the numbers". Inoltre, vista la centralità assunta dalle battaglie contro la discriminazione, la qualificazione di un algoritmo come "algoritmo fair", fornisce una legittimazione degli stessi dispositivi di governo anche quando questi dispositivi operano a partire da condizioni strutturali in cui sono presenti forti discriminazioni, riproducendole.

L'emergenza storica del concetto di discriminazione, frutto di un percorso tortuoso ed in particolare di conflitti attorno ai temi del razzismo e delle discriminazioni di genere, non è oggetto di questa trattazione, ci preme però ripetere che questo tema raggruppa una serie di significati originati storicamente, che non possono essere ridotti alla loro formalizzazione aritmetica. In questo senso è utile ricordare l'esempio della "fairness come unawareness" di cui abbiamo parlato nella sezione 3: se assumiamo assiomaticamente questa definizione di fairness, chiaramente è possibile eliminare completamente la discriminazione algoritmica, semplicemente costruendo database che non contengono gli attributi sensibili; questo avviene a costo di una separazione radicale tra il significato storico della discriminazione e la sua formalizzazione. Questo processo di separazione è presente, in diverse maniere, dentro ogni modellizzazione formale, e quindi le interpretazioni degli algoritmi come dispositivi neutri di analisi oggettiva della realtà, attuano effettivamente un'amputazione del processo storico contenuto nel concetto di discriminazione. Guardare il tema della discriminazione attraverso strumenti come gli algoritmi di decisione, dal nostro punto di vista significa piuttosto capire in che

modo le forme di discriminazione strutturale interagiscono con questi sempre più diffusi strumenti di governo.

Ci sembra importante porre l'accento anche su alcune scelte terminologiche che indirizzano la percezione collettiva del problema: è inadeguato riferirsi alla ricerca di algoritmi privi di bias o *unbiased*, perché queste locuzioni suggeriscono la presenza di contesti di selezione completamente "giusti" a cui ogni algoritmo può avvicinarsi per approssimazione. Il nostro interesse è piuttosto orientato sull'organizzazione di senso (la visione del mondo) specifica che viene codificata in ogni definizione di fairness, nei rapporti concreti di relazione e frizione che vengono a formarsi tra modelli distinti, e con la realtà sociale.

Oggi ogni opzione di governo iscrive le sue motivazioni ideologiche anche all'interno delle modellizzazioni matematiche proprie alla "governance by the numbers", lo studio della discriminazione strutturale deve dunque porre il problema del significato politico di queste modellizzazioni per poter portare una critica profonda al ruolo degli algoritmi. Per lo stesso motivo queste analisi sono anche uno strumento irrinunciabile di quei movimenti politici e culturali che portano una critica generale ad ogni sistema fondato su forme di discriminazione strutturale.

6. CONCLUSIONI

La narrazione oggi più diffusa attorno ai processi di decisione algoritmica pone il problema della fairness come l'insieme degli strumenti di calcolo necessari ad ottimizzare alcune funzioni che misurerebbero il livello di discriminazione oggettiva. La nostra proposta di ricerca, che abbiamo provato embrionalmente a sviluppare in questo lavoro, suggerisce invece di ricercare il rapporto tra fairness ed algoritmi nella complessa rete di compatibilità ed incompatibilità che esistono tra i modelli matematici con cui è possibile descrivere la discriminazione, e le strutture sociali in cui la discriminazione strutturale è emersa storicamente e come problema politico, soprattutto in conseguenza delle lotte messe in atto dai soggetti oppressi. Ci siamo qui concentrati in particolare su un caso di studio che evidenzia il rapporto tra diversi modelli di fairness applicabili al campo degli algoritmi giuridici, e la presenza di una discriminazione razziale strutturale nelle pratiche di polizia; questo stesso approccio potrebbe essere generalizzato a numerosi campi come la selezione scolastica e lavorativa, l'accesso al credito bancario, l'analisi del linguaggio, il

gender salary gap, etc.

Il nostro lavoro si fonda sull'assunzione che ogni analisi efficace della discriminazione debba saper individuare in che modo i dispositivi di governo vigenti si iscrivono in un quadro strutturalmente discriminante, ed in che modo lo riproducono. In un'epoca in cui i modelli statistici e gli algoritmi sono centrali per le pratiche di governo, la nostra ricerca vuole anche essere uno strumento per portare nel campo delle decisioni automatizzate una critica ai meccanismi che veicolano e riproducono ogni forma di discriminazione.

RIFERIMENTI BIBLIOGRAFICI

AGAMBEN, G. (2006). *Che cos'e un dispositivo?*. Roma: Nottetempo.

ANGWIN, J., LARSON, J., MATTU, S., KIRCHNER, L. (2016). *Machine bias. there's software used across the country to predict future criminals. and it's biased against blacks*. *Propublica*, May 23.

BAROCAS, S., SELBST, A. (2016). Big data's disparate impact. *California Law Review*, 104, 671-732.

CHEN, L., MA, R., HANNAK, A., WILSON, C. (2018). Investigating the impact of gender on rank in resume search engines. *Proceedings of the 2018 chi conference on human factors in computing systems*, 651, 1-14.

CORBETT-DAVIES, S., PIERSON, E., FELLER, A., GOEL, S., HUQ, A. (2017). Algorithmic decision making and the cost of fairness. *KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797-806.

COURTLAND, R. (2018). Bias detectives: the researchers striving to make algorithms fair. *Nature, 558*, 357-360.

CRAMMOND, B., CAREY, G. (2016). What do we mean by 'structure' when we talk about structural influences on the social determinants of health inequalities?. *Social Theory & Health*, 15(1), 1-15.

CURCIO, A., MELLINO, M. (2012). *La razza al lavoro*. Roma: Manifestolibri.

DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O., ZEMEL, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-226.

FRIEDLER, S. A., SCHEIDEGGER, C. E., VENKATASUBRAMANIAN, S. (2016). On the (im)possibility of fairness. *arXiv.org*, 1609.07236

HARDT, M., PRICE, E., SREBRO, N. (2016). Equality of opportunity in supervised learning. *Proceedings of the 30th international conference*

on neural information processing systems, 3323-3331.

KAMIRAN, F., CALDERS, T. (2009, Feb). Classifying without discriminating. *2009 2nd international conference on computer, control and communication*, 1-6.

KAMIRAN, F., CALDERS, T. (2010). Classification with no discrimination by preferential sampling. *Informal proceedings of the 19th annual machine learning conference of belgium and the netherlands (Benelearn'10, Leuven, Belgium, may 27-28, 2010)*, 1-6.

KATZ, H. (2017). *Manufacturing an Artificial Intelligence Revolution: Neoliberalism and the 'new' big data Yarden Katz*. Harvard: Harvard University.

KLEINBERG, J., LUDWIG, J., Mullainathan, S., Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings*, 108, 22-27.

LONGO, G. (2014). Le conseguenze della filosofia. In R. Lanfredini (a cura di), *A Plea for Balance in Philosophy. Essays in Honour of Paolo Parrini. Vol.2: New Contributions and Replies* (pp. 17-44). Pisa: ETS.

MAZZOTTI, M. (2015). Per una sociologia degli algoritmi. *Rassegna Italiana di Sociologia*, 3-4, 465-478.

PEDRESCHI, D., RUGGIERI, S., TURINI, F. (2008). Discrimination-aware data mining. *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining*, 560-568.

PEDRESCHI, D., RUGGIERI, S., TURINI, F. (2009). Measuring discrimination in socially-sensitive decision records. *Proceedings of the SIAM International Conference on Data Mining Sdm*, 581-592.

ROMEI, A., RUGGIERI, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5), 582-638.

SONNAD, N. (2017). Google translate's gender bias pairs "he" with "hardworking" and "she" with lazy, and other examples. *Quartz*, November 29.

SUPIOT, A. (2012). *The spirit of Philadelphia: Social justice vs. the total market*. London: Verso.

VAGNARELLI, G. (2017). Foucault e i confini del governo: la governamentalità. *Heteroglossia*, 15, 145-164.

ŽLIOBAITĖ, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv.org*, 1511.00148.

Numero chiuso il 30 marzo 2019

ULTIMI NUMERI

2018/2 (aprile-giugno):

1. ILARIA IANNUZZI, L'ebraismo nella formazione dello spirito capitalistico. Un excursus tra le opere di Werner Sombart;
2. NICOLÒ PENNUCCI, Gramsci e Bourdieu sul problema dello Stato. Dalla teoria della dominazione alla sociologia sto-rica;
3. ROSSELLA REGA, ROBERTA BRACCIALE, La self-personalization dei leader politici su Twitter. Tra professionalizzazione e intimizzazione;
4. STEFANO SACCHETTI, Il mondo allo specchio. La seconda modernità nel cinema di Gabriele Salvatores;
5. GIULIA PRATELLI, La musica come strumento per osservare il mutamento sociale. Dylan, Mozart, Mahler e Toscanini;
6. LUCA CORCHIA, Sugli inizi dell'interpretazione sociologica del rock. Alla ricerca di un nuovo canone estetico;
7. LETIZIA MATERASSI, Social media e comunicazione della salute, di Alessandro Lovari.

2018/3 (luglio-settembre):

1. RICARDO A. DELLO BUONO, Social Constructionism in Decline. A "Natural History" of a Paradigmatic Crisis;
2. MAURO LENCI, L'Occidente, l'altro e le società multiculturali;
3. ANDREA BORGHINI, Il progetto dei Poli universitari penitenziari tra filantropia e istituzionalizzazione;
4. EMILIANA MANGONE, Cultural Traumas. The Earthquake in Italy: A Case Study;
5. MARIA MATTURRO, MASSIMO SANTORO, Madre di cuore e non di pancia. Uno studio empirico sulle risonanze emotive della donna che si accinge al percorso adottivo;
6. PAULINA SABUGAL, Amore e identità. Il caso dell'immigrazione messicana in Italia;
7. FRANCESCO GIACOMANTONIO, Destino moderno. Jürgen Habermas. Il pensiero e la critica, di Antonio De Simone.
8. VINCENZO MELE, Critica della folla, di Sabina Curti.

2018/4 (ottobre-dicembre):

1. ENRICO CAMPO, ANTONIO MARTELLA, LUCA CICCARESE, Gli algoritmi come costruzione sociale. Neutralità, potere e opacità;
2. MASSIMO AIROLDI, DANIELE GAMBETTA, Sul mito della neutralità algoritmica;
3. CHIARA VISENTIN, Il potere razionale degli algoritmi tra burocrazia e nuovi idealtipi;
4. MATTIA GALEOTTI, Discriminazione e algoritmi;
5. BIAGIO ARAGONA, CRISTIANO FELACO, La costruzione socio-tecnica degli algoritmi;
6. ANIELLO LAMPO, MICHELE MANCARELLA, ANGELO PIGA, La (non) neutralità della scienza e degli algoritmi;
8. LUCA SERAFINI, Oltre le bolle dei filtri e le tribù online;
9. COSTANTINO CARUGNO, TOMMASO RADICIONI, Echo chambers e polarizzazione;
10. IRENE PSAROUDAKIS, Mario Tirino, Antonio Tramontana (2018), I riflessi di «Black Mirror»;
11. JUNIO AGLIOTI COLOMBINI, Daniele Gambetta (2018), Datacrazia;
12. PAOLA IMPERATORE, Safiya Umoja Noble (2018), Algorithms of Oppression;
13. DAVIDE BERALDO, Cathy O'Neil (2016), Weapons of Math Destruction;
14. LETIZIA CHIAPPINI, John Cheney-Lippold (2017), We Are Data.