Intelligenza artificiale e diritti delle donne: siamo dinanzi ad un algoritmo maschilista?

Susanna Viggiani *

ARTIFICIAL INTELLIGENCE AND WOMEN'S RIGHTS: ARE WE UP AGAINST A SEXIST ALGORITHM?
ABSTRACT: The algorithms used by AI systems are developed by scientists and computer scientists, who train the algorithm on the basis of a set of data, which - sometimes - can hide their own biases and errors. One of the most widespread and pervasive prejudices is that of gender, which tends to manifest itself in a variety of contexts, from automated decision-making processes in human resources and credit access systems to its use in so-called 'revenge porn', whose victims are, in fact, mostly women. Such forms of discrimination result in a violation of the principles and freedoms that underpin our democracies. Therefore, to struggle against the perpetuation of such stereotypes and prejudices, it is important to be awareness-raising and to address diversity as a fundamental priority of policies and institutional structures.

KEYWORDS: Algorithms; artificial intelligence; principle of equality; discrimination; women.

ABSTRACT: Gli algoritmi impiegati dai sistemi di AI sono elaborati da scienziati ed informatici, i quali addestrano l'algoritmo sulla base di una serie di dati, che – talvolta - possono nascondere i loro stessi pregiudizi ed errori. Uno dei pregiudizi maggiormente diffuso e pervasivo è quello di genere, il quale tende a manifestarsi in svariati contesti, dai processi decisionali automatizzati nelle risorse umane, nei sistemi di accesso al credito sino ai suoi utilizzi nel c.d. "*Revenge porn*", le cui vittime sono, infatti, per la maggioranza donne. Tali forme di discriminazione si traducono in una violazione dei principi e delle libertà che sono alla base delle nostre democrazie. Per questi motivi, per combattere il perpetuarsi di tali stereotipi e pregiudizi, è necessario esserne consapevoli e affrontare la diversità come priorità fondamentale delle politiche e delle strutture istituzionali.

PAROLE CHIAVE: Algoritmi; intelligenza artificiale; principio di uguaglianza; discriminazioni; donne.

* *Consulente Legale e Privacy, specializzata SPISA. Mail: viggianisusanna@gmail.com.* *Contributo sottoposto a doppio referaggio anonimo.*

ISSN 2284-4503

1. Il divieto di discriminazione

Il divieto di discriminazione affonda le proprie radici nell'idea più generale di uguaglianza, quale pilastro fondamentale dello Stato di diritto[1]. Il diritto antidiscriminatorio è il risultato dell'incontro di norme di diritto nazionale, di norme di recepimento di direttive europee[2] e di norme primarie dell'Unione Europea. A livello nazionale, il divieto di discriminazione trova la sua consacrazione nel principio di uguaglianza, di cui all'art. 3 commi 1 e 2 della Costituzione, nella duplice forma di uguaglianza formale e sostanziale. Tali principi sostengono non solo che tutti i cittadini hanno pari dignità dinanzi alla legge senza discriminazione di sesso, razza, religione, ma altresì il divieto di adottare comportamenti differenziati in situazioni eguali. Il divieto di discriminazione cui si riferisce la nostra Costituzione non prevede, però, una parificazione assoluta e indiscriminata, nel senso che il divieto si rivolge a tutte quelle caratteristiche immutabili del soggetto o scelte personalissime dello stesso che ne rafforzano la dignità[3]. Ne consegue, dunque, che il *Leitmotiv* del divieto di discriminazione si traduce in comportamenti volti ad impedire distinzioni produttive di diseguaglianze.

Nel contesto del diritto europeo, il divieto di discriminazioni ha assunto caratteri particolari, dettati dalle specificità delle competenze e delle funzioni di cui l'Unione Europea è investita. Innanzitutto, l'art. 14 CEDU sancisce il divieto di discriminazione: «il godimento dei diritti e delle libertà riconosciuti nella presente Convenzione deve essere assicurato senza discriminazione alcuna, di sesso, di razza, di colore, di lingua, di religione, di opinione politica o di altro genere, di origine nazionale o sociale, di appartenenza a una minoranza nazionale, di ricchezza, di nascita o di altra condizione». A partire dagli anni 2000, all'art. 14 CEDU viene affiancato il Protocollo Addizionale n. 12, il quale riconosce il divieto di discriminazione ancorandolo non più ai soli diritti sanciti dalla Convenzione, ma a tutti i diritti previsti a livello nazionale[4]. Tuttavia, la legittimazione del diritto antidiscriminatorio si avrà solo con la Carta dei diritti fondamentali e specificatamente all'art. 21. Il testo dell'art. 21 dispone, infatti, che «è vietata qualsiasi forma di discriminazione fondata, in particolare, sul sesso, la razza, il colore della pelle o l'origine etnica o sociale, le caratteristiche genetiche, la lingua, la religione o le convinzioni personali, le opinioni politiche o di qualsiasi altra natura, l'appartenenza ad una minoranza nazionale, il patrimonio, la nascita, la disabilità, l'età o l'orientamento sessuale». Il divieto di discriminazione concretamente, quindi, si declina in svariati comportamenti, potenzialmente in grado di

[1] G. Dodaro, *Uguaglianza e diritto penale. Uno studio sulla giurisprudenza costituzionale*, Milano, 2012, 9.

[2] Art. 2, § 2, Dir. 2000/43/CE in materia di discriminazioni per razza e origine etnica; Art. 2, § 2, Dir. 2000/78/CE in materia di discriminazioni per religione, convinzioni personali, handicap, età, tendenze sessuali; Art. 2, § 2, Dir. 2006/54/CE in materia di discriminazioni di genere. Nella disciplina interna di recepimento: art. 2 D.lgs. n. 215/2003 in materia di discriminazioni per razza e origine etnica; art. 2 D.lgs. n. 216/2003 in materia di discriminazioni per religione, convinzioni personali, handicap, età, tendenze sessuali; art. 25 D.lgs. n. 198/2006 in materia di discriminazioni di genere.

[3] S. Rodotà, *Il diritto di avere diritti*, Roma-Bari, 2012, 184 ss.; M. Militello, *Principio di uguaglianza e di non discriminazione tra Costituzione italiana e Carta dei diritti fondamentali dell'Unione Europea*, in *Biblioteca "20 Maggio"*, 2010, 1, 158, «*in virtù del legame esistente tra la tutela antidiscriminatoria e la connotazione sociale e storica dei divieti si ricava una lettura obbligata per cui i divieti di discriminazione, più che sancire l'irrilevanza di determinate qualità soggettive, sono destinati ad impedire che esse si traducano in distinzioni produttive di diseguaglianze*».

[4] Nel Preambolo si parla di un «*principio fondamentale, secondo il quale tutte le persone sono uguali innanzi alla legge e hanno diritto alla stessa protezione da parte della legge*».

ISSN 2284-4503

generare una disuguaglianza rilevante. L'elaborazione dell'idea di discriminazione nelle sue diverse accezioni si deve alle Direttive di matrice europea, che hanno consentito di approfondire la nozione di discriminazione distinguendo, al suo interno, tra discriminazioni dirette ed indirette, molestie, molestie sessuali, ritorsioni e ordini di discriminare[5]. In particolare, la discriminazione diretta[6] si realizza in tutte quelle ipotesi in cui un soggetto è trattato meno favorevolmente di quanto sia, sia stata o sarebbe trattata un'altra persona in una situazione analoga. Al contrario, si verifica una discriminazione indiretta[7] se una disposizione o una prassi apparentemente neutra contribuisce a mettere un certo soggetto in una posizione di particolare svantaggio rispetto ad altre persone. Più di recente ai concetti cardine di discriminazione diretta e indiretta si è aggiunto il concetto di discriminazione organizzativa, che, ai sensi del Codice delle Pari Opportunità – d.lgs. 198/2006 – all'art. 25 comma 2*bis* chiarisce che costituisce discriminazione ogni trattamento o modifica dell'organizzazione delle condizioni e dei tempi di lavoro che, in ragione del sesso, dell'età anagrafica, delle esigenze di cura personale o familiare, dello stato di gravidanza nonché di maternità o paternità, anche adottive, ovvero in ragione della titolarità e dell'esercizio dei relativi diritti, pone o può porre il lavoratore in almeno una delle seguenti condizioni: a) posizione di svantaggio rispetto alla generalità degli altri lavoratori; b) limitazione delle opportunità di partecipazione alla vita o alle scelte aziendali; c) limitazione dell'accesso ai meccanismi di avanzamento e di progressione nella carriera.

Tali fenomeni risultano oggi accentuati dall'ingresso della tecnologia nell'agire pubblico e privato, e più segnatamente, dall'impiego di strumenti di intelligenza artificiale (d'ora in avanti IA o AI). I sistemi di intelligenza artificiale stanno contribuendo all'emersione di maggiori criticità a fronte di trattamenti differenti operati non più solo dall'azione umana, ma influenzati altresì dal funzionamento di una macchina[8] capace di utilizzare algoritmi di apprendimento automatico, in grado di analizzare grandi volumi di dati di addestramento per identificare correlazioni, schemi e metadati. Gli algoritmi che permettono all'intelligenza artificiale di funzionare sono algoritmi talvolta artefatti, poiché risentono

[5] F. Amato, *Le nuove direttive comunitarie sul divieto di discriminazione. Riflessioni e prospettive per la realizzazione di una società multietnic*a, in *Lavoro e diritto*, 2003, 127 ss.

[6] CGUE, *causa C-507/18, sentenza della Grande Camera del 23 aprile 2020, su rinvio pregiudiziale relativo al caso NH c. Associazione Avvocatura per i diritti LGBTI – Rete Lenford*: L'Associazione avvocatura per i diritti LGBTI aveva convenuto in giudizio davanti al Tribunale di Bergamo l'avvocato NH per alcune sue dichiarazioni considerate contrarie al divieto di non discriminazione dei lavoratori in base agli orientamenti sessuali. A seguito della decisione del suddetto Tribunale che dichiarava illecite le sue dichiarazioni, l'Avvocato NH ha presentato ricorso giungendo fino alla Corte di Cassazione italiana che ha effettuato il rinvio pregiudiziale. La Corte di giustizia nella sentenza afferma che nella nozione di "condizioni di accesso all'occupazione e al lavoro" rientrano anche le dichiarazioni rese da una persona durante una trasmissione audiovisiva in base alle quali egli non assumerebbe mai nella sua impresa una persona con un determinato orientamento sessuale. Ciò vale anche nel caso in cui in quel momento non sia in corso alcuna selezione, sempre che vi sia un collegamento non ipotetico tra dette dichiarazioni e le condizioni di accesso all'occupazione all'interno di detta impresa.

[7] *Corte di cassazione, sezione lavoro, ordinanza 21 aprile 2020 n.7982*: In tema di requisiti per l'assunzione, sussiste una discriminazione indiretta qualora sia previsto come requisito una statura minima identica per uomini e donne, in contrasto con il principio di uguaglianza, presupponendo erroneamente la non sussistenza della diversità di statura mediamente riscontrabile tra uomini e donne.

[8] L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Milano, 2017: «L'impatto dell'intelligenza artificiale supera, come noto, la dimensione più circoscritta e propria del fenomeno discriminatorio, tanto da aver indotto autorevole dottrina a coniare l'espressione ormai famosa di realtà «on-life» a enfatizzare come la realtà virtuale si sta oppure si è ormai imposta al fianco di quella materiale e concreta».

dei significati, dei concetti, delle idee, dei giudizi e dei pregiudizi che l'essere umano apprende sin dalla nascita. Ne deriva che, certe decisioni vengono adottate mediante il supporto di sistemi di IA che sbagliano, perché non sono in grado di profilare attendibilmente per via di dati incompleti, obsoleti o *biased*, di errori nella costruzione degli algoritmi o di limitazioni al loro uso. Si parla, in tali casi, di discriminazioni algoritmiche, perché la determinazione o la pratica che definisce svantaggi per taluni soggetti è adottata o attuata (anche) mediante l'impiego di algoritmi, compresi quelli dell'IA[9].

2. Tipologie di algoritmi e lavoro

L'avvento dell'intelligenza artificiale sta rivoluzionando molti settori della nostra esistenza, tra i quali, in via principale, tutto ciò che attiene al mondo del lavoro. Il tema delle discriminazioni dettate dagli algoritmi, infatti, sta prendendo sempre più spazio nel mondo del lavoro, sia rispetto all'accesso al mercato del lavoro sia in relazione alle condizioni contrattuali. Pensati e scritti dall'essere umano, gli algoritmi si rivelano potenzialmente in grado di riprodurre nella sfera digitale i pregiudizi e gli stereotipi già esistenti nella realtà. Per questo motivo, è fondamentale la qualità dei *dataset* utilizzati, i quali devono essere sufficientemente completi e ampi da non ricreare i pregiudizi e le discriminazioni già presenti nella realtà sociale. Tuttavia, non tutti gli algoritmi operano nello stesso modo, per cui, risulta necessario, anzitutto, comprendere cosa siano gli algoritmi e, perché e come il loro utilizzo possa autorizzare processi decisionali discriminatori. Con riferimento alle differenti tipologie di algoritmi, è possibile distinguere tra Algoritmi *rule-based* e Algoritmi di *machine learning.* Entrambi rientrano all'interno della definizione di AI prevista dal Regolamento sull'IA all'art. 3 e al Cons. n. 12[10]. La distinzione è rilevante soprattutto per capire le modalità e i differenti gradi con cui si possono presentare i

[9] G. Gometz, *Intelligenza artificiale, profilazione e nuove forme di discriminazione,* in *Il lato oscuro della legge* (a cura di F. Mancuso e V. Giordano), *Teoria e storia del diritto privato,* www.teoriaestoriadeldirittoprivato.com

[10] Art. 3 "Definizioni"- Reg. UE 1689/2024 – AI Act: «sistema di IA": un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali.».
Cons. 12- Reg. UE 1689/2024 – AI Act: «La nozione di "sistema di IA" di cui al presente regolamento dovrebbe essere definita in maniera chiara e dovrebbe essere strettamente allineata al lavoro delle organizzazioni internazionali che si occupano di IA al fine di garantire la certezza del diritto, agevolare la convergenza internazionale e un'ampia accettazione, prevedendo nel contempo la flessibilità necessaria per agevolare i rapidi sviluppi tecnologici in questo ambito. Inoltre, la definizione dovrebbe essere basata sulle principali caratteristiche dei sistemi di IA, che la distinguono dai tradizionali sistemi software o dagli approcci di programmazione più semplici, e non dovrebbe riguardare i sistemi basati sulle regole definite unicamente da persone fisiche per eseguire operazioni in modo automatico. Una caratteristica fondamentale dei sistemi di IA è la loro capacità inferenziale. Tale capacità inferenziale si riferisce al processo di ottenimento degli output, quali previsioni, contenuti, raccomandazioni o decisioni, che possono influenzare gli ambienti fisici e virtuali e alla capacità dei sistemi di IA di ricavare modelli o algoritmi, o entrambi, da input o dati. Le tecniche che consentono l'inferenza nella costruzione di un sistema di IA comprendono approcci di apprendimento automatico che imparano dai dati come conseguire determinati obiettivi e approcci basati sulla logica e sulla conoscenza che traggono inferenze dalla conoscenza codificata o dalla rappresentazione simbolica del compito da risolvere. La capacità inferenziale di un sistema di IA trascende l'elaborazione di base dei dati consentendo l'apprendimento, il ragionamento o la modellizzazione. Il termine "automatizzato" si riferisce al fatto che il funzionamento dei sistemi di IA prevede l'uso di macchine.»

rischi per i diritti dei lavoratori, con particolare riguardo alle necessità che si pongono nella costruzione di garanzie di trasparenza ed eliminazione di ogni forma di discriminazione.

Gli algoritmi del primo tipo si qualificano come sistemi basati sulla logica, statici perché si alimentano di una serie di istruzioni fisse e modificabili solo in fase di programmazione e il cui risultato è prevedibile *ex ante*, perché tutte le variabili e i risultati possibili sono già programmati nell'algoritmo. Nel contesto lavoristico, un esempio di tale algoritmo si rinviene nell'algoritmo *Frank*, conosciuto per il suo utilizzo nel caso *Deliveroo*. Tale algoritmo si basava su un sistema di apprendimento automatico, impiegato per valutare le prestazioni dei lavoratori e classificarli sulla base dei parametri di affidabilità e partecipazione e dare, quindi, la precedenza sugli ordini ai migliori in classifica. Sul punto è intervenuto il Tribunale di Bologna[11], il quale ha accertato la natura discriminatoria ex art. 2 d.lgs. 216/2003 della condotta di *Deliveroo* in relazione alle condizioni di accesso alla prenotazione delle sessioni di lavoro, in quanto l'algoritmo non prendeva in considerazione la legittimità delle motivazioni di astensione dal lavoro – quali malattia, stato di necessità o esercizio del diritto di sciopero – e penalizzava ingiustamente i lavoratori tramite un abbassamento delle statistiche, a cui conseguiva una riduzione delle occasioni lavorative e quindi retributive.

Altro esempio di questo tipo, si può ravvisare, nel caso *Foodinho*[12], il quale impiegava una piattaforma, tramite la quale ai *riders* venivano assegnati "punteggi di eccellenza" in ragione di specifici criteri di produttività, tra cui il numero di consegne e la disponibilità nelle fasce orarie ad alta richiesta e nel fine settimana, tenendo conto, inoltre, della mancata presentazione negli slot prenotati. Sulla base di tali punteggi, infatti, la piattaforma consentiva ai lavoratori di scegliere in anticipo la collocazione delle successive prestazioni. Nel caso di specie, il Tribunale di Palermo ha ritenuto sussistente una discriminazione indiretta, in quanto prevedeva l'attribuzione di un punteggio negativo ai *riders* nelle ipotesi di c.d. *late cancellation*, ossia di cancellazione o annullamento della prenotazione di uno slot con un preavviso inferiore alle 24 ore, senza valutare le motivazioni che avevano dato luogo alla cancellazione[13].

Sono, invece, algoritmi del secondo tipo, tutti quegli algoritmi basati su un metodo statistico/probabilistico, dinamici, perché l'insieme delle istruzioni è calcolato nel tempo in modo automatizzato nella fase di addestramento e il cui risultato non è prevedibile *ex ante*, perché deriva da correlazioni di natura probabilistica, ma può essere, almeno teoricamente, compreso e spiegato solo *ex post*. Il risultato ottenuto, però, nella maggior parte dei casi, risulta poco trasparente, poiché pienamente comprensibile solo a un soggetto esperto, con la conseguenza della necessità di un intervento successivo volto a facilitarne la comprensione da parte di un soggetto non esperto[14].

Nel contesto lavorativo, un esempio potrebbe ravvisarsi nel sistema impiegato da *Amazon* per la selezione del personale: l'algoritmo era stato addestrato a stilare la graduatoria dei migliori candidati

[11] Tribunale di Bologna, sez. Lavoro, ordinanza 31 dicembre 2020, (discriminazione algoritmica di lavoratori), https://giurcost.org/casi_scelti/GM/TribunaleBologna31dicembre2020.pdf

[12] Tribunale di Palermo, sez. Lavoro, sentenza 17 novembre 2023, https://onelegale.wolterskluwer.it/document/10SE0002795397

[13] A. Perulli, *La discriminazione algoritmica: brevi note introduttive a margine dell'Ordinanza del Tribunale di Bologna*, in *Lavoro Diritti Europa*, 1, 2021, 2.

[14] M. Peruzzi, *Il diritto antidiscriminatorio al test di intelligenza artificiale*, in *Lab. & Law Issues*, 2021, 1, pp. 50 ss.

per posizioni di ingegneri informatici osservando i *curricula* ricevuti nei dieci anni precedenti. La maggior parte degli stessi proveniva da uomini, maggiormente impiegati in settori tecnologici. Sulla base di queste statistiche, l'algoritmo aveva "insegnato" a sé stesso a preferire i candidati uomini rispetto alle candidate donne, nel senso che, pur essendo stato addestrato a non utilizzare direttamente il sesso come criterio selettivo, era riuscito a riconoscerlo da altre informazioni, comunque presenti nei *curricula*, utilizzando poi questi indici di genere come criteri utili ad effettuare la selezione. In alcuni casi, veniva favorito, addirittura, chi utilizzava alcuni termini - come verbi in forma attiva - che, nel campione storico di *curricula* alla base del modello decisionale algoritmico, erano statisticamente usati più dagli uomini che dalle donne.

Nel mercato del lavoro automatizzato, infatti, si riproducono potenzialmente gli stessi atteggiamenti discriminatori che si riscontrano nei lavori tradizionali, con riguardo a tutti i fattori di discriminazione, poiché le menti che programmano gli algoritmi sono menti umane[15]. I sistemi di IA discriminano, infatti, non perché il sistema sia di per sé maligno, ma perché eredita comportamenti sbagliati che poi ripete. Ciò che è noto, è che le discriminazioni algoritmiche assumono oggi una portata drasticamente pervasiva, capace di determinare conseguenze distruttive sulla società. Simili algoritmi, infatti, se guidati da dati imprecisi, parziali o non rappresentativi del fenomeno a cui si applicano, possono produrre risultati non trasparenti e distorti e condurre, perciò, a varie forme di discriminazione. Come affermato rispettivamente dai Tribunali di Bologna[16] e di Palermo[17], l'incoscienza della macchina e l'applicazione indifferenziata dei parametri di valutazione non giustificano la discriminazione, proprio per il particolare svantaggio che tali parametri implicano nei confronti dei portatori di determinati fattori di rischio, quali anzitutto l'affiliazione sindacale, il genere, la religione e la disabilità.

3. Fattori scatenanti la discriminazione

Gli algoritmi funzionano secondo la logica *garbage in – garbage out*, secondo cui dati incongrui, inesatti o non aggiornati possono generare solamente risultati decisionali inaffidabili. A differenza dei tradizionali sistemi informatici, l'AI, infatti, non si limita ad eseguire istruzioni predefinite, ma impara e genera contenuti basandosi sui dati forniti. Per questo, alimentare un sistema AI con dati non accurati porta a risultati imprevedibili e potenzialmente rischiosi.

Il primo fattore rilevante che contribuisce al perpetuarsi delle discriminazioni algoritmiche è sicuramente rappresentato dalla componente umana. Sebbene gli algoritmi operino sempre di più in modo autonomo, il ruolo degli esseri umani resta comunque cruciale per il loro sviluppo e funzionamento. È la persona, immersa in una realtà non scevra di pregiudizi, a fornire i dati alla macchina e, pure a fronte di tecniche che palesano un livello di progressiva autonomia dalla decisione o programmazione originaria, esse rimangono pur sempre il prodotto di un'azione umana, non sempre imparziale. Non a caso, infatti, l'AI Act[18] adotta un approccio basato sul rischio: maggiore è il rischio che l'applica-

[15] C. Alessi, *Lavoro tramite piattaforma e divieti di discriminazione nell'UE*, in C. Alessi, M. Barbera, L. Guaglianone (a cura di) *Impresa, lavoro e non lavoro nell'impresa digitale*, Bari, 2019, 663 ss.

[16] Tribunale di Bologna, sez. Lavoro, ordinanza 31 dicembre 2020.

[17] Tribunale di Palermo, sez. Lavoro, sentenza 17 novembre 2023.

[18] https://artificialintelligenceact.eu/

zione dell'AI può causare per i diritti e le libertà fondamentali degli interessati, più rigidi sono gli obblighi di sicurezza e trasparenza previsti sia in capo ai produttori che agli utilizzatori. L'AI Act classifica i sistemi di AI secondo quattro livelli di rischio e associa ad ognuno di essi delle salvaguardie che ne compensino la pericolosità. In particolare, vengono individuati sistemi a rischio inaccettabile, in tale ipotesi vi rientrano i sistemi o le applicazioni di IA che influenzano in maniera significativa gli utenti, distorcendone il comportamento, mediante tecniche manipolative, ingannevoli e/o sfruttandone le diversità e vulnerabilità. Pratiche di questo genere possono causare una lesione dei diritti fondamentali delle persone e, di conseguenza, lo sviluppo e la diffusione degli stessi è vietata all'interno dell'UE; sistemi a rischio alto, si fa riferimento a tutti quei sistemi appositamente identificati[19], che potrebbero comportare delle conseguenze sulla salute, sulla sicurezza o sui diritti fondamentali delle persone, e che per essere ammessi all'interno dell'UE, dovranno soddisfare requisiti rigorosi previsti dagli artt. 8 - 49 dell'AI Act. Per questa particolare tipologia di sistemi, si prevede che la sorveglianza umana debba essere affidata a persone fisiche differenti, allo scopo di verificare in che misura il sistema elaborato rischia di incidere sui diritti fondamentali dei cittadini, procurandone anche potenzialmente discriminazioni. L'art. 14 AI Act, nello specifico, teorizza il principio etico-giuridico di *Human-In-The-Loop*[20] e cioè quel principio in base al quale è necessario garantire che gli individui siano informati delle scelte che li riguardano e dell'impiego di strumenti di AI e più nel dettaglio è quella metodologia secondo cui gli esseri umani etichettano i dati, il che aiuta il modello a ottenere dati di addestramento di alta qualità e in quantità elevata. Il modello si basa sul connubio tra le capacità delle macchine e l'intelligenza umana, le quali attraverso varie interazioni contribuiscono ad alimentare il modello di apprendimento automatico. In altre parole, l'intelligenza umana dovrebbe intervenire quando la macchina ha difficoltà a risolvere un problema.

Secondo il Gruppo di Esperti[21], i sistemi di AI dovrebbero essere progettati per aumentare, integrare e potenziare le abilità cognitive, sociali e culturali umane, senza sostituirsi completamente ad esso. L'apporto del supervisore umano sulla macchina può manifestarsi in varie forme, non solo come previsto dal c.d. *Human In The Loop*, ma altresì nelle sue declinazioni di *Human On The Loop* che assicura un controllo minimo, in fase di progettazione o di monitoraggio e *Human In Command*, che consente un monitoraggio costante sul sistema e sui suoi effetti, lasciando ampia discrezionalità al supervisore con la conseguenza, però, che il supervisore umano potrebbe decidere di ignorare la decisione

[19] Capo III "Sistemi di IA ad alto rischio" - Reg. UE 1689/2024- artt. 6 e ss. Si definisce ad alto rischio se sono soddisfatte entrambe le condizioni seguenti: a) il sistema di IA è destinato ad essere utilizzato come componente di sicurezza di un prodotto, o il sistema di IA è esso stesso un prodotto, disciplinato dalla normativa di armonizzazione dell'Unione elencata nell'Allegato I dell'AI Act; b) il prodotto, il cui componente di sicurezza a norma della lett. a) è il sistema di IA, o il sistema di IA stesso in quanto prodotto, è soggetto a una valutazione della conformità da parte di terzi ai fini dell'immissione sul mercato o della messa in servizio di tale prodotto ai sensi della normativa di armonizzazione dell'Unione elencata nell'Allegato I. Oltre ai sistemi di IA ad alto rischio di cui sopra, sono considerati ad alto rischio anche i sistemi di IA di cui all'Allegato III dell'AI Act.

[20] Art. 14 par. 2 "*Sorveglianza Umana*" «La sorveglianza umana mira a prevenire o ridurre al minimo i rischi per la salute, la sicurezza o i diritti fondamentali che possono emergere quando un sistema di IA ad alto rischio è utilizzato conformemente alla sua finalità prevista o in condizioni di uso improprio ragionevolmente prevedibile, in particolare qualora tali rischi persistano nonostante l'applicazione di altri requisiti di cui alla presente sezione».

[21] High-Level Expert Group on AI, *Ethics guidelines for trustworthy AI*, https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

assunta mediante l'AI. Non a caso, il paragrafo 4, lettera b) dell'art. 14 AI Act stabilisce che le persone, preposte alla sorveglianza, devono restare consapevoli «della possibile tendenza a fare automaticamente affidamento o a fare eccessivo affidamento sull'output prodotto da un sistema di IA ad alto rischio ("*automation bias*[22]"), in particolare per i sistemi di IA ad alto rischio utilizzati per fornire informazioni o raccomandazioni per le decisioni che devono essere prese da persone fisiche». Da qui l'esigenza dell'intervento di un'altra persona fisica, in grado di verificare la logicità e la legittimità delle scelte dettate in fase di input e degli esiti e intervenire sulla decisione automatizzata[23].

Le altre due tipologie di rischio vengono individuate in sistemi di AI a rischio di trasparenza, cioè quei sistemi destinati ad interagire direttamente con le persone fisiche, o che generano o manipolano immagini, contenuti, audio o video, e che possono comportare specifici rischi di furti di identità, manipolazioni o inganni (es. *chatbots* o *deep fakes*). Per queste ipotesi, sono previsti specifici obblighi di informazione e trasparenza. Gli utenti devono essere informati quando interagiscono con un sistema di Intelligenza Artificiale o quando un contenuto è stato generato da un'IA, al fine di consentire loro di prendere decisioni informate e interagire con la tecnologia in modo consapevole.

I sistemi a rischio minimo, infine, sono quei sistemi che presentano rischi minimi o nulli per i diritti e/o la sicurezza dei cittadini. Tali sistemi sono attualmente esenti da obblighi specifici, tuttavia, potranno essere introdotti dei codici di buone pratiche.

Altro fattore produttivo di discriminazione si rinviene nella qualità e nella difficile comprensibilità dei dati utilizzati nel processo decisionale algoritmico che, se non quantitativamente o qualitativamente adeguati, possono viziarlo e replicare discriminazioni insite nel loro stesso processo di programmazione e addestramento. L'opacità o l'effetto scatola nera, infatti, rende difficoltoso determinare dove si trova la radice della discriminazione: sistemi decisionali automatizzati o software semiautomatizzati possono contenere *bias* non intenzionali introdotti da loro programmatori, o, se intenzionali, possono essere nascosti o mascherati in un codice molto complesso e di non facile comprensione. Motivo per cui, diviene più difficile per le vittime di tali discriminazioni rendersi conto delle stesse. L'art. 13 dell'AI Act, al fine di scongiurare questi pericoli, si concentra sugli aspetti di trasparenza e sulle informazioni da procurare dai fornitori specialmente di sistemi di IA definiti ad alto rischio. L'AI Act sviluppa tre concetti fondamentali: trasparenza, interpretabilità e spiegabilità, al fine di evitare qualsiasi forma di discriminazione e rendere più facilmente comprensibile il funzionamento del sistema algoritmico impiegato. Per trasparenza si intende che i sistemi di AI devono esser progettati e sviluppati in modo tale da rendere i fornitori in grado di interpretare l'output del sistema e usarlo in maniera appropriata. L'interpretabilità si riferisce alla capacità di una persona di comprendere il funzionamento interno di un sistema di IA e nello specifico si traduce in un funzionamento del modello sufficientemente trasparente per permettere agli utenti di discernere come gli input siano trasformati in output. La spiegabilità, invece, si concentra sulla capacità di articolare gli esiti di un sistema di IA in termini comprensibili all'uomo. Può avvalersi dell'impiego di strumenti e metodi supplementari volti a fornire chiarimenti su come il sistema di IA giunga a determinate decisioni con l'obiettivo di colmare

[22] Per *automation bias* si intende la tendenza da parte dell'essere umano, coinvolto nell'interazione con la macchina, ad affidarsi ai suoi output, fino a trascurare o ignorare altre informazioni che derivano da fonti diverse.

[23] G. Lo Sapio, *La trasparenza sul banco di prova dei modelli algoritmici*, in *Federalismi.it*, 11, 2021, 242 ss.

il divario tra la complessità dell'IA e la comprensione umana, consentendo agli utenti di apprendere le motivazioni dietro le decisioni dell'IA.
Infine, un ultimo elemento produttivo di discriminazione sarebbe ravvisabile, almeno nei c.d. *algoritmi machine learning*, nell'utilizzo di un *proxy*, un indicatore statistico di un'altra caratteristica a cui vengono ricollegati effetti sfavorevoli, che risulta, però, più difficilmente percepibile anche da chi programma ed eventualmente supervisiona il funzionamento dell'algoritmo stesso.

4. Proxy discriminations di genere

Un *proxy* è un elemento utilizzato da un sistema di intelligenza artificiale per fare distinzioni tra individui e/o gruppi sociali. La *proxy discrimination* – letteralmente discriminazione "per delega" – è da intendersi quale discriminazione provocata da un criterio apparentemente neutro, basato su una caratteristica (*proxy*) strettamente collegata ad un fattore protetto dalla normativa antidiscriminatoria. La discriminazione si verifica, quindi, ogni volta che il *proxy* perpetua pregiudizi influenzando in modo negativo determinati individui e gruppi, senza fondare la distinzione sui fattori classici della discriminazione, ma, piuttosto, basandosi su loro correlazioni presumibilmente non discriminatorie. Già nel caso *Coleman* del 2008, una madre aveva sostenuto di aver subito un trattamento discriminatorio sul posto di lavoro in ragione del fatto che il figlio fosse disabile. I giudici di Lussemburgo avevano rilevato che la tutela offerta dalla normativa dell'Unione rispetto al motivo della disabilità non andava riferita esclusivamente al figlio, ma poteva essere estesa anche alla madre, poiché il trattamento discriminatorio da essa subito era stato comunque posto in essere in ragione di quella disabilità[24]. In tali casi, il problema principale si rinviene nella prova della discriminazione: se la lavoratrice, ad esempio, riesce a dimostrare che un determinato criterio – basato su un certo *proxy* – adottato dal datore di lavoro produce un effetto pregiudizievole su tutti i lavoratori appartenenti ad una determinata categoria protetta, il giudice potrà anche dedurre l'esistenza di una discriminazione[25].
Tuttavia, valutare questi elementi non è così semplice, poiché gli utenti non sempre hanno gli strumenti per comprendere le modalità e i dati statistici circa i gruppi di utenti raggiunti o esclusi da un determinato trattamento. Invero, oggi sia le pubbliche amministrazioni che le aziende private stanno impiegando sempre più frequentemente algoritmi progettati per aiutare o sostituire le persone incaricate di prendere decisioni. Sistemi informatici costruiti, però, sulla base di pregiudizi discriminano sistematicamente e ingiustamente, negando opportunità o beni ovvero attribuendo un risultato indesiderato sulla base di motivazioni irragionevoli o inappropriate[26]. Le decisioni basate su algoritmi non sufficientemente trasparenti e addestrati possono avere un impatto di vario tipo sui diritti umani e le

[24] Corte di giustizia (Grande Sezione), sentenza del 17 luglio 2008, *causa C-303/06, Coleman*, EU:C:2008:415.

[25] A.E.R. Prince, D. Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, in *Iowa Law Review*, 2020, 105.

[26] B. Friedman, H. Nissembaum, *Bias in Computer Systems*, ACM Transactions on Information Systems, 3/1996, 332: «Accordingly, we use the term bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate».

libertà fondamentali, in particolare – ai fini della trattazione - sugli stereotipi di genere[27]. Le vittime delle discriminazioni algoritmiche risultano essere in prevalenza donne. Da sempre, le donne sono una delle categorie di individui vulnerabili che subiscono l'impatto delle trasformazioni sociali. Secondo i risultati del Global Gender Gap Report 2024 del World Economic Forum[28], le donne stanno scontando a caro prezzo gli squilibri sistemici del mercato del lavoro. Questi squilibri non solo significano che ci sono meno donne in ruoli di *leadership*, ma anche che quando ci sono *shock* economici, le donne sono più colpite.

L'intelligenza artificiale non rispecchia egualmente entrambi i sessi e, anzi, si traduce, talvolta, in una discriminazione strutturale ai danni delle donne. Tale affermazione è supportata da analisi statistiche che convergono nell'attestare la natura non *gender-neutral* dei modelli di AI impiegati nella grande maggioranza delle attività[29]. In particolare, in tale ipotesi si realizza una *proxy discrimination*[30], poiché il problema nasce dalla presenza nei *data-sets* di *redundant encodings*, ovverosia si cerca di mascherare l'appartenenza ad un determinato sesso o stato (stato di gravidanza o di invalidità) in altri dati associati alla medesima categoria. Invero, come noto, le tecniche di intelligenza artificiale funzionano sulla base delle associazioni tra dati, per cui la macchina riesce a selezionare tutti quegli elementi che consentono di raggiungere più facilmente il risultato desiderato anche ricorrendo ad altri fattori capaci di determinare, direttamente o indirettamente, l'affiliazione ad una categoria protetta.

Il caso *Amazon* - sopra esposto - è considerato una pietra miliare delle problematiche connesse all'*inclusive recruitment,* dal momento che ha consentito di denunciare – forse per la prima volta - la mancata inclusione di persone diverse tra loro all'interno dei team, nel rispetto delle pari opportunità[31]. Eppure, appare opportuno ricordare, che la stessa Unione europea ha posto la parità tra uomo e donna tra i suoi principi fondanti e come guida per i lavori dell'Eurofound[32], soprattutto per ciò che attiene alla parità di genere in ogni ambito lavorativo, inclusa la parità retributiva. Esemplare, ancora, è il caso di LinkedIn, secondo cui gli algoritmi impiegati dai motori di pubblicazione di annunci di lavoro e, quindi, utilizzati per abbinare i candidati alle rispettive opportunità, producevano risultati distorti, privilegiando candidati uomini rispetto alle donne: gli uomini risultavano maggiormente propensi a cercare nuove opportunità rispetto alle donne. Le principali società di reclutamento online abbinano candidati qualificati con le posizioni disponibili. Molte piattaforme, però, per pianificare il *matching* tra le posizioni disponibili e i candidati, utilizzano algoritmi c.d. di raccomandazione, che elaborano le informazioni ricevute dalle organizzazioni e da chi cerca lavoro per stilare un elenco di soggetti in li-

[27] M. D'AMICO, *Una parità ambigua. Costituzione e diritti delle donne*, Milano, 2020.

[28] https://www.weforum.org/publications/global-gender-gap-report-2024/.

[29] Si richiamano il *Progetto Gender Shades* sul carattere discriminatorio ai danni delle donne, in particolare afroamericane di alcuni sistemi di riconoscimento facciale (su cui, anche,); la vicenda di Amazon in tema di reclutamento, su cui J. DUSTIN, *Amazon scraps secret AI recruiting tool that showed bias against wome*n, in *Reuters*, 11 ottobre 2018; J. LAURET, *Amazon's sexist AI recruiting tool: how did it go so wrong?,* in *Becominghuman.ai*, 16 agosto 2019.

[30] A.E.R. PRINCE, D. SCHWARCZ, cit.

[31] C. DELLA GIUSTINA, *Quando il datore di lavoro diviene un algoritmo: la trasformazione del potere del datore di lavoro in algocrazia. Quale spazio per l'applicazione dei principi costituzionali*?, in *Media Laws,* 2021, 2.

[32] https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/search-all-eu-institutions-and-bodies/eurofound_it.

nea con la posizione lavorativa ricercata, «raccomandando» appunto solamente determinate categorie di soggetti, in prevalenza uomini. Anche, il rapporto dell'Unesco sugli impatti dell'IA nella vita lavorativa delle donne[33] denuncia, ad esempio, la riduzione della capacità delle donne africane di accedere al credito a causa dell'uso di sistemi di *credit scoring* che valutano l'impronta digitale di un individuo. Le differenze nell'uso e nell'accesso delle donne africane a Internet - il cosiddetto *digital divide* – diventano, in questo senso, un fattore discriminante rispetto alla possibilità di ottenere finanziamenti, che potrebbero, invece, rivelarsi di grande utilità per il riscatto di queste donne.

Le discriminazioni algoritmiche di genere, infatti - come sopra accennato- non si limitano solo all'ambito lavorativo: si pensi ai motori di ricerca come Google che associano la parola «infermiera» a una donna e la parola «dottore» a un uomo, confinando determinate categorie di lavori e di potere in capo ai solo soggetti uomini. Questi episodi di discriminazione da parte degli strumenti di intelligenza artificiale sono accompagnati dal rischio costante per le donne di essere esposte a violenza online, *cyber-stalking* o bullismo. Secondo quanto denunciato dal rapporto dell'Unesco[34], la tecnologia vocale con voce femminile dà troppo spesso risposte dal tono sottomesso rispetto alle domande. La maggior parte degli assistenti vocali presenta una voce femminile, trasmettendo un segnale di donne garanti, docili e desiderose di aiutare, sempre disponibili al solo e semplice tocco di un pulsante o con un comando vocale. La sottomissione con cui interagiscono, influenza il modo in cui la gente reagisce alle voci femminili e come le donne rispondono alle richieste e si esprimono. Questo rafforza i pregiudizi di genere e restituisce un'immagine di donna sottomessa e tollerante nei confronti di trattamenti inadeguati. Il rischio è che questa scelta progettuale perpetui uno stereotipo discriminatorio e trasmetta il messaggio, soprattutto alle generazioni più giovani, che alle donne vada attribuito un ruolo di subordinazione e incondizionata disponibilità, creando i presupposti per un'ulteriore violenza di genere.

5. Donne e molestie online: deep fake, cyberstalking, deep nude e revenge porn

Esistono diverse forme di violenza virtuale contro le donne, fra cui *cyberstalking*, pornografia non consensuale, molestie basate sul genere, stigmatizzazione a sfondo sessuale, stupro e minacce di morte, pubblicazione online di informazioni personali e private. La violenza virtuale contro le donne può manifestarsi come violenza sessuale, psicologica ed economica, in cui l'attuale o futura occupazione della vittima potrebbe esser compromessa da informazioni pubblicate *online*.

A seguito dell'avvento delle nuove tecnologie, lo *stalking* è oggi un fenomeno ancora più insidioso e invasivo: si parla di *cyberstalking* quando mediante l'uso di dispositivi di comunicazione elettronica si intende molestare un'altra persona. I comportamenti dello *stalker* che potrebbero connotare l'attività di *cyberstalking* sono legati: alla sorveglianza *online* nei confronti della vittima, mediante attivazione delle funzioni di geo- localizzazione; alla ricerca di contatto, attraverso pedinamento elettronico e aggiramento compiuto anche collegandosi ad amicizie sui *social network;* ad un controllo, ad esempio della posta elettronica o dei suoi profili social o conto correnti bancari, spesso all'insaputa della vittima. In Italia il *cyberstalking* ricade nella fattispecie disciplinata dall'art. 612-bis c.p. che regola-

[33] The effects of AI on the working lives of women, https://unesdoc.unesco.org/ark:/48223/pf0000380861

[34] https://www.unesco.org/en/forum-against-racism-discrimination.

menta il delitto di *stalking*, il quale sancisce che è punito con la reclusione da un anno a sei anni e sei mesi «chiunque, con condotte reiterate, minaccia o molesta taluno in modo da cagionare un perdurante e grave stato di ansia o di paura ovvero da ingenerare un fondato timore per l'incolumità propria o di un prossimo congiunto o di persona al medesimo legata da relazione affettiva ovvero da costringere lo stesso ad alterare le proprie abitudini di vita». Con riferimento specifico al *cyberstalking* viene in rilievo l'aggravante prevista dal secondo comma che contempla un aumento della pena allorché «il fatto è commesso attraverso strumenti informatici o telematici», incluso Whatsapp come chiarito dalla Corte di Cassazione[35].

Per *deepfake* – secondo la definizione offerta dal Garante per la protezione dei dati personali[36] – si intendono tutte quelle foto, video e audio creati grazie a software di intelligenza artificiale che, partendo da contenuti reali quali immagini e audio, riescono a modificare o ricreare, in modo estremamente realistico, le caratteristiche e i movimenti di un volto o di un corpo e ad imitare fedelmente una determinata voce. In particolare, il legame tra *deepfake* e furto d'identità è molto stretto. Il reato di furto d'identità è regolato dall'art. 494 c.p. secondo cui «chiunque, al fine di procurare a sé o ad altri un vantaggio o di arrecare ad altri un danno, induce taluno in errore, sostituendo la propria all'altrui persona, o attribuendo a sé o ad altri un falso nome, o un falso stato, ovvero una qualità a cui la legge attribuisce effetti giuridici, è punito, se il fatto non costituisce un altro delitto contro la fede pubblica, con la reclusione fino a un anno[37]».

Il furto d'identità assume una gravità particolare quando si collega al contesto sessuale. La produzione di immagini fittizie con connotazioni sessuali è comunemente denominata *deepnude.* Il *deepnude* è una tecnica che permette di manipolare e spogliare artificialmente le figure femminili - sembra funzionare solo con queste - trasformandole in foto di nudo in relazione alla corporatura del soggetto. La semplicità con cui tali software possono essere installati e utilizzati è stata posta in rilievo nel caso di cronaca che coinvolgeva un gruppo di studenti di una scuola media di Latina. Gli studenti mediante l'impiego di un'applicazione denominata "*BikiniOff*", avevano posto in essere la manipolazione di fotografe di cinque studentesse e di una docente. Il *deepnude* della docente, nel caso specifico, era risultato così convincente da comparire su rinomati siti pornografici[38]. Nonostante, infatti, le immagini siano elaborate artificialmente, è innegabile che queste - considerato quanto esse si presentano realistiche - possano intaccare la dignità di una persona che si ritrovi a sua insaputa letteralmente spogliata sul web. Al momento nel nostro ordinamento non esiste una specifica tutela. In verità, sul punto era stata anche presentata nel 2021 una proposta di Legge, mai portata a compimento, volta a contrastare tale fenomeno con l'obiettivo di introdurre un ulteriore comma all'art. 612 del codice penale, finalizzato a prevedere una multa e la reclusione da due a sette anni per chi «invia, cede, pub-

[35] Cassazione penale, sez. V, sentenza 28/01/2019 n° 3989, https://www.altalex.com/documents/news/2019/02/07/stalking

[36] Garante per la Protezione dei Dati Personali "*Deepfake: dal Garante una scheda informativa sui rischi dell'uso malevolo di questa nuova tecnologia*", https://garanteprivacy.it/home/docweb/-/docweb-display/docweb/9512278.

[37] Il furto d'identità non è il solo reato perpetrabile creando, usando o condividendo un *deepfake*. Ad esempio, se il contenuto d*eepfake* va a ledere anche la reputazione dell'individuo, al reato di furto d'identità si aggiunge quello di diffamazione (art. 595 c.p.).

[38] D. Barbera, T*utti i rischi di usare Bikinioff, il chatbot che spoglia le donne*, 19 aprile 2023.

blica o diffonde immagini manipolate di nudo appartenenti a persone fisiche riconoscibili, attraverso l'utilizzo di strumenti tecnologici e di applicazioni, allo scopo di trarre in inganno l'osservatore.»

Sempre legato al profilo sessuale, si registra un fenomeno denominato *Porno Deepfake*: una tecnica, rientrante nell'idea di *Revenge porn*[39], che attraverso l'uso dell'intelligenza artificiale rielabora immagini o video ritraenti persone reali al fine di trasformarli in materiali multimediali a carattere pornografico, falso ma altamente realistico[40]. Tali prodotti manipolati sono poi diffusi *online* attraverso i siti porno, i social network e le app di messaggistica istantanea. In tale ipotesi, però, l'ordinamento italiano prevede una tutela, in quanto il *Revenge porn* è considerato reato ai sensi dell'art. 612 ter del codice penale consistente nella diffusione in rete di immagini sessualmente esplicite, senza il consenso della persona raffigurata. La vittima è solitamente una donna, mentre il reato viene realizzato spesso dagli ex partner mediante la diffusione di video o immagini. La Cassazione ha recentemente avuto modo di precisare che per configurare il reato in questione, la divulgazione può riguardare non solo immagini o video che ritraggono atti sessuali ovvero organi genitali, ma anche altre parti erogene del corpo umano in condizioni e contesti tali da evocarne la sessualità[41].

Nel 2023, ad esempio, *The Guardian*[42] ha pubblicato un'indagine approfondita, che descrive come molti algoritmi presentino dei *bias* di genere che comporterebbero la diffusione di innumerevoli foto con corpi femminili. Le foto di donne vengono classificate come più spinte o sessualmente suggestive rispetto a foto analoghe di uomini. Anche i pancioni delle donne incinte diventano problematici per questi strumenti di Intelligenza Artificiale, in quanto ad esempio l'algoritmo di Google ha valutato la foto come molto probabile che contenga contenuti scabrosi e quello di Microsoft era convinto al 90% che l'immagine fosse di natura sessualmente suggestiva. Proprio allo scopo di limitare la diffusione non consensuale di tali contenuti pornografici e a tutela della dignità delle donne, il Garante per la protezione dei dati personali già nel 2022 con cinque Provvedimenti[43], aveva comminato in via d'urgenza a Facebook, Instagram e Google di adottare immediatamente tutte le misure necessarie ad impedire la diffusione sulle relative piattaforme del materiale (video, foto) segnalato all'Ufficio del Garante da alcune persone che ne temevano la messa *online*.

[39] Si tratta di un fenomeno della pornografia non consensuale, consiste nella diffusione di immagini pornografiche o sessualmente esplicite a scopo vendicativo (ad esempio per "punire" l'ex partner che ha deciso di porre fine ad una relazione) o per denigrare pubblicamente, bullizzare e molestare la persona cui si riferiscono.

[40] G. NATALE, *Intelligenza artificiale, neuroscienze, algoritmi. aggiornato al nuovo Regolamento Europeo AI Act*, Pisa, 2024, 243.

[41] Cass. pen., Sez. V, sent. n. 14927 del 22 febbraio 2023, https://www.sistemapenale.it/pdf_contenuti/1696488184_sentenza-612-ter-oscurata.pdf

[42] https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies

[43] https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9775414; https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9775327; https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9775401; https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9775948; https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9775932.

6. Prevenzione delle discriminazioni algoritmiche

L'intelligenza artificiale si sta rivelando uno strumento diretto a rendere più acuta le vulnerabilità, ai fini della trattazione delle donne, e di conseguenza del divario di genere, producendo questi effetti su scala globale. Eppure, il ricorso a procedimenti automatizzati o semi automatizzati potrebbe di per sé migliorare sia la prevenzione che la repressione delle discriminazioni in ambito lavorativo e personale. Tuttavia, ciò è possibile solo in presenza di regole utili a far emergere e arginare il rischio discriminazione. È auspicabile, infatti, grazie alle prescrizioni contenute all'interno dell'AI Act, che l'utilizzo di questi sistemi venga subordinato al rispetto delle garanzie e dei limiti stabiliti dalle vigenti disposizioni in materia di protezione delle persone fisiche, anche riguardo al trattamento dei dati personali e, più in generale, che venga operato un congruo bilanciamento tra gli interessi coinvolti.

Nel dicembre 2020, il Consiglio d'Europa e il *Committee on Artificial Intelligence* (CAHAI) hanno adottato il documento *Feasibility study on a legal framework on AI design, development and application based on CoE standards*[44], al fine di dare seguito alle criticità emerse circa le specificità della *AI-derived discrimination*. Invero, l'art. 4 lett. m) del Regolamento allegato al Quadro relativo agli aspetti etici dell'intelligenza artificiale, della robotica e delle tecnologie correlate di cui alla Risoluzione del Parlamento europeo del 20 ottobre 2020 definiva già la discriminazione, come «qualsiasi trattamento differenziato di una persona o di un gruppo di persone per un motivo privo di giustificazione obiettiva o ragionevole e, pertanto, vietato dal diritto dell'Unione». Impostazione ripresa anche dall'AI Act, il quale affronta il problema del possibile utilizzo degli algoritmi con finalità discriminatorie prevedendo una differenziazione degli obblighi basata sul criterio del rischio, come sopra illustrato. L'obiettivo è minimizzare il rischio di discriminazione algoritmica, nel rispetto di quanto previsto all'art. 21 della Carta dei diritti fondamentali dell'Unione Europea. Il Regolamento fa proprio, quindi, il principio di equità e non discriminazione, sostenendo che le organizzazioni sono tenute a garantire che i loro sistemi non pregiudichino o perpetuino discriminazioni basate su caratteristiche personali quali sesso, genere, razza e/o origine etnica. Invero, al fine di evitare simili distorsioni, l'AI Act riprende, in parte, quel *risk approach* tipico del Regolamento Europeo n. 679/2016 in materia di protezione dei dati personali (d'ora in avanti GDPR). L'attenzione al rischio di discriminazione algoritmica emerge già dal Considerando 71 GDPR, ove si legge che il titolare del trattamento dei dati deve garantirne la sicurezza e impedire effetti discriminatori nei confronti di persone fisiche sulla base della razza o dell'origine etnica, delle opinioni politiche, della religione o delle convinzioni personali, dell'appartenenza sindacale, dello status genetico, dello stato di salute o dell'orientamento sessuale. Sempre al fine di proteggere le persone dalla discriminazione, l'art. 22 GDPR vieta determinate decisioni completamente automatizzate con effetti significativi. Si legge, infatti, al primo comma, che «l'interessato ha il diritto di non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato, compresa la profilazione, che produca effetti giuridici che lo riguardano o che incida in modo analogo significativamente sulla sua persona». Nei casi nei quali il trattamento automatizzato è, invece, consentito in base al secondo comma dell'art. 22[45], il titolare del trattamento è tenuto ad attuare misure appro-

[44] Il testo integrale dello studio può essere letto al link: https://rm.coe.int/cahai-2020-23-final-engfeasibility-study-/1680a0c6da.

[45] Art. 22 comma 2 – GDPR - «Il paragrafo 1 non si applica nel caso in cui la decisione: a) sia necessaria per la conclusione o l'esecuzione di un contratto tra l'interessato e un titolare del trattamento; b) sia autorizzata dal

priate per tutelare i diritti, le libertà e i legittimi interessi dell'interessato, garantendo almeno il diritto di ottenere l'intervento umano da parte del titolare del trattamento, il diritto di esprimere la propria opinione insieme al diritto di contestarne la decisione. Non a caso, il GDPR reca in sé i principi fondamentali di legalità algoritmica, quali principio di non esclusività della decisione algoritmica, di conoscibilità e di non discriminazione algoritmica.

Il principio di non esclusività della decisione algoritmica stabilisce che nelle ipotesi in cui una decisione algoritmica produca effetti giuridici o incida significativamente sulla persona, l'interessato ha il diritto che questa non sia basata unicamente su un trattamento automatizzato, ivi compresa la profilazione, ma deve comunque sempre essere garantito un intervento umano. In tal modo, il soggetto titolare della decisione, pur potendo avvantaggiarsi dello strumento informatico idoneo a fornirgli la soluzione apparentemente migliore, mantiene il controllo della decisione.

Il principio di conoscibilità prevede che ognuno ha il diritto di conoscere l'esistenza di processi decisionali automatizzati, che lo riguardino, ai sensi dell'art. 15, comma 1, lett. h). Il principio di conoscibilità è strettamente correlato al principio di comprensibilità, secondo cui l'interessato ha anche il diritto di ottenere «informazioni significative sulla logica utilizzata». Nell'attribuire un rilievo centrale al principio di conoscibilità dell'algoritmo, la giurisprudenza assume come riferimento normativo gli artt. 13, comma 2, lett. f), e 14, comma 2, lett. g), del GDPR, i quali, impongono al titolare del trattamento l'obbligo di fornire indicazioni circa «l'esistenza di un processo decisionale automatizzato», nonché di procurare «informazioni significative sulla logica utilizzata[46] ».

Infine, il principio di non discriminazione, in virtù del quale «la legittimità dell'azione non è garantita dalla sola presenza di un algoritmo conoscibile e comprensibile, oggetto di controllo e validazione da parte di un funzionario, ma occorre che lo stesso non assuma carattere intrinsecamente discriminatorio[47]». L'attuazione del principio di non discriminazione algoritmica richiede dunque, come evidenziato dalla dottrina[48] e giurisprudenza[49], da un lato, la verifica della correttezza, dell'affidabilità e della qualità dei dati di input, al fine di evitare che gli eventuali profili di errore influenzino il risultato decisionale e producano un effetto discriminatorio; dall'altro, coinvolge la responsabilità organizzativa e preventiva nella fase iniziale di configurazione dei procedimenti automatizzati e delle regole algoritmiche da utilizzare.

Sulla scia del GDPR e allo scopo di arginare le preoccupanti dimensioni della diffusione di *deep fake*, come sopra accennato, il Titolo IV dell'AI Act[50] introduce precisi obblighi di trasparenza. Questi dovranno applicarsi ai sistemi che interagiscono con gli esseri umani, che rilevano emozioni, che stabili-

diritto dell'Unione o dello Stato membro cui è soggetto il titolare del trattamento, che precisa altresì misure adeguate a tutela dei diritti, delle libertà e dei legittimi interessi dell'interessato; c) si basi sul consenso esplicito dell'interessato».

[46] E. Carloni, *Algoritmi su carta. Politiche di digitalizzazione e trasformazione digitale delle amministrazioni*, in Dir. pubbl., 2, 2019, 363 ss.

[47] E. Carloni, *AI, algoritmi e pubblica amministrazione in Italia, in IDP. Revista de Internet, Derecho y Política*, 1, 2020.

[48] E. Carloni, *I principi della legalità algoritmica. Le decisioni automatizzate di fronte al giudice amministrativo*, in *Dir. amm.,* 2, 2020, 298-299.

[49] Cons. St., sez. VI, 13 dicembre 2019, n. 8472.

[50] M. Colonna, *Sezione I – I sistemi ad alto rischio,* in AIRIA Associazione per la Regolazione dell'Intelligenza Artificiale (a cura di), *Navigare l'European AI Act*, Milano, 2024, 71-82.

scono associazioni con categorie sociali sulla base di dati biometrici, o che, appunto, generano o manipolano contenuti. Il considerando 134[51] precisa che i fornitori dovranno adottare soluzioni tecniche per i *deep fake* e *deloyer*, in modo tale da render chiaro che il contenuto è stato manipolato artificialmente. Nel contesto dell'IA Act, la trasparenza dei provider presuppone che vi sia a monte una fiducia da parte degli utenti finali che essi effettivamente mettano a disposizione le informazioni utili per conoscere e comprendere i sistemi immessi sul mercato. Laddove è impossibile avere contezza dei dati di addestramento, dei modelli, delle infrastrutture informatiche dei sistemi di IA, viene imposta al provider l'obbligo di una comunicazione chiara e comprensibile, secondo standard di ragionevolezza che, però, lascia ampi spazi di valutazione a chi deve raccontare e quindi selezionare le informazioni. L'art. 50 che apre il Capo IV dedicato agli obblighi di trasparenza prevede che i fornitori devono garantire che le soluzioni tecniche adottate per la marcatura siano «efficaci, interoperabili, solide e affidabili nella misura in cui ciò sia tecnicamente possibile, tenendo conto delle specificità e dei limiti dei vari tipi di contenuti, dei costi di attuazione e dello stato dell'arte generalmente riconosciuto, come eventualmente indicato nelle pertinenti norme tecniche». Analoga disposizione – art.50 paragrafo 4 - è prevista anche per i sistemi che generano *deep fake* con video, immagini, musica; o testi linguistici su questioni di pubblico interesse. In particolare, il Considerando 120[52] sancisce l'obbligo di dichiarare l'origine artificiale di un contenuto allo scopo di individuare i rischi sistemici che possono derivare dalla diffusione di contenuti manipolati e causare forme di discriminazione verso soggetti maggiormente vulnerabili. La sfida più importante, quindi, si svolge sul terreno della trasparenza "co-

[51] Cons. 134 – AI Act: «Oltre alle soluzioni tecniche utilizzate dai fornitori del sistema di IA, i *deployer* che utilizzano un sistema di IA per generare o manipolare immagini o contenuti audio o video che assomigliano notevolmente a persone, oggetti, luoghi, entità o eventi esistenti e che potrebbero apparire falsamente autentici o veritieri a una persona (*deep fake*), dovrebbero anche rendere noto in modo chiaro e distinto che il contenuto è stato creato o manipolato artificialmente etichettando di conseguenza gli output dell'IA e rivelandone l'origine artificiale. L'adempimento di tale obbligo di trasparenza non dovrebbe essere interpretato nel senso che l'uso del sistema di IA o dei suoi output ostacola il diritto alla libertà di espressione e il diritto alla libertà delle arti e delle scienze garantito dalla Carta, in particolare quando il contenuto fa parte di un'opera o di un programma manifestamente creativo, satirico, artistico, fittizio, o analogo fatte salve le tutele adeguate per i diritti e le libertà dei terzi. In tali casi, l'obbligo di trasparenza per i *deep fake* di cui al presente regolamento si limita alla rivelazione dell'esistenza di tali contenuti generati o manipolati in modo adeguato che non ostacoli l'esposizione o il godimento dell'opera, compresi il suo normale sfruttamento e utilizzo, mantenendo nel contempo l'utilità e la qualità dell'opera. È inoltre opportuno prevedere un obbligo di divulgazione analogo in relazione al testo generato o manipolato dall'IA nella misura in cui è pubblicato allo scopo di informare il pubblico su questioni di interesse pubblico, a meno che il contenuto generato dall'IA sia stato sottoposto a un processo di revisione umana o di controllo editoriale e una persona fisica o giuridica abbia la responsabilità editoriale della pubblicazione del contenuto».

[52] Cons. 120 – AI Act: «Inoltre, gli obblighi imposti dal presente regolamento ai fornitori e ai deployer di taluni sistemi di IA, volti a consentire il rilevamento e la divulgazione del fatto che gli output di tali sistemi siano generati o manipolati artificialmente, sono molto importanti per contribuire all'efficace attuazione del regolamento (UE) 2022/2065. Ciò si applica specialmente agli obblighi per i fornitori di piattaforme online di dimensioni molto grandi o motori di ricerca online di dimensioni molto grandi di individuare e attenuare i rischi sistemici che possono derivare dalla diffusione di contenuti generati o manipolati artificialmente, in particolare il rischio di impatti negativi effettivi o prevedibili sui processi democratici, sul dibattito civico e sui processi elettorali, anche mediante la disinformazione».

municata" e sull'auspicio che questa percorra senza troppi ostacoli tutta la catena dai fornitori di sistemi di AI, a chi li utilizza a chi ne subisce l'impatto per effetto di singole decisioni[53].

7. Il dovere delle Istituzioni

Alla luce delle considerazioni svolte, è possibile affermare che l'impiego delle nuove tecnologie sta determinando, e continuerà a determinare, un significativo cambiamento nella nostra vita. Nelle pagine che precedono, si è cercato di illustrare come l'impiego delle nuove forme tecnologiche, in particolare dei sistemi di AI, ponga gli interpreti del diritto dinanzi a nuove sfide e a nuovi interrogativi, ancora in attesa di un'unanime e puntuale risposta. Il punto più problematico pare rintracciarsi nell'esigenza di evitare che la complessità e l'incertezza, insieme con le strumentazioni e la progettazione di natura strettamente tecnica, si qualifichino come fattori di giustificazione delle discriminazioni.

Occorre, invece, intervenire, innanzitutto, a livello di c.d. morale soggettiva, ossia delle conoscenze, della cultura e dei codici di comportamento degli individui. Si rivela determinante l'azione sul piano educativo, mediante quella che gli studiosi definiscono tecnica di tutela by *education*[54]*,* al fine di avere un'IA in grado di rispettare i valori umani fondamentali che promuovono l'inclusione, l'equità, l'uguaglianza di genere e le diversità linguistiche e culturali, nonché di rispettare opinioni ed espressioni plurali. L'UNESCO ha già invitato, infatti, la comunità internazionale a riflettere sulle implicazioni di questa tecnologia sul lungo periodo in termini di conoscenza, insegnamento, apprendimento e valutazione, e ha offerto raccomandazioni concrete ai decisori politici e alle istituzioni educative su come l'uso degli strumenti di IA possa essere progettato per proteggere l'azione umana.

È compito del legislatore e delle autorità di regolamentazione, quindi, non sottovalutare l'impatto che la discriminazione ha sulle persone vulnerabili, perché un tale errore di calcolo potrebbe danneggiare lo spazio di libertà e di diritto che è alla base dell'Unione europea. Le istituzioni, pertanto, a parere di chi scrive, dovranno sempre più assumere un ruolo guida nello sviluppo di standard internazionali etici e di linee guida in grado di proteggere i diritti e le libertà degli interessati, soprattutto dei più vulnerabili. Questo richiederà sicuramente lo sviluppo di meccanismi di sorveglianza e audit che consentano di identificare tempestivamente eventuali pregiudizi insiti nei sistemi di AI. La cooperazione tra le istituzioni dei diversi Paesi risulta essere fondamentale per sviluppare standard comuni e garantire un utilizzo e una tutela uniforme. Inoltre, è necessario che vengano reperiti e fornite risorse umane capaci di regolamentare e sorvegliare sull'AI, quindi, in possesso non solo di competenze tecnico-informatiche, ma altresì di competenza legali ed etiche per identificare il rischio di discriminazione.

In particolare, compito primario delle istituzioni- alla luce degli obblighi imposti dal Regolamento – sarà quello di promuovere la trasparenza, quale architrave della progettazione e conseguente uso dei sistemi di AI. Solo così, attraverso la cultura della trasparenza, si potranno scongiurare i rischi di di-

[53] G. Lo Sapio, *L'Artificial Intelligence Act e la prova di resistenza per la legalità algoritmica*, in *Federalismi.it*, 16, 2024.

[54] A. Simoncini, S. Suweis, *Il cambio di paradigma nell'intelligenza artificiale e il suo impatto sul diritto costituzionale*, in *Rivista di filosofia del diritto*, 1, 2019, 87 ss.

scriminazione algoritmica. A tal proposito, però, sarà necessario che il pubblico di utenti sia educato ai rischi della discriminazione che potrebbero essere insiti nel sistema stesso.

In conclusione, quindi, si ritiene che minare la discriminazione prodotta dall'impiego di questi strumenti di AI rappresenti la vera sfida attuale. L'AI Act, quale complesso normativo di regolamentazione dell'intelligenza artificiale rappresenta solo un primo passo verso la prevenzione dei rischi. Il suo successo dipenderà, infatti, dalla capacità dei governi, delle istituzioni e della società di individuare e sfruttare gli enormi benefici dell'AI, senza sacrificare i diritti fondamentali della persona.

ISSN 2284-4503