DigitCult | Scientific Journal on Digital Cultures

# Equità degli algoritmi e democrazia

Antonio Santangelo
Università di Torino

## Abstract

È sempre più comune, in diverse sfere della nostra vita quotidiana, imbatterci in algoritmi che vengono utilizzati per prendere decisioni al posto di un essere umano o per aiutare qualcuno a prenderle. Per fare in modo che queste tecnologie non producano ingiustizie, sta prendendo piede una branca della *computer science* denominata "*fairness*". L'idea di fondo è che una specifica visione della giustizia debba essere formalizzata con criteri statistici, che vengono poi utilizzati per realizzare gli strumenti informatici di cui ci serviamo. Così facendo, si dovrebbe essere certi che questi mezzi tecnici saranno automaticamente "giusti". Però, analizzando i criteri più attestati, nell'ambito della *fairness*, ci si rende conto che la maggior parte di essi si basa su una visione specifica della giustizia, che è quella liberale, nei termini che verranno descritti nell'articolo. Solo pochissimi esulano da questa concezione. Tutto ciò può produrre degli effetti sulla qualità della nostra vita democratica, visto che gli algoritmi che ne derivano vengono poi utilizzati per decidere se i carcerati possono ottenere la libertà vigilata o l'abbreviazione della pena, se qualcuno può ricevere un prestito da una banca o una carta di credito, se qualcun altro può accedere a un colloquio di lavoro, a una borsa di studio all'università, eccetera. Nelle pagine che seguono, verrà condotta una panoramica dei criteri di *fairness* degli algoritmi più comuni e verranno sollevati alcuni problemi filosofici, a proposito del tipo di democrazie che stiamo costruendo, facendo ricorso a strumenti informatici che supponiamo essere "intelligenti" e "giusti", nei vari ambiti della nostra società.

## Algorithmic Fairness and Democracy

It is becoming common, in many spheres of our everyday life, to bump into algorithms that are used to take decisions in our place or to help us take decisions. To prevent these technologies from generating injustices, a new branch of computer science is taking place, called algorithmic fairness. The idea is that a certain vision of justice must be formalized into some statistical criteria, that are put at the core of the computer tools we recur to. Doing so, we can be sure that those instruments will automatically be fair, every time we use them. However, if we study the most common algorithmic fairness criteria, we find out that they are generally based on a very specific vision of justice, which is liberal, in a sense that will be described. Only a few of them derive from another concept of justice. This may have some effects on the quality of our democratic life, as the algorithms that are projected coherently with those principles are used to decide whether to free people from jail or not, to give them a loan, a job, the possibility to study in a university or in another, etc. This article makes an overview of the field of algorithmic fairness and raises some philosophical problems about the kind of democracies we are building, by recurring to the computer tools that are more and more diffused in our societies.

*Published* 2 May 2021

Correspondence should be addressed to Antonio Santangelo, Università di Torino. Email: antonio.santangelo@unito.it

*DigitCult, Scientific Journal on Digital Cultures* is an academic journal of international scope, peer-reviewed and open access, aiming to value international research and to present current debate on digital culture, technological innovation and social change. ISSN: 2531-5994. URL: http://www.digitcult.it

 doi:10.53136/979125994120634

## Algoritmi liberali

Questo articolo prende le mosse da un lavoro condotto da chi scrive, presso il Nexa Center for Internet & Society del Dipartimento di Automatica e Informatica del Politecnico di Torino, in collaborazione con la Fondazione Bruno Kessler di Trento, insieme con un team interdisciplinare di studiosi[1], di cui fanno parte ingegneri informatici, statistici, data scientist e filosofi. Sull'argomento che qui si intende approfondire, questo gruppo di ricerca ha già prodotto un *paper*[2], a cui si farà riferimento, soprattutto per le nozioni tecniche e statistiche a supporto del discorso filosofico. L'obiettivo è di riflettere sul rapporto tra i criteri statistici che, poggiando su diverse concezioni dell'idea di giustizia, vengono utilizzati per stabilire l'equità (*fairness*) degli algoritmi, e alcune specifiche visioni della vita democratica, al fine di provare a esplicitare in quali tipi di democrazie viviamo e vivremo, nell'epoca in cui le tecnologie digitali per il processamento dei big data sono divenute parte integrante del nostro tessuto sociale.

Il motivo per cui si è deciso di portare avanti questo studio è legato alle preoccupazioni che, da più parti, si sollevano, a proposito dell'utilizzo sempre più massiccio di strumenti informatici basati sull'analisi di grandi moli di dati, sia a supporto delle nostre decisioni, sia per operare scelte al posto nostro, nei più svariati ambiti: giudiziario, carcerario, creditizio, assicurativo, sanitario, formativo, informativo, pubblicitario, eccetera. Qualcuno, come Kathy O'Neil (2016), chiama gli algoritmi che determinano il funzionamento di queste tecnologie "armi di distruzione matematica" e sostiene che essi «promettendo efficienza ed equità, distruggono l'istruzione superiore, fanno aumentare il debito, incentivano la carcerazione di massa, bistrattano i poveri in ogni maniera possibile e minacciano la democrazia» (O'Neil 2017[2016], 287). Nelle prossime pagine, si riporterà qualche esempio di come questo possa avvenire, ma per il momento è sufficiente sottolineare che se, come afferma Bobbio (1984), la vita democratica si dispiega in ogni settore della società – non solo nei parlamenti, dunque, ma anche nelle aziende, nelle scuole o negli ospedali – e se in ognuno di questi contesti i mezzi informatici contribuiscono a determinare ciò che è equo e ciò che non lo è, influenzando le nostre azioni e le condizioni in cui ci veniamo a trovare, affermazioni come quella di Kathy O'Neil non appaiono fuori luogo e devono essere prese in seria considerazione.

Proprio per evitare di produrre armi di distruzione matematica, nel contesto in cui vengono progettati gli algoritmi di cui si sta discutendo, si è sviluppato un filone di ricerche molto interessante, decisamente collegato – come si cercherà di dimostrare – con la filosofia morale e politica: quello, menzionato sopra, sull'equità (*fairness*) di questi ultimi. Si tratta di scongiurare l'eventualità che le tecnologie a cui essi danno origine operino ingiustizie, automatizzandole e rendendole opache al nostro sguardo, visto che molto spesso i loro meccanismi di funzionamento vengono legalmente protetti, al fine di precluderne il pubblico scrutinio. A questo scopo, i principi di equità dei loro autori vengono formalizzati sotto forma di criteri statistici, in modo che chiunque li condivida li possa inserire all'interno dei prodotti del proprio ingegno, per mezzo di formule e funzioni facilmente replicabili.

Ciò che si sta verificando, però, è che, nei dibattiti sulla *fairness* algoritmica, le definizioni a proposito di ciò che è giusto e di ciò che non lo è sembrano oggi derivare, per la maggior parte, da una visione tipicamente liberale del concetto di giustizia, incentrata sulla tutela dei diritti individuali e sull'eguale probabilità assegnata a ognuno di vederli rispettati dagli strumenti informatici. È molto più raro imbattersi in definizioni pensate per portare avanti meccanismi di giustizia focalizzati, piuttosto, sul tentativo di realizzare perequazioni là dove le differenze iniziali tra le persone – di nascita, censo, eccetera – impediscono loro di godere appieno delle stesse opportunità che sono garantite agli altri. In questo senso, l'idea di equità su cui è improntata la maggior parte degli algoritmi attuali sembra molto affine a quella su cui si basano le democrazie liberali, nelle loro varie sfumature, mentre appare molto meno simile a concezioni della giustizia di tipo distributivo (Rawls 1971) più tipiche delle social-democrazie o di quelle che Bobbio (1995) definirebbe "democrazie egualitarie".

Poiché i mezzi informatici, come tutte le tecnologie, sono forme di vita (Winner 1983), artefatti che ci impongono di utilizzarli e di relazionarci tra di noi in un certo modo, in questo articolo si

[1] Su temi affini a quelli che qui verranno trattati, questo team ha già prodotto alcuni lavori: Beretta, Vetrò, Lepri, De Martin (2019); Vetrò, Santangelo, Beretta, De Martin (2019); Beretta, Vetrò, Lepri, De Martin (2021).

[2] Beretta, Santangelo, Lepri, Vetrò, De Martin (2019).

cercherà di esplicitare quale tipo di società e quale idea della convivenza civile traspaiono, dietro al funzionamento degli algoritmi che, per l'appunto, danno forma alla nostra vita democratica[3]. Visto che, in fondo, è la nostra visione di noi stessi e del mondo, che si imprime negli oggetti che produciamo, la speranza è che questo lavoro possa risultare utile per capire meglio chi siamo e chi vogliamo diventare, in un presente e in un futuro che, sempre più, appaiono condizionati dal nostro rapporto con le macchine.

## Le ingiustizie perpetrate dagli algoritmi

Per rendere più comprensibile il problema che qui si intende affrontare, può essere utile portare qualche esempio, a proposito delle ingiustizie che vengono perpetrate oggi, servendosi di programmi informatici. Come anticipato, il libro di Kathy O'Neil – dal titolo significativo, in inglese, di *Weapons of math destruction*, un gioco di parole che accomuna le odierne tecnologie digitali per il processamento dei big data con le armi di distruzione di massa – è pieno di casi esplicativi. Uno dei primi a essere affrontati da questa autrice – come del resto avviene spesso, in questo ambito di studi – solleva la questione, tipica del liberalismo, della tutela delle libertà individuali. Si tratta dell'utilizzo che viene fatto, negli Stati Uniti, dell'LSI-R (*Level of Service Inventory-Revised*), un modello informatizzato di cui i giudici si servono, per stabilire il rischio di recidiva criminale da parte degli imputati nei processi che presiedono, al fine di comminare le loro pene detentive. Questo strumento assegna un punteggio a ogni individuo, sulla base delle risposte che egli fornisce a domande sul suo passato, legate al numero di reati che ha commesso, ma anche alle volte in cui lui, i suoi parenti o i suoi amici hanno avuto a che fare con la legge, per motivi di rilevanza penale, oppure solo per un banale controllo. O'Neil (op. cit., 36-41) sottolinea come, in questi casi, sia molto più probabile che un alto punteggio venga totalizzato da persone povere, di quartieri malfamati, cresciute loro malgrado in contesti problematici e che magari sono state solo fermate per strada dalla polizia perché sospette, pur essendo alla fine risultate innocenti (come avviene nel novanta per cento dei casi, secondo la stessa O'Neil). Questo accade molto meno a chi proviene da famiglie benestanti, che molto probabilmente delinque per la prima volta e non ha mai fatto esperienza di storie di reati tra i propri parenti, né di controlli o perquisizioni da parte di un agente. Le differenze di nascita, di censo e di provenienza sociale entrano dunque a far parte del modello dell'LSI-R, che si dimostra doppiamente ingiusto, per questa ragione e perché tiene conto anche di gesti devianti compiuti da altri, ma non da chi deve essere giudicato.

Non va meglio, nell'ambito della ricerca di un'occupazione, un'altra tematica molto delicata, per la nostra vita democratica. Oggi, molte aziende si servono di programmi informatici, per scremare le migliaia di curricula che ricevono e questi strumenti si basano su algoritmi che tengono conto di dati come la distanza del domicilio dei candidati dalla sede delle aziende stesse, poiché questo elemento è un fattore importante, che risulta correlato con la probabilità di abbandono del proprio impiego, un'eventualità che i datori di lavoro vorrebbero scongiurare fin dall'inizio. Così facendo, però, visto che spesso gli uffici si trovano in zone centrali e ricche delle città, decidere di tenere conto di questa variabile significa discriminare chi non si può permettere un appartamento costoso nelle loro vicinanze, favorendo allo stesso tempo chi, probabilmente, ha già un'occupazione di rilievo, oppure ha il privilegio di provenire da una famiglia agiata (O'Neil, op. cit., 174).

C'è poi il problema del credito, che negli Stati Uniti viene assegnato dalle banche servendosi, ancora una volta, di strumenti informatici per la schedatura automatica delle persone. Questi mezzi tengono conto di punteggi denominati "*e-scores*", simili a quelli di cui si è scritto sopra, a proposito del rischio di recidiva dei comportamenti criminali. Per stabilire l'affidabilità di chi richiede un prestito o una carta di credito, essi utilizzano dei dati vicarianti, indiretti. Ancora una volta, se qualcuno vive in un quartiere povero, popolato di gente che fatica a rifondere i propri debiti, riceve meno soldi e a tassi di interesse più elevati, a prescindere dalle proprie effettive capacità di pagatore. Tutto ciò appare ingiusto per diverse ragioni. Innanzitutto, perché non è corretto, per questo genere di attività, non tenere conto delle caratteristiche peculiari degli individui, accomunandoli ad altri che, pur essendo loro vicini di casa, possono avere problemi molto diversi. Inoltre, visto che vivere in un certo luogo significa spesso appartenere a un gruppo etnico ben preciso, almeno negli Stati Uniti, così facendo, quest'ultimo viene discriminato (O'Neil, op. cit., 207-215). Infine, chi fatica ad accedere al credito può diventare facilmente un cattivo

[3] A questo proposito, si pensi anche al concetto di "infosfera" e al suo funzionamento, così come sono descritti in Floridi (2014).

 doi:10.53136/979125994120634

pagatore, con tutte le nefaste conseguenze che questo comporta, soprattutto in Paesi come il Nord America, in cui i dati che attestano di essere in regola con le bollette o con le altre incombenze economico-finanziarie della vita quotidiana vengono utilizzati da altri algoritmi, come per esempio, ancora una volta, quelli che decidono le assunzioni nelle aziende, visto che questa variabile viene interpretata come un indice della responsabilità e della serietà dei candidati. Quindi, se qualcuno che ha sempre rispettato le proprie scadenze non riesce a ottenere un lavoro, può facilmente contrarre debiti che non riuscirà a ripagare, dando così ragione, col passare del tempo, a chi lo ha schedato ingiustamente tra i soggetti inaffidabili, dentro a un algoritmo per l'erogazione di prestiti. O'Neil (op. cit., 41) chiama questo perverso meccanismo "ciclo di feedback negativo".

## Definizioni algoritmiche del concetto di equità

Si potrebbe proseguire ancora a lungo, mostrando che in ogni ambito della nostra società, ormai, operano programmi informatici potenzialmente ingiusti. Per porre rimedio a questo problema, si è fatto cenno agli studi sull'equità (*fairness*) degli algoritmi. Qui di seguito, sono riportate le più ricorrenti definizioni di tale equità, volte a fare in modo che gli algoritmi stessi vengano progettati per realizzare una certa idea di giustizia (Beretta et al., op. cit., 4). Innanzitutto, si possono individuare le categorie che, dall'inglese, si possono tradurre come "equità di gruppo", "individuale", "controfattuale", "basata sulla preferenza" e "legata all'inconsapevolezza". L'equità di gruppo e quella basata sulla preferenza, a loro volta, hanno diverse sfumature.

Ognuna di queste definizioni può essere descritta con criteri statistici, in modo che chiunque la condivida e intenda utilizzarla, se ne possa servire nei propri algoritmi. Per esempio, se denominiamo "C" la regola che ci consente di classificare come ad alto o a basso rischio di recidiva un soggetto che ha violato la legge; se gli attribuiamo il valore 0 quando il soggetto a cui si riferisce è a basso rischio, mentre gli attribuiamo 1 se quest'ultimo è ad alto rischio; e se, infine, chiamiamo "a" il gruppo delle persone di razza bianca e "b" quello dei neri, possiamo allora scrivere che la parità statistica (Dwork et al. 2012, 214–226), intesa come una categoria specifica dell'equità di gruppo, si può definire con la seguente condizione: Pa (C=0) = Pb (C=0) e Pa (C=1) = Pb (C=1). Per esempio, questo significa che, con un algoritmo progettato per perseguire questa forma di giustizia, una volta stabilite le logiche che determinano la pericolosità di un individuo (quante volte ha avuto problemi con la legge, in che contesto è cresciuto, eccetera) i bianchi e i neri hanno le stesse probabilità di essere classificati come soggetti a basso o ad alto rischio di recidiva. Un fatto, questo, che alla luce delle considerazioni del paragrafo precedente, sarebbe importantissimo, per evitare le discriminazioni che sono state introdotte nel sistema giudiziario statunitense, per via dell'adozione di programmi informatici come il già citato LSI-R.

Andando avanti con le definizioni dei diversi tipi di equità degli algoritmi – ma senza eccedere con la loro descrizione statistica, che in questo contesto non è rilevante e per la quale si rimanda a Beretta et al. (op. cit., 3-7) – si può dunque fare riferimento alle varie tipologie dell'equità stessa, come segue:

- parità di accuratezza (*accuracy parity*)[4]: sia i bianchi, sia i neri devono avere la medesima probabilità di essere correttamente classificati come soggetti a basso rischio di recidiva, se davvero lo sono, e di essere correttamente classificati come soggetti ad alto rischio, se sono davvero ad alto rischio;

- parità dei falsi positivi (*false positive parity*)[5]: sia i bianchi, sia i neri che sono davvero ad alto rischio di recidiva devono avere le stesse probabilità di essere scorrettamente classificati come soggetti a basso rischio (tasso di falsi positivi);

- parità di classificazione positiva (*positive rate parity*)[6]: sia i bianchi, sia i neri devono avere le medesime probabilità di essere classificati scorrettamente come soggetti a basso rischio (tasso di falsi positivi) e di essere classificati correttamente come soggetti a basso rischio (tasso di veri positivi);

[4] Dieterich, Mendoza e Brennan (2016).
[5] Corbett-Davies et al. (2017); Gajane e Pechenizkiy (2018).
[6] Hardt, Price e Srebro (2016); Zafar et al. (2017a); Binns (2018, 149-159).

- parità predittiva (*predictive parity*)[7]: sia i bianchi, sia i neri classificati come soggetti a basso rischio di recidiva devono avere la stessa probabilità di appartenere veramente a questa classe di individui;

- parità di valore predittivo (*predictive value parity*)[8]: sia i bianchi, sia i neri che sono classificati come soggetti ad alto rischio di recidiva devono avere le stesse probabilità di appartenere davvero alla classe dei soggetti ad alto rischio; allo stesso tempo, sia i bianchi, sia i neri che sono classificati come soggetti a basso rischio di recidiva devono avere le stesse probabilità di appartenere davvero alla classe dei soggetti a basso rischio;

- pari opportunità (*equal opportunity*)[9]: sia i bianchi, sia i neri che, nella realtà, sono soggetti a basso rischio di recidiva devono avere la medesima probabilità di essere scorrettamente classificati come individui ad alto rischio (falsi negativi) e di essere correttamente classificati come persone a basso rischio;

- pari soglia (*equal threshold*)[10]: sia i bianchi, sia i neri devono avere un medesimo punteggio soglia sotto il quale vengono classificati come soggetti a basso rischio e sopra il quale sono classificati come soggetti ad alto rischio;

- buona calibratura (*well-calibration*)[11]: sia i bianchi, sia i neri con i medesimi punteggi devono essere trattati allo stesso modo, per ciò che riguarda la loro classificazione come soggetti ad alto o a basso rischio di recidiva, invece di trattarli in maniera differente, a seconda del gruppo etnico a cui appartengono;

- bilanciamento per la classe positiva (*balance for positive class*)[12]: sia i bianchi, sia i neri che, nella realtà, sono soggetti a basso rischio di recidiva devono potersi aspettare di vedersi assegnato il medesimo valore del classificatore “C”, vale a dire che non deve accadere che il processo di classificazione di queste persone sia sistematicamente meno accurato nell’assegnare punteggi di alto rischio a chi appartiene a uno dei due gruppi;

- bilanciamento per la classe negativa (*balance for negative class*)[13]: sia i bianchi, sia i neri che, nella realtà, sono soggetti ad alto rischio di recidiva devono potersi aspettare di vedersi assegnato il medesimo valore del classificatore “C”, vale a dire che non deve accadere che il processo di classificazione di queste persone sia sistematicamente meno accurato nell’assegnare punteggi di basso rischio a chi appartiene a uno dei due gruppi.

Nell’equità individuale (*individual fairness*)[14], che è una categoria a parte, rispetto alle varie sfumature dell’equità di gruppo viste sin qui, due individui devono essere classificati allo stesso modo, se sono considerati simili dall’algoritmo, rispetto a qualche tipo di condizione, come per esempio quella di aver commesso lo stesso numero di reati, di essere stati fermati lo stesso numero di volte dalla polizia, eccetera. Nell’equità controfattuale (*counterfactual fairness*)[15], invece, una decisione presa da un algoritmo si può considerare giusta se verrebbe presa allo stesso modo a partire dai dati che provengono dal mondo reale e da un mondo controfattuale in cui un individuo non è contrassegnato da un attributo protetto, come per esempio la provenienza etnica. Vale a dire che, per esempio, egli verrebbe riconosciuto come un soggetto a basso o alto rischio di recidiva, a prescindere dal fatto che sia bianco, nero, ispanico, eccetera.

Le definizioni di equità degli algoritmi basate sulla preferenza (*preference based*) sono piuttosto ispirate alla teoria dei giochi e alla giusta divisione dei beni in economia:

[7] Simoiu, Corbett-Davies e Goel (2017); Chouldechova (2017).
[8] Berk, Heidari, Jabbari, Kearns e Roth (2021).
[9] Hardt, Price e Srebro (*ibidem*); Chouldechova (*ibidem*); Kusner et al. (2017).
[10] Hardt, Price e Srebro (*ibidem*); Chouldechova (*ibidem*).
[11] Kleinberg, Mullainathan, Raghavan (2017).
[12] Kleinberg, Mullainathan, Raghavan (*ibidem*).
[13] Kleinberg, Mullainathan, Raghavan (*ibidem*).
[14] Dwork et al. (*ibidem*).
[15] Kusner et al. (*ibidem*).

 doi:10.53136/979125994120634

- trattamento preferito (*preferred treatment*)[16]: in questo caso, ogni individuo preferisce essere trattato dall'algoritmo come un soggetto appartenente a una certa classe, poiché i vantaggi che ne derivano sono maggiori, rispetto al fatto di appartenere a un'altra classe. Questo significa che i neri, per esempio, possono preferire di essere riconosciuti come tali, perché questo può comportare, nei loro confronti, un trattamento privilegiato, ma considerato giusto;

- impatto preferito (*preferred impact*)[17]: un soggetto appartenente a una certa classe preferisce essere riconosciuto come tale perché l'impatto della serie di decisioni che l'algoritmo prende su di lui è migliore di quello delle decisioni che avrebbe preso, se egli fosse appartenuto a un'altra classe. Anche in questo caso, quindi, un nero può preferire di essere riconosciuto come tale, perché l'impatto delle decisioni prese su di lui si fa preferire, rispetto a quello delle decisioni che verrebbero prese, se non appartenesse a questa categoria di individui.

Infine, l'equità attraverso l'inconsapevolezza (*fairness through unawareness*)[18] è concepita in modo che un algoritmo non utilizzi un certo classificatore, per prendere le proprie decisioni, essendo, appunto, inconsapevole, a proposito del fatto che esso possa essere collegato alle persone di cui si deve occupare. In pratica, dunque, può essere preferibile che un attributo come la provenienza etnica non venga utilizzato dall'algoritmo stesso, al fine di evitare che la sua previsione, circa le possibilità di recidiva criminale, abbiano a che vedere col fatto di essere bianchi, neri o ispanici. L'algoritmo deve essere inconsapevole dell'etnia dei soggetti che deve valutare.

Questa lunga carrellata evidenzia l'ampiezza del dibattito sull'equità degli algoritmi e le sedici definizioni qui riportate non esauriscono la ricchezza delle soluzioni immaginate per rendere giusti gli strumenti informatici di cui ci serviamo nella vita quotidiana. Se riflettiamo bene sui principi di giustizia che sono alla base del pensiero dei loro progettisti, però, ci rendiamo conto che sono solo due. Il primo, predominante, è incentrato sul tentativo di fare sì che tutti abbiano, in partenza, le stesse probabilità di essere classificati in un modo o in un altro – come potenziali criminali recidivi, cattivi pagatori, bravi lavoratori, eccetera – a prescindere da alcune loro caratteristiche che li accomunano ad altre persone e di cui non sarebbe corretto tenere conto, come la provenienza etnica, il genere o il censo. Queste caratteristiche possono essere decise di volta in volta, nella "calibratura" dell'algoritmo. Il secondo principio di giustizia, invece, è l'opposto. Esso è rappresentato solo dalle due definizioni basate sulla preferenza (*preference based*), secondo le quali deve, per l'appunto, risultare preferibile appartenere a una certa classe di individui poiché, evidentemente, essendo svantaggiati, nei loro confronti verranno operate delle perequazioni, per fare in modo che siano trattati correttamente.

## Gli algoritmi e gli ideali di giustizia nei diversi sistemi democratici

Per le ragioni di cui si è scritto nei primi due paragrafi di questo lavoro, le scelte di chi progetta gli algoritmi di cui ci serviamo quotidianamente appaiono molto rilevanti, per stabilire la qualità della nostra vita democratica. Per fortuna, comunque, non mancano i tentativi di costruire tecnologie informatiche che ci appaiano "giuste". A giudicare dallo sbilanciamento tra le posizioni in campo, però, sembra assente una certa consapevolezza, a proposito del tipo di democrazia che si dimostra di voler realizzare, portando avanti le definizioni del concetto di giustizia di cui si è scritto sopra. Oppure, più semplicemente, è possibile che nei contesti in cui si sviluppa il dibattito sulla *fairness* algoritmica, la posizione prevalente sia quella liberale, per una ragione squisitamente storica e culturale.

Il motivo per cui le idee di giustizia descritte nel paragrafo precedente sembrano tipiche della democrazia liberale è presto detto. Quest'ultima, infatti, è notoriamente incentrata sulla difesa dei diritti e delle libertà fondamentali – autodeterminazione, scelta dei propri rappresentanti, pensiero, espressione, eccetera – da ogni forma di potere coercitivo (Dunn, 1979; Held, 2006), dunque anche da quello, sicuramente temibile, degli algoritmi che, come si è visto, possono contribuire a metterli in discussione. Secondo Bobbio, però, i fautori di questa forma di democrazia, negano la

[16] Zafar et al. (2017b).
[17] Zafar et al. (*ibidem*).
[18] Dwork et al. (*ibidem*); Hardt, Price e Srebro (*ibidem*).

massima dell'egualitarismo, secondo cui «tutti gli uomini debbono essere, al limite, uguali in tutto» (Bobbio, 1995: 36). Essi, piuttosto, ammettono l'eguaglianza di tutti soltanto in qualche cosa, vale a dire, appunto, nei «cosiddetti diritti fondamentali, o naturali, o, come si dice oggi, umani. Questi diritti altro non sono che le varie forme di libertà personale, civile e politica» (Bobbio, *ibidem*). Una volta garantita questa forma di uguaglianza, le altre disuguaglianze, come per esempio quelle economiche, non hanno nulla di scandaloso, ma sono viste come la conseguenza delle normali differenze tra i singoli. È in questo senso che il liberalismo viene spesso giudicato come una forma di filosofia politica di stampo individualistico (Bobbio, op. cit., 38).

I concetti appena tratteggiati sono alla base del dibattito sull'equità degli algoritmi. Una volta ottenuto che non ci siano discriminazioni, nelle probabilità di chiunque di vedersi riconosciuti, dalle macchine e da chi se ne serve, i diritti e le libertà di cui si è scritto sopra, nulla viene pensato, a proposito di ciò che dovrebbe essere compiuto a monte, per assicurarsi che tutti vengano messi nelle migliori condizioni di partenza per goderne appieno. In altre parole, riprendendo un esempio riportato in queste pagine, se ci si assicura che ogni soggetto che è effettivamente in grado di rifondere un prestito lo possa ottenere, non è necessario preoccuparsi di chi poi vi avrà accesso: uomini, donne, bianchi, neri, americani o messicani. Magari le donne, i messicani e i neri che potranno legittimamente usufruire del denaro saranno di meno, rispetto ai rappresentanti delle altre categorie di individui, ma non sarebbe giusto operare per favorirli. Lo stesso dicasi per chi deve essere valutato nella ricerca di un lavoro: a parità di capacità, nessuno deve essere avvantaggiato nel reperirlo, nemmeno se proviene da un contesto sociale in cui è più difficile procurarsi quelle competenze. Qualcosa di simile, infine, deve valere anche per chi è condannato a un certo numero di anni di carcere, in base al rischio di recidiva nel compiere un reato: chi ottiene il medesimo punteggio nel test che stabilisce la sua pericolosità, deve essere giudicato nello stesso modo di chiunque altro, anche se la probabilità che uno come lui arrivi a trovarsi in una situazione del genere è più alta.

Un altro modello liberale molto affine ai principi di giustizia che si utilizzano del dibattito sull'equità degli algoritmi è quello della democrazia agonistica (Mouffe 2009, 745-758), che si contrappone logicamente alla democrazia deliberativa (Elster 1997, 3-34). Se quest'ultima è intesa come una procedura per prendere decisioni collettive, basate sul confronto razionale tra cittadini liberi e uguali, volto alla ricerca di soluzioni per soddisfare il bene di tutti, i sostenitori della democrazia agonistica, invece, non ritengono che le persone possano raggiungere un accordo sul bene comune. Spesso, infatti, le posizioni in gioco sono basate su passioni, interessi di parte, istanze che non si vuole o che non si può contrattare. La democrazia, allora, deve consistere in un insieme di procedure che consentano a chi porta avanti le varie istanze di esprimersi e di confrontarsi, al fine di poter prevalere. Questo, naturalmente, deve avvenire sapendo che gli avversari, perso un confronto, avranno la possibilità di farsi valere nelle occasioni successive. L'uguaglianza, in questo caso, è simile a quella che intercorre, appunto, in un contesto agonistico, in cui i partecipanti alla gara hanno, in partenza, le medesime opportunità di affermarsi.

Riprendendo le definizioni di equità degli algoritmi riportate sopra, la maggior parte di esse persegue proprio l'obiettivo di rendere le regole del "gioco" democratico uguali per tutti. È in questo senso che vanno letti i principi di giustizia che danno origine alle diverse sfumature dell'equità di gruppo, individuale, controfattuale e incentrata sull'inconsapevolezza. Si tratta, in pratica, di fare in modo che chiunque abbia le medesime probabilità di accedere a un prestito, di ottenere un lavoro o di essere giudicato nella maniera più consona. In questa sorta di "partita", è ovvio che qualche individuo o qualche gruppo sarà dotato in partenza di talenti superiori, oppure muoverà da condizioni di privilegio che lo metteranno nella condizione di poter vincere più facilmente, ma questo non appare ingiusto, ai sostenitori delle suddette definizioni della *fairness* algoritmica. Così come, nel gioco del calcio professionistico attuale, chi è più ricco costruisce una squadra più forte e si trova avvantaggiato, senza che questo consenta di affermare che sia favorito, visto che gli arbitri sono imparziali e chiunque sia in grado di arricchirsi altrettanto è più che benvenuto, nella lega che organizza gli incontri del campionato, allo stesso modo gli algoritmi di cui qui si sta scrivendo sono progettati, nella stragrande maggioranza dei casi, per non tenere conto, per esempio, del fatto che potrebbe essere giusto premiare chi proviene da una famiglia povera, rispetto a chi è nato in una famiglia ricca, per aver saputo procurarsi certe competenze lavorative, nonostante le condizioni sfavorevoli da cui è partito. Questa visione della giustizia è liberale e, naturalmente, se gli strumenti informatici che utilizziamo sono coerenti con essa, non possono che rafforzare l'istituzione, nelle nostre società, di una forma di democrazia, per l'appunto, liberale.

 doi:10.53136/979125994120634

Ma, utilizzando ancora la metafora dello sport, ci sono alcuni giochi, come il golf, in cui fa parte delle regole complessive il fatto che chi si trova in una condizione iniziale di privilegio – perché più talentuoso, per esempio – si debba confrontare con gli altri venendo gravato da un "handicap", che funziona come una sorta di meccanismo perequativo: se si comincia da una posizione di vantaggio, si deve partire svantaggiati nel punteggio e completare il medesimo percorso in un minor numero di colpi. In questo modo, viene favorita la socialità, poiché chiunque può giocare con chiunque altro. Questo modello è molto simile a quello che, seguendo i ragionamenti di Bobbio (op. cit., 26-38)[19], potremmo definire della "democrazia egualitaria". Quest'ultima si basa sull'ideale di garantire un'uguaglianza di fatto, che si configura, secondo lo stesso Bobbio (op. cit., 27), come un'eguaglianza di tipo economico. In pratica, la democrazia egualitaria cerca di ridurre la distanza tra chi, in partenza, ha di meno e chi ha di più, servendosi di strumenti come le tasse progressive, i sussidi ai poveri, eccetera. Se vogliamo, dunque, il principio su cui essa poggia è l'opposto di quello della democrazia agonistica, nel senso che i suoi sostenitori riconoscono le differenze tra le persone, più che ciò che le accomuna, e cercano di favorire chi è svantaggiato. Una volta messi tutti nelle medesime condizioni per giocare, allora diventa importante che l'arbitro sia imparziale e che venga garantita la normale alternanza dei vincitori.

Forse perché, come sostiene Bobbio, l'ugualitarismo parte da ragionamenti di tipo economico, i modelli di equità degli algoritmi che lo utilizzano sono quelli che provengono da questo ambito di studi, incentrati sulla preferenza di trattamento e di impatto. Come anticipato, l'idea è che debba essere preferibile fare parte di una certa classe di individui, per via dei vantaggi che questo comporta. Dunque, per fare un esempio, a parità di competenze lavorative, dovrebbe risultare più vantaggioso essere donna, perché le tecnologie informatiche che selezionano i curricula di chi cerca un'occupazione prenderanno in considerazione prima le candidature femminili. Questo, magari, perché chi deve progettare questo genere di strumenti digitali ritiene che, generalmente, il contesto sociale del mercato del lavoro sia, per qualche ragione, ingiusto nei confronti delle donne, svantaggiandole in partenza. Lo stesso dicasi per la concessione dei prestiti, che secondo questi modelli dovrebbero essere assegnati prima a quella tipologia di persone che di solito ne riceve di meno, e per la decisione su quanti anni di pena comminare ai soggetti potenzialmente recidivi, che dovrebbe considerare come attenuante il fatto che essi provengano da contesti difficili.

## Quali algoritmi per le nostre democrazie

Come è evidente dagli esempi appena riportati, la scelta su come tarare gli algoritmi di cui ci serviamo nella vita di tutti i giorni non è facile. Sicuramente, agli occhi di chi concepisce l'equità da un punto di vista liberale può apparire ingiusto svantaggiare qualcuno che ha le medesime competenze professionali di qualcun altro, solo perché il suo potenziale datore di lavoro si serve di un mezzo informatico progettato ritenendo che, per chi proviene da una famiglia ricca, perdere una certa opportunità di lavoro sia meno grave, rispetto al caso di chi è povero. Inoltre, dividere in maniera così netta, come si è fatto in queste pagine, il campo dei liberali da quello egualitario può risultare fuorviante: il principio della giustizia distributiva di Rawls, per esempio, pur se di matrice liberale, si basa sull'idea che «le ineguaglianze economiche e sociali, come quelle di ricchezza e di potere, sono giuste soltanto se producono benefici compensativi per ciascuno, e in particolare per i membri meno avvantaggiati della società» (Rawls, op. cit. [2008], 35).

Ciò che però è chiaro, anche solo per una considerazione numerica, è che le definizioni dell'equità degli algoritmi che oggi vengono dibattute e danno forma alle nostre tecnologie digitali, tendono a trascurare i principi dell'egualitarismo, con tutte le conseguenze che questo può comportare, per la nostra vita democratica. È necessario esserne consapevoli e decidere che si vuole andare in questa direzione, ma non perché pochi studiosi, nelle loro discussioni un po' criptiche, di difficile accesso per via della natura tecnica dei loro discorsi, ritengono giuste queste posizioni, bensì avviando una riflessione collettiva. L'obiettivo di questo articolo è proprio quello di chiarire i termini della questione, rendendoli accessibili a tutti. In questo modo, si intende perseguire l'ideale di un'altra forma di democrazia, quella repubblicana (Bozdag e Van Den Hoven 2015), notoriamente incentrata sull'idea che tutti i cittadini devono essere liberi dall'esercizio

[19] Bobbio individua storicamente queste forme di democrazia in quelle incarnate dagli Stati interventisti e dirigisti, contrapposti a quelli solo garantisti, che limitano al massimo la loro ingerenza sul funzionamento dell'economia.

arbitrario del potere da parte di qualcuno e che, per questo, devono essere in grado di controllare l'attività di chi li governa, avendo modo di contestarla, se non la condividono. Sono molto importanti, a questo scopo, la trasparenza e la pubblicità delle azioni dei governanti. Queste, però, possono risultare difficili da perseguire, se chi prende le decisioni più importanti per la nostra vita quotidiana si avvale di algoritmi che, con i loro automatismi, formalizzati in un linguaggio tecnico non conosciuto da tutti e spesso protetti dal pubblico scrutinio, ci appaiono oscuri. Come sostiene Pasquale (2015), c'è la possibilità che queste tecnologie ci conducano verso una "società della scatola nera"[20], su cui è necessario gettare luce, per evitare trovarsi, senza nemmeno rendersene conto, in un mondo dominato dall'informatica, che all'improvviso ci può sembrare ingiusto o, peggio ancora, che ci sembra giusto, anche se non lo è.

## Bibliografia

Beretta, E., A. Vetrò, B. Lepri B. e J.C. De Martin. "Ethical and socially aware data labels, in Information management and big data." In *Annual international symposium on information management and big data*, 320-327. Cham: Springer, 2019.

Beretta, E., A. Santangelo, B. Lepri, A. Vetrò e J.C. De Martin. "The invisible power of fairness. How machine learning shapes democracy." In *Proceedings of the 32nd Canadian Conference on Artificial Intelligence*, 238-250. Cham: Springer, 2019.

Beretta, E., A. Vetrò, B. Lepri B. e J.C. De Martin. "Detecting discriminatory risk through data annotation based on Bayesian inferences." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 794-804. New York: Association for Computing Machinery, 2021.

Berk, R., H. Heidari, S. Jabbari, M. Kearns e A. Roth. "Fairness in criminal justice risk assessments: the state of the art." *Sociological Methods & Research* 50.1 (2021): 3-44.

Binns, R. "Fairness in machine learning: lessons from political philosophy." *Proceedings of Machine Learning Research* 81 (2018): 149-159.

Bobbio, N. *Il futuro della democrazia*. Torino: Einaudi, 1984.

Bobbio, N. *Eguaglianza e libertà*. Torino: Einaudi, 1995.

Bozdag, E. e J. Van Den Hoven. "Breaking the filter bubble: democracy and design." *Ethics and Information Technology* 17 (2015): 249-265.

Chouldechova, A. "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments." *Big Data* 5 (2017): 153-163.

Corbett-Davies S., E. Pierson, A. Feller, S. Goel e A. Huq. "Algorithmic decision making and the cost of fairness." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797-806. New York: Association for Computing Machinery, 2017.

Dieterich W., C. Mendoza C. e T. Brennan. *Compas risk scales: demonstrating accuracy equity and predictive parity*, Tech. rep., Northpointe Inc., 2016.

Dunn, J. *Western political theory in the face of the future. Vol. 3*. Cambridge: Cambridge University Press, 1979.

[20] Il titolo del suo libro è, appunto, *The black box society*.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold e R. Zemeln. “Fairness through awareness.” In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-226, 2012.

Elster, J. “The market and the forum: Three varieties of political theory.” In Bohman J. e Rehg W., eds, *Deliberative democracy: Essays on reason and politics*, 3-34. Cambridge: The MIT Press, 1997.

Floridi, L. *The fourth revolution. How the infosphere is reshaping human reality*. Oxford: Oxford University Press, 2014.

Gajane, P. e M. Pechenizkiy. “On formalizing fairness in prediction with machine learning”. *arXiv* 1710.03184 (2018).

Hardt, M., E. Price e N. Srebro. “Equality of opportunity in supervised learning.” In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, December 2016*, 3323–3331. New York: Association for Computing Machinery, 2016.

Held, D. *Models of democracy*. Palo Alto: Stanford University Press, 2006.

Kleinberg, J., S. Mullainathan e M. Raghavan. “Inherent trade-offs in the fair determination of risk scores.” In *Proceedings of Innovations in Theoretical Computer Science*, 2017.

Kusner, M.J., J.R. Loftus, C. Russell e R. Silva. “Counterfactual fairness”. *Advances in Neural Information Processing Systems* 30 (2017).

Mouffe, C. “Deliberative democracy or agonistic pluralism?” *Social Research* 66 (1999): 745-758.

Mouffe, C. *The democratic paradox*. London: Verso, 2009.

O’Neil, K. *Weapons of math destruction. How big data increases inequality and threatens democracy*, Penguin, New York, 2016; trad. it. *Armi di distruzione matematica. Come i big data aumentano la disuguaglianza e minacciano la democrazia* Milano: Bompiani, 2017.

Pasquale, F. *The Black Box Society. The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press, 2015.

Rawls, J. *A theory of justice*. Cambridge: Harvard University Press, 1971; trad. It. *Una teoria della giustizia*. Milano: Feltrinelli, 2008.

Simoiu, C., S. Corbett-Davies e S. Goel. “The problem of infra-marginality in outcome tests for discrimination.” Annals of Applied Statistics (2017): 1193-1216.

Vetrò, A., A. Santangelo, E. Beretta e J.C. De Martin. “AI: from rational agents to socially responsible agents.” In *Digital policy, regulation and governance*, 291-304. London: Emerald Group Publishing, 2019.

Winner, L. *Technologies as forms of life*. In Cohen R. S. e Wartofsky M. W., eds, *Epistemology, Methodology and the Social Sciences*. Amsterdam, Kluwer Academic Publishers, 1983.

Zafar, M.B., I. Valera, M.G. Rodriguez e K.P. Gummadi. “Fairness beyond disparate treatment and disparate impact: learning classification without disparate mistreatment.” In *Proceedings of the 26th International Conference on World Wide Web*, 2017a.

Zafar, M.B., I. Valera, M.G. Rodriguez, K.P. Gummadi e A. Weller. “From parity to preference-based notions of fairness in classification.” In *Proceedings of the 31stConference on Neural Information Processing Systems* (NIPS 2017), 2017b.