T

Benedetta Giovanola, Simona Tiribelli

Equità e decisioni algoritmiche

Introduzione

L'enorme sviluppo, negli ultimi decenni, dei sistemi di intelligenza artificiale (IA) e, nello specifico, di algoritmi di *machine learning* (ML) e *deep learning* (DL) ha dato origine a un crescente dibattito nell'ambito dell'etica degli algoritmi[1], volto a mettere in luce, in particolare, i potenziali rischi insiti nei cosiddetti processi decisionali automatizzati, basati – appunto – su algoritmi di ML e DL. Le capacità probabilistiche degli algoritmi di ML e DL nel processare enormi quantità di dati e scoprire modelli e correlazioni preziose hanno comportato un loro utilizzo esteso, dando impulso a un inedito fenomeno di delega a sistemi algoritmici di compiti, scelte e decisioni, prima esclusivamente umani, in ambiti fondamentali, quali l'educazione, la medicina, la giustizia e la difesa nazionale. Tuttavia, questo iniziale entusiasmo, dovuto principalmente alla presunta neutralità, accuratezza e affidabilità dei modelli algoritmici, ha lasciato presto il posto a una serie di critiche sul loro uso nei processi decisionali, poiché gli algoritmi si sono spesso dimostrati difettosi e, soprattutto, iniqui nei risultati generati, piuttosto che esatti e imparziali.

Queste scoperte hanno messo in luce la centralità dell'equità nei processi decisionali algoritmici, stimolando numerose iniziative sul tema, nonché una grande produzione scientifica, di matrice sia filosofica, sia tecni-

[1] A. Tsamados *et al.*, *The ethics of algorithms: Key problems and solutions*, in «AI & Society», 37 (2022), pp. 215-230. Sul tema dell'etica degli algoritmi e, più in generale, dell'intelligenza artificiale, si vedano: A. Fabris, *Etica per le tecnologie dell'informazione e della comunicazione*, Carocci, Roma 2018; L. Floridi, *Etica dell'intelligenza artificiale*, Raffaello Cortina, Milano 2022; P. Benanti, *Human in the loop*, Mondadori, Milano 2022.

 DOI: 10.4454/teoria.v42i2.160

ca[2]. Tuttavia, nonostante l'equità sia riconosciuta come un tema centrale sia nelle riflessioni nell'ambito dell'etica degli algoritmi, sia negli studi più tecnici, il concetto di equità implicato dai processi decisionali algoritmici non è stato ancora indagato in modo adeguato e appare, anzi, piuttosto vago e opaco.

Lo scopo del nostro articolo è colmare questa lacuna, integrando la riflessione sull'equità condotta nell'ambito degli studi sull'etica degli algoritmi e della letteratura tecnica con una riflessione propriamente filosofico-morale sul tema. Nello specifico, mostreremo che un'indagine filosofico-morale sul concetto di equità è necessaria sia per chiarire il significato dell'equità nei processi decisionali algoritmici, sia per individuare i criteri che dovrebbero orientarne la progettazione.

L'articolo è suddiviso in tre paragrafi. Nel primo paragrafo ricostruiamo lo stato dell'arte del dibattito sull'equità nei processi decisionali algoritmici e mostriamo che questo presuppone un concetto di "equità negativa", ovvero di equità come assenza di discriminazione: mostriamo anche che la discriminazione, a sua volta, viene intesa come semplice assenza di distorsioni e pregiudizi (*bias*) nei set di dati con cui processi decisionali algoritmici vengono allenati; sosteniamo, infine, che il concetto di "equità negativa" non è adeguato e argomentiamo la necessità di elaborare un concetto "equità positiva" capace di andare oltre la sola non discriminazione e la considerazione dei *bias*. Nel secondo paragrafo sviluppiamo il concetto di "equità positiva" grazie agli strumenti offerti dalla riflessione filosofico-morale: in particolare mostriamo che equità e non discriminazione non coincidono e individuiamo le dimensioni e componenti costitutive dell'equità. Nella nostra rielaborazione concettuale dell'equità prendiamo le mosse da una originale riflessione sul concetto di rispetto, che riconosce il ruolo del rispetto per le persone in quanto persone, ma include anche il rispetto per le persone in quanto individui particolari. Infine, nel terzo paragrafo mostriamo come la nostra indagine filosofico-morale e la nostra rielaborazione del concetto di equità ci permettano di individuare i criteri che dovrebbero orientare il *design* degli algoritmi, rendendo i processi decisionali realmente più equi.

[2] Si vedano, tra gli altri, C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown, New York 2016; D. Shin, Y.J. Park, *Role of fairness, accountability, and transparency in algorithmic affordance*, in «Computers in Human Behavior», 98 (2019), pp. 277-284.

1. *"Equità negativa" e decisioni algoritmiche: non discriminazione e assenza di* bias

L'equità è un tema centrale nell'etica degli algoritmi e, in particolare, nelle riflessioni sui processi decisionali algoritmici[3]. Oltre a essere riconosciuta come un valore fondamentale da integrare nel *design* etico – o *value sensitive* – delle tecnologie basate su sistemi algoritmici[4], l'equità è l'unico tra i principi etici adottati nell'ambito dell'IA a essere riconosciuto in tutti i principali documenti che offrono linee giuda a livello globale per orientare in modo affidabile lo sviluppo degli algoritmi di ML e DL[5].

Questa crescente attenzione per l'equità si spiega anche in risposta a una serie di esiti iniqui prodotti dai processi decisionali algoritmici in vari ambiti, che vanno dalla pubblicità e dal marketing all'accesso al mercato del lavoro e al credito, alla giustizia e alla sanità[6]. Tra gli esempi più rilevanti si possono citare i *bias* di tipo etnico rilevati nell'algoritmo decisionale di COMPAS, un sistema di valutazione del rischio utilizzato nella *criminal justice* statunitense per prevedere il tasso di recidiva degli indagati, denunciato nel 2016 dall'agenzia giornalistica ProPublica perché profondamente discriminante nei confronti degli afroamericani. Un caso simile ha coinvolto, nel 2018, il sistema algoritmico decisionale alla base del software di assunzione utilizzato dalla *big tech* Amazon, rivelatosi discriminante nei confronti dei candidati in base al loro genere, a causa di *bias* presenti nei dati di formazione del sistema. Infine, nello stesso anno, due studiose statunitensi, Timnit Gebru e Joy Buolamwini, hanno denunciato una combinazione problematica di *bias* di genere ed etnici negli algoritmi alla base di alcuni dei software più utilizzati per l'identificazione delle persone tramite riconoscimento facciale. Distorsioni simili sono state scoperte,

[3] A. Tsamados *et alia*, *art. cit.*

[4] S. Umbrello, I. van de Poel, *Mapping value sensitive design onto AI for social good principles*, in «AI Ethics», 1, 3 (2021), pp. 1-14.

[5] Jobin A. *et al*., *Artificial intelligence: the global landscape of ethics guidelines*, in «Nature Machine Intelligence», 1 (2019), pp. 389-399.

[6] C. O'Neil, *op. cit*.; R. Benjamin, *Race after technology: abolitionist tools for the new Jim code*, Polity, Medford 2019. V. Eubanks, *Automating inequality. How high-tech tools profile, police, and punish the poor*, St Martin's Publishing, New York 2018. S.U. Noble, *Algorithms of oppression: how search engines reinforce racism*, New York University Press, New York 2019; B. Giovanola, S. Tiribelli, *Beyond Bias and Discrimination. Redefining the AI Ethics Principle of Fairness in Healthcare Machine-Learning Algorithms*, in «AI & Society», Special Issue "AI4People", (2022), https://doi.org/10.1007/s00146-022-01455-6.

più tardi, anche nel funzionamento di vari sistemi algoritmici utilizzati in ambito sanitario per compiti quali l'identificazione di patologie e l'attribuzione di priorità nell'ordine di accesso dei pazienti a programmi di cura speciali o agevolati[7].

Questi eventi hanno condotto numerosi ricercatori, tecnologi e attivisti a denunciare pubblicamente l'utilizzo dei sistemi basati su processi decisionali algoritmici, accusati di essere strumenti di ingiustizia e, nello specifico, di "discriminazione algoritmica"[8] a causa della loro propensione sia a replicare sia a esacerbare in modo invisibile e silenzioso discriminazioni e pregiudizi, incorporati nella forma di distorsioni o *bias* nei set di dati di formazione e allenamento dei modelli algoritmici.

Ad alimentare ulteriormente le preoccupazioni e le critiche sull'uso dei processi decisionali algoritmici è anche la difficoltà di rintracciare le distorsioni o i *bias* menzionati, soprattutto a causa del basso livello di scrutabilità e/o di intelligibilità degli stessi algoritmi. Alla base di questa difficoltà vi sono due fattori principali: in primo luogo, i modelli algoritmici più utilizzati sono proprietari e, dunque, coperti da segreto commerciale; in secondo luogo, questi modelli spesso includono nel loro funzionamento anche algoritmi di DL, ovvero complesse architetture di reti neurali che, pur consentendo una maggiore capacità di predizione, producono risultati non basati su nessi causali, bensì su correlazioni indotte dai dati, che spesso rendono i processi decisionali algoritmici delle vere e proprie "scatole nere"[9], ovvero modelli opachi e non esplicabili.

La necessità di contrastare e prevenire gli esiti discriminanti prodotti dall'impiego degli algoritmi a fini decisionali ha generato una crescente attenzione sul tema dell'equità, rendendo il *design* e lo sviluppo di processi decisionali algoritmici equi una delle sfide più urgenti e importanti nel settore dell'IA[10]. Tuttavia, nonostante l'importanza dell'equità sia ormai ampiamente riconosciuta[11], il concetto di equità nei processi decisionali

[7] Z. Obermeyer *et al.*, *Dissecting racial bias in an algorithm used to manage the health of populations*, in «Science», 366 (2018), pp. 447-453.

[8] O'Neil, *op. cit.*

[9] F. Pasquale, *The black box society: the secret algorithms that control money and information*, Harvard University Press, Cambridge 2015.

[10] D. Shin, Y.J. Park, *art. cit.*

[11] J. Kleinberg *et al.*, *Human decisions and machine predictions*, in «Quarterly Journal of Economics», 133, 1 (2018), pp. 237-293; R. Overdorf *et al.*, *Questioning the assumptions behind fairness solutions*, in «NeurIPS», (2018), pp. 1-7; P. Wong, *Democratizing algorithmic fairness*, in «Philosophy &Technology», 33, 2 (2020), pp. 225-244.

algoritmici non è stato ancora indagato in modo soddisfacente e appare, anzi, piuttosto vago e opaco[12].

Analizzando le riflessioni condotte nell'ambito dell'etica degli algoritmi e degli studi tecnici sull'IA, emerge un concetto di equità come assenza di discriminazione e un'accezione di quest'ultima come assenza di *bias*. Considerando le quattro definizioni di equità principalmente adottate nella letteratura sugli algoritmi[13], l'equità è definita tramite metodi di misura matematica come: l'*anti-classificazione*, secondo cui l'equità nei processi algoritmici si ottiene evitando l'uso di termini che si riferiscono a categorie protette (quali l'etnia, la religione e il genere); la *parità di classificazione*, secondo cui un processo decisionale algoritmico è equo se le misure della sua *performance* predittiva sono uguali tra gruppi protetti; la *calibrazione*, secondo la quale l'equità di un sistema decisionale algoritmico è data dalla misura di quanto un algoritmo è calibrato tra gruppi protetti; la *disparità statistica*, secondo cui l'equità di un modello corrisponde a una stima di probabilità media uguale nei risultati per tutti i membri dei gruppi protetti.

Queste definizioni, oltre a rivelarsi incompatibili tra di loro[14], tendono a fornire metriche per misurare l'equità basate sulla considerazione del trattamento da parte del sistema decisionale algoritmico di gruppi o categorie protetti, facendo coincidere, dunque, l'idea di un sistema decisionale algoritmico equo con quella di un sistema non discriminante[15]. La discriminazione algoritmica prodotta dai sistemi algoritmici è a sua volta principalmente ricondotta alla presenza di distorsioni e, nello specifico, a due tipologie di *bias*: i *bias di automazione*, che si verificano quando i processi decisionali algoritmici riproducono su larga scala pregiudizi sociali e culturali incorporati nei dati di formazione del sistema[16]; e i *bias by proxy*, che si verificano quando determinate informazioni inferibili dai dati fungono da *proxy* per l'identificazione di caratteristiche riconducibili a gruppi protetti. Di conseguenza, l'idea diffusa è che un processo decisionale algorit-

[12] N. Saxena *et al.*, *How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness*, in «AI Ethics and Society», (2019), pp. 1-12.

[13] Wong, *art. cit.*

[14] Si veda la critica di Kleinberg *et al.*, *art. cit.*: gli autori sottolineano come la rimozione di alcuni termini che si riferiscono a categorie protette non è sempre auspicabile; si pensi, ad esempio, al settore della salute, dove fattori come il genere e l'etnia svolgono un ruolo cruciale per la predizione di determinate patologie e, di conseguenza, per il *design* di sistemi decisionali algoritmici accurati, oltre che equi.

[15] S. Barocas, A.D. Selbst, *Big data's disparate impact*, in «California Law Review», 104, 671 (2016), pp. 671-732.

[16] Noble, *op. cit.*; Benjamin, *op. cit.*

mico equo, ovvero non discriminante, sia un sistema esente da distorsioni o *bias*[17].

Questa concettualizzazione dell'equità emerge anche nei principali quadri etici elaborati a livello globale per lo sviluppo dei sistemi basati sull'IA, come mostrato da una recente analisi condotta su 84 documenti[18]. Un esempio emblematico sono gli "Orientamenti etici per un'IA affidabile" pubblicati nel 2018 dalla Commissione Europea, che costituiscono uno dei documenti cardine nel panorama internazionale per la progettazione etica dei sistemi algoritmici. Qui il principio di equità è uno tra i cinque principi chiave per il *design* etico dei sistemi basati su algoritmi e prescrive l'impegno ad assicurare l'eliminazione di *bias* che acuiscono forme di discriminazione sociale.

In sintesi, il concetto di equità che emerge dalla letteratura sui processi decisionali algoritmici e dai principali quadri etici di riferimento per l'IA può essere definito come "equità negativa"[19], poiché l'equità è concepita come assenza di discriminazione, e quest'ultima è a sua volta definita come assenza di *bias*. In altre parole, un processo è equo se i suoi risultati ed effetti non producono discriminazione nel trattamento di individui appartenenti a categorie protette e questo è possibile se vengono eliminati i *bias* del sistema algoritmico.

Tuttavia, nonostante la necessità di mitigare i *bias* nei processi decisionali algoritmici sia innegabile, dobbiamo chiederci se la rimozione di *bias*, da sola, possa garantire sistemi algoritmici equi. Per rispondere a questa domanda, dobbiamo chiederci se il concetto di "equità negativa" emergente del dibattito sul tema sia adeguato oppure se non sia necessario sviluppare un'indagine più approfondita su un concetto complesso come quello di equità[20], chiarendone i presupposti teorici[21].

Nel paragrafo seguente svilupperemo questa indagine, offrendo una rielaborazione concettuale dell'equità. Grazie alla riflessione filosofico-morale, proporremo un concetto di "equità positiva" capace di andare oltre la non discriminazione e la considerazione dei *bias* e ne individueremo le

[17] Benjamin, *op. cit.*; Noble, *op. cit.*; O' Neil, *op. cit*.

[18] Jobin *et al.*, *art. cit.*, p. 8.

[19] In questa definizione prendiamo spunto, come è evidente, dalla nota distinzione tra libertà negativa e libertà positiva introdotta da I. Berlin (cfr. I. Berlin, *Quattro saggi sulla libertà*, Feltrinelli, Milano 1989).

[20] A. Rajkomar *et al.*, *Ensuring fairness in machine learning to advance health equity*, in «Annals of Internal Medicine», 16 (2018), pp. 866-872.

[21] R. Overdorf *et al.*, *art. cit.*

dimensioni e componenti costitutive, finora trascurate nel dibattito sull'equità nei processi decisionali algoritmici.

2. *"Equità positiva": equa eguaglianza di opportunità, diritto alla giustificazione, equa eguaglianza di relazione*

Per sviluppare la nostra rielaborazione concettuale dell'equità chiariremo, in primo luogo, la differenza tra equità e (non) discriminazione; argomenteremo poi l'importanza dell'equità positiva, soffermandoci sulle sue dimensioni e componenti costitutive, e mostrando che essa consente di rispettare le persone sia in quanto persone, sia in quanto individui particolari.

Il rapporto tra equità e discriminazione è stato ampiamente riconosciuto dalla riflessione filosofica, specialmente nel contesto delle teorie della giustizia. L'argomento ricorrente è che la discriminazione ostacola l'equità, poiché si fonda sul mancato riconoscimento dell'eguaglianza morale delle persone[22] e implica che alcune di esse vengano trattate in modo crudele o umiliante[23], dunque profondamente irrispettoso. Tuttavia, grazie agli strumenti offerti dalla riflessione filosofico-morale, possiamo evidenziare che l'equità, pur essendo strettamente collegata alla (non) discriminazione, non coincide con questa e include pure altre dimensioni e componenti costitutive[24].

Un primo elemento costitutivo dell'equità è l'*equa eguaglianza di opportunità*, argomentata in modo efficace nelle teorie della giustizia di matrice liberal-egualitaria, a partire da quella Rawlsiana[25]. L'equa eguaglianza di opportunità regola la distribuzione dei benefici e degli oneri della cooperazione sociale e la gestione delle diseguaglianze socio-economiche in modo non solo da prevenire la discriminazione, ma da creare anche le condizioni che consentano l'esercizio dell'agency individuale e l'auto-realizzazione.

[22] Cfr. S. Scheffler, *what is egalitarianism?*, in «Philosophy and public affairs», 31 (2003), n. 1, pp. 5-39; E. Anderson, *What is the point of equality?*, in «Ethics», 109 (1999), pp. 289-337.

[23] A. Sangiovanni, *Humanity without dignity. moral equality, respect, and human rights*, Harvard University Press, Cambridge (MA) 2017.

[24] Per un ulteriore approfondimento di questi temi cfr. B. Giovanola, S. Tiribelli, *Weapons of moral construction? On the value of fairness in algorithmic decision-making*, in «Ethics and Information Technology», 24 (2022), n. 3. 10.1007/s10676-022-09622-5

[25] J. Rawls, *Una teoria della giustizia*, Feltrinelli, Milano 2004. Rawls, come è noto, oltre al principio di equa eguaglianza di opportunità, individua il principio di differenza e il principio di eguale libertà. Non potendoci soffermare su tali principi, in questa sede ne sottolineiamo comunque la coerenza con la nostra discussione sull'equità.

L'equa eguaglianza di opportunità mostra una dimensione distributiva della giustizia, fondata sul bisogno di rispettare le persone sia come destinatarie della distribuzione, sia come soggetti capaci di agency morale.

Un secondo elemento costitutivo dell'equità è il *diritto alla giustificazione*, rivendicato con forza da studiosi come Rainer Forst. Il diritto alla giustificazione esprime la pretesa etica che non vi siano relazioni e strutture intersoggettive "che non possono essere adeguatamente giustificate nei confronti di coloro che vi sono coinvolti"[26]; esso esprime, dunque, un principio di giustificazione reciproca, fondato sull'importanza di rispettare ogni persona in quanto persona, ovvero in quanto soggetto capace di (e titolato a) offrire e richiedere giustificazione. Di conseguenza, la questione del diritto alla giustificazione è anche una questione di potere, ovvero la questione di chi decide cosa[27]. Il diritto alla giustificazione fa emergere una dimensione socio-relazionale dell'equità, la quale mostra sia l'importanza del riconoscimento reciproco, sia la necessità di mitigare le asimmetrie di potere a livello decisionale.

Sia l'equa eguaglianza di opportunità, sia il diritto alla giustificazione sono componenti costitutive dell'equità e ne mostrano, rispettivamente, la dimensione distributiva e la dimensione socio-relazionale[28]. L'individuazione di queste componenti consente di superare l'accezione "negativa" dell'equità come assenza di discriminazione e di mostrare, piuttosto, l'importanza di un'accezione "positiva" dell'equità, fondata sul riconoscimento dell'eguaglianza e del valore morale delle persone e capace di promuovere attivamente l'eguale *rispetto per le persone in quanto persone.*

Tuttavia, il valore morale delle persone non riguarda solo la loro (astratta) capacità di agency morale. Come è stato opportunamente argomentato, esso richiede anche di tenere in considerazione le persone in quanto "individui particolari" [29], che concretamente esercitano la loro agency in modi differenti. Riconoscere questo significa superare l'attenzione esclusiva sull'eguale rispetto per le persone in quanto persone e considerare anche il *rispetto per le persone in quanto individui particolari.*

[26] R. Forst, *Two pictures of justice*, in *Justice, democracy and the right to justification. Rainer Forst in dialogue*, Bloomsbury. London 2014, pp. 3-26, qui p. 6.

[27] *Ivi*, p. 24.

[28] Per un'argomentazione più dettagliata della compresenza di dimensione distributiva e dimensione socio-relazionale dell'equità e, più in generale, della giustizia sociale, cfr. B. Giovanola, *Giustizia sociale. Rispetto ed eguaglianza nelle società diseguali*, il Mulino, Bologna 2018.

[29] R. Noggle, *Kantian respect and particular persons*, in «Canadian Journal of Philosophy», 29 (1999), pp. 449-477.

Un passo in questa direzione può essere rintracciato nella nota distinzione, introdotta da Darwall, tra rispetto come riconoscimento (*recognition respect*) e rispetto come stima (*appraisal respect*): se il primo è fondato sul riconoscimento dell'eguaglianza morale delle persone in quanto persone, il secondo consiste in un "apprezzamento positivo" delle persone "in quanto impegnate in uno specifico compito" o dotate di "caratteristiche che si ritiene manifestino la loro eccellenza"[30]. Valorizzando questa distinzione, possiamo affermare che rispettare realmente le persone richiede *anche* di considerare i "progetti fondativi" che danno senso alla loro vita[31] e li rendono dei sé concreti, piuttosto che astratti[32]. Il riferimento al rispetto per gli individui particolari esprime proprio la necessità di considerare le persone in quanto aventi specifici obiettivi, affiliazioni, valori, impegni, e non solo di trattarle come se fossero "opache", evitando di guardare dentro di loro, e astenendoci "dal guardare oltre l'esteriorità" che esse "presentano a noi in quanto agenti morali"[33]. Comporta, insomma, la necessità di considerare le persone non sono come astratti eguali morali, ma anche come individui che esercitano concretamente la propria agency in modi diversi[34].

Valorizzare il rispetto per le persone in quanto individui particolari consente di individuare un terzo elemento costitutivo dell'equità, che possiamo chiamare *equa eguaglianza di relazione*. L'equa eguaglianza di relazione mette in luce l'importanza delle relazioni nel processo di formazione di obiettivi, affiliazioni, valori, impegni degli individui particolari. Le relazioni, infatti, sono alla base delle nostre affiliazioni e impegni condivisi[35], e questi a loro volta sono centrali per definire i nostri valori e obiettivi[36]. Al contempo va rilevavo che molte relazioni, oggi, sono sempre più mediate

[30] Darwall, *Two kinds of respect*, in «Ethics», 88 (1977), pp. 36-49, qui pp. 38-39, traduzione nostra.

[31] B. Williams, *Persons, character and morality*, in *Moral luck: philosophical papers 1973-1980*, Cambridge University Press, Cambridge 1981, pp. 1-19.

[32] Cfr. M. Sandel, *The procedural republic and the unencumbered self*, in «Political theory», 12 (1984), pp. 81-96.

[33] I. Carter, *Il rispetto e le basi dell'eguaglianza*, in I. Carter, A.E. Galeotti, V. Ottonelli (a cura di), *Eguale rispetto*, Feltrinelli, Milano 2008, pp. 54-77, qui p. 66.

[34] In questa direzione cfr. L. Valentini, *Respect for persons and the moral force of socially constructed norms*, in «Noûs», (2019), pp. 1-24. https://doi.org/10.1111/nous.12319.

[35] M. Gilbert, *A theory of political obligation: membership, commitment, and the bonds of society*, Oxford University Press, New York 2006.

[36] C. Calhoun, *What good is commitment?*, in «Ethics», 119 (2009), n. 4, pp. 613-641. https://doi.org/10.1086/605564.

da tecnologie basate su sistemi algoritmici[37]. Tuttavia queste tecnologie creano spesso bolle[38] o eco camere[39]: basti pensare, a titolo esemplificativo, alle tecniche di personalizzazione alla base dei social media, che spesso tendono a restringere anziché espandere le nostre relazioni, spingendoci verso coloro che sono più simili a noi, limitando così le nostre alternative di scelta e favorendo la creazione di gruppi chiusi, con possibili rischi in termini di estremizzazione, polarizzazione e conflitto[40]. Inoltre, come mostrato dai più recenti studi sulle distorsioni cognitive ed emotive alla base delle nostre motivazioni e convinzioni, percependosi come membri di gruppi chiusi in conflitto con altri gruppi chiusi, gli individui tendono a erodere inconsapevolmente la loro capacità di percepirsi come parte di un progetto condiviso e di avere obiettivi comuni[41].

Questi esempi mostrano che le relazioni possono estremizzare i valori e gli obiettivi degli individui particolari e possono restringere, anziché ampliare, le loro affiliazioni e i loro impegni condivisi. L'equa eguaglianza di relazione è richiesta proprio affinché ciò non avvenga e consiste nel rivendicare con forza l'importanza, per gli individui, di relazioni *genuine*, ovvero radicate in una reale libertà di scelta, che esprime la nostra agency[42] e autonomia. Solo a partire da una equa eguaglianza di relazione, gli individui particolari possono riconoscersi reciprocamente come eguali eppure diversi e rispettarsi in virtù dei reciproci obiettivi e affiliazioni.

La disamina finora condotta ci ha consentito di individuare le dimensioni e componenti costitutive dell'equità, e di fare emergere un'accezione positiva, non solo negativa, di questo concetto, finora ignorata nel dibattito sui processi decisionali algoritmici. Nel paragrafo seguente mostreremo come la nostra rielaborazione del concetto di equità ci permetta di indivi-

[37] B. Giovanola, *Justice, emotions, socially disruptive technologies*, in «Critical review of international social and political philosophy», (2021), pp. 1-16. https://doi.org/10.1080/13698230.2021.18932552021

[38] E. Pariser, *The filter bubble*, Penguin, London 2011.

[39] C. Sunstein, *Democracy and the internet*, in J. van den Hoven, J. Weckert (eds.), *Information Technology and moral philosophy*, Cambridge University Press, Cambridge 2008, pp. 93-110.

[40] Parsell, *Pernicious virtual communities: identity, polarisation and the web 2.0*, in «Ethics and information technology», 10 (2008), n. 1, p. 43.

[41] B. Giovanola, R. Sala, *The reasons of the unreasonable: is political liberalism still an option?*, in «Philosophy and social criticism», (2021), pp. 1-21, DOI: https://doi.org/10.1177/01914537211040568

[42] Valentini, *op. cit.*, p. 7.

duare i criteri che dovrebbero orientare il *design* dei processi decisionali algoritmici, rendendoli realmente più equi.

3. *"Equità positiva" e decisioni algoritmiche: criteri per un* design *etico*

Come abbiamo evidenziato sopra, l'equità comprende tre componenti principali: l'equa eguaglianza di opportunità, il diritto alla giustificazione e l'equa eguaglianza di relazione.

L'*equa eguaglianza di opportunità* indica un criterio fondamentale da rispettare per far sì che i processi decisionali algoritmici garantiscano una distribuzione delle risorse e delle opportunità realmente equa. Soddisfare il criterio di equa eguaglianza di opportunità richiede, dunque, di progettare strumenti compensativi che non si limitino solo a correggere i *bias* nei set di dati di allenamento degli algoritmi, ma che siano pensati e utilizzati per mitigare le disparità sociali esistenti. Questo implica tenere conto nel *design* dei sistemi algoritmici delle profonde disuguaglianze socio-economiche esistenti, integrando tali sistemi con strumenti specifici pensati per compensarle.

Il *diritto alla giustificazione* indica un secondo criterio fondamentale per operazionalizzare l'equità nei processi decisionali algoritmici. Tale criterio richiede il rispetto di ogni persona come soggetto che può offrire e richiedere una giustificazione, ovvero richiede il rispetto dello status di eguale decisore di ogni persona. Come accennato in precedenza, i risultati dei sistemi decisionali algoritmici sono l'esito di processi probabilistici spesso opachi che operano processando enormi quantità di dati al fine di definire correlazioni che permettono la realizzazione di determinati obiettivi in modo efficiente. Sulla base di queste correlazioni, le opzioni alternative disponibili a ogni persona sono pre-determinate in modi che possono minarne le relative possibilità di scelta e azione – e dunque: lo *status* di eguale decisore. Il diritto alla giustificazione, dunque, prescrive la tutela dell'eguale diritto di ogni persona di richiedere una giustificazione per il trattamento decisionale algoritmico a cui è sottoposta e richiede che i *designer* considerino questa richiesta in modi accessibili agli utenti; richiede, cioè, di progettare sistemi in cui le inferenze e/o correlazioni utilizzate per elaborare un determinato risultato siano rese esplicabili in modo tale che le persone soggette a un risultato algoritmico possano esercitare il loro diritto di conoscere, valutare e/o contestare i parametri alla base del risultato del

processo decisionale stesso, mitigando così anche le asimmetrie di potere a livello decisionale.

Infine, l'*equa uguaglianza di relazione* indica un terzo criterio da tener presente nella progettazione di sistemi decisionali algoritmici equi. Questo criterio richiede di tutelare la possibilità di ogni persona di potersi impegnare in relazioni che esprimano la propria capacità di agire in modo genuino, che favoriscano, dunque, lo sviluppo di affetti, impegni, valori e obiettivi genuini. In realtà, la maggior parte delle tecniche attualmente impiegate nei sistemi decisionali algoritmici utilizza metodi di profilazione degli utenti che si basano su macro-correlazioni standardizzanti (quali il filtro collaborativo), ovvero sulla scoperta di macro-caratteristiche comuni tra gli utenti; questi metodi facilitano infatti la categorizzazione di individui diversi in gruppi di persone "simili" ai quali poi proporre contenuti specifici prestabiliti, pre-determinandone così l'esposizione sia informazionale sia socio-relazionale. Tali profili, tuttavia, non considerano e dunque non rispettano le persone come individui particolari, prediligendo la loro considerazione come aggregati probabilistici di caratteristiche comuni. Garantire il rispetto delle persone come individui particolari e, dunque, soddisfare il criterio di equa uguaglianza di relazione richiede un *design* dei sistemi algoritmici capace di combinare l'apprendimento continuo (tipico del ML e del DL) con strumenti volti a favorire l'interazione specifica tra i sistemi decisionali algoritmici e gli utenti, in modo che questi ultimi possano essere informati sulla loro considerazione algoritmica, ovvero sul modo in cui sono profilati e categorizzati, e a loro volta siano nella posizione di informare attivamente il sistema decisionale sui loro reali affetti, impegni, valori e fini e quindi, in altre parole, di partecipare attivamente al funzionamento del sistema nel plasmare la loro esposizione alle informazioni e alle relazioni che sono significative per sviluppare ed esprimere la loro *agency*.

Conclusioni

In questo articolo abbiamo affrontato uno dei rischi più significativi insiti nei cosiddetti processi decisionali algoritmici, ovvero il rischio di essere ingiusti e promuovere risultati iniqui.

Abbiamo preso le mosse da un'analisi critica del concetto di equità emergente nel dibattito sui processi decisionali algoritmici, per proporre poi una nostra rielaborazione concettuale dell'equità, sviluppata grazie agli strumenti della filosofia morale. Abbiamo così fatto emergere l'importanza

del concetto di "equità positiva", di cui abbiamo messo in luce dimensioni e componenti fondamentali. Infine, abbiamo mostrato le implicazioni della nostra ridefinizione dell'equità sui criteri che dovrebbero orientare il *design* di sistemi decisionali algoritmici equi, suggerendo anche alcuni strumenti per implementarli.

Speriamo di aver contribuito, così, a chiarire alcuni presupposti teorici del dibattito sull'equità nei sistemi decisionali algoritmici, che possono rivelarsi utili anche nella concreta implementazione di sistemi decisionali algoritmici equi.

English title: Fairness and algorithmic decision-making

Abstract

The paper focuses on one of the most urgent risks of artificial intelligence, and more specifically of algorithmic decision-making (ADM), that is, the risk of being unfair. In the first section we provide an overview of the discussion on fairness in ADM and show its shortcomings; in the second section we pursue an ethical inquiry into the concept of fairness, and identify its main dimensions and components, drawing insight from a renewed reflection on respect, which goes beyond the idea of equal respect to include respect for particular individuals too. In the third section we show how our conceptual re-elaboration of fairness can help identify the criteria that ought to steer the ethical design of ADM-based systems to make them really fair.

Keywords: Artificial Intelligence; algorithmic decision-making; fairness; ethical design.

Benedetta Giovanola
Università di Macerata
benedetta.giovanola@unimc.it

Simona Tiribelli
Università di Macerata
simona.tiribelli@unimc.it