Le discriminazioni algoritmiche: un approccio multidisciplinare

Alfredo Di Giorgio

Il presente lavoro cerca di presentare il fenomeno delle discriminazioni algoritmiche attraverso una prospettiva multidisciplinare, integrando prospettive tecniche, etiche e giuridiche. Esplora la non-neutralità intrinseca dei dati e degli algoritmi, precisando come i *bias* emergono e si manifestano in vari settori critici. Il *contributo* esamina criticamente i quadri giuridici esistenti, come il GDPR, ed evidenzia il potenziale trasformativo dell'*AI Act* e della recentissima *Legge n. 132/2025*, entrata in vigore in Italia il 10 ottobre 2025. Approfondisce nuove forme di discriminazione, come la "discriminazione per *proxy*" e il "trattamento disparato inconsapevole", e il loro impatto su gruppi marginalizzati, sia tradizionali che emergenti. Vengono inoltre proposte alcune strategie di valutazione e mitigazione, inclusi l'*auditing* algoritmico e la supervisione umana, e si analizzano le profonde sfide affrontate nella valutazione giudiziaria attraverso una rassegna di casi emblematici. In ultima analisi, viene proposto un approccio normativo responsabile e incentrato sull'essere umano per garantire che l'intelligenza artificiale serva l'umanità, sostenendo al contempo i principi fondamentali di eguaglianza, giustizia e dignità umana. Il presente contributo si colloca nel solco della riflessione sulle narrazioni come strutture generative di senso — e, al tempo stesso, come veicoli di discriminazione. Se le *Giornate di Studio sul Razzismo* hanno indagato le logiche nascoste dei razzismi interrogando il ruolo delle narrazioni nella costruzione della percezione sociale delle differenze, queste pagine intendono estendere tale interrogazione al dominio algoritmico: i sistemi di decisione automatizzata, lungi dall'essere neutrali elaboratori di dati, incorporano e perpetuano narrazioni implicite — schemi classificatori, gerarchie di rilevanza, modelli di normalità — che, proprio in quanto sottratte alla visibilità discorsiva, risultano particolarmente resistenti alla critica e alla contestazione.

1. Introduzione: L'onnipresenza degli algoritmi e l'impatto dei sistemi decisori algoritmici

L'era digitale è indissolubilmente legata alla crescente onnipresenza degli algoritmi e allo sviluppo di sistemi di IA che agiscono come il motore silenzioso di innumerevoli processi decisionali. Questi sistemi modellano le esperienze individuali e collettive in settori critici come la finanza, la sanità, l'informazione e la giustizia penale. La capacità di prendere decisioni automatizzate ha trasformato radicalmente le modalità di interazione tra individui, istituzioni e società civile, promettendo efficienza e

DOI Code: 10.1285/i31034888n2p183

oggettività. Tuttavia, la loro crescente influenza solleva serie preoccupazioni riguardo al potenziale intrinseco di discriminazione.

La pervasività dei sistemi decisori algoritmici automatizzati, che si estende dalla concessione di prestiti all'assunzione di personale, dalla profilazione dei consumatori alla gestione della giustizia predittiva, ha introdotto nuove sfide etico-giuridiche. La complessità di questi sistemi, in particolare la loro capacità di apprendere da vasti *set* di dati (*big data*) e di operare con un'autonomia apparente, ha sollevato interrogativi fondamentali sulla giustizia, la trasparenza e la responsabilità. La dicotomia tra l'efficienza promessa dagli algoritmi e la loro capacità di generare risultati discriminatori è un paradosso intrinseco. I meccanismi progettati per ottimizzare i processi possono, involontariamente, diventare veicoli per ingiustizie sistemiche, evidenziando che il mero avanzamento tecnologico non garantisce di per sé esiti equi.

Secondo la percezione comune, lo sviluppo dell'intelligenza artificiale generativa e degli algoritmi procederebbe lungo una traiettoria lineare e progressiva verso l'Intelligenza Artificiale Generale (*Artificial General Intelligence,* AGI). Si tratta però di una rappresentazione fuorviante, sostenuta più dalle logiche di mercato e dalla retorica dell'"inevitabile crescita" che da un'analisi critica dei limiti effettivi. L'idea che sia sufficiente "scalare" indefinitamente — aumentando potenza computazionale e dimensioni dei modelli — per raggiungere l'AGI appare infatti sempre più come un'illusione. A suggerirlo è proprio l'attuale rallentamento nella curva di miglioramento dei modelli più avanzati di intelligenza artificiale generativa, che sembra indicare l'incontro con vincoli strutturali piuttosto che meramente contingenti[1]. Questo *plateau* tecnologico invita dunque a ridefinire le metriche con cui misuriamo il successo dell'IA, spostando l'attenzione verso il problema fondamentale: colmare l'anello mancante tra linguaggio e conoscenza. Se i modelli attuali eccellono nella manipolazione formale di simboli linguistici, resta aperta la questione cruciale di come questa competenza possa tradursi in una genuina comprensione del mondo. Questa distanza tra elaborazione sintattica e accesso epistemico al reale non è un semplice problema tecnico, ma solleva interrogativi filosofici profondi sulla natura stessa della conoscenza artificiale.

Due elementi critici aiutano a comprendere questo passaggio: la fragilità epistemica dei modelli attuali e la natura intrinsecamente non neutrale dei dati e degli algoritmi, che amplificano i *bias* sociali preesistenti.

[1] J. Sevilla, T. Besiroglu, B. Cottier, J. You, E. Roldán, P. Villalobos, E. Erdil, *Can AI scaling continue through 2030?*, 2024, in «epoch.ai», link: https://epoch.ai/blog/can-ai-scaling-continue-through-2030, ultima consultazione: 12.02.2026.

1.1 Dalla fluidità alla fragilità: la sfida epistemica

Gli algoritmi operano attraverso l'identificazione di *pattern* statistici nei dati di addestramento, apprendendo correlazioni che riflettono le regolarità — e le distorsioni — della produzione umana che li ha generati. Questi sistemi non applicano un ragionamento causale o la logica deduttiva, ma ottimizzano funzioni che tendono a massimizzare la corrispondenza con i *pattern* osservati. La realtà empirica mostra che l'aumento dell'intensità di calcolo e della quantità di dati non è sufficiente per cambiare la natura intrinseca di questi sistemi: si ottiene un perfezionamento incrementale delle prestazioni, ma non una trasformazione qualitativa delle loro capacità epistemiche. Nonostante il miglioramento delle prestazioni, la natura del sistema non cambia. Si ottiene una calcolatrice statistica più potente, ma non una mente capace di produrre conoscenza fondata.

Questa fragilità epistemica, ovvero la condizione in cui la capacità di ottimizzare metriche di *performance* si disconnette dalla capacità di produrre conoscenza fondata, emerge con particolare evidenza e chiarezza nei *Large Language Models* (LLM). Un LLM è essenzialmente un riflesso statistico della produzione linguistica umana, un simulatore sofisticato che campiona dalla distribuzione probabilistica appresa dai dati di *training*. Non costruisce modelli causali del mondo, non verifica proposizioni rispetto a una base di conoscenza strutturata, non ragiona su relazioni logiche in senso proprio: proietta *pattern* statistici appresi da enormi *corpora* testuali, generando sequenze di *token* che massimizzano la probabilità dato il contesto. Poiché l'LLM è addestrato per prevedere la parola successiva più probabile, esso produce testi che "sembrano" autorevoli e coerenti. L'utente percepisce una base di conoscenza genuina dove in realtà esiste solo una "plausibilità linguistica". L'*output* appare affidabile non perché sia verificato, ma perché rispetta perfettamente le strutture del linguaggio umano apprese statisticamente.

Per comprendere appieno questa discontinuità, occorre partire dalla definizione specifica di *epistemia* nel contesto degli LLM: si tratta dell'illusione di conoscenza che emerge quando la plausibilità linguistica sostituisce la verifica, un fenomeno in cui l'*output* generato dal sistema appare fondato ed affidabile pur mancando di genuine basi epistemiche[2]. Sono dunque due facce della stessa medaglia: l'*epistemia* è il fenomeno osservabile, la manifestazione superficiale del problema, mentre la fragilità epistemica è la radice strutturale, l'architettura difettosa che genera sistematicamente quell'illusione.

Le "allucinazioni" degli LLM rappresentano il sintomo più evidente di questa dinamica. Non si tratta semplicemente di errori occasionali o correggibili, ma di una conseguenza sistemica dell'architettura stessa: il modello genera affermazioni che

[2] Cfr. E. Loru, J. Nudo, N. Di Marco, A. Santirocchi, R. Atzeni, M. Cinelli, V. Cestari, C. Rossi-Arnaud, W. Quattrociocchi, *The simulation of judgment in LLMs*, in «Proceedings of the National Academy of Sciences», 2025, vol. 122, n. 42, e2518443122. DOI: 10.1073/pnas.2518443122, pp. 1-24: 1-2.

massimizzano la plausibilità linguistica dato il contesto, anche quando tali affermazioni sono epistemicamente false. Un LLM può inventare citazioni bibliografiche perfettamente formattate ma completamente inesistenti, può attribuire scoperte scientifiche a ricercatori che non le hanno mai compiute, può descrivere eventi storici mai accaduti, tutto con un livello di dettaglio, coerenza testuale e autorevolezza stilistica che "suona vero". Questa è l'*epistemia* in tutta la sua portata: l'illusione di conoscenza generata dalla perfetta simulazione delle forme linguistiche associate alla conoscenza vera, in assenza della base epistemica che dovrebbe sostenerle.

Questa fragilità è ulteriormente aggravata dal principio del «garbage in, garbage out» (GIGO)[3], che nel contesto dei *big data* assume una dimensione particolarmente insidiosa. L'accumulo di più dati non è che l'accumulo di più frammenti di una realtà sociale parziale che può contenere *bias* ed essere distorta[4]. L'idea intuitiva che più dati portino inevitabilmente a modelli migliori si rivela pericolosamente semplicistica quando i dati stessi riflettono prospettive limitate, stereotipi sociali, pregiudizi culturali, discriminazioni storiche. L'accumulo quantitativo non neutralizza le distorsioni qualitative, al contrario le amplifica e le consolida. Se i dati di *training* sovra-rappresentano certe voci e sotto-rappresentano altre, se codificano sistematicamente visioni parziali della realtà, il modello cristallizza e riproduce questi *pattern* su scala industriale. La massa stessa dei dati crea un'ulteriore illusione epistemica: con miliardi di parametri addestrati su terabyte di testo, il sistema appare oggettivo, neutrale, completo. Ma la verità non è una funzione della quantità di dati, è una proprietà della loro qualità epistemica, della loro capacità di rappresentare accuratamente la realtà in tutta la sua complessità e pluralità.

Durante l'addestramento, gli LLM assorbono e interiorizzano non solo i *pattern* linguistici, ma anche le strutture ideologiche e i pregiudizi presenti nei testi. Stereotipi di genere, pregiudizi razziali e *bias* economici si cristallizzano nei parametri del modello, riemergendo negli *output* in forme sistematiche e spesso impercettibili. La sofisticazione linguistica di questi sistemi li rende particolarmente efficaci nel naturalizzare e diffondere forme sottili di discriminazione, conferendo loro un'aura di autorevolezza che può renderle più insidiose delle forme esplicite di pregiudizio.

[3] R. Stenson, *Is This the First Time Anyone Printed,'Garbage In, Garbage Out'?*, in «Atlas Obscura», 14.03.2016, link: https://www.atlasobscura.com/articles/is-this-the-first-time-anyone-printed-garbage-in-garbage-out, ultima consultazione: 12.02.2.026.

[4] Cfr. I.O. Gallegos, R.A. Rossi, J. Barrow, M. Mehrab Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N.K. Ahmed, *Bias and Fairness in Large Language Models: A Survey*, in «Computational Linguistics», 2024, vol. 50, n. 3, pp. 1097-1179, DOI: 10.1162/coli_a_00524.

1.2 Il dato come costrutto umano: «Datificazione», «Datum», «Capta»

I dati non sono entità neutre o oggettive preesistenti nel mondo in attesa di essere scoperte, ma rappresentano costrutti culturali e umani profondamente radicati in specifici contesti storici, sociali e politici. L'idea stessa di dati "grezzi" (*raw data*) costituisce un ossimoro epistemologico fondamentale: ogni dato è sempre già il risultato di un processo intenzionale e teoricamente orientato di indagine, osservazione, selezione e classificazione[5]. Questo processo, che i *critical data studies* definiscono "datificazione" — la trasformazione di fenomeni sociali in forma quantificata e computabile — riflette invariabilmente decisioni, interessi e assunzioni specifiche di chi lo compie[6]. Come sostiene Donna Haraway nella sua critica al "dio trucco" dell'oggettività scientifica, ogni forma di conoscenza è "situata" — emerge cioè da una specifica posizione nel mondo sociale, geografico e temporale[7].

La sociologia della scienza ha dimostrato come anche le pratiche scientifiche più rigorose producano risultati che sono inevitabilmente plasmati dai contesti in cui operano[8]. La stessa etimologia del termine "dato" nasconde questa complessità costruttiva sotto un'accezione ingannevole di certezza e passività. Il participio passato latino *datum* (letteralmente "ciò che è dato, consegnato") veicola l'immagine di un'informazione semplicemente ricevuta, come un dono che preesiste all'atto della sua acquisizione. In contrapposizione critica a questa semantica fuorviante, Johanna Drucker propone il termine alternativo "capta" (dal latino *capere*, catturare, afferrare) per sottolineare esplicitamente l'atto umano intenzionale che sottende ogni processo di selezione, misurazione e registrazione[9]. Daston e Galison dimostrano come l'ideale stesso di "oggettività meccanica" sia un'invenzione storica relativamente recente[10] .

La critica alla neutralità dei dati sfida frontalmente la concezione positivista del dato come specchio fedele della realtà. I metodi di campionamento possono sistematicamente escludere popolazioni intere, perpetuando la loro invisibilità statistica e politica[11]. Le stesse categorie di classificazione utilizzate nei *dataset* — genere, razza, età, occupazione — non sono descrizioni naturali ma costrutti storici e sociali che riflettono e perpetuano specifici ordini di potere. Come mostrano gli studi critici sulle infrastrutture di classificazione, ogni sistema tassonomico incorpora

[5] Cfr. *"Raw data" is an oxymoron*, a cura di L. Gitelman, MIT Press, Cambridge (MA) 2013.

[6] Cfr. V. Mayer-Schönberger, K. Cukier, *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt, Boston 2013.

[7] Cfr. D. Haraway, Situated knowledges: *The science question in feminism and the privilege of partial perspective*, in «Feminist Studies», 1988, vol. 14, n. 3, pp. 575-599.

[8] Cfr. B. Latour, S. Woolgar, *Laboratory life: The construction of scientific facts,* Princeton UP, 1979.

[9] Cfr J. Drucker, *Graphesis: Visual Forms of Knowledge Production,* Harvard UP, Cambridge (MA) 2014, pp. 128-129.

[10] Cfr. L. Daston, P. Galison, *Objectivity*, Zone Books, New York 2007.

[11] G.C. Bowker, S.L. Star, *Sorting things out: Classification and its consequences*, MIT Press, Cambridge (MA) 1999, pp. 5-6.

giudizi di valore su cosa sia simile o diverso, normale o deviante, centrale o marginale[12]. Come dimostrano in maniera incontrovertibile alcuni *feminist data studies*, i *dataset* riproducono sistematicamente *bias* di genere[13]. Gli studi decoloniali mostrano invece come i *big data* contemporanei riproducano forme di "colonialismo dei dati"[14].

Se accettiamo che il dato è *capta* piuttosto che *datum* — catturato intenzionalmente piuttosto che semplicemente dato — le implicazioni per l'intelligenza artificiale sono profonde e ineludibili. La "conoscenza" prodotta dai sistemi di IA non può essere considerata oggettivamente vera o neutra, ma rappresenta necessariamente un riflesso dei *bias*, delle assunzioni e delle esclusioni incorporati nei dati di addestramento[15].

2. Il concetto di "*bias* algoritmico": origini e manifestazioni

Il *bias* algoritmico rappresenta uno dei fenomeni più complessi e sfaccettati dell'era digitale contemporanea, la cui comprensione richiede un'analisi che vada oltre le semplici categorizzazioni per abbracciare la natura profondamente sistemica e multidimensionale di questa problematica. Esso si configura innanzitutto come una forma di distorsione non casuale, ma piuttosto sistemica e ripetibile, che emerge dall'interazione complessa tra dati, modelli matematici, contesti applicativi e dinamiche sociali preesistenti. Non si tratta semplicemente di "errori" nel senso tradizionale del termine, ma di *pattern* di comportamento degli algoritmi che riflettono, amplificano o addirittura creano nuove forme di discriminazione e iniquità. La distinzione è cruciale: mentre un errore casuale si distribuisce uniformemente e può essere corretto con maggiore precisione tecnica, il *bias* algoritmico è asimmetrico, colpisce in modo sproporzionato certi gruppi rispetto ad altri, e spesso resiste ai semplici aggiustamenti tecnici perché è radicato in problematiche più profonde.

Un aspetto fondamentale per comprendere il *bias* algoritmico è riconoscere che gli algoritmi, contrariamente alla percezione comune di oggettività e neutralità che li circonda, sono profondamente incorporati nel tessuto sociale, culturale e storico che li produce. Ogni sistema algoritmico è il risultato di scelte umane: dalla selezione dei dati di addestramento, alla definizione degli obiettivi da ottimizzare, dalla scelta delle variabili da includere o escludere, fino alle metriche utilizzate per valutare le *performance*. Gli algoritmi apprendono dai dati storici, e quando questi dati riflettono pratiche discriminatorie del passato, il rischio è che tali discriminazioni vengano non

[12] Id., pp. 195-225.
[13] Cfr. C. D'Ignazio, L.F. Klein, *Data feminism*, MIT Press, Cambridge (MA) 2020, p. 34.
[14] Cfr. N. Couldry, U.A. Mejias, *The costs of connection: How data is colonizing human life and appropriating it for capitalism*, Stanford University Press, Stanford 2019, pp. ix-xiv, 37-68.
[15] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown, New York 2016.

solo replicate, ma anche normalizzate e legittimate attraverso l'apparente oggettività della matematica[16].

Ma il *bias* algoritmico non è solo una forma di replica del passato. In molti casi, gli algoritmi possono amplificare distorsioni preesistenti o generarne di nuove attraverso meccanismi di *feedback* che creano circoli viziosi. Quando un algoritmo di polizia predittiva indirizza maggiori risorse di controllo verso certi quartieri sulla base di dati storici di arresti, questo porta inevitabilmente a un maggior numero di arresti in quelle zone, che a loro volta alimentano l'algoritmo rafforzando la predizione iniziale e creando un fenomeno noto come *feedback loop* o circolo vizioso algoritmico, indipendentemente dai tassi reali di criminalità[17]. Il *bias* algoritmico si manifesta anche attraverso quella che potremmo chiamare una "falsa universalità": gli algoritmi sono spesso progettati con l'ambizione di creare soluzioni universali, ma questa pretesa nasconde il fatto che i modelli sono invariabilmente ottimizzati per certi gruppi più che per altri. Un algoritmo di riconoscimento vocale addestrato principalmente su voci maschili di parlanti nativi inglesi funzionerà peggio per donne, bambini, o persone con accenti diversi[18]. Un sistema di diagnosi medica basato su dati provenienti principalmente da popolazioni europee può essere meno accurato per persone di altre etnie[19].

È importante comprendere che il *bias* algoritmico non è semplicemente una questione tecnica che può essere "risolta" attraverso migliori tecniche di programmazione o *dataset* più ampi. Esso solleva questioni profondamente normative ed etiche su cosa significhi equità, giustizia e pari trattamento. Esistono infatti molteplici concezioni di *fairness*, spesso matematicamente incompatibili tra loro[20], e la scelta di quale adottare non è un problema tecnico ma una decisione valoriale che riflette priorità etiche e politiche. Dovremmo mirare a garantire che gruppi diversi ricevano *outcome* positivi in proporzioni uguali? O dovremmo invece assicurare che, a parità di merito, le probabilità di successo siano le stesse? Queste domande non hanno risposte univoche.

La letteratura scientifica ha identificato varie tipologie di tassonomia per identificare i principali *bias* algoritmici. Tra queste, due particolarmente importanti ed incisive sono la tassonomia proposta da Barocas e Selbst, che si concentra prevalentemente

---

[16] Cfr. J. Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, in «Reuters», 11.10.2018, link : https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/, ultima consultazione: 12.02.2026.

[17] Cfr. K. Lum, W. Isaac, *To predict and serve?*, in «Significance», 2016, vol. 13, n. 5, pp. 14-19.

[18] Cfr. A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J.R. Rickford, D. Jurafsky, S. Goel, *Racial disparities in automated speech recognition*, in «Proceedings of the National Academy of Sciences», 2020, vol. 117, n. 14, pp. 7684-7689.

[19] Cfr. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, *Dissecting racial bias in an algorithm used to manage the health of populations*, in «Science», 2019, vol. 366, n. 6464, pp. 447-453.

[20] Cfr. A. Chouldechova, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, in «Big Data», 2017, vol. 5, n. 2, pp. 153-163.

sulle fasi tecniche del processo di *data mining*[21], e quella di Loza de Siles, che presenta un grado di granularità impressionante, identificando oltre cinquanta distinti *bias* algoritmici organizzati in sei categorie[22]. La tassonomia qui proposta — articolata in *bias* di progettazione, *bias* ereditati dai dati, *bias* di interazione, di soglia e intersezionalità e *bias* architetturali — nasce dall'esigenza di bilanciare rigore analitico e intelligibilità teorica, cercando di identificare i meccanismi generativi fondamentali della discriminazione algoritmica piuttosto che catalogarne le manifestazioni empiriche.

2.1 *Bias* di progettazione

Il *bias* di progettazione emerge dalle scelte effettuate durante la fase di sviluppo dell'algoritmo. Le decisioni relative alla selezione delle variabili, alla definizione delle funzioni obiettivo e alla strutturazione delle regole decisionali incorporano inevitabilmente le assunzioni, i valori e i pregiudizi dei progettisti[23]. Questo tipo di *bias* è particolarmente insidioso poiché si annida nelle fondamenta stesse del sistema, risultando spesso invisibile agli utenti finali e persino agli stessi sviluppatori.

Una delle manifestazioni più insidiose riguarda l'utilizzo di "variabili proxy", ovvero attributi apparentemente neutri che funzionano come indicatori indiretti di caratteristiche protette dalla legge come razza, genere, etnia o classe sociale. Barocas e Selbst analizzano in profondità questo fenomeno, evidenziando come variabili tecnicamente "pulite" come il codice postale, il nome dell'università frequentata, gli hobby dichiarati o persino le abitudini di acquisto possano fungere da *proxy* perfetti per categorie protette, aggirando di fatto le tutele anti-discriminatorie[24]. Il codice postale, ad esempio, incorpora informazioni sulla composizione etnica ed economica dei quartieri a causa della segregazione residenziale storica[25], quello che Noble definisce «technological redlining», un'evoluzione digitale delle pratiche discriminatorie immobiliari del Novecento[26].

---

[21] Cfr. S. Barocas, A.D. Selbst, *Big data's disparate impact*, in «California Law Review», CIV (2016), n. 3, pp. 671-732: 677-693.

[22] Cfr. E. Loza de Siles, *Disaggregating artificial intelligence biases: a law and systems engineering approach for AI governance and regulation*, *Research Handbook on the Law of Artificial Intelligence*, 2ª ed., a cura di in W. Barfield, U. Pagallo, Edward Elgar Publishing, Cheltenham-Northampton 2025, pp. 274-295.

[23] Cfr. B. Friedman, H. Nissenbaum, *Bias in computer systems*, in «ACM Transactions on information systems» (TOIS), 1996, vol. 14, n. 3, pp. 330-347.

[24] Cfr. S. Barocas, A.D. Selbst, *Big data's disparate impact*, cit., pp. 677-693.

[25] Cfr. R. Richardson, J.M. Schultz, K. Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, in «New York University Law Review Online», 2019, vol. 94, pp. 15-55.

[26] Cfr. S.U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press, New York 2018, p. 17.

La pericolosità dei *proxy* risiede nella loro capacità di eludere sia i controlli legali sia i meccanismi di *fairness* tecnica: eliminare variabili protette dai modelli (approccio noto come *fairness through unawareness*) non garantisce equità se persistono correlazioni statistiche con altre variabili. Sweeney ha documentato empiricamente come nomi associati alla popolazione afroamericana generassero pubblicità online che suggerivano precedenti penali con frequenza significativamente maggiore[27]. Citron e Pasquale descrivono una *scored society*, dove le persone vengono valutate attraverso punteggi opachi derivanti da procedure non trasparenti[28]. Kilbertus et al. introducono una distinzione fondamentale tra "l'attributo protetto" in sé – come il genere o l'appartenenza razziale – e le sue manifestazioni osservabili nel mondo, ciò che gli autori chiamano «proxy osservabili»[29].

La distinzione tra attributi protetti e *proxy* è fondamentale per comprendere come evitare discriminazioni nei sistemi algoritmici. Sebbene gli algoritmi possano non avere accesso diretto agli attributi protetti, i *proxy* possono comunque portare a decisioni discriminatorie. Questo accade quando il sistema utilizza segnali che correlano con gli attributi protetti per fare previsioni o scelte, senza comprendere che questi segnali potrebbero favorire un gruppo a scapito di un altro. Ad esempio, se un sistema di assunzione di personale utilizza il nome come *proxy* per determinare il genere o l'origine etnica, potrebbe finire per selezionare o scartare candidati sulla base di pregiudizi impliciti legati a questi attributi, anche senza che il sistema abbia accesso esplicito alle informazioni sul genere o sull'etnia. La discriminazione algoritmica si verifica proprio quando i *proxy* vengono utilizzati senza consapevolezza del loro legame con gli attributi protetti. Questo crea una situazione in cui gli algoritmi riflettono e amplificano *bias* storici e pregiudizi, anche se non c'è un'intenzione esplicita di discriminare.

Un altro aspetto critico del *bias* di progettazione riguarda le "funzioni di ottimizzazione". Quando un algoritmo viene progettato, uno degli aspetti cruciali è la funzione obiettivo, ovvero la metrica che guida l'algoritmo nel suo apprendimento e miglioramento. Ad esempio, un modello di riconoscimento facciale può essere ottimizzato per ridurre al minimo l'errore nel riconoscere le identità, mentre un algoritmo di assunzione del personale potrebbe essere ottimizzato per massimizzare la probabilità di assumere persone che avranno successi professionali.

Tuttavia, la scelta di quale metrica ottimizzare non è mai neutra. Ogni scelta di metrica implica delle assunzioni normative su cosa sia considerato un "buon risultato",

[27] Cfr. L. Sweeney, *Discrimination in Online Ad Delivery*, in «Communications of the ACM», 2013, vol. 56, n. 5, pp. 44-54.

[28] Cfr. D.K. Citron, F.A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, in «Washington Law Review», 2014, vol. 89, n. 1, pp. 1-34.

[29] Cfr. N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, *Avoiding Discrimination through Causal Reasoning*, in Advances in *Neural Information Processing Systems 30*, (NIPS 2017), Curran Associates, Long Beach (CA) 2017, pp. 656-666.

riflettendo inevitabilmente delle valutazioni su cosa sia giusto o desiderabile. Per esempio, se l'obiettivo di un algoritmo di assunzione è quello di massimizzare il numero di assunzioni da una determinata università prestigiosa, l'algoritmo potrebbe finire per favoreggiare persone provenienti da un gruppo sociale o economico specifico, anche se ciò non era un obiettivo esplicito. Questo crea un potenziale *bias*: anche se il modello ottimizza perfettamente la metrica scelta, potrebbe finire per produrre risultati discriminatori o ingiusti, semplicemente perché la metrica stessa ha delle implicazioni sociali o etiche non affrontate. A tale proposito, sono state sviluppate le cosiddette *model cards*, che sono un *framework* ideato per evidenziare e analizzare questi aspetti in modo più trasparente[30]. Le *model cards*, mettendo in evidenza non solo le prestazioni tecniche del modello, ma anche le decisioni progettuali che influenzano i risultati e le implicazioni etiche e sociali delle sue applicazioni, mirano a migliorare la trasparenza nei modelli di AI, rendendo chiari gli assunti sottostanti, le scelte normative e i rischi di *bias*. In pratica, una *model card* è una sorta di "carta d'identità" di un modello di *machine learning*, che include diverse informazioni.

La questione della progettazione algoritmica si complica quando le funzioni di ottimizzazione vengono utilizzate per perseguire obiettivi sociali complessi e talvolta controversi. In molti casi, infatti, queste funzioni agiscono come *proxy* di ideali sociali che non possono essere ridotti a semplici formule matematiche. Per esempio, un algoritmo che ottimizza l'allocazione delle risorse sanitarie per "massimizzare gli anni di vita salvati" (misurati in *Quality-Adjusted Life Years*, o QALYs) finisce per discriminare le persone anziane e disabili, dato che tende a favorire chi è più giovane e in buona salute, sacrificando il benessere di categorie più vulnerabili. Un altro esempio è un algoritmo educativo progettato per "predire il successo accademico futuro", che, pur ottimizzando tecnicamente il proprio obiettivo, perpetua le disuguaglianze sociali, premiando chi proviene da ambienti privilegiati e avvantaggiando coloro che già partono con un vantaggio[31].

In entrambi i casi, gli algoritmi si trovano a ridurre obiettivi sociali complessi e sfaccettati a metriche misurabili, dimenticando le dimensioni qualitative, contestuali e normative che definiscono concetti come la giustizia. Come sottolineato da Selbst e colleghi, i progettisti di sistemi algoritmici spesso cadono nell'errore di ridurre questioni sociali complesse a funzioni matematiche facilmente ottimizzabili, senza considerare che la giustizia non può essere ridotta a un semplice problema di ottimizzazione numerica.

---

[30] Cfr. S. Mitchell, E. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji, T. Gebru, *Model Cards for Model Reporting*, in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT 2019, Atlanta, GA, USA, January 29-31*, 2019, pp. 220-229.

[31] Cfr. A.D. Selbst, D. Boyd, S.A. Friedler, S. Venkatasubramanian, J. Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT 2019, Atlanta, GA, USA, January 29-31*, cit., pp. 59-68.

Il caso della giustizia predittiva è un esempio paradigmatico di come la scelta di una funzione obiettivo diventi inevitabilmente politica. Gli algoritmi utilizzati in questo contesto sono progettati per minimizzare la probabilità che un individuo rilasciato dal carcere commetta nuovamente un crimine. Tuttavia, l'algoritmo deve affrontare una difficile bilanciatura tra due tipi di errori: i falsi positivi e i falsi negativi. Da un lato, c'è il rischio che l'algoritmo identifichi erroneamente una persona come pericolosa (falso positivo), provocando detenzioni ingiuste; dall'altro, c'è il rischio che l'algoritmo rilasci un individuo che, in realtà, potrebbe tornare a commettere reati (falso negativo), mettendo a rischio la sicurezza pubblica.

Purtroppo, non è possibile eliminare entrambi i tipi di errore contemporaneamente. Se l'algoritmo tenta di ridurre i falsi negativi per proteggere la società, inevitabilmente aumenterà i falsi positivi, privando ingiustamente qualcuno della sua libertà. Al contrario, cercare di ridurre i falsi positivi comporta un rischio maggiore per la sicurezza pubblica. Di fatto, l'ottimizzazione di tale algoritmo non consiste nel minimizzare gli errori in senso assoluto, ma nel trovare un compromesso tra la protezione della libertà individuale e la sicurezza pubblica.

Questa tensione irrisolvibile tra i vari concetti di equità e giustizia è formalmente dimostrata nei "teoremi di impossibilità della *fairness*", elaborati indipendentemente da Chouldechova[32] e Kleinberg et al.[33]. Questi teoremi rappresentano uno dei risultati più significativi nella ricerca sull'equità algoritmica, poiché evidenziano le difficoltà strutturali insite nel cercare di applicare concetti morali e sociali complessi a un contesto tecnico e numerico. In altre parole, essi dimostrano rigorosamente che, in presenza di determinate condizioni statistiche, è matematicamente impossibile soddisfare simultaneamente diverse definizioni intuitive di equità, e dunque, la scelta di quale metrica di equità privilegiare diventa una decisione politica e normativa. Questa irriducibile dimensione politica della progettazione algoritmica richiede un cambiamento epistemologico che riconosca i limiti della *fairness through formalization*[34].

2.2 *Bias* ereditati dai dati

I *bias* ereditati dai dati rappresentano la forma più documentata di discriminazione algoritmica. Gli algoritmi di *machine learning* apprendono *pattern* dai dati storici,

[32] Cfr. A. Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, cit.

[33] Cfr. J. Kleinberg, S. Mullainathan, M. Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, in*8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 2017, pp. 1-23.

[34] Cfr. C. Barabas, M. Virza, K. Dinakar, J. Ito, J. Zittrain, *Interventions over predictions: reframing the ethical debate for actuarial risk assessment*, in *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, a cura di S.A. Friedler, C. Wilson, ACM, New York (NY) 2018, pp. 62–76.

riproducendo e talvolta amplificando le discriminazioni pregresse contenute nei *dataset* di addestramento. Come ha osservato Ruha Benjamin, i codici postali americani costituiscono «l'output delle politiche Jim Crow e l'input dei nuovi Jim Codes»[35], evidenziando, in questo modo, come le disuguaglianze storiche si sedimentino nei dati e vengano perpetuate attraverso i sistemi algoritmici. Il gioco di parole adottato da Benjamin è estremamente efficace e si basa su un'analogia storica e geografica che illustra perfettamente come i pregiudizi sociali vengano "codificati" nei sistemi tecnologici. Il riferimento è alle *leggi di Jim Crow*, il sistema di leggi e politiche che, dal tardo XIX secolo fino alla metà del XX secolo, istituì la segregazione razziale e la discriminazione legale negli Stati Uniti meridionali. Benjamin afferma che i codici postali americani (*zip codes*) sono l'«output delle politiche Jim Crow»; questo significa che la disparità di opportunità, ricchezza, qualità dei servizi e sorveglianza tra i diversi quartieri non è casuale, ma è il risultato diretto delle politiche discriminatorie e segregazioniste del passato. Il codice postale non è solo un indirizzo, ma una variabile che incorpora e riflette un intero secolo di ingiustizie strutturali. Quello che Benjamin chiama ”New Jim Code” sono proprio queste nuove tecnologie che riflettono e riproducono le disuguaglianze esistenti, pur essendo promosse e percepite come più obiettive rispetto ai sistemi discriminatori del passato. Comprendere la discriminazione algoritmica richiede di superare la percezione di un "errore" software e di concentrarsi sui meccanismi strutturali che assicurano la riproduzione delle disuguaglianze sociali. La questione è molto più profonda e riguarda il modo stesso in cui i dati vengono raccolti e interpretati. Per capirlo veramente, dobbiamo esplorare due meccanismi fondamentali che operano spesso in modo invisibile: il problema dei *selection bias* e quello del *labeling bias*.

- Il *selection bias* si manifesta quando i dati di addestramento non riflettono l'intera popolazione *target*. Non è sufficiente avere “molti dati”: ciò che conta è la loro rappresentatività rispetto alle diverse sotto-popolazioni che il sistema dovrà servire.
- Il *labeling bias* emerge invece quando le etichette assegnate ai dati — ovvero le categorizzazioni che definiscono la “verità” che l'algoritmo deve apprendere — incorporano e cristallizzano pregiudizi, stereotipi e discriminazioni sedimentati nelle pratiche sociali e istituzionali umane. Quando i dati di addestramento riflettono discriminazioni storiche, l'algoritmo non solo apprende questi *pattern* discriminatori, ma li codifica come relazioni statistiche valide, riproducendoli e amplificandoli[36].

[35] R. Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*, Polity Press, Cambridge 2019, pp. 49-76.
[36] Cfr. G.C. Bowker, S.L. Star, *Sorting things out: Classification and its consequences*, MIT Press, Cambridge (MA) 1999, pp. 5-6.

2.3 *Bias* di interazione, soglia e intersezionalità

La società contemporanea si regge su una fitta trama di interazioni umane che non si limitano a veicolare informazioni, ma costituiscono il luogo primario in cui la realtà sociale viene attivamente costruita e negoziata. Proprio perché le interazioni svolgono questa funzione costitutiva, le distorsioni che le attraversano non restano confinate al piano psicologico individuale: esse plasmano direttamente la struttura del mondo sociale. La psicologia sociale ha mostrato come i processi interpretativi e i condizionamenti ambientali possano alterare in modo significativo la percezione della realtà e gli scambi relazionali, dando origine a quelli che possiamo definire *bias* di interazione. Questi non sono semplici errori casuali del pensiero, ma meccanismi adattivi e strutturali che influenzano il comportamento umano a livello intrapersonale, interpersonale e collettivo.

È precisamente attraverso tali meccanismi che le disuguaglianze sociali si riproducono nel tempo. Per comprendere questo processo, occorre abbandonare l'idea delle disuguaglianze come dati statici e riconoscerle invece come dinamiche stratificate, che operano tanto nelle interazioni quotidiane quanto nella configurazione delle traiettorie di vita individuali. Sul piano più concreto, questi meccanismi si traducono in micro-comportamenti osservabili nelle relazioni dirette: un tono di voce differente, interruzioni più frequenti, aspettative predefinite. Pur non essendo sempre espliciti, questi comportamenti condizionano le opportunità che si aprono alle persone nel momento stesso in cui si presentano agli altri, trasformando le aspettative sociali in realtà vissute.

Tali micro-eventi, tuttavia, non restano isolati: si accumulano nel tempo e cristallizzano in barriere sistematiche. È qui che emerge il *bias* di soglia, secondo cui chi appartiene a gruppi marginalizzati deve superare standard di valutazione più severi per ottenere riconoscimenti o avanzamenti — come percorrere un cammino con un peso invisibile, dove ogni successo richiede uno sforzo straordinario. Il *bias* di soglia si manifesta in particolare quando decisioni binarie, come approvazione o rifiuto, si applicano attraverso criteri che penalizzano sistematicamente alcuni gruppi.

Per comprendere la complessità di queste esperienze, non possiamo analizzare una sola dimensione dell'identità alla volta. È qui che entra in gioco l'"intersezionalità", un concetto sviluppato da Kimberlé Crenshaw[37], che sottolinea come genere, razza, classe e altre categorie si intreccino, dando vita a esperienze uniche di svantaggio o privilegio. La natura processuale delle disuguaglianze si complica ulteriormente quando si considera la loro dimensione intersezionale. La metafora del "crocevia" risulta particolarmente efficace per descrivere come le diverse appartenenze

[37] Cfr. K. Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*, University of Chicago Legal Forum, vol. 1989, n. 1, 1989, pp. 139-167.

categoriali — razza, genere, classe sociale — non si sommino semplicemente, ma si sovrappongano generando svantaggi specifici irriducibili a una singola dimensione.

L'IA crea inoltre "nuove minoranze" basate su criteri completamente inediti rispetto alle tradizionali categorie giuridiche: quartiere, codice postale, preferenze *online*, *pattern* comportamentali. Un esempio significativo di intersezionalità nei sistemi tecnologici è fornito dagli studi di Buolamwini e Gebru[38], che hanno dimostrato che i sistemi di riconoscimento facciale avevano un'accuratezza superiore al 99% per gli uomini bianchi, ma scendevano drasticamente al 65% per le donne nere. Questo evidenzia come le discriminazioni non siano solo un fenomeno isolato, ma si intrecciano lungo molteplici assi, creando esperienze di svantaggio profonde e complesse.

Le disuguaglianze sociali sono processi complessi e stratificati che richiedono un'analisi approfondita e multidimensionale. Solo attraverso l'intersezionalità possiamo iniziare a comprendere come le diverse identità si sovrappongano, influenzando le esperienze individuali di svantaggio e privilegio. È qui che l'intersezionalità diventa la lente indispensabile, illuminando come genere, etnia, classe e altre categorie si intreccino creando esperienze uniche e specifiche di svantaggio o privilegio. Una donna nera, ad esempio, non affronta semplicemente la somma del sessismo e del razzismo, ma un insieme peculiare di *bias* di interazione (come essere percepita in modo diverso sia dagli uomini neri che dalle donne bianche) e di *bias* di soglia (come dover dimostrare una competenza iperbolica per essere considerata semplicemente adeguata), che nascono proprio dalla sovrapposizione delle sue identità. In questo modo, i tre concetti si illuminano a vicenda: l'intersezionalità ci fornisce la mappa della complessità identitaria, mentre il *bias* di interazione e il *bias* di soglia ci mostrano i meccanismi attraverso cui quella complessità si traduce in disuguaglianze concrete e perpetuate.

Ciò che rende la discriminazione algoritmica particolarmente insidiosa è la sua capacità di autoalimentarsi attraverso altri meccanismi interconnessi:

- Il "loop di feedback positivo" rappresenta uno dei meccanismi più insidiosi attraverso cui la discriminazione algoritmica si autoalimenta nel tempo, trasformando una disparità iniziale anche minima in una divergenza strutturale crescente. Il processo si innesca quando un algoritmo produce *output* discriminatori che riducono sistematicamente le opportunità per determinati gruppi di accedere a risorse, visibilità o servizi. Si crea così una spirale discriminatoria discendente: meno visibilità → meno *engagement* → conferma algoritmica della "bassa qualità" → ulteriore riduzione della visibilità. Questo processo non è lineare ma esponenziale: ogni iterazione del ciclo amplifica la disparità precedente.

[38] Cfr. J. Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, cit. pp. 77-91.

- Il "bias di conferma algoritmico" costituisce una forma peculiare di rigidità epistemica: il sistema interpreta l'assenza di *feedback* positivi da gruppi discriminati come conferma che quei gruppi siano effettivamente "meno rilevanti".
- Gli "effetti di rete"**:** la discriminazione iniziale viene amplificata dalla tendenza degli utenti a interagire preferenzialmente con profili simili ai loro, esacerbando la polarizzazione. Gli effetti di rete introducono una dimensione ulteriore e cruciale nella dinamica della discriminazione algoritmica, mostrando come la discriminazione iniziale del sistema venga non solo rafforzata internamente (attraverso i *feedback loop*) ma anche amplificata socialmente attraverso le interazioni degli utenti stessi.

2.4 *Bias* architetturali

I sistemi di intelligenza artificiale rappresentano insiemi più ampi che inglobano algoritmi, dati, interfacce e processi decisionali in architetture integrate. I già citati teoremi di impossibilità elaborati da Kleinberg, Mullainathan e Raghavan e da Chouldechova, successivamente organizzati in una cornice teorica unitaria da Corbett-Davies e Goel nella loro rassegna critica della letteratura sulla *fairness* algoritmica[39], portano con sé implicazioni di notevole portata per la *governance* dei sistemi di intelligenza artificiale. Il nucleo concettuale di tali implicazioni può essere così formulato: qualsiasi sistema predittivo che non raggiunga la perfezione assoluta nelle proprie previsioni — condizione, va da sé, irrealizzabile nella pratica — si trova necessariamente costretto a compiere scelte allocative tra diverse e reciprocamente incompatibili concezioni di equità. Non è possibile, in altre parole, soddisfare simultaneamente tutte le definizioni ragionevoli di trattamento giusto quando le previsioni del sistema presentano margini di errore e quando i gruppi sociali di riferimento differiscono nei tassi base del fenomeno che si intende predire.

La conseguenza teorica di questa impossibilità matematica è di natura radicale: la discriminazione cessa di configurarsi come un difetto tecnico suscettibile di correzione mediante affinamenti successivi del modello, e si rivela invece come una proprietà strutturale e ineliminabile di qualunque sistema decisionale automatizzato che operi in contesti sociali caratterizzati da disuguaglianze preesistenti. La scelta non è dunque "se" discriminare, ma "come" e "chi" discriminare — ovvero quale concezione di giustizia privilegiare a scapito delle altre: ogni sistema predittivo imperfetto deve

[39] Cfr. S. Corbett-Davies, H. Nilforoshan, R. Shroff, S. Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, in «Journal of Machine Learning Research», 2023, vol. 24, n. 1, pp. 1-117.

operare un *trade-off* tra diverse concezioni di giustizia. La ricerca di Blattner e Nelson[40] ha dimostrato, ad esempio, che i punteggi di credito sono sistematicamente meno accurati per le minoranze e le famiglie a basso reddito, creando forme di discriminazione che non derivano da intento ma dal rumore informativo maggiore.

Le piattaforme social costituiscono il livello più complesso dell'architettura tecnologica. Il concetto di "bolla di filtro" (*filter bubble*), introdotto da Eli Pariser[41], descrive come gli algoritmi di personalizzazione creino universi informativi individualizzati e segregati. Alcuni studi hanno invece documentato come gli algoritmi di X (Twitter fino al 2022), tendano ad amplificare contenuti politicamente estremi[42]. Un'area particolarmente documentata riguarda la discriminazione nella distribuzione degli annunci di lavoro: Lambrecht e Tucker[43] hanno scoperto significative disparità di genere negli annunci per carriere STEM, mentre un'indagine di Global Witness[44] ha rilevato che il 91% delle persone esposte ad annunci per posizioni di meccanico si identificava come uomo, nonostante non vi fossero targeting intenzionali di genere.

Il "trattamento disparato inconsapevole" rappresenta una delle manifestazioni più eloquenti del carattere costitutivo della discriminazione algoritmica, e per questa ragione trova la sua collocazione più appropriata all'interno della categoria dei *bias* architetturali. Il fenomeno è noto: un sistema algoritmico può "ricostruire" attributi protetti anche quando questi vengono esplicitamente esclusi dall'insieme delle variabili di *input*. Il codice postale, il *pattern* di navigazione, la sequenza temporale degli acquisti, persino le scelte lessicali in un curriculum diventano *proxy* sufficientemente informativi da permettere all'algoritmo di inferire — e quindi di utilizzare come base decisionale — caratteristiche come razza, genere, orientamento sessuale o appartenenza religiosa che i progettisti avevano intenzionalmente rimosso dal modello. Tecnicamente, il problema deriva dalla capacità degli algoritmi di apprendimento automatico di identificare correlazioni latenti nello spazio

---

[40] Cfr. L. Blattner, S. Nelson, *How Costly is Noise? Data and Disparities in Consumer Credit*, Stanford Graduate School of Business, Working Paper n. 3978, 2021.

[41] Cfr. E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Press, New York 2011.

[42] Cfr. F. Huszár, S.I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, M. Hardt, *Algorithmic Amplification of Politics on Twitter*, in «Proceedings of the National Academy of Sciences», 2022, vol. 119, n. 1.

[43] Cfr. A. Lambrecht, C. Tucker, *Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads*, in «Management Science», 2019, vol. 65, n. 7, pp. 2966-2981.

[44] Cfr. Global Witness, *Facebook's Gender Bias in Job Ads*, Global Witness Report, 2023. Comunicato stampa, giugno 2023. Disponibile presso: https://globalwitness.org/en/press-releases/facebook-accused-of-gender-discrimination-new-research-finds-bias-in-advertising-algorithm/, ultima consultazione: 12.02.2026. L'indagine è stata condotta tra marzo e aprile 2023 pubblicando annunci per posizioni lavorative reali su Facebook in sei paesi (Regno Unito, Paesi Bassi, Francia, India, Irlanda e Sud Africa). Gli annunci utilizzavano l'obiettivo "Traffic/Link Clicks" dell'algoritmo di ottimizzazione di Facebook, senza *targeting* manuale basato su genere.

multidimensionale delle *features*, correlazioni che sfuggono alla percezione umana ma che riflettono le strutture profonde della stratificazione sociale.

A prima vista, il "trattamento disparato inconsapevole" potrebbe apparire come un caso paradigmatico di discriminazione contingente: un difetto correggibile attraverso tecniche di *fairness-aware machine learning*, *auditing* algoritmico, rimozione iterativa dei *proxy*. E in effetti, una parte significativa della letteratura tecnica si concentra precisamente su questi rimedi. Tuttavia, una riflessione più attenta rivela come questo fenomeno tocchi la dimensione costitutiva della discriminazione algoritmica. La capacità di ricostruire attributi protetti non è un'anomalia del funzionamento algoritmico, ma una conseguenza diretta della sua architettura cognitiva. Gli algoritmi di apprendimento automatico sono progettati precisamente per scoprire *pattern* statistici latenti nei dati — questa è la loro funzione, la loro ragion d'essere. Impedire loro di identificare correlazioni che codificano indirettamente attributi protetti equivale a limitare artificialmente la loro capacità predittiva, introducendo un trade-off strutturale tra accuratezza e *fairness* che i teoremi di impossibilità hanno dimostrato essere irrisolvibile in termini generali. Il punto cruciale è che la "scoperta" degli attributi protetti non avviene per errore o negligenza dei progettisti, ma emerge dal modo stesso in cui l'algoritmo elabora l'informazione.

Qui si manifesta la frattura ontologica tra giudizio umano e funzionamento algoritmico. Un decisore umano che applica consapevolmente criteri discriminatori può essere rieducato, sanzionato, sostituito; uno che discrimina inconsapevolmente può essere reso consapevole dei propri *bias* impliciti attraverso formazione e riflessione. Ma un algoritmo che "ricostruisce" attributi protetti non opera né consapevolmente né inconsapevolmente in senso proprio — opera secondo una logica che è strutturalmente cieca alla distinzione tra correlazione statistica e rilevanza normativa. La discriminazione che ne risulta non è il frutto di un'intenzione (neppure inconscia) né di una negligenza correggibile, ma di un'architettura cognitiva che tratta ogni regolarità statistica come potenzialmente predittiva, indipendentemente dal suo significato sociale o dalla sua ammissibilità giuridica.

Collocare il trattamento disparato inconsapevole nei *bias* architetturali significa quindi riconoscere che esso non rappresenta un fallimento dell'implementazione ma una conseguenza strutturale del modo in cui i sistemi algoritmici producono conoscenza e decisioni. I rimedi tecnici — dalla rimozione delle *features* correlate ai vincoli di *fairness* nelle funzioni obiettivo — possono attenuare il problema ma non eliminarlo, precisamente perché intervengono sui sintomi senza poter modificare l'architettura cognitiva che li genera. In questo senso, il trattamento disparato inconsapevole diventa un caso di studio privilegiato per comprendere la natura costitutiva della discriminazione algoritmica: una forma di discriminazione che emerge non malgrado, ma in virtù del funzionamento ordinario del sistema. Ma le nuove forme di discriminazione non si limitano a replicare le categorie tradizionali.

L'infrastruttura algoritmica sta generando nuove forme di marginalità che sfuggono alle classificazioni protette consolidate.

In un'economia predittiva, l'assenza di tracce digitali diventa svantaggiosa: chi non ha storia creditizia, profili social, o *pattern* di consumo tracciabili viene classificato come "non valutabile" e quindi rischioso[45]. Paradossalmente, l'assenza dai sistemi di sorveglianza commerciale diventa fonte di penalizzazione. Gli algoritmi sono addestrati su *pattern* comportamentali maggioritari e quindi accade che anche chi si discosta dalla norma statistica — per ragioni culturali, economiche, o semplicemente idiosincratiche — tende a essere mal classificato. Un consumatore che non segue i *pattern* d'acquisto tipici della sua fascia demografica può essere addirittura segnalato come potenzialmente fraudolento!

3. Il quadro giuridico e normativo

Il panorama giuridico della discriminazione algoritmica si configura oggi come un sistema multilivello in rapida evoluzione, che intreccia normative generali sulla protezione dei dati personali con strumenti specificamente dedicati all'intelligenza artificiale. Il GDPR (*General Data Protection Regulation*) rappresenta la prima pietra angolare di questo edificio normativo: sebbene non pensato originariamente per governare i sistemi di IA, esso introduce principi fondamentali. A questo si affianca la normativa antidiscriminatoria tradizionale, che vieta discriminazioni basate su caratteristiche protette. Tuttavia, l'applicazione di questi principi ai sistemi algoritmici rivela rapidamente i limiti di un approccio giuridico pensato per comportamenti umani: come identificare, ad esempio, l'intenzionalità discriminatoria in un sistema che apprende autonomamente da pattern nei dati? L'*AI Act* europeo segna una svolta paradigmatica nel tentativo di colmare questo vuoto regolatorio. La Legge italiana n. 132/2025 rappresenta un ulteriore momento di innovazione normativa.

3.1 Analisi critica del GDPR

Il GDPR[46] introduce un *corpus* di principi fondamentali: minimizzazione dei dati, liceità, correttezza e trasparenza. Tuttavia, emergono limiti significativi per la prevenzione della discriminazione algoritmica. L'articolo 22, che disciplina il processo decisionale automatizzato, stabilisce che «l'interessato ha il diritto di non essere

[45] Cfr. D.K. Citron, F.A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, in «Washington Law Review», cit.

[46] *Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati)*, in GU L 119 del 4.5.2016.

sottoposto a una decisione basata unicamente sul trattamento automatizzato che produca effetti giuridici significativi». Il divieto risulta però circoscritto ai soli sistemi completamente automatizzati, lasciando fuori i sistemi ibridi *human-in-the-loop* sempre più diffusi e quei sistemi che si avvalgono di variabili *proxy*.

Le eccezioni previste dall'articolo 22 contribuiscono ulteriormente a limitare l'efficacia del divieto, specialmente l'eccezione basata sul consenso esplicito dell'interessato, problematica nel contesto dei rapporti di forza asimmetrici che caratterizzano molte interazioni digitali: un cittadino che richiede un prestito difficilmente può rifiutare il trattamento automatizzato senza compromettere le proprie possibilità di ottenere quanto richiesto. Una criticità fondamentale riguarda l'impianto stesso della tutela offerta dal GDPR, che rimane ancorata a un paradigma individualista. Il *Regolamento* è costruito attorno ai diritti del singolo interessato: diritto di accesso, alla rettifica, alla cancellazione, alla portabilità. Questo approccio si rivela inadeguato per affrontare la discriminazione sistemica, che per sua natura colpisce interi gruppi sociali. Emerge pertanto la necessità di sviluppare una sfera di protezione collettiva che includa meccanismi di *class action*, sistemi di audit algoritmico indipendente e obblighi di trasparenza aggregata.

3.2 Il Regolamento (UE) 2024/1689 (*AI Act*)

L'*AI Act*, entrato in vigore il 1° agosto 2024[47]. rappresenta un pilastro normativo fondamentale nella regolamentazione globale dell'intelligenza artificiale. Il *Regolamento* europeo è il primo tentativo organico e sistematico di affrontare i rischi generati dai sistemi di IA attraverso un quadro giuridico vincolante, adottando un approccio innovativo basato sulla stratificazione del rischio. Questa metodologia *risk-based* consente di calibrare l'intensità degli obblighi normativi in funzione della pericolosità potenziale di ciascun sistema. L'architettura del *Regolamento* si articola su quattro livelli principali di rischio: inaccettabile, alto, limitato e minimo.

Per i sistemi considerati a rischio inaccettabile, l'*AI Act* introduce un divieto assoluto e incondizionato. Rientrano in questa categoria i sistemi di "punteggio sociale" (*social scoring*) da parte di governi o imprese[48], i sistemi di categorizzazione biometrica che deducono dati sensibili, e l'uso non mirato di sistemi di identificazione biometrica in tempo reale in spazi pubblici.

[47] *Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio del 13 giugno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale e che modifica i regolamenti (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (regolamento sull'intelligenza artificiale)*, in GU L del 12.7.2024.

[48] Cfr. Id., articolo 5, paragrafo 1, lettere a), b) e d).

Una disciplina particolarmente articolata è riservata ai sistemi di IA ad alto rischio[49], che comprendono biometria, istruzione, occupazione, *credit scoring*, servizi di emergenza, giustizia e processi democratici. Per questi sistemi è richiesto un sistema di gestione del rischio iterativo: il *Regolamento* impone un *corpus* articolato di requisiti che coprono l'intero ciclo di vita del sistema, dalla progettazione alla dismissione. In primo luogo, i fornitori devono istituire un sistema di gestione del rischio che identifichi e analizzi i rischi noti e ragionevolmente prevedibili, inclusi quelli relativi a potenziali *bias*, e adotti misure appropriate di gestione e mitigazione[50]. Inoltre viene introdotto l'obbligo che i *set* di dati siano «pertinenti, rappresentativi, esenti da errori e completi»[51]. Un elemento di particolare rilievo innovativo introdotto dal *Regolamento* europeo sull'intelligenza artificiale riguarda l'obbligo, posto in capo ai soggetti che impiegano sistemi di IA ad alto rischio (i cosiddetti *deployer*), di condurre una Valutazione d'Impatto sui Diritti Fondamentali — indicata con l'acronimo FRIA, dall'inglese *Fundamental Rights Impact Assessment* — prima che il sistema venga effettivamente messo in servizio[52].

Questa previsione segna un cambio di paradigma rispetto agli approcci regolatori precedenti, che tendevano a concentrare gli obblighi di conformità quasi esclusivamente sui produttori (*provider*) dei sistemi. L'*AI Act* riconosce invece che anche chi utilizza un sistema di IA in contesti operativi concreti — un'azienda che adotta un algoritmo di selezione del personale, un ente pubblico che impiega strumenti di *scoring* per l'accesso a prestazioni sociali, un'istituzione sanitaria che si avvale di sistemi diagnostici automatizzati — contribuisce in modo determinante a definire l'impatto effettivo di quella tecnologia sui diritti delle persone coinvolte.

La FRIA richiede al *deployer* di analizzare sistematicamente, e documentare in modo tracciabile, i rischi che l'impiego del sistema può comportare per i diritti fondamentali tutelati dalla *Carta* dell'Unione europea: dalla non discriminazione alla protezione dei dati personali, dalla dignità umana al diritto a un ricorso effettivo. Questa valutazione deve essere condotta *ex ante*, ossia prima dell'effettiva attivazione del sistema, configurandosi così come uno strumento di prevenzione piuttosto che di mera reazione a danni già verificatisi.

[49] Cfr. Id., allegato III.

[50] Cfr. Id., articolo 9. Si veda anche: European Commission, *High-Level Expert Group on Artificial Intelligence - Ethics Guidelines for Trustworthy AI*, 2019.

[51] Cfr. articolo 10 del *Regolamento (UE) 2024/1689*. Sul tema della qualità dei dati, si veda S. Barocas, M. Hardt, A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, MIT Press, 2023.

[52] Cfr. articolo 27 del *Regolamento (UE) 2024/1689*.

3.3. La *Legge n. 132/2025* (Italia)

La *Legge n. 132/2025,* entrata in vigore il 10 ottobre 2025[53], rappresenta il primo intervento normativo organico con cui l'ordinamento italiano affronta le sfide poste dall'intelligenza artificiale. Il provvedimento si qualifica come proattivo sotto un duplice profilo: da un lato, l'Italia è stata tra i primi Stati membri a tradurre in disciplina nazionale i principi del *Regolamento europeo 2024/1689* (il già precedentemente analizzato *AI Act*); dall'altro, il legislatore non si è limitato a un mero recepimento, ma ha cercato di elaborare una cornice regolatoria che declina quei principi nelle specificità del sistema giuridico e produttivo nazionale.

L'architettura della legge poggia su un'impostazione dichiaratamente antropocentrica[54]: i sistemi di intelligenza artificiale vengono concepiti come strumenti al servizio della persona, mai sostitutivi del giudizio e della responsabilità umana. Questo orientamento si traduce in prescrizioni puntuali per settori strategici — dalla sanità al lavoro, dalle professioni intellettuali alla pubblica amministrazione — nei quali l'impiego dell'IA deve coniugarsi con trasparenza, non discriminazione e tutela dei diritti fondamentali.

Sul piano della *governance*, la legge istituisce un sistema duale: all'Agenzia per l'Italia Digitale spettano funzioni di promozione e sviluppo, mentre l'Agenzia per la Cybersicurezza Nazionale assume compiti di vigilanza e sicurezza. La norma, tuttavia, presenta una struttura programmatica che affida a futuri decreti legislativi la definizione di aspetti cruciali, dalla responsabilità civile alla disciplina dell'addestramento algoritmico, configurandosi così come un punto di partenza più che come un approdo definitivo nel governo della trasformazione digitale.

Essa si pone in rapporto di complementarietà con l'*AI Act* e il GDPR. Nell'ambito sanitario, l'articolo 7 introduce un «divieto esplicito e rafforzato di discriminazione» e riafferma il principio del *human-in-command*[55]: il medico non può essere ridotto a "esecutore" delle indicazioni di un sistema di IA, ma deve mantenere la piena responsabilità decisionale e la possibilità di discostarsi dalle raccomandazioni algoritmiche.

Per i contratti pubblici, l'articolo 5[56] introduce criteri vincolanti: sicurezza, trasparenza, non discriminazione e proporzionalità. La legge consolida la dottrina dell'"amministrazione algoritmica responsabile": obbligo di motivazione aumentata

[53] *Legge n. 132/2025*, *Disposizioni e deleghe del governo in materia di intelligenza artificiale*, in GU Serie Generale n. 237 del 10.10.2025.

[54] Cfr. articolo 1, comma 2 della Legge n. 132/2025. L'approccio antropocentrico è coerente con la *Carta dei diritti fondamentali dell'Unione europea*.

[55] Articolo 7 della *Legge n. 132/2025*. Sul principio di *human-in-command* in ambito sanitario cfr. Council of Europe, *Recommendation CM/Rec(2019)2 on the protection of health-related data*, 2019.

[56] Cfr. articolo 5, comma 1, lettera d) della *Legge n. 132/2025*.

per le decisioni algoritmiche[57] e trasparenza algoritmica come condizione essenziale di giustiziabilità. L'approccio della *Legge 132/2025*, richiamando la *Costituzione* e in particolare gli articoli 2, 3 e 32, rivela come la discriminazione algoritmica debba essere riconosciuta come ostacolo concreto all'eguaglianza sostanziale che la Repubblica è chiamata a rimuovere.

4. Strategie per valutare e mitigare le discriminazioni algoritmiche

La mitigazione efficace della discriminazione algoritmica richiede un ripensamento radicale degli approcci tradizionali, tipicamente frammentati lungo linee disciplinari che separano artificialmente ciò che nella realtà si presenta come un fenomeno unitario e complesso. Il *bias* algoritmico, come abbiamo evidenziato, non è un difetto tecnico isolabile e correggibile attraverso interventi puntuali sul codice o sui dati; è piuttosto l'emergenza computazionale di dinamiche sociali, economiche e istituzionali stratificate nel tempo, che trovano negli algoritmi un nuovo medium di espressione e, potenzialmente, di amplificazione.

Questa consapevolezza impone di superare la tentazione riduzionistica che caratterizza molte proposte di intervento. Gli informatici tendono a concepire il problema come una questione di ottimizzazione sotto vincoli, risolvibile attraverso la scelta della metrica di equità appropriata e l'implementazione di tecniche di *debiasing*. I giuristi lo inquadrano prevalentemente come una questione di conformità normativa e responsabilità civile. I sociologi lo analizzano come manifestazione di disuguaglianze strutturali. Gli economisti lo esaminano attraverso la lente dell'efficienza allocativa e delle esternalità negative. Ciascuna di queste prospettive cattura aspetti genuini del fenomeno, ma nessuna, presa isolatamente, ne esaurisce la complessità.

Un approccio autenticamente unitario richiede invece di riconoscere che il *bias* algoritmico è un fenomeno socio-tecnico nel senso più pieno del termine: emerge dall'interazione ricorsiva tra artefatti tecnologici e contesti sociali, in modo tale che né la dimensione tecnica né quella sociale possono essere comprese indipendentemente l'una dall'altra. Gli algoritmi sono progettati da esseri umani situati in specifici contesti organizzativi e culturali, addestrati su dati che riflettono pratiche sociali pregresse, implementati in istituzioni con proprie logiche e dinamiche di potere, e utilizzati da operatori che interpretano e mediano i loro *output* secondo schemi cognitivi e normativi preesistenti. Intervenire efficacemente richiede di agire simultaneamente su tutti questi livelli, riconoscendo le interdipendenze che li collegano.

[57] Cfr. D.U. Galetta, *Algoritmi, procedimento amministrativo e garanzie: brevi riflessioni, anche alla luce della esperienza spagnola*, in «Rivista italiana di diritto pubblico comunitario», 2020, n. 4, pp. 501-516.

Gli interventi tecnici più sofisticati possono rivelarsi inefficaci se non accompagnati da trasformazioni organizzative che modifichino le culture aziendali e le dinamiche di potere che determinano come gli algoritmi vengono progettati, implementati e governati. È proprio in questa consapevolezza della complessità del fenomeno che si articolano le strategie chiave per valutare e mitigare le discriminazioni algoritmiche:

- Miglioramento della qualità e rappresentatività dei dati. Esso rappresenta la prima strategia fondamentale, radicata nella consapevolezza espressa dall'adagio informatico *garbage in, garbage out* (GIGO). I meccanismi includono tecniche di pulizia dei dati, bilanciamento dei *dataset* attraverso sottocampionamento (*undersampling*) o sovracampionamento (*oversampling*), e stratificazione demografica. Tali pratiche trovano esplicito riconoscimento normativo nei requisiti di *data quality* stabiliti dall'*AI Act* (articolo 10).
- La progettazione di *algoritmi fairness-aware i*ntegra l'equità direttamente nel design algoritmico. Questo approccio si avvale di metriche come la *demographic parity* (che richiede proporzioni uguali di classificazioni positive tra gruppi), l'*equal opportunity* (che richiede uguali tassi di veri positivi) e l'*equalized odds* (che richiede sia uguali tassi di veri positivi sia di falsi positivi). I teoremi di impossibilità dimostrano che queste diverse nozioni sono matematicamente incompatibili, rendendo inevitabile che la scelta privilegi alcuni valori normativi a scapito di altri. *Framework* come la ISO/IEC 42001:2023[58] forniscono linee guida per l'implementazione di pratiche di AI responsabile.
- L'auditing algoritmico e la supervisione umana. Essi garantiscono tracciabilità e *accountability* attraverso esame sistematico e indipendente di molteplici dimensioni: qualità e rappresentatività dei dati, logica computazionale, risultati prodotti su diverse sotto-popolazioni. Per condurre audit efficaci sono cruciali le tecniche di *Explainable AI* (XAI) come LIME e SHAP che permettono di identificare quali variabili influenzano maggiormente le decisioni algoritmiche. La supervisione umana (*Human-in-the-Loop* o *Human-on-the-Loop*) trova riconoscimento nell'*AI Act* (articoli 14 e 29)[59].
- La formazione e sensibilizzazione è cruciale per sviluppare un'*algorithmic literacy* diffusa. È necessario formare sviluppatori, *data scientist*, utilizzatori dei sistemi e operatori del diritto su rischi e strategie di mitigazione del *bias* algoritmico. L'*AI Act* prevede esplicitamente obblighi di formazione per gli operatori dei sistemi ad alto rischio[60].

[58] ISO/IEC 42001:2023, *Information technology — Artificial intelligence — Management system*, International Organization for Standardization, 2023.
[59] *Regolamento (UE) 2024/1689*, artt. 14 e 29.
[60] Id., art. 4.

5. Conclusioni: verso una regolamentazione responsabile e centrata sull'uomo

L'intelligenza artificiale ha operato una trasformazione radicale nella fenomenologia della discriminazione, introducendo modalità in cui l'ingiustizia non è più solo il prodotto di pregiudizi umani espliciti, ma emerge dalla materialità stessa dei sistemi computazionali o da complesse interazioni tra agire umano e processamento artificiale. Questa trasformazione richiede di ripensare categorie consolidate: non siamo più di fronte esclusivamente alla discriminazione come atto intenzionale di un soggetto identificabile, ma a forme di ingiustizia che possono radicarsi nell'architettura stessa dei sistemi algoritmici.

L'immagine dell'algoritmo come "ombra umana" — che obbedisce rigorosamente alle istruzioni predefinite dai suoi programmatori — rappresenta una metafora epistemologicamente potente ma che necessita di essere problematizzata. Se l'algoritmo fosse mera ombra, seguirebbe fedelmente i contorni di chi lo ha creato, amplificando e distorcendo i *bias* preesistenti secondo una logica di semplice riproduzione. Tuttavia, questa concezione rischia di oscurare la dimensione propriamente costitutiva della discriminazione algoritmica. Gli algoritmi non si limitano a riflettere passivamente le strutture discriminatorie della società: essi le trasformano, le consolidano secondo logiche proprie, le rendono operative in modi che sfuggono spesso alla consapevolezza stessa dei loro creatori. La metafora dell'ombra coglie la continuità tra *bias* umano e *bias* algoritmico, ma non rende conto della novità qualitativa introdotta dalla mediazione computazionale: l'opacità dei processi decisionali, la scala dell'impatto, l'impossibilità di ricondurre univocamente gli esiti a intenzioni umane specifiche.

Riconoscere questa complessità sposta decisamente il problema da quella che potrebbe apparire come un'inevitabilità tecnologica a una questione fondamentale di responsabilità umana e sistemica. Non è sufficiente identificare il *bias* come residuo tecnico eliminabile attraverso migliori procedure di training dei modelli. La discriminazione algoritmica ci confronta con una forma di responsabilità diffusa: essa coinvolge certamente i progettisti e gli sviluppatori, ma anche le organizzazioni che implementano i sistemi, le strutture sociali che producono i dati su cui gli algoritmi si addestrano, le scelte politiche ed economiche che orientano lo sviluppo tecnologico. La responsabilità diventa così una categoria che deve essere ricostruita per comprendere forme di *agency* distribuita tra attori umani e non-umani.

L'emergere della discriminazione algoritmica si caratterizza inoltre per la sua frequente invisibilità rispetto all'individuo che ne subisce gli effetti. A differenza della discriminazione tradizionale, dove la vittima può almeno in linea di principio riconoscere l'atto discriminatorio e identificarne l'autore, la discriminazione algoritmica opera spesso silenziosamente, attraverso decisioni automatizzate che sfuggono alla percezione e al controllo dei soggetti interessati. È precisamente in risposta a questa sfida che assume rilevanza cruciale il principio del "comando

umano" (*human-in-command*). Non si tratta semplicemente di mantenere un essere umano "nel ciclo" decisionale (*human-in-the-loop*), presenza che può ridursi a ratifica formale di decisioni già predeterminate algoritmicamente. Il comando umano implica che gli esseri umani mantengano il controllo ultimo, sostanziale e non meramente procedurale, sulle decisioni che incidono significativamente sulle loro vite.

La lotta contro il *bias* algoritmico non può dunque limitarsi a soluzioni meramente tecniche, a una logica che potremmo definire di "soluzionismo tecnologico" — l'illusione che ogni problema sociale possa trovare risposta in un migliore design algoritmico, in tecniche più raffinate di *debiasing*, in metriche di *fairness* più sofisticate. Pur riconoscendo l'importanza di questi interventi tecnici, è necessario comprendere che il *bias* algoritmico non è un'anomalia correggibile attraverso migliori procedure ingegneristiche, ma il sintomo di ingiustizie strutturali radicate nelle nostre società. Gli algoritmi apprendono da dati che incorporano le disuguaglianze esistenti, vengono sviluppati in contesti organizzativi che riflettono squilibri di potere, sono implementati in sistemi sociali già attraversati da discriminazioni sistemiche.

Serve quindi un ripensamento profondo delle strutture sociali e culturali che generano discriminazione. La diversificazione dei *team* di sviluppo non è una mera questione di rappresentanza simbolica: *team* più inclusivi portano prospettive epistemiche diverse, sono più capaci di identificare *bias* che risulterebbero invisibili a gruppi omogenei. Il coinvolgimento significativo delle comunità impattate non è solo un esercizio di consultazione, ma deve tradursi in forme di co-progettazione che possano ridistribuire il potere decisionale sullo sviluppo tecnologico. La promozione di un'alfabetizzazione algoritmica diffusa (*algorithmic literacy*) è condizione necessaria perché i cittadini possano esercitare un controllo democratico su sistemi che sempre più mediano l'accesso a diritti e opportunità fondamentali.

Le normative recenti, come il *Regolamento europeo sull'intelligenza artificiale* (*AI Act*) e la *Legge italiana n. 132/2025,* rappresentano in questa prospettiva passi cruciali verso un nuovo paradigma di governance dell'IA. Questi strumenti normativi non si limitano a sanzionare *ex post* le pratiche discriminatorie, ma impongono vincoli di equità proattivi: obblighi di valutazione d'impatto sui diritti fondamentali, requisiti di trasparenza e documentazione, meccanismi di supervisione umana, procedure di monitoraggio continuo. Creano un quadro di *governance* che, almeno nelle sue intenzioni, mira a trasformare profondamente il modo in cui i sistemi di IA vengono concepiti, sviluppati e implementati. L'obiettivo non è soffocare l'innovazione attraverso vincoli burocratici, ma indirizzarla verso forme di responsabilità sociale, trasformando la non-discriminazione da costo da minimizzare a elemento di differenziazione competitiva e di costruzione di fiducia pubblica.

In definitiva, la sfida che ci troviamo ad affrontare non è semplicemente tecnica o regolatoria, ma profondamente politica e culturale: si tratta di costruire sistemi di intelligenza artificiale che riflettano autenticamente i valori fondamentali di eguaglianza, giustizia e dignità umana che caratterizzano le democrazie costituzionali

contemporanee. Questo richiede che l'IA sia progettata e governata in modo da espandere le opportunità per tutti, di ampliare gli spazi di libertà e autodeterminazione, di rafforzare piuttosto che erodere i diritti fondamentali.