CAMMINO DIRITTO
Rivista di informazione giuridica
https://rivista.camminodiritto.it

Articolo Scientifico

IL BIAS ALGORITMICO: NUOVE FRONTIERE DI VIOLAZIONE DEI DIRITTI FONDAMENTALI

Il bias algoritmico costituisce un rischio strutturale dei sistemi di intelligenza artificiale, capace di generare effetti discriminatori e compromettere i diritti fondamentali. L'analisi distingue tra le diverse tipologie di bias a cui si può incorrere quando si opera con sistemi di machine learning e che, se connesse all'opacità algoritmica, possono rappresentare fonti di distorsioni capaci di incidere direttamente sui diritti dei destinatari, soprattutto nei casi di scelte automatizzate. Le strategie di bias mitigation offrono soluzioni parziali, senza eliminare del tutto le distorsioni, facendo emergere l'esigenza di affiancare misure tecniche a garanzie legislative e giurisdizionali per preservare trasparenza, dignità e pluralismo democratico.

di **Andrea Racca**
SPS/08 - SOCIOLOGIA DEI PROCESSI CULTURALI E COMUNICATIVI
Estratto dal n. 11/2025 - ISSN 2532-9871

Direttore responsabile
Alessio Giaquinto

Pubblicato, Venerdì 7 Novembre 2025

CamminoDiritto.it
ISSN 2532-9871

Abstract ENG

Algorithmic bias is a structural risk inherent in artificial intelligence systems, capable of generating discriminatory effects and undermining fundamental rights. This study distinguishes between different types of bias that may arise when operating with machine learning systems and which, when linked to algorithmic opacity, can represent sources of distortion directly affecting individuals' rights, especially in cases of automated decision-making. Bias mitigation strategies provide only partial remedies, without fully eliminating distortions. From a legal and constitutional perspective, there is a need to complement technical measures with legislative and judicial safeguards in order to preserve transparency, dignity, and democratic pluralism.

Sommario: 1. Premesse; 2. Il bias algoritmico come fattore di rischio; 3. Opacità algoritmica e rischio discriminatorio; 4. Strategie di Bias Mitigation; 5. Riflessioni interlocutorie; 6. Conclusioni.

1. Premesse

L'adozione crescente di sistemi di intelligenza artificiale e di algoritmi di machine learning, in moltissime applicazioni di uso quotidiano e ad ampia accessibilità, sta determinando un ricorso indiscriminato a questi applicativi, che, tuttavia, non sono esenti da rischi. Se, da un lato, promettono efficienza, rapidità e capacità predittiva, dall'altro sono in grado di riprodurre e anche amplificare disuguaglianze e classificazioni discriminanti, che possono risultare in contrasto con i diritti fondamentali così come arduamente conquistati. Un'approfondita analisi dei criteri di calcolo di questi software rileva, infatti, la presenza di distorsioni sistematiche (bias algoritmici), che tendono a trattare i dati personali acquisiti in enormi archiviazioni come fattori slegati dalle individualità a cui appartengono, mettendo in serio pericolo la dignità umana.

Analizzare il fenomeno del c.d. bias algoritmico significa, quindi, muoversi su un duplice crinale: da un lato, comprendere le logiche tecniche che ne determinano l'insorgenza e le difficoltà di mitigazione delle distorsioni di calcolo; dall'altro, interrogarsi sugli strumenti giuridici e istituzionali capaci di prevenirne o contenerne gli effetti discriminatori. Questa prospettiva, comparata e interdisciplinare, rappresenta un imprescindibile punto di vista critico sul sempre più massiccio utilizzo dei sistemi di intelligenza artificiale, al fine di evitare che le nuove frontiere dell'informatica applicata possano tradursi in inconsapevoli (o meno) violazioni dei diritti fondamentali.

CamminoDiritto.it
ISSN 2532-9871

2. Il bias algoritmico come fattore di rischio

Da questo punto di vista, occorre premettere che la crescente complessità delle tecnologie informatiche impone una revisione approfondita delle valutazioni di impatto sulla protezione dei dati, rendendole più sofisticate e preventive, in modo da minimizzare i rischi per i diritti e le libertà fondamentali dell'individuo[1]. L'innovazione dovrebbe essere accompagnata, infatti, da un'etica della telematica che garantisca sempre il rispetto della dignità umana, anche in ambiti dominati dall'intelligenza artificiale e dai sistemi di automazione decisionale[2]. In effetti, l'ambito relativo alla tutela delle informazioni assume sempre più un valore strategico, proprio in conseguenza del crescente "valore" dei dati archiviati negli enormi server delle Big Tech companies, tanto che, non infondatamente, è stato osservato come, ormai, in una scala di importanza, esse rivestano non di rado una valenza perfino superiore a quella attribuibile al possesso di determinate materie prime, potendo essere considerate alla stregua del nuovo "oro nero"[3]. In tale contesto, risulta imprescindibile promuovere l'inclusione digitale e l'accesso responsabile alle tecnologie, per evitare nuove forme di esclusione sociale e garantire a tutti una reale partecipazione democratica.

Il ruolo delle tecnologie virtuali e predittive apre, infatti, nuove dimensioni dell'identità, la quale dovrà essere tutelata con la stessa attenzione ai principi di libertà e di autodeterminazione. Come evidenziato dalla sociologa statunitense Shoshana Zuboff, è necessario regolamentare attentamente le potenti capacità di profilazione e di controllo sociale derivanti da queste applicazioni, poiché è immanente il rischio di una perdita di sovranità individuale, con conseguente nuova forma di schiavitù, attraverso un giogo virtuale[4]. L'offerta di strumenti, fino a pochi anni fa inconcepibili, pone, infatti, il rischio di nuove dipendenze da questi applicativi, tanto che recentemente si parla di FoMO (Fear of Missing out), quale sindrome di dipendenza dai social media. La pervasività degli applicativi, unita alla funzione predittiva, possono così incidere direttamente sui processi decisionali; occorrono pertanto norme di trasparenza, spiegabilità e controllo umano rafforzate, isolando, tra l'altro, i rischi di sistemi autonomi che operino senza adeguata supervisione, in scenari apocalittici "alla Matrix".

Se si condividono le riflessioni della matematica statunitense Cathy O'Neil, che avverte sui pericoli di algoritmi opachi e non regolamentati, proponendo principi di responsabilità e accountability mirati alla tutela dell'etica e della dignità umana[5], anche nel c.d. metaverso, risulta quanto mai necessario comprendere ciò che, dal punto di vista tecnico, viene definito come bias algoritmico, ovvero la presenza di una distorsione sistematica - dunque non casuale - che altera l'output prodotto dal calcolo algoritmico, determinando effetti discriminatori o comunque ingiusti nei confronti di determinati soggetti o categorie di soggetti, che ne subiscono il risultato, anche senza che questo determini a sua volta decisioni automatizzate. Il concetto di bias rappresenta, dunque, uno scostamento

CamminoDiritto.it
ISSN 2532-9871

sistematico tra la stima prodotta da un campione e il valore reale della popolazione, che nel contesto del machine learning - ovvero degli algoritmi di intelligenza artificiale - designa un insieme di fenomeni che generano distorsioni negli output (ovvero nei risultati dei processi automatizzati), indipendentemente dall'affidabilità formale degli algoritmi utilizzati. In sostanza, si producono risultati non in linea con l'addestramento, con la definizione delle variabili, o persino con l'impiego operativo del modello. Dunque, il bias consiste nella riproduzione di un errore non occasionale del processo di calcolo, che si contraddistingue a seconda degli effetti che produce:

Il data bias: si verifica quando i dataset (ovvero le raccolte di dati) di addestramento non sono rappresentativi della popolazione di riferimento o risultano sbilanciati nella distribuzione di alcune variabili. Questo fenomeno è frequente nei sistemi di visione artificiale e nel riconoscimento facciale, dove la prevalenza di immagini di soggetti caucasici ha prodotto un aumento significativo degli errori di classificazione nel riconoscimento di individui con tonalità di pelle più scura o di donne rispetto agli uomini. L'acquisizione dei dati e il successivo addestramento sono avvenuti in un contesto non omogeneo e quindi il risultato risulta conseguentemente disorto dall'input (non omogeneo), riflettendo effetti discriminatori nei risultati. Alcuni autori come Buolamwini e Gebru hanno dimostrato che alcuni dei sistemi commerciali più diffusi raggiungevano tassi di errore inferiori all'1% nel riconoscimento di volti maschili caucasici, ma oltre il 30% per volti femminili con pelle scura[6]. In sostanza, il problema non risiede nell'algoritmo in sé, bensì nella distribuzione dei dati utilizzati in ingresso, che riflettono una realtà parziale e sbilanciata. Il prejudice bias**:** si manifesta quando il trasferimento, intenzionale o meno, di pregiudizi umani all'interno dei modelli, determina una categorizzazione discriminatoria. Ciò avviene soprattutto nella fase di etichettatura o di selezione delle variabili. Se, ad esempio, un dataset viene annotato con categorie preconcette, l'algoritmo ne assimilerà la logica, anche se il calcolo formalmente risulta neutrale. Questo tipo di bias è particolarmente evidente nei sistemi di natural language processing, laddove i modelli di linguaggio hanno dimostrato di interiorizzare stereotipi di genere: termini come "uomo" risultano più frequentemente associati a ruoli professionali di prestigio, mentre "donna" tende a comparire in contesti domestici o familiari[7]. I pregiudizi culturali incorporati nei dati testuali finiscono così per tradursi in associazioni semantiche, che il modello riproduce ed amplifica a seconda delle operazioni cui viene applicato. Il measurement bias**:** rappresenta, invece, una distorsione che emerge dalla scelta di variabili surrogate o da proxy inadeguati. Questo significa che, nei sistemi di analisi algoritmica, molte volte non risulta possibile misurare direttamente la caratteristica d'interesse e si ricorre a indicatori indiretti, che possono costituire distorsioni sistematiche. Un esempio noto è rappresentato dai sistemi di valutazione del rischio recidivale, nei quali il luogo di residenza o il numero di arresti precedenti sono stati utilizzati come variabili predittive. Tali scelte, pur statisticamente correlate al fenomeno da stimare, non riflettono necessariamente le cause reali e possono portare a sovrastimare il rischio per determinati gruppi. Questo tipo di bias è tecnicamente complesso e

CamminoDiritto.it
ISSN 2532-9871

rischioso, proprio perché non dipende dall'asimmetria dei dati, ma dal modo in cui le grandezze sono rappresentate e trattate secondo schemi non direttamente oggettivabili. L'aggregation bias: si verifica quando un modello ignora la struttura eterogenea dei dati alla base e cerca di adattare un'unica funzione predittiva a popolazioni che presentano sottogruppi distinti con caratteristiche specifiche. Ad esempio, nella diagnostica medica, se il machine learning viene allenato a ricercare una determinata patologia utilizzando un dataset molto ampio, ove i pazienti non sono omogenei per genere o età (ad esempio, con una maggioranza di uomini), il modello presenterà performance migliori sul dato con quantità maggioritaria, mentre la qualità del risultato sarà peggiore sulla popolazione di minor presenza. Questo tipo di bias emerge dall'errata ipotesi di omogeneità dei dati in input, che porta a costruire modelli "medi" poco efficaci nelle periferie della popolazione di calcolo[8]. L'evaluation bias: molto simile all'aggregation, riguarda la mancata considerazione dell'eterogeneità del dataset nella fase di test e di validazione dei modelli, trasponendo negli output una falsa percezione di affidabilità, in quanto le metriche globali nascondono differenze significative nelle prestazioni[9], ovvero le operazioni di calcolo non tengono conto di tutte le sottovariabili. Il problema si radica, dunque, sul trattamento algoritmico, in quanto, un modello che raggiunge una buona accuratezza globale può presentare forti distorsioni di performance su categorie minoritarie, offrendo risultati completamente inattendibili sui dati lontani dal dato medio. Se le problematiche di natura metrica, ovvero di calcolo su ampi dataset, possono offrire già rischi discriminatori, ad aggravare il sistema si aggiungono poi i c.d. bias storico-sociali, in particolare se si ritiene che i dati in input possano presentare già al loro interno diseguaglianze e discriminazioni consolidate nella società. Ad esempio, un sistema di selezione del personale che si basa su dati storici tenderà a privilegiare candidati simili a quelli assunti in passato, replicando modelli già acquisiti (historical bias). Se, invece, la costruzione del modello riflette valori culturali, normativi o linguistici di un contesto di riferimento, l'estensione dell'applicativo al di fuori del contesto di progettazione offrirà valori totalmente discriminanti, posto che il modello non risulta adeguato a dataset acquisiti fuori dal contesto di progettazione (societal bias). Queste forme sono particolarmente evidenti nei dataset linguistici, nei quali le strutture grammaticali e semantiche di una lingua riflettono visioni del mondo peculiari, difficilmente trasferibili a sistemi globali di IA[10].

A tutto ciò, occorre ancora aggiungere i bias derivanti dalle funzioni di filtraggio acquisitivo, ovvero i c.d. emergent bias, che rappresentano distorsioni derivanti dalla fase di apprendimento, ove il sistema non risulta in grado di scartare i dati distonici che emergono dall'interazione tra il sistema e l'ambiente sociale in cui opera. Il caso maggiormente emblematico è sicuramente quello della chatbot Tay di Microsoft, lanciata su Twitter nel 2016: il programma di language processing ideato per apprendere funzioni linguistiche dai dialoghi con gli utenti, venne disattivato dopo appena 16 ore di attività, poiché, a seguito di input razzisti, misogini, complottisti e violenti da parte di gruppi di utenti di Twitter (troll, provocatori, ma anche comunità coordinate), l'applicativo iniziò a riprodurli con ridondanza[11]. Questo fenomeno mostra come l'apprendimento in tempo

CamminoDiritto.it
ISSN 2532-9871

reale, se non controllato, possa trasformarsi in una spirale di bias crescente, con effetti difficilmente prevedibili al momento della progettazione[12].

In definitiva, il bias algoritmico è un fenomeno multidimensionale, che può manifestarsi in tutte le fasi del ciclo di vita dei modelli di intelligenza artificiale e che non si esime neppure nei sistemi più sofisticati. Si tratta, infatti, di operazioni, che trovano i propri limiti sia nella fase di acquisizione, ovvero di input dei dati, sia nella fase di calcolo, che, per quanto complessa, non riesce mai a replicare appieno la complessità del reale. In definitiva, non esiste un'unica causa del bias algoritmico, ma una combinazione di fattori legati ai dati, alle scelte di progettazione, ai metodi di valutazione e all'interazione con l'ambiente. La natura sistematica e persistente del bias ne fa un elemento strutturale, non eliminabile ma, semmai, solo mitigabile attraverso strategie di controllo dei dati, di validazione su sottogruppi e di monitoraggio continuo dei sistemi in esercizio, che possono presentare sempre il rischio di non adeguatezza, con conseguente effetto discriminatorio per i destinatari delle operazioni di calcolo, tanto più se queste vengono utilizzate per attuare decisioni automatizzate. Questo rischio risulta, per di più duplice, poiché può condurre sia a una discriminazione diretta - ovvero quando un sistema utilizza variabili sensibili (genere, razza, etnia, orientamento religioso) come criterio di profilazione - sia a una discriminazione indiretta, quando variabili apparentemente neutrali (es. codice postale o cronologia di consumo) fungono tuttavia da proxy di categorie protette[13].

La mancata comprensione dell'effetto discriminatorio nelle operazioni di machine learning, che riproducono ripetitivamente operazioni di calcolo, può avere effetti devastanti, alimentando il bias lungo tutto il processo decisionale e amplificando disuguaglianze già esistenti; la dottrina configura, in questi casi, quello che rappresenta il rischio più grave di discriminazione sistemica.

3. Opacità algoritmica e rischio discriminatorio

Dopo aver delineato, dunque, il concetto di bias algoritmico, occorre soffermarci sui casi in cui i risultati di un machine learning non risultano comprensibili, accessibili o interpretabili da parte di utenti, ricercatori, autorità di controllo o persino dagli stessi sviluppatori. In questi casi si parla espressamente di "opacità algoritmica", che può manifestarsi in tre dimensioni: i) l'opacità intenzionale, derivante dalla scelta di proteggere il codice sorgente e i dataset per ragioni economiche e di segreto industriale; ii) l'opacità tecnica, determinata dalla crescente complessità dei modelli di deep learning, che operano come vere e proprie "scatole nere" matematiche; e, infine, iii) l'opacità interpretativa, dovuta al divario cognitivo che separa gli sviluppatori dagli utenti comuni [14].

CamminoDiritto.it
ISSN 2532-9871

Queste forme di oscurità non sono meri limiti tecnici, ma diventano fattori di rischio discriminatorio, se presenti in sistemi di decisione automatizzata. Da questo punto di vista, va rilevato che la Corte di giustizia dell'Unione europea, nel caso SCHUFA[15] (C-634/21, 2023), ha espressamente riconosciuto che l'uso determinante di uno score creditizio costituisce "decisione automatizzata" ai sensi dell'art. 22 GDPR, evidenziando la necessità di presidiare i rischi discriminatori delle operazioni di calcolo attuate. Laddove un algoritmo produca, infatti, output distorti a causa di dataset sbilanciati o variabili proxy inadatte (bias), l'oscurità impedisce di accorgersi di tali distorsioni. Quello che rimane invisibile risulta, di fatto, incontestabile: se un candidato viene escluso da un processo di selezione automatizzato o un imputato viene classificato ad "alto rischio" da un software predittivo, l'assenza di trasparenza rende quasi impossibile contestare la legittimità della valutazione[16]. La dottrina ha sottolineato come l'uso di algoritmi in ambiti come giustizia penale e welfare rischi di trasformare il cittadino in "oggetto di calcolo" privo di possibilità di contestazione effettiva[17].

La connessione tra oscurità e discriminazione emerge con particolare chiarezza nel caso del software COMPAS[18], utilizzato negli Stati Uniti per valutare la probabilità di recidiva degli imputati, che ha dimostrato come un algoritmo possa classificare in maniera sistematicamente più severa gli afroamericani rispetto ai bianchi. La caratteristica più preoccupante non era soltanto la presenza di bias nei dati di addestramento, ma l'impossibilità, per imputati e giudici, di accedere ai criteri decisionali, in quanto protetti da segreto industriale. L'oscurità ha, dunque, amplificato l'impatto discriminatorio, sottraendo la verifica dei dati al controllo democratico. Allo stesso modo, sistemi di riconoscimento facciale hanno mostrato tassi di errore significativamente più alti per le donne e per le persone con pelle scura rispetto agli uomini caucasici. Anche qui, l'oscurità algoritmica ha reso difficile comprendere se tali distorsioni derivassero da dataset squilibrati o da limiti architetturali dei modelli. La conseguenza è risultata che, in sistemi non open, nelle metodologie applicative o comunque difficilmente verificabili, i dati riferibili a gruppi già marginalizzati nella società determinano, nelle operazioni di calcolo o nei risultati, output fortemente discriminanti[19].

Sul piano tecnico, l'oscurità algoritmica ostacola anche l'applicazione di metriche di fairness. Strumenti come statistical parity, equalized odds o predictive parity richiedono accesso completo ai dati in ingresso e agli output per verificare eventuali effetti discriminatori. Se tali informazioni sono inaccessibili o incomprensibili, diventa impossibile garantire un controllo indipendente. In altre parole, l'oscurità priva la società degli strumenti necessari per misurare e correggere i bias.

La relazione tra oscurità e discriminazione si rafforza così in chiave dinamica, laddove maggiore è la prima, più aumenta il rischio della seconda. Nei sistemi che apprendono

CamminoDiritto.it
ISSN 2532-9871

continuamente dai dati generati durante l'uso, come nel caso delle chatbot, si possono infatti verificare cicli di retroazione che amplificano pregiudizi iniziali. Se il modello è opaco, tali cicli rimangono invisibili fino a quando non producono effetti macroscopici, come è accaduto con la chatbot Tay di Microsoft, rapidamente deviata verso contenuti razzisti e sessisti dalle interazioni con gli utenti. L'oscurità, in questo senso, agisce come un velo che copre processi di radicalizzazione algoritmica. Per questi motivi, dottrina e ricerca tecnica hanno sottolineato la necessità di strumenti capaci di ridurre il rischio discriminatorio connesso all'oscurità. Le tecniche di esplicitazione dei processi (Explainable AI) cercano di attribuire un "peso" alle variabili che hanno condotto a un certo output, offrendo così una base per verificare l'eventuale presenza di pregiudizi. Gli audit indipendenti, obbligatori in alcune giurisdizioni, consentono di misurare l'impatto dell'uso di determinati software su diversi gruppi e di imporre correzioni in caso di disparità sistematiche. In vista dell'impossibilità di eliminare totalmente gli errori di calcolo, si sono diffuse le c.d. strategie di bias mitigation che, come si esaminerà nel paragrafo successivo, mirano a bilanciare i dataset, integrare vincoli di equità nell'addestramento e correggere gli output finali.

Tuttavia, nessuna di queste soluzioni elimina completamente il rischio discriminatorio: l'oscurità algoritmica, in quanto strutturale, continuerà a produrre sempre spazi di incomprensione tra macchina e uomo, che possono celare discriminazioni. La sfida, pertanto, non è quella di eliminare l'opacità - obiettivo probabilmente irrealizzabile - ma di garantire che essa non si traduca in una barriera insormontabile contro la rilevazione e la correzione delle ingiustizie. Ciò richiede politiche legislative che integrino competenze tecniche e garanzie giurisdizionali, affinché l'algoritmo non diventi uno strumento di discriminazione invisibile.

4. Strategie di Bias Mitigation

La consapevolezza che i sistemi di intelligenza artificiale incorporino inevitabilmente forme di bias ineliminabili ha condotto la ricerca scientifica a sviluppare tecniche di mitigazione, poiché i dati storici riflettono disuguaglianze sociali, pregiudizi culturali e discriminazioni già esistenti, che l'algoritmo non può risolvere, semmai solo filtrare per contenerne l'impatto e ridurne gli effetti. Per altro verso, i modelli algoritmici riflettono anche scelte progettuali, volte a perseguire determinati scopi, che denotano una non neutralità originaria del sistema. Occorre, quindi, ritenere che l'affidabilità dell'applicativo non dipenda dal grado di riduzione dell'effetto discriminatorio del risultato, bensì dalla precisione del calcolo, che, trattandosi di un'operazione matematica, implica inevitabilmente dei compromessi tra precisione ed equità. Su questa considerazione la dottrina ha effettuato una distinzione tra gli approcci volti a mitigare gli effetti negativi del bias algoritmico, individuando tre tipologie di intervento mitigatorio[20], a seconda del momento in cui esso interviene.

CamminoDiritto.it
ISSN 2532-9871

Il pre-processing opera sulla fase di preparazione dei dati, partendo dalla considerazione che la principale fonte di bias derivi da dataset sbilanciati o scarsamente rappresentativi, che riflettono diseguaglianze storiche e culturali. Tecniche come il ricampionamento (re-sampling), che bilancia i campioni di gruppi sovra o sottorappresentati, o la riponderazione (re-weighting), che assegna pesi differenziati ai dati per compensare le disparità, sono pensate per ridurre tali squilibri. Un'ulteriore strategia è la data augmentation, che genera artificialmente esempi appartenenti a categorie minoritarie, ampliando così la copertura del dataset[21]. Questi metodi hanno il vantaggio di essere relativamente semplici da sviluppare e di non incidere sull'architettura del modello. Tuttavia, presentano limiti significativi: manipolare i dati rischia di introdurre artefatti statistici e di generare pattern che non corrispondono alla realtà empirica. In sostanza, si inseriscono dati correttivi negli input, modificando la raccolta originaria delle informazioni al fine di rendere più equo il dataset.

L'in-processing si colloca, invece, nella fase di addestramento dell'algoritmo, intervenendo direttamente sulla funzione matematica per incorporare vincoli di equità. Tecniche come i fairness constraints obbligano il modello a rispettare determinate metriche, quali l'equalized odds[22] o la demographic parity[23], mentre approcci come l'adversarial debiasing[24] utilizzano una seconda funzione algoritmica (in gergo anche rete neurale) che cerca di predire l'attributo sensibile (es. genere, etnia) a partire dal modello principale: l'obiettivo è dunque minimizzare tale predizione, costringendo così l'algoritmo a non basare le proprie decisioni su caratteristiche protette. L'in-processing ha il pregio di integrare la dimensione dell'equità nel cuore del modello, ma comporta compromessi non trascurabili. Da un lato, la performance predittiva può risultare penalizzata, poiché appesantita; dall'altro, la scelta della "fairness desiderata" non è un dato neutrale, bensì una decisione algoritmica che riflette valori e priorità sociali più che criteri puramente tecnici. In sostanza, si introduce uno strumento di mitigazione che, a sua volta, costituisce un effetto discriminatorio volto a correggere un criterio discriminante, secondo una regola di fairness pre-individuata sul risultato principale, a sua volta non neutrale. Questo strumento, oltre a essere tecnicamente complicato e oneroso, posto che richiede una doppia programmazione, riduce l'accuratezza del risultato principale, che si allontana dal dataset e può essere fonte di maggiore opacità algoritmica (maggiore complessità delle formule).

Il post-processing agisce sugli output prodotti dal modello, mediante tecniche che dovrebbero rendere i risultati maggiormente equi. Tecniche comuni sono l'adjustment of thresholds[25], che applica soglie differenti per gruppi diversi al fine di bilanciare i tassi di errore, oppure la calibration, che modifica le probabilità per garantire una distribuzione coerente tra i gruppi[26]. L'aspetto positivo di questo approccio è la sua applicabilità anche a modelli già sviluppati, senza necessità di accesso ai dati originari o all'algoritmo.

CamminoDiritto.it
ISSN 2532-9871

Tuttavia, le correzioni ex post rischiano di creare una percezione di "discriminazione inversa", in quanto prevedono trattamenti differenziati e si limitano a compensare gli effetti finali, senza incidere sulle cause strutturali del bias. In sostanza, si riconosce che il risultato finale è discriminato e si cercano degli arrotondamenti per cercare di riportare il risultato del calcolo a maggiore equità, che, tuttavia, è solo convenzionale.

La questione cruciale risulta, dunque, che tutti questi approcci condividono limiti strutturali, poiché quando trattiamo dati sensibili, provenienti da acquisizioni su scala sociale, l'eteronomia del dato e le modalità di acquisizione integrano differenze intrinseche difficilmente riproducibili mediante operazioni matematiche. Inoltre, la successiva introduzione di vincoli di fairness su base matematica non fa che ridurre la precisione complessiva del modello. In secondo luogo, la pluralità di metriche di equità disponibili (parità statistica, equalized odds, predictive parity) genera un problema di incompatibilità: come dimostrato nel caso COMPAS, non è possibile soddisfare simultaneamente tutte le definizioni di equità, cosicché la scelta di una metrica implica inevitabilmente un'opzione di carattere politico e valoriale[27].

In definitiva, permane sempre il rischio di bias latenti, non osservabili nei dati o nascosti dall'opacità dei modelli, che sfuggono alle tecniche di mitigazione e che possono riemergere in nuove forme nel corso del ciclo di vita del sistema. La preoccupazione permane laddove la previsione di operazioni di bias mitigation risulti un mero esercizio tecnico di costruzione di regole di fairness, finalizzate a fornire una legittimazione esteriore all'uso di algoritmi potenzialmente discriminatori, piuttosto che rappresentare buone prassi di progettazione e di applicazione. Si parla, a questo proposito, di fairwashing, ossia della strategia di adottare metriche di equità in modo superficiale od opportunistico, senza un reale impegno a modificare le diseguaglianze sottostanti[28]. A questo punto diviene inevitabile una considerazione: se il bias non è eliminabile e le operazioni algoritmiche di mitigation non sono altro che processi di "aggiustamento" dell'operazione di calcolo, non sarebbe meglio riconoscere che questi strumenti non sono idonei a effettuare scelte automatizzate, ma costituiscono piuttosto validi strumenti di calcolo statistico? Se un medico esperto, nella sua carriera, può comparare i risultati di una lastra su un'esperienza personale di qualche migliaio, mentre un algoritmo può effettuare una comparazione diagnostica su alcuni milioni di campioni, si possono di certo avere diagnosi più approfondite, ma la presenza del medico che cura il paziente rimane, in ogni caso, irrinunciabile.

5. Riflessioni interlocutorie

L'excursus tra le tecniche di funzionamento degli applicativi di intelligenza artificiale rappresenta - a parere di chi scrive - una presa di coscienza del fatto che questi strumenti

CamminoDiritto.it
ISSN 2532-9871

non sono perfetti e anzi integrano al loro interno rilevanti problemi discriminatori, latenti nei processi di profilazione e decisione automatizzata, tanto da richiedere precise cornici giuridiche che ne limitino l'uso e i possibili effetti negativi. Le strategie di mitigation sviluppate dalla tecnica informatica, volte a combattere le varie forme di bias, costituiscono risposte parziali a una sfida strutturale, ovvero adattare la tecnica informatica alle varie esigenze concrete, senza che questa possa prevaricare le eteronomie funzionali del vivere sociale. Il divieto previsto dall'art. 22 del Regolamento generale sulla protezione dei dati (GDPR) di non essere sottoposti a decisioni unicamente automatizzate che producano effetti giuridici significativi sull'individuo rappresenta anche il limite all'hypernudging, inteso come versione tecnologicamente amplificata del nudging comportamentale, che, attraverso architetture dinamiche e personalizzate delle offerte, compromette l'autenticità del consenso mantenendo al contempo l'apparenza di una decisione volontaria[29]. I principi tipici dell'ordinamento comunitario di trasparenza, correttezza e non discriminazione obbligano, infatti, i titolari del trattamento a prevenire e correggere squilibri derivanti dall'uso di algoritmi. Il recente AI Act conferma questo orientamento, introducendo una classificazione dei sistemi di intelligenza artificiale ad alto rischio (tra cui quelli destinati a occupazione, credito e giustizia), imponendo obblighi di trasparenza, gestione dei dati e valutazioni di impatto sui diritti fondamentali. In tale cornice, tuttavia, le tecniche di mitigazione non sono concepite solo come buone pratiche, ma come strumenti funzionali al rispetto di obblighi legali vincolanti.

Sul piano giuridico-istituzionale, la disciplina europea dell'AI Act, pur improntata a un modello di regolazione risk-based, evidenzia la necessità di garantire trasparenza e scrutinabilità delle decisioni automatizzate, poiché l'opacità di alcuni sistemi, in particolare quelli di deep learning, rischia di celare bias con effetti discriminatori che, se non rilevabili né contestabili, comprometterebbero il principio di uguaglianza e la tutela effettiva dei diritti fondamentali[30]. Ne consegue che il contrasto al rischio discriminatorio richiede tanto soluzioni tecniche di bias mitigation, quanto un rafforzamento delle garanzie processuali e costituzionali, affinché la tecnologia non diventi un alibi per mascherare nuove forme di esclusione sociale sotto una veste di apparente oggettività scientifica.

Nuovamente risulta necessaria una meta-sensibilità ai diritti fondamentali, che da un lato induca in via volontaria le imprese ad adottare tecniche di bias mitigation come strumento reputazionale per dimostrare affidabilità e attenzione etica, dall'altro orienti le tecniche di mitigazione a una funzione di conformità giuridica. L'impatto dei sistemi algoritmici non si esaurisce, infatti, nella dimensione individuale, ma investe anche le strutture collettive della convivenza politica, sollevando quello che la dottrina definisce "rischio democratico"[31]. L'opacità algoritmica, derivante tanto dalla complessità tecnica quanto dalla segretezza commerciale, priverebbe i cittadini della possibilità di comprendere e controllare le decisioni che li riguardano. Quando determinate scelte rilevanti vengono

CamminoDiritto.it
ISSN 2532-9871

affidate a processi automatizzati incomprensibili, si realizza uno spostamento di potere dalle istituzioni democratiche ai detentori delle tecnologie[32]. Gli algoritmi che regolano la circolazione delle informazioni online - come quelli dei social network o dei motori di ricerca - selezionano e gerarchizzano i contenuti, con l'effetto di amplificare polarizzazioni, diffondere disinformazione e alimentare filter bubbles, che riducono la pluralità informativa alla base del dibattito democratico. Gli scandali legati a Cambridge Analytica hanno mostrato come i dati personali possano essere ampiamente sfruttati per operazioni di micro-targeting politico in grado di incidere profondamente sulla libertà di scelta[33].

Il rischio democratico si manifesta, così, anche nel campo della governance pubblica: sistemi di predictive policing o di gestione algoritmica delle risorse sociali trasferiscono in capo a macchine con bias non eliminabili processi decisionali tipicamente riservati alle istituzioni democratiche. In assenza di adeguati meccanismi di controllo, si rischia così di introdurre criteri di gestione automatizzata della sicurezza e del welfare che sfuggono al vaglio pubblico e parlamentare, con conseguente lesione dei principi di responsabilità e legittimazione democratica[34].

Il rischio democratico va, dunque, inteso come fenomeno multidimensionale: non solo l'algoritmo può discriminare individui o gruppi, ma può alterare i presupposti stessi della democrazia, minando trasparenza, pluralismo e responsabilità. La sfida non consiste nel bandire l'uso degli algoritmi, ma nel garantire che essi operino all'interno di un quadro istituzionale in grado di preservare anni di lotta ai diritti e di evitare che i risultati delle operazioni di calcolo algoritmico possano attuare effetti discriminatori. In assenza di ciò, l'algoritmo rischia di trasformarsi da strumento di efficienza in fattore di opacità e concentrazione del potere, con effetti corrosivi per le istituzioni rappresentative. Sistemi di selezione del personale, di concessione del credito o di giustizia predittiva hanno mostrato come l'opacità algoritmica possa tradursi in discriminazioni effettive, che incidono sui diritti individuali di soggetti classificati in base a variabili statistiche. L'impossibilità di controllare e contestare le logiche interne dei processi di calcolo impedisce, di fatto, di verificare l'adeguatezza dei modelli alle norme giuridiche codificate.

La dignità, come fondamento delle costituzioni europee e della Carta dei diritti fondamentali dell'Unione europea, può così essere messa in discussione dall'uso di sistemi che classificano gli individui in base a variabili cataloganti con effetto discriminatorio. L'autodeterminazione informativa - elaborata dalla giurisprudenza costituzionale tedesca e recepita nel diritto europeo con il GDPR - rischia di essere svuotata se l'individuo non può conoscere né controllare i processi di trattamento automatizzato che lo riguardano. In questa prospettiva, l'algoritmo diventa un elemento di reificazione, riducendo la persona a "dato" e minando la centralità della dignità come

CamminoDiritto.it
ISSN 2532-9871

nucleo irrinunciabile degli ordinamenti giuridici liberaldemocratici.

Il ricorso ad algoritmi nei procedimenti giudiziari, come nel caso del risk assessment penale, solleva poi questioni legate al giusto processo. La gestione di dati sensibili in piattaforme in grado di catalogare i procedimenti mediante sistemi algoritmici non accessibili o non spiegabili, mina alla base il diritto di difesa. La giurisprudenza statunitense (caso State v. Loomis, 2016) ha, del resto, riconosciuto i limiti di trasparenza di sistemi informatici in grado di calcolare il rischio di recidiva (COMPAS), pur consentendone l'uso come mero strumento ausiliario; ciò conferma la tensione tra efficienza degli applicativi e garanzie processuali. L'algoritmo, dunque, non solo può discriminare individui o gruppi, ma può alterare i presupposti stessi della democrazia, compromettendo trasparenza, pluralismo e accountability. La sfida non consiste, dunque, nell'eliminare l'uso degli algoritmi, ma nell'assicurarne l'integrazione entro un quadro istituzionale che preservi i valori democratici e impedisca gli effetti discriminatori sotto la parvenza di una falsa oggettività scientifica.

La selezione algoritmica dei contenuti può, infatti, influenzare sensibilmente la visibilità delle opinioni, creando invisibili filtri che limitano il pluralismo informativo. Il rischio democratico si traduce, così, in un potenziale deficit della sovranità popolare, che incide direttamente sulla libertà di espressione, condizionando la visibilità delle opinioni e generando nuove forme di concentrazione del potere informativo. Paradossalmente, se il controllo sull'accesso all'informazione e sulla distribuzione delle risorse pubbliche resta affidato a logiche algoritmiche opache, il processo decisionale rischia di spostarsi dal dominio delle istituzioni rappresentative a quello delle grandi piattaforme tecnologiche. In tal senso, le ricadute costituzionali dell'impiego di strumenti di intelligenza artificiale dimostrano che la questione non attiene soltanto alla tecnica o al diritto dei dati personali, ma riguarda il cuore stesso delle garanzie democratiche.

Ne consegue che il governo della tecnica algoritmica non può ridursi a un problema di mera fairness computazionale, ma deve essere ricondotto entro le categorie del diritto costituzionale e del diritto delle fonti[35]. La partita in gioco assume una rilevanza fondamentale, posto che il potere algoritmico rischia, infatti, di introdurre forme di governance privata non controllata democraticamente[36]. La tecnica algoritmica esige, dunque, di essere ricondotta entro una cornice giuridica-istituzionale che preservi i valori democratici fondamentali, altrimenti il rischio è che essa si trasformi da strumento di efficienza in fattore di opacità sistemica, capace di minare i principi di uguaglianza, dignità e pluralismo, ossia il cuore stesso dei diritti fondamentali inviolabili.

6. Conclusioni

CamminoDiritto.it
ISSN 2532-9871

Il percorso analitico appena condotto evidenzia come il bias algoritmico non sia un accidente marginale, ma un rischio strutturale e immanente ai sistemi di intelligenza artificiale. Esso si manifesta lungo tutte le fasi del ciclo di vita degli algoritmi - dall'acquisizione dei dati all'addestramento, dalla validazione fino all'implementazione - e, quando connesso all'opacità tecnica e interpretativa, si traduce in un potenziale fattore di discriminazione individuale e collettiva.

Le strategie di bias mitigation rappresentano strumenti utili, ma non definitivi: agiscono come soluzioni correttive parziali che, se non accompagnate da solide garanzie giuridiche e istituzionali, rischiano di scivolare in meri esercizi di tecnica. La vera sfida, dunque, non consiste nell'illusione di eliminare completamente il rischio, bensì nel costruire un quadro regolatorio che renda la tecnica verificabile, contestabile e compatibile con i valori fondativi degli ordinamenti democratici.

Ne emerge un punto fermo: la tecnologia algoritmica non può essere considerata neutrale, né tantomeno autonoma rispetto ai principi di uguaglianza, dignità e giusto processo. Essa deve essere governata entro una cornice normativa che integri trasparenza, accountability e tutela effettiva dei diritti fondamentali. Solo in tal modo sarà possibile coniugare l'innovazione tecnologica con la salvaguardia dei valori costituzionali ed evitare che l'algoritmo, da strumento di progresso, si trasformi in un mezzo di discriminazione invisibile e di concentrazione del potere. Il governo dell'intelligenza artificiale non è questione meramente tecnica, ma eminentemente politica e giuridica: richiede un bilanciamento tra efficienza ed equità, tra calcolo e giustizia, affinché le nuove frontiere dell'informatica applicata non si traducano in una regressione delle libertà, ma in un'opportunità di rafforzamento dello Stato di diritto e della democrazia.

CamminoDiritto.it
ISSN 2532-9871

Note e riferimenti bibliografici

[1] G. DE MINICO, Algoritmi e diritti fondamentali. Il nuovo volto della sovranità digitale, Giappichelli, Torino, 2019, p. 27 ss.

[2] K. DAVIS, D. PATTERSON, Ethics of Big Data: Balancing Risk and Innovation, O'Reilly Media, 2012, cfr. pp. 64-82.

[3] F. VITALI, L'oro nero dei dati, in Limes, 2014, n. 7, cfr. p. 29 ss.

[4] S. ZUBOFF, The Age of Surveillance Capitalism, Luiss University Press, 2019, cfr. p. 175 ss.

[5] C. O'NEIL, Weapons of Math Destruction, Penguin Books Ltd, 2016, cfr. p. 68.

[6] J. BUOLAMWINI, T. GEBRU, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in Proceedings of Machine Learning Research, 2018, p. 77 ss.

[7] T. BOLUKBASI et al., Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, in Advances in Neural Information Processing Systems, 2016, p. 4349 ss.

[8] J. ANGWIN, J. LARSON, S. MATTU, L. KIRCHNER, Machine Bias, ProPublica, 2016.

[9] S. MITCHELL, Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions, in arXiv preprint, 2018, p. 12 ss.

[10] K. CRAWFORD, T. PAGLEN, Excavating AI: The Politics of Training Sets for Machine Learning, in AI Now Institute Report, 2019.

[11] A. BINNS, Fairness in Machine Learning: Lessons from Tay, in Philosophical Transactions of the Royal Society A, 2018.

[12] J. VINCENT, Microsoft's AI chatbot is making racist tweets, in The Verge, 24 marzo 2016.

[13] S. BAROCAS, A.D. SELBST, Big Data's Disparate Impact, in California Law Review, vol. 104, 2016, p. 684.

[14] J. BURRELL, How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms, in Big Data & Society, vol. 3, n. 1, 2016, p. 3 ss.

[15] CGUE, 7 dicembre 2023, causa C-634/21, OQ v. Land Hessen, SCHUFA Holding AG, §72 ss.

[16] F. PASQUALE, The Black Box Society. The Secret Algorithms That Control Money and Information, Harvard University Press, Cambridge (MA), 2015, p. 19 ss.

[17] D. K. CITRON, F. PASQUALE, The Scored Society: Due Process for Automated Predictions, in Washington Law Review, vol. 89, 2014, p. 13 ss.

[18] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) è un software di gestione dei casi e di supporto alle decisioni sviluppato e di proprietà di Northpointe (ora Equivant), utilizzato dai tribunali statunitensi per valutare la probabilità che un imputato diventi un recidivo. COMPAS è stato utilizzato dagli stati americani di New York, Wisconsin, California, Broward County in Florida e altre giurisdizioni.

[19] J. BUOLAMWINI, T. GEBRU, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in Proceedings of Machine Learning Research, 2018, p. 77 ss.

[20] S. BAROCAS, M. HARDT, A. NARAYANAN, Fairness and Machine Learning, MIT Press (draft), 2019, p. 112 ss.

[21] C. DWORK, Fairness Through Awareness, in Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012, p. 214 ss.

CamminoDiritto.it
ISSN 2532-9871

[22] Equalized Odds si intende una metrica di equità algoritmica secondo cui un classificatore binario deve garantire uguali tassi di errore tra i diversi gruppi protetti (es. uomini/donne, bianchi/neri). In termini tecnici, un algoritmo soddisfa l'equalized odds quando: il True Positive Rate (ossia la sensibilità: proporzione di positivi correttamente identificati) e il False Positive Rate (ossia la proporzione di negativi erroneamente classificati come positivi) sono uguali per tutti i gruppi sensibili considerati.

[23] Per Demographic Parity si intende una condizione di equità secondo la quale un classificatore algoritmico deve garantire che la probabilità di predire un esito positivo (es. essere assunti, ricevere un prestito, ottenere la libertà condizionata) sia uguale per tutti i gruppi sensibili, indipendentemente dal loro reale stato o dalla variabile di verità.

[24] L'adversarial debiasing è una tecnica di mitigazione che si ispira all'architettura delle reti antagoniste generative (Generative Adversarial Networks, GANs). L'idea di base è addestrare due modelli in competizione: il modello predittivo principale produce l'output desiderato, mentre il modello antagonista, si pone in conflitto con il modello principale cercando di verificare gli errori predittivi del modello principale. L'addestramento procede con un obiettivo duplice: il modello principale cerca di massimizzare la performance predittiva sul compito primario, ma allo stesso tempo deve "confondere" l'avversario, riducendo la sua capacità di indovinare l'attributo sensibile. In questo modo, il modello principale viene indotto a rimuovere o attenuare le informazioni correlate all'attributo protetto, producendo decisioni meno influenzate dal bias.

[25] Per Adjustment of Thresholds si intende una strategia di mitigazione del bias che interviene dopo l'addestramento del modello, modificando i criteri decisionali utilizzati per classificare gli individui nei diversi gruppi protetti.

[26] F. KAMISHIMA, S. AKIYAMA, H. SAITO, Fairness-Aware Learning through Regularization Approach, in 2011 IEEE International Conference on Data Mining Workshops, p. 643 ss.

[27] J. ANGWIN, J. LARSON, S. MATTU, L. KIRCHNER, Machine Bias, ProPublica, 2016.

[28] M. KLEINBERG, J. LUDWIG, S. MULLAINATHAN, Inherent Trade-Offs in the Fair Determination of Risk Scores, in Proceedings of the National Academy of Sciences, vol. 116, n. 51, 2019.

[29] A. JACI, La responsabilità algoritmica nei contratti con i consumatori: verso un nuovo paradigma di equilibrio contrattuale, in Cammino Diritto, n. 8/2025, ISSN 2532-9871, p. 8.

[30] G. DE MINICO, Giustizia e intelligenza artificiale: un equilibrio mutevole, in Rivista AIC, n. 2/2024, p. 86 ss.

[31] F. PASQUALE, New Laws of Robotics. Defending Human Expertise in the Age of AI, Harvard University Press, Cambridge (MA), 2020, p. 78 ss.

[32] J. BURRELL, How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms, in Big Data & Society, vol. 3, n. 1, 2016, p. 3 ss.

[33] D. CADWALLADR, E. GRAHAM-HARRISON, Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica, in The Guardian, 17 marzo 2018.

[34] V. EUBANKS, Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor, St. Martin's Press, New York, 2018, p. 113 ss.

[35] G. DE MINICO, Le fonti del diritto: un argine all'intelligenza artificiale?, in «Rivista AIC», n. 3/2025, ISSN 2039-8298, pubblicato il 21 luglio 2025, disponibile in www.rivistaaic.it

[36] F. PASQUALE, New Laws of Robotics. Defending Human Expertise in the Age of AI, Harvard University Press, Cambridge (MA), 2020, p. 102 ss.

---

* Il simbolo {https/URL} sostituisce i link visualizzabili sulla pagina:
https://rivista.camminodiritto.it/articolo.asp?id=11367