Dis-incarnare genere e religioni

Intelligenza Artificiale, bias, discriminazioni

Ilaria Valenzi

1. Intelligenza Artificiale e Big Data nel contesto dei diritti fondamentali

L'affermazione «sono musulmana» contiene una componente offensiva? Dichiararsi «buddista» produce le medesime conseguenze che dichiararsi ebreo? Il grado di offensività dell'una e dell'altra affermazione si equivalgono? Che ruolo svolge il contesto, analogico o digitale, in cui queste frasi vengono pronunciate per determinarne natura e valore nei confronti delle soggettività coinvolte?

Lo spazio disincarnato delle nuove tecnologie costituisce senza dubbio un ambito innovativo per intessere relazioni umane. La traslazione nel digitale, talora nel virtuale, di parti cospicue delle nostre identità personali apre scenari nuovi, che si inseriscono nella possibilità di sperimentare situazioni favorevoli all'interscambio umano e sociale, ma che si concretizzano anche nei rischi per i diritti della personalità. A compiere questo atto di dis-incarnazione dei corpi è l'attività informativa che si fonda sull'elaborazione di dati. Rappresentazioni originarie di fenomeni, fatti o eventi mediante simboli, i dati sono il motore dello sviluppo tecnologico e, in quanto tali, costituiscono un bene il cui valore supera la mera sfera economica. La digitalizzazione dei processi e le continue interazioni che avvengono via web si basano e a loro volta generano dati e informazioni, materiale indispensabile per produrre scambi di servizi, creare nuove aree di attività economica, sviluppare tecnologie utili agli individui. L'era dei Big Data è quello spazio temporale in cui il comportamento umano, nella sua relazione con dispositivi tecnologici *smart*, produce una quantità di dati e informazioni superiori a quanto mai prodotto in precedenza e genera attività tecnologiche che plasmano le interazioni tra esseri umani[1]. Viviamo, più o meno consapevol-

[1] L'impatto dei Big Data nell'ambito della ricerca supera la sfera delle scienze computazionali e si diffonde nei maggiori ambiti del sapere. Senza pretesa di esaustività, si vedano: L. Lombi, *La ricerca sociale al tempo dei Big Data: sfide e prospettive*, in «Studi di sociologia», 53, 2015, 2, pp. 215-227;

DOI Number: 10.61015/ Corpi_anticorpi_202308

mente, in questa era, un tempo di dati e tecnologie che plasmano il nostro stare insieme. Senza dubbio l'utilizzo dei dati a fini tecnologici ha consentito di compiere un salto qualitativo epocale in un particolare campo di ricerca scientifica, i cui risultati in termini di impatto sulle nostre vite è senza precedenti. L'Intelligenza Artificiale è ovunque e i dati, che costituiscono la materia prima per il suo funzionamento, raccontano una storia su ognuna e ognuno. Questo assunto sta a significare che l'Intelligenza Artificiale non si limita a entrare nelle nostre vite come un corpo esterno, ma vive di noi, della proiezione digitale e disincarnata delle nostre esistenze. La tensione tra corpo e macchina, tra umano e digitale, appare più sfumata di quanto possa sembrare a un primo approccio e richiede un'attenzione particolare nei confronti del rispetto dei diritti fondamentali della persona umana.

Con questa consapevolezza l'Unione Europea ha da tempo iniziato un percorso per dotarsi di una propria politica nel campo della regolamentazione dell'Intelligenza Artificiale, con ciò affermando concretamente la volontà di rivestire un ruolo guida a livello mondiale nella definizione del *global gold standard*[2] in materia. In questo senso, la strategia europea sull'Intelligenza Artificiale si compone, a partire dal 2018, di tappe in cui fasi di studio, consultazione e politiche attive si susseguono, tutte dirette alla produzione di atti e documenti dal valore normativo. In questa ideale linea del tempo[3], merita di essere ricordata la consultazione pubblica del febbraio 2020 avente a oggetto il *White Paper* della Commissione sull'approccio europeo all'eccellenza e alla fiducia in tema di Intelligenza Artificiale[4]. Il Centro per le Scienze Religiose della Fondazione Bruno Kessler ha contribuito al dibattito con un suo *Policy paper* sul coinvolgimento e ruolo degli attori religiosi e di convinzione nel dar forma al cambiamento derivante dall'Intelligenza Artificiale[5]. In quell'occasione si è rilevato come lo sguardo religioso sull'Intelligenza Artificiale non si limiti a osservazioni inerenti agli

---

V. Ferrari, *Note socio-giuridiche introduttive per una discussione su diritto, intelligenza artificiale e big data*, in «Sociologia del diritto», 47, 2020, 3, pp. 9-32; V. Zeno-Zencovich, *Big Data e epistemologia giuridica*, in G. Resta - V. Zeno-Zencovich (edd), *Governance of/trough Big Data*, Roma, RomaTrE-Press, 2023, pp. 439-448.

[2] Così si esprime la Commissione europea nell'ambito della politica europea sull'Intelligenza Artificiale, a partire dal Piano di Coordinamento sull'Intelligenza Artificiale varato nel 2021: https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review.

[3] Le tappe della strategia europea sull'Intelligenza Artificiale sono consultabili al seguente link: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.

[4] Il libro bianco è consultabile in lingua italiana al presente link: https://commission.europa.eu/document/d2ec4039-c5be-423a-81ef-b9e44e79825b_it.

[5] Il *Policy paper* è consultabile al seguente link: https://isr.fbk.eu/wp-content/uploads/2021/12/ISR_Policy_Paper_2021.pdf.

aspetti etici delle questioni poste dall'innovazione tecnologica, ma prenda forma in un coinvolgimento più o meno diretto delle religioni nell'utilizzo delle stesse tecnologie. Attualmente il processo istituzionale dell'Unione Europea per dotarsi di strumenti politici e giuridici in tema di Intelligenza Artificiale ha raggiunto uno stadio avanzato, con l'approvazione da parte del Parlamento europeo dell'*AI Act*[6], il Regolamento europeo sull'Intelligenza Artificiale, primo strumento legislativo di tale portata a livello globale. Una visione centrata sull'umano, che obbliga le parti in gioco al rispetto dei diritti e dei valori fondamentali dell'Unione Europea e che vuole combattere la cosiddetta «discriminazione digitale». Un aspetto che, infatti, appare assodato riguardo all'utilizzo dell'Intelligenza Artificiale concerne il rischio reale di proiezione (e, con ogni probabilità, amplificazione) nel mondo digitale di comportamenti discriminatori che caratterizzano la nostra esistenza incarnata. La legislazione che promana dalle istituzioni europee mostra di aver chiara la centralità della questione, introducendo il divieto di usi intrusivi e discriminatori di tecnologie potenzialmente rischiose per i diritti fondamentali di individui e collettività. Ciò è più vero per alcuni profili dell'identità particolarmente meritevoli di protezione, sui quali si concentrano gli sforzi di tutela. Vi rientrano a pieno titolo le caratteristiche personali legate all'etnia, alla religione professata, al genere, ma anche l'orientamento politico o l'appartenenza a un sindacato. I profili qui indicati costituiscono pertanto un ideale confine non oltrepassabile da prassi e strumenti che fanno uso delle nuove tecnologie, a partire dai rilievi biometrici utili per le identificazioni e per la categorizzazione degli individui, che possono comportare l'esclusione dall'accesso ai servizi, o la compressione degli spazi di libertà personale, come la libertà di movimento o l'esercizio del diritto di voto. In questo quadro di riconoscimento di tutela, non deve stupire l'affermazione dell'Agenzia dell'Unione Europea per i Diritti Fondamentali secondo la quale, in un contesto in cui la capacità decisionale diviene sempre più automatizzata, assume un significato vitale che la tecnologia lavori per e non contro gli esseri umani, o contro alcuni di essi[7]. Appare chiaro che l'incontro tra l'umano e la macchina, sia esso volontario o eterodiretto, non ammette una sottovalutazione del suo potenziale impatto sull'esistenza di individui e collettività.

---

[6] https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRSBRI(2021)698792_EN.pdf.

[7] Il riferimento è al report della European Union Agency for Fundamental Rights, *Bias in Algorithms. Artificial Intelligence and Discrimination*, pubblicato nel dicembre del 2022. Il report in lingua inglese è consultabile al seguente link: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf.

Sebbene le implicazioni etiche nel campo dell'applicazione dell'Intelligenza Artificiale costituiscano un campo di riflessione essenziale per la creazione di tecnologie giuste e l'etica dell'«IA» sia senza dubbio una delle nuove frontiere di analisi del presente, è importante rilevare che l'approccio europeo alla regolamentazione dell'Intelligenza Artificiale trova il suo primo referente nell'applicazione dei valori sociali che fondano l'Unione e il suo principale fondamento nei diritti umani. Non si tratta cioè di fondare le norme della comune regolamentazione su cosa è ritenuto buono e cosa sbagliato, di lasciare alle regole della morale la disciplina dei comportamenti umani; si tratta bensì di comprendere quando sia necessario l'intervento del diritto per evitare che i principi fondamentali vengano violati[8]. Nel rapporto tra essere umano e tecnologie sta infatti una nuova «età dei diritti»[9] e nell'individuazione di una risposta ai nuovi bisogni una nuova frontiera di tutela[10].

2. Intelligenza Artificiale e discriminazione algoritmica

Nel quadro fin qui disegnato, tra discriminazione e Intelligenza Artificiale sembra delinearsi un rapporto diretto, difficilmente scindibile. Ma con quali strumenti e modalità il potenziale comportamento discriminatorio si realizza nell'universo digitale? Senza dubbio lo strumento algoritmico rappresenta il canale privilegiato per la realizzazione dei nostri bisogni come per la concretizzazione dei timori per i nostri diritti. Gli algoritmi basati sull'Intelligenza Artificiale costituiscono il principale veicolo di trasmissione di informazioni e le tecnologie che si fondano su di essi svolgono attività predittive o, addirittura, sono in grado di compiere attività decisionali completamente automatizzate. Riferirsi all'algoritmo conferendo a esso l'attributo della soggettività è prassi diffusa nel linguaggio comune. Senza voler indagare la natura filosofica di una tale attività di riconoscimento, è sufficiente ai nostri fini ricordare che l'algoritmo si concretizza in una successione determinata e organizzata di istruzioni, utili a definire le operazioni che portano a ottenere un risultato finale[11]. L'ampia diffusione di una consapevolezza circa il ruolo e il potere degli algoritmi è in larga parte

---

[8] Sul punto cfr. S. Rodotà, *La vita e le regole. Tra diritto e non diritto*, Milano, Feltrinelli, 2006.

[9] Il riferimento è all'omonima opera di N. Bobbio, *L'età dei diritti*, Torino, Einaudi, 1990.

[10] S. Rodotà, *Il diritto di avere diritti*, Roma - Bari, Laterza, 2012.

[11] Nella sconfinata bibliografia in materia, segnaliamo, per continuità con il tema affrontato in questo saggio, T. Numerico, *Big data e algoritmi. Prospettive critiche*, Roma, Carocci, 2021; R. Marmo, *Algoritmi per l'intelligenza artificiale. Progettazione dell'algoritmo, dati e machine learning, neural network, deep learning*, Milano, Hoepli, 2020; D. Harel - Y. Feldman, *Algoritmi. Lo spirito dell'informatica*, Milano, Springer, 2008.

dovuta alla popolarità dell'utilizzo dei motori di ricerca e dei social media. L'applicazione degli algoritmi in questi ambiti consente, ad esempio, di offrire all'utente contenuti su misura, studiando il suo comportamento sul web. In tale campo, la tecnologia costituisce un evidente facilitatore delle nostre vite, ma l'interscambio di dati, richiesto per il funzionamento dei sistemi informativi, sembra non essere percepito nella sua globalità come continua cessione del patrimonio di informazioni personali di cui l'essere umano è intrinsecamente dotato. Una funzione sociale, pertanto, quella dell'algoritmo, che si svolge grazie alla fornitura continua di materiale da elaborare e restituire sotto forma di servizi, inviti al consumo, analisi della propensione agli acquisti e alla direzionabilità delle scelte individuali in ogni campo della vita. L'attività fin qui descritta non può certamente dirsi esaustiva delle possibili applicazioni algoritmiche, né sufficiente a illustrare il loro potenziale impatto sulla collettività[12]. Luci e ombre da sempre caratterizzano l'analisi dei sistemi intelligenti e le seconde, in particolare, non possono essere sottovalutate[13]. In tale ambito l'attenzione ai diritti fondamentali della persona svolge un ruolo essenziale. La riflessione pare concentrarsi principalmente intorno a due macro questioni. La prima attiene alla raccolta di dati personali afferenti alle sfere più sensibili dell'identità. Se il dato nasce di per sé come rappresentazione non interpretata di fenomeni, fatti o eventi, la connotazione personalissima delle informazioni raccolte può riverberare in una particolare visione del mondo fornita dalla macchina. Tale meccanismo è imputabile ad algoritmi che presentano una qualche tipologia di bias, una distorsione cioè nel funzionamento o nel risultato dell'algoritmo, che rivela un'inclinazione al pregiudizio nei confronti di particolari caratteristiche protette di singoli o categorie di persone. In questo caso, il bias algoritmico potrà portare a un risultato discriminatorio qualora i parametri codificati o i dati di addestramento contengano un riferimento diretto agli specifici *grounds* oggetto di tutela legale in materia[14]. Esemplificando, qualora nell'elaborazione informatica i dati trattati attengano all'etnia, alla religione o al genere, il risultato finale dell'elaborazione ha una qualche probabilità di produrre un trattamento differenziato non giustificabile delle persone cui l'informazione è riferibile.

---

[12] Si veda, tra gli altri, F. Fossa - V. Schiaffonati - G. Tamburrini (edd), *Automi e persone: introduzione all'etica dell'intelligenza artificiale e della robotica*, Roma, Carocci, 2021.

[13] Tra i più noti contributi si veda S. Zuboff, *Il capitalismo della sorveglianza*, Roma, Luiss University Press, 2019.

[14] Il riferimento è al più ampio ambito del diritto antidiscriminatorio di matrice europea e nazionale. Tra gli altri, si veda M. Barbera (ed), *Il nuovo diritto antidiscriminatorio. Il quadro comunitario e nazionale*, Milano, Giuffrè, 2007; M. Barbera - A. Guariso (edd), *La tutela antidiscriminatoria. Fonti strumenti interpreti*, Torino, Giappichelli, 2020.

La rilevanza giuridica delle categorie protette si rivela, pertanto, essenziale affinché un dato modello algoritmico possa considerarsi discriminatorio. Lo stesso risultato non è, al contrario, rinvenibile qualora la differenziazione operata dall'algoritmo avvenga sulla base di informazioni su dati, abitudini o stili di vita non meritevoli di tutela secondo il diritto antidiscriminatorio. Si pensi all'ipotesi di un diniego di accesso a servizi motivata sulla base dell'origine etnica di una persona o, diversamente, sulla base del possesso o meno da parte di quest'ultima di un animale domestico. Solo nel primo caso il bias algoritmico produrrà una discriminazione vietata dalla legge, mentre nel secondo caso il pregiudizio, per essere considerato illegittimo, necessiterà di un qualche aggancio con altre previsioni di tutela offerte dall'ordinamento. Se ciò corrisponde a una stretta logica di garanzia di tutela delle categorie protette, in un tempo di rapidi mutamenti della sensibilità sociale e collettiva intorno ai beni giuridici meritevoli di protezione occorre chiedersi quale tipo di tutela garantire a quelle posizioni giuridiche soggettive nuove e all'ipotetica nascita di gruppi di individui considerati, secondo il senso comune, svantaggiati (si pensi al caso delle persone che giocano online o ai cosiddetti *sad teenagers*)[15]. La seconda questione attiene all'utilizzo dei dati personali per mettere in atto politiche mirate, finalizzate ad attività di controllo sociale e repressione dei reati. In tale ambito gli algoritmi affetti da pregiudizi possono condurre a errori gravi, con conseguenze sulle persone interessate e sugli equilibri sociali nel loro complesso. Il timore di un tale risultato non costituisce un caso di scuola, ma si fonda su situazioni reali. Si pensi a quanto accaduto nel 2020, allorché il governo olandese è stato condannato per violazione dei diritti umani a causa dell'utilizzo di Syri, acronimo di *System Risk Indication*, un algoritmo finalizzato alla creazione di profili di rischio di commissione di frode nell'accesso ai sussidi di Stato. In quel caso, specifiche categorie di cittadini sono state sottoposte al controllo dell'algoritmo in quanto residenti in quartieri ad alta densità abitativa e a basso reddito, generalmente appartenenti a minoranze etniche e con retroterra migratorio[16]. Il risultato della discriminazione algoritmica verificatasi ha coinvolto oltre ventiseimila persone, ingiustamente accusate di frode e opportunamente risarcite per quanto subito.

---

[15] Tutti gli esempi fin qui proposti si riferiscono al report *Bias in Algorithms – Artifical Intelligence and Discrimination*, pubblicato nel dicembre del 2022. Qui il link: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf.

[16] Il testo della sentenza della Corte Distrettuale dell'Aja può essere consultato al seguente link:https://www.ohchr.org/en/press-releases/2020/02/landmark-ruling-dutch-court-stopsgovernment-attempts-spy-poor-un-expert?LangID=E&NewsID=25522.

Il caso di discriminazione diretta cui abbiamo fatto riferimento costituisce certamente un'ipotesi eclatante di cattivo funzionamento dell'algoritmo. Non sempre, tuttavia, siamo in grado di comprendere come e perché gli strumenti di Intelligenza Artificiale violino i diritti fondamentali. La necessità di prove in tale ambito è un'urgenza cui occorre fornire risposta.

3. La ricerca applicata: alcuni casi di studio

Con l'intento di rispondere alla suddetta esigenza, l'Agenzia europea dei diritti fondamentali[17], struttura autonoma e indipendente per la promozione e la tutela dei diritti umani nell'ambito dell'Unione Europea, ha condotto uno studio teorico e applicativo sul legame tra Intelligenza Artificiale e bias[18]. Facendo seguito all'assunto già indagato in precedenti report, per cui soltanto un approccio basato sui diritti è in grado di garantire più alti livelli di protezione contro possibili danni connessi all'utilizzo delle nuove tecnologie[19], in questa analisi l'Agenzia affronta due *case studies* che esemplificano alcune ipotesi di manifestazione di bias discriminatori. Si tratta, rispettivamente, dell'utilizzo di algoritmi nei sistemi di polizia predittiva e dei pregiudizi etnici e di genere nel rilevamento dei discorsi offensivi. A tal fine, l'Agenzia si è dotata di strumenti di analisi tecnico-informatici e di competenze analitiche in materia di tutela dei diritti e *policy-making*. In questa sede non è possibile procedere a una disamina dettagliata dei risultati della ricerca, per i quali si rimanda al report. Prima di soffermarci con più attenzione sull'ipotesi dei pregiudizi nell'*offensive speech*, è opportuno rilevare che lo studio della distorsione in chiave pregiudiziale etnica dello strumento di polizia predittiva vuole essere una risposta scientifica al verificarsi di casi concreti come quello precedentemente illustrato. L'utilizzo della polizia predittiva comporta, cioè, rischi reali cui è importante contrapporre risposte tecnicamente motivate. Queste ultime trovano sede nello studio dei cosiddetti *feedback loops*, vale a dire sull'influenza che gli algoritmi producono su altri algoritmi. Le previsioni algoritmiche sull'insorgenza di comportamenti criminogeni riverberano cioè sulla condotta

---

[17] Informazioni maggiori sulla vita e l'attività dell'Agenzia sono consultabili nel sito web istituzionale https://fra.europa.eu/it. In particolare, tra i suoi settori di impegno, segnaliamo la protezione dei dati personali, della vita privata, e nuove tecnologie, Intelligenza Artificiale, megadati.

[18] Si tratta del già citato report *Bias in Algorithms – Artifical Intelligence and Discrimination*.

[19] Il riferimento è al report *Getting the Future Right – Artificial Intelligence and Fundamental Rights*, pubblicato nel 2021 e consultabile al seguente link: https://fra.europa.eu/en/publication/2021/getting-future-right-artificial-intelligence-and-fundamental-rights-summary.

della polizia, a sua volta in grado di influenzare il rilevamento di reati con il proprio comportamento pregiudiziale. Un maggior rilevamento di reati così causato determinerà la convinzione di una più forte necessità di controllo di polizia in alcuni quartieri e su alcuni gruppi etnici, ingenerando la convinzione che gli stessi siano di fatto più inclini a comportamenti criminogeni. L'influenza dell'algoritmo sulla criminalizzazione di parti della popolazione è pertanto evidente e necessita di essere contrastata.

Il caso dei bias etnico-religiosi e di genere nella rilevazione del linguaggio offensivo presenta senza dubbio profili di particolare interesse per gli studi sui diritti fondamentali nelle società plurali. L'urgenza di una riflessione in tale ambito è resa evidente da alcuni dati forniti dalla stessa Agenzia europea. Il riferimento è al report del 2018 sull'antisemitismo, che ha rilevato come il più alto tasso di incidenza di molestie individuali antisemite sia di tipo informatico[20]. Del pari, i dati riferiti al 2012 sulla tutela di genere nel contesto dell'Unione Europea rilevano che il tasso di donne che dichiarano di aver subito molestie informatiche è pari a 1 su 20[21]. Il mondo del web e dei «social» costituisce il terreno per eccellenza dei comportamenti molesti; il dato è di immediata percezione per chiunque abbia una frequentazione anche solo saltuaria della rete. I numeri forniti dall'Agenzia europea con il report 2022 qui in analisi offrono un'idea concreta del fenomeno. Nel corso del 2021 Twitter ha rimosso più di cinque milioni di contenuti offensivi, ciò sebbene l'Intelligenza Artificiale che governa la piattaforma non proceda alla moderazione automatica dell'incitamento all'odio e non sia in grado di riconoscere modelli di comunicazione tossica che non abbia precedentemente registrato. Dall'altro lato, le politiche di *under-* o *over-blocking* messe in atto dalle piattaforme online rischiano di rivelarsi misure di limitazione aprioristica di contenuti, come tali a rischio a loro volta di violazione di diritti fondamentali, primo tra tutti la violazione della libertà di espressione.

A partire da tali presupposti, l'Agenzia europea per i diritti fondamentali ha sviluppato un proprio test sul funzionamento degli algoritmi di rilevamento dei discorsi d'odio. Il test si fonda sulla costruzione di tre *dataset*, rispettivamente in lingua inglese, tedesca e italiana, composti da informazioni raccolte sui più diffusi *social media* e catalogate dal team di ricerca come offensive o non offensive. La scelta plurilinguistica così connotata si deve alla necessità di prevedere la declinazione del linguaggio differenziato nel

---

[20] https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-experiences-and-perceptions-of-antisemitism-survey-summary_it.pdf.

[21] https://eige.europa.eu/sites/default/files/documents/ti_pubpdf_mh0417543itn_pdfweb_20171026164002.pdf.

genere, risultato che non avrebbe potuto ottenersi limitando l'utilizzo di dati, ad esempio, esclusivamente alla lingua inglese. Il comparto dei dati è stato poi completato con l'introduzione di un *dataset* aggiuntivo per la rilevazione di specifici bias. Tale *dataset* è composto da affermazioni differenziate secondo la valutazione di offensività (affermazioni neutre come «Io sono …», o offensive come «Io odio …» dove la rispettiva connotazione è popolata da termini identitari come «musulmano» o «buddista») o secondo l'utilizzo di verbi o aggettivi particolarmente qualificanti («odiare» o «amare»; «grande, forte, disgustoso» ecc.).

Al di là di un'analisi completa dei risultati della ricerca, occorre rilevare come il semplice utilizzo di alcuni termini riferiti all'identità personale renda gli algoritmi propensi a catalogare un commento come offensivo e tali risultati variano a seconda dei termini, genere e lingua utilizzati. Se si osservano, ad esempio, i risultati in lingua inglese relativi all'affermazione di appartenenza religiosa, si rileva come a fronte di una bassa probabilità di offensività di affermazioni relative all'appartenenza religiosa cristiana o buddista si contrappongono forti probabilità che frasi come «I am jewish» o «I am muslim» siano generalmente registrate dall'algoritmo come offensive. Se, ancora, le medesime affermazioni vengono analizzate in lingua italiana, la declinazione di genere comporta una più alta probabilità di previsione di offensività. Così, affermare di essere «musulmana» è più offensivo che dirsi «cristiana»; dichiararsi «ebrea» è meno offensivo del suo omologo al maschile[22]. Risultati rilevanti provengono ancora dall'analisi condotta secondo il metodo della «library lime», del sistema cioè di assegnazione di un valore numerico a ogni parola, indicando quanto quest'ultima abbia contribuito alla previsione algoritmica di offensività. Al riguardo, affermazioni connotate secondo l'identità religiosa o sessuale (come «amo tutti i musulmani» o «affermo di essere lesbica») contengono un'altissima probabilità di offensività, laddove il grado di quest'ultima è quasi interamente attribuibile all'utilizzo della connotazione identitaria prescelta[23].

Quanto fin qui analizzato ci consente di giungere ad alcune conclusioni provvisorie. Sebbene non sempre le affermazioni analizzate siano nella realtà dei fatti offensive, l'utilizzo di particolari termini legati all'identità personale, prima di tutto religiosa e/o di genere, contribuisce di *default* a classificare queste ultime come tali. In questo senso, distorsioni di tal tipo possono essere contrastate dal lavoro, tutto umano, di revisione del

---

[22] Cfr. la tabella 1, a p. 64 del già citato report *Bias in Algorithms – Artifical Intelligence and Discrimination*.

[23] Cfr. i grafici 12 e 13 a p. 66 del report.

risultato algoritmico, discernendo dove la previsione di offensività si riveli concretamente tale e dove essa nasconda un bias di funzionamento della macchina. Sebbene, come analizzato in precedenza, i bias non contengano necessariamente componenti discriminatorie, i risultati della ricerca confermano il peso del pregiudizio umano sul funzionamento tecnologico. Così la connotazione negativa che emerge dall'applicazione dell'algoritmo restituisce un'immagine dell'approccio alle differenze identitarie stereotipato e tendenzialmente pregiudizievole per le categorie protette. Un'immagine, cioè, di generale respingimento delle logiche sociali e delle politiche inclusive che sottostanno all'affermazione dei diritti fondamentali della persona umana.

4. Oltre la neutralità: allargare le competenze

L'analisi delle ricerche condotte dall'Agenzia europea per i diritti fondamentali fornisce dati concreti per comprendere se e come il legame tra bias e algoritmi determini il rischio di un utilizzo discriminatorio dell'Intelligenza Artificiale. Alcune osservazioni conclusive si rivelano essenziali per individuare alcune prime possibili risposte ai problemi posti dall'utilizzo delle nuove tecnologie.

Il primo ordine di riflessione attiene alla modalità di valutazione dei risultati algoritmici. Al riguardo, appare necessario prendere in considerazione la più che probabile sussistenza di bias qualora l'oggetto dell'indagine coinvolga informazioni connotate identitariamente secondo il fattore religioso e/o di genere. Così, se non inclusi nei dati di addestramento degli algoritmi, i riferimenti a tali categorie, separate o in intersezione tra loro, possono portare a previsioni distorte. Una risposta a tali problemi non sembra risiedere, tuttavia, nell'adozione di una logica di neutralità. I dati di addestramento privi di connotazioni identitarie potrebbero, ad esempio, ridurre la capacità della macchina di intercettare contributi problematici e le strategie di mitigazione dell'offensività basati sulla neutralità potrebbero, a loro volta, emarginare le voci dei gruppi più vulnerabili. Allo stesso tempo, optare per forme assolute di protezione che vietino la raccolta di dati sulle caratteristiche protette degli individui, potrebbe condurre a una sottorappresentazione di interi *target* di popolazione, con ciò rendendo i risultati dell'analisi non veritieri. Anche puntare interamente sulla risoluzione della questione sulla base di una previsione assoluta di offensività non appare una soluzione soddisfacente. Cosa accadrebbe, ad esempio, entro le specifiche comunità di appartenenza se il solo utilizzo del termine «ebreo» o di espressioni tipiche dello *slang* afro-americano comportasse il blocco di un'utenza

social? È di tutta evidenza come gli individui rischierebbero di subire una penalizzazione nell'accesso ai servizi di comunicazione sulla sola base di un uso incrementale di termini descrittivi della propria condizione personale.

Che fare, dunque, per mitigare i rischi di bias nell'utilizzo dell'Intelligenza Artificiale? Certamente un insieme di strategie combinate sembra la strada più funzionale. Si tratta, cioè di lavorare sull'incremento di conoscenze, consapevolezza e risorse per la verifica del cattivo funzionamento degli algoritmi. Ma si tratta, altresì, di migliorare la qualità dei dati a disposizione e di diversificare le competenze scientifiche all'interno delle équipe di valutazione. In tal senso, uno sguardo interculturale sulla costruzione della tecnologia e dell'algoritmo, come pure sulla comprensione plurale delle società, appare un punto imprescindibile. L'enfasi sui diritti umani rende i sistemi di Intelligenza Artificiale non solo più affidabili, ma anche più attraenti. Decidere di non investire in questo ambito non costituisce un'ipotesi accettabile.