

**ALGORITMI
+ INCLUSIVI**

Intelligenza artificiale e inclusione delle persone LGBTQIA+

Report

Aprile 2026



LGBTQIA+ LEADERS FOR CHANGE

Sommario

Introduzione	3
Premessa (1)	3
I Bias Algoritmici	4
Gli usi dell'Intelligenza Artificiale e la discriminazione	7
Intelligenza Artificiale e regolazione	9
Per un approccio etico orientato alla soluzione di problemi concreti	10
1. L'intelligenza artificiale in prospettiva tecnica	15
1.1 I processi dell'intelligenza artificiale e del machine learning e le aree di rischio	16
1.1.1 Intelligenza artificiale e machine learning: definizioni e concetti di base	16
1.1.2 Le tipologie di machine learning e i relativi bias potenziali	19
1.1.3 Algoritmi di machine learning: tipologie e meccanismi di utilizzo, applicazioni pratiche	21
1.1.4 Gli step nel ciclo di vita degli algoritmi di machine learning	26
1.1.5 Le tipologie di bias nell'IA	27
2. Dati, algoritmi, minoranze e discriminazioni	31
2.1 Fairness e algoritmi	35
2.2 Spiegabilità dei modelli di machine learning: classificazione dei metodi di spiegazione	38
2.3 Strategie per affrontare i bias	40
2.4 Soluzioni per applicazioni specifiche	41
2.4.1 L'IA per l'elaborazione del linguaggio naturale e le deviazioni discriminatorie	42
2.4.2 L'IA per lo studio delle immagini e le possibili deviazioni discriminatorie	44
3. Industry	46
3.1. La consapevolezza, tra compliance e competitività	47
3.2 Principi e governance	49
3.3 Aziende sviluppatrici e aziende utenti di sistemi di IA	53
3.4 Gestione dei dati	54
3.5. I sistemi di recruiting e le risorse umane (HR)	56
3.6 Contenuti social e di comunicazione	58
4. Humanities	60
4.1. La prospettiva delle scienze umane e sociali	60

4.2. I (veri) pericoli dell'intelligenza artificiale	61
4.3. La problematizzazione culturale dei bias e delle discriminazioni	66
4.4. La datificazione dell'identità sessuale	68
4.5. Il trade-off tra efficienza e spiegabilità	70
5. Policy e regolamentazione	71
5.1. Che cos'è la discriminazione	72
5.2. Discriminazione e IA	73
5.3. Come combattere la discriminazione: alcune caratteristiche auspicabili dell'IA	75
5.3.1 Spiegabilità	75
5.3.2 Giustificabilità	77
5.3.3 Contestabilità	78
5.4. Verso una IA non discriminatoria: l'Artificial Intelligence Act europeo	79
5.5. Il GDPR e il DSA	84
5.6. La partecipazione degli stakeholder come strumento di policy	86
Bibliografia	87

Introduzione

Premessa (1)

Nel 2021 [EDGE LGBTI+ Leaders for Change](#) ha iniziato ad interessarsi dell'impatto che l'intelligenza artificiale avrebbe potuto avere sulla discriminazione e sulla inclusione delle persone LGBTQIA+ così come sulla discriminazione e sull'inclusione delle persone appartenenti ad altri gruppi vulnerabili.

Le notizie sui casi di discriminazione legati all'utilizzo di strumenti di intelligenza artificiale (IA) iniziavano a circolare con riguardo, per esempio, alla discriminazione delle persone non bianche nel riconoscimento facciale [Buolamwini, J., & Gebru, T. (2018) (2)] o delle donne nelle graduatorie automatizzate e nei sistemi di raccomandazione [Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016), Aylin Caliskan, Joanna Bryson, Arvind Narayanan (2019) (3)] o, ancora, degli appartenenti alle minoranze nella valutazione del rischio di recidiva in ambito penale (4).

Certamente la relazione fra IA e fattori quali orientamento sessuale, identità di genere, espressione di genere e altre caratteristiche sessuali delle persone (c.d. fattori SOGIESC) ha molto in comune con altri fattori di discriminazione, come il genere, l'origine etnica, l'età, etc... eppure ci è parso necessario uno sguardo specifico, che porti alla luce non solo le peculiarità sotto il profilo tecnico ma soprattutto l'impatto sulle persone e la necessità che le diversità di cui, in quanto persone, siamo portatrici, siano tenute in considerazione nello sviluppo e nell'implementazione dei sistemi di IA, con una attenzione peculiare ai citati fattori SOGIESC.

Abbiamo, quindi, deciso di lanciare un progetto di studio e ricerca finalizzato alla divulgazione e sensibilizzazione per comprendere se e come il mondo della ricerca, quello dell'impresa e le istituzioni stiano affrontando il tema, con attenzione agli aspetti legati all'orientamento sessuale e all'identità di genere, così da verificare se sia possibile stimolare soluzioni tecniche concrete.

Il Progetto complessivo si chiama "[A+I Algoritmi + inclusivi](#)", si articola in diversi filoni di attività e fa leva su un gruppo di lavoro multidisciplinare. Fra le diverse attività, abbiamo condotto interviste e dialoghi con esperti, ricercatori, professionisti e protagonisti dello sviluppo dell'AI oltre ad una serie di dialoghi svolti in conferenze e incontri di lavoro.

Ne è scaturito questo **Report** che prende le mosse da quelle interviste, i cui contenuti abbiamo deciso di arricchire con una illustrazione dei concetti di base, cercando di realizzare

una presentazione più organica dei temi di maggiore impatto sul rapporto fra IA e discriminazione.

Quel che ci è parso chiaro dalle interviste e dalle nostre letture è la marginalità del tema della discriminazione e dell'inclusione delle persone LGBTQIA+ tanto nella ricerca quanto nella riflessione sull'IA.

Si tratta di un dato confermato dalle fasi successive della ricerca. Basti, per esempio, osservare che nel corso del progetto una ricerca sul database [Scopus](#) con le parole chiave *"artificial intelligence"*, *"gender"*, *"sexuality"*, *"sexual"* e *"fairness"* ha prodotto 230 risultati. Solo nove contengono nell'abstract le espressioni *"sexual preference"*, *"sexual orientation"*, *"sexual minorities"*, *"gender identity"*, *"queer"*.

Di questi nove solo tre includono *"gender identity"*, uno *"queer"* e nessuno "LGBT".

Si potrebbe immaginare che - quanto meno in una prospettiva europea - la limitatezza degli studi sul rapporto fra IA e discriminazione o inclusione delle persone LGBTQIA+ dipenda dalla assai limitata possibilità di raccogliere ed utilizzare dati relativi a orientamento sessuale, identità di genere, espressione di genere e altre caratteristiche sessuali delle persone.

È tuttavia una spiegazione che non convince, non solo perché i sistemi di intelligenza artificiale sono capaci di inferire caratteristiche protette delle persone da altri elementi o correlando dati, ma anche perché sono molti gli aspetti che hanno impatto diretto e peculiare legato ai fattori SOGIESC (Orientamento Sessuale, Identità di Genere, Espressione di Genere e Caratteristiche Sessuali).

Eclatanti sono i progetti che mirano esplicitamente ad individuare l'orientamento sessuale delle persone, ad es. tramite il riconoscimento facciale [Wang e Kosinski (2018)].

E non importa quanto accurato sia il risultato giacché l'impatto sulle persone deriva dall'uso dei sistemi algoritmici e ciò sia in termini problematici e di rischio sia in termini di innovazione e benefici.

I Bias Algoritmici

Certamente la riflessione non può che partire dal tema dei c.d. ***bias algoritmici***, espressione dalle sfumature diverse ma che viene usata in senso esteso per indicare non solo i *bias algoritmici* in senso stretto ma in generale la discriminazione algoritmica, l'insieme di comportamenti ed effetti dei sistemi di intelligenza artificiale e del loro uso che replicano e amplificano stereotipi e generano output discriminatori.

Una [campagna di Wired del gennaio 2023](#) mostrava le risposte di un sistema di generazione di immagini a *prompt* che chiedevano di immaginare “*due amanti*”, “*manager*”, “*atleta*”, “*genitori*” e così via. Il risultato era ampiamente stereotipato. Per quel che riguarda il nostro tema, le coppie e le famiglie erano sempre eterosessuali. Il video di Wired consente una percezione immediata del problema. In quel caso un problema di rappresentazione e di invisibilizzazione delle persone che lo stereotipo esclude. A distanza di due anni dalla campagna di Wired, i sistemi di generazione di immagini restituiscono risultati più sfaccettati ma ancora non risulta possibile averne un pieno dominio tramite il *prompting*. Per altri esempi basta provare ad interrogare modelli linguistici di generazione di contenuti privi di filtro, in grado di riprodurre stereotipi, fake news, espressioni offensive e via dicendo.

Debora Nozza, Federico Bianchi e Dirk Hovy, del Dipartimento di Computer Science dell'Università Bocconi, hanno studiato il modello linguistico **BERT**, sviluppato da Google tra il 2018 e il 2019 [Nozza, Bianchi, Hovy (2021)].

BERT è (era) un modello usato da Google per comprendere il significato di una parola in base al contesto. I ricercatori hanno chiesto a BERT di completare alcune frasi scritte in lingue diverse, misurando la probabilità di completamento con linguaggio offensivo in generale e in particolare per donne e comunità LGBTQIA+.

BERT ha prodotto risultati offensivi per tutte le identità, ma è emerso un particolare *bias* di genere, poiché il 4% delle frasi a soggetto maschile a differenza del 9% di quelle a soggetto femminile vengono completate usando riferimenti alla sfera sessuale. Se invece la frase è associata a un soggetto LGBTQIA+, i completamenti che si riferiscono alla sfera sessuale o che risultano offensivi raggiungono una media del 13% con un massimo dell'87%. Numerosi altri studi hanno analizzato e confermato l'associazione fra termini con riferimento ai fattori di discriminazione e termini offensivi o, comunque, con approcci discriminatori.

I modelli di IA, infatti, incorporano gli stereotipi dai dati a partire dai quali vengono creati e allenati. Incorporano spesso - senza che nessuna riflessione preliminare al riguardo venga svolta su questo aspetto - anche gli stereotipi delle persone che partecipano alla modellazione, all'allenamento e alla verifica.

Il problema tuttavia non è limitato alla semplice riproduzione dei *bias* nei risultati dei sistemi algoritmici giacché sono gli stessi modelli ad amplificare i **bias** presenti nei dati. In questa prospettiva è rilevante il c.d. effetto «*black box*» cioè la limitata capacità di comprendere il funzionamento dei modelli di IA e in particolare il rapporto fra l'input e l'output del modello.

Le conversazioni avute nel corso della preparazione del Report ci indicano che il problema dei *bias* sia sentito dalla comunità degli sviluppatori di IA come un problema anzitutto tecnico di accuratezza dei risultati.

Resta, tuttavia, delicato l'intervento sul problema, che implica spesso scelte che risentono inevitabilmente delle diverse visioni del mondo, delle conoscenze, sensibilità e competenze degli attori del processo.

Numerosi sono i tentativi di individuare soluzioni tecniche in grado di ridurre l'impatto dei *bias* contenuti nei dati senza diminuire in maniera apprezzabile l'accuratezza dei modelli. **Un dato rilevante è, tuttavia l'attuale limitatezza delle competenze etiche all'interno dei team di soggetti che sviluppano AI [Griffin, Green, Welie (2024) (5)].**

Qui vale la pena ricordare che l'emersione amplificata dei *bias* veicolati mediante l'uso del linguaggio consente anche un'analisi del linguaggio stesso e una evidenziazione dei *bias* nell'uso comune degli strumenti di comunicazione verbale e non verbale. In altre parole, lo studio degli effetti di amplificazione dei *bias* consente, a ritroso, di andare ad analizzare le caratteristiche culturali del linguaggio con cui i modelli sono stati addestrati.

Nel riflettere sull'impatto dell'IA sulla discriminazione e sull'inclusione delle persone non va infatti dimenticato che non esiste una sola intelligenza artificiale ma che esistono tanti modelli e tanti sistemi di IA, con approcci e funzioni diverse, e ancora che esistono molteplici usi. Una tassonomia sviluppata dal **Joint Research Center (JRC) (6)** della Commissione Europea nel 2025 indica i principali sotto-domini di sviluppo dell'IA attraverso un elenco che è utile per guardare il panorama nella sua interezza: rappresentazione della conoscenza, ragionamento automatizzato, pianificazione, ricerca, ottimizzazione, apprendimento automatizzato, elaborazione del linguaggio naturale, visione computerizzata, elaborazione dei suoni, sistemi multi-agenti, robotica e automazione, veicoli a guida autonoma (oltre ad altri meno rilevanti qui), a cui dovrebbe essere aggiunta la più ampia produzione di contenuti. Bisogna poi considerare la combinazione di questi ambiti.

Dunque, le riflessioni che abbiamo illustrato poc'anzi con riferimento all'intelligenza artificiale generativa, che abbiamo imparato a conoscere e usare da fine 2022, devono essere estese ad ogni sistema algoritmico, da quelli che riconoscono le immagini a quelli che fanno moderazione di contenuti e così via. Di grande impatto risultano in particolare i sistemi utilizzati in ambito digitale, specie nella raccomandazione e nella moderazione di contenuti.

Un esempio è utile. Un importante social network alcuni anni fa ha impedito per un certo periodo di tempo la pubblicazione di baci tra uomini. Ciò perché in fase di preparazione e allenamento del sistema di IA che gestiva in maniera automatizzata la moderazione dei

contenuti - fase di produzione gestita in un contesto culturale diverso dal nostro - l'immagine di un bacio fra due uomini era stato etichettato (in senso tecnico, *labelled*) come inappropriata. Il blocco sarebbe stato poi rimosso dall'azienda ma solo dopo molte segnalazioni degli utenti.

Un approccio simile è stato attribuito, questa volta come azione volontaria, a TikTok dagli sviluppatori delle app di dating Surge e Zoe, che hanno [pubblicamente contestato](#) la chiusura permanente del loro account dopo la pubblicazione di una foto che ritraeva un bacio tra due donne che avrebbe violato la policy sul "*intimate kissing*", rilevando come contenuti analoghi che ritraggono baci eterosessuali non vengano rimossi ed evidenziando un doppio standard nella moderazione che andava a discriminare le relazioni fra persone dello stesso sesso.

Le casistiche dei *bias* dell'IA si collocano, quindi, su uno spettro che va dall'incorporazione di *bias* impliciti nelle fonti dei dati (come ad es. i *bias* sistemici) a *bias* dovuti alla struttura dei dati o alle modalità di raccolta degli stessi (come i *bias* di misura o di campionamento) a *bias* direttamente trasfusi dalle persone che lavorano ad un modello sia in relazione ai dati immessi nelle diverse fasi del processo sia in relazione ai parametri più specifici dell'algoritmo e ai suoi obiettivi.

Gli usi dell'Intelligenza Artificiale e la discriminazione

L'enorme potenziale dei sistemi algoritmici fa sì che essi abbiano o possano avere impatti estremamente estesi in termini di ampiezza geografica e sociale e in termini di pervasività.

I sistemi algoritmici possono incorporare, con maggiore o minore consapevolezza di chi li sviluppa o impiega, stereotipi e approcci discriminatori e l'esistenza di questi *bias* può essere considerata più o meno impattante sulla scelta di impiegare il sistema algoritmico in questione. **Ci sono però casi in cui l'AI è usata in maniera consapevolmente finalizzata alla discriminazione.**

Come abbiamo visto, il mondo digitale è uno di quelli in cui le decisioni automatizzate sono già pervasive e possono impattare specificamente sulle persone LGBTQIA+.

Uno [studio recente](#) nel contesto cinese ha analizzato il modo in cui i contenuti pubblicati sui social da uomini gay cinesi vengono sottoposti al cd. ***shadowban***, cioè a un tipo di moderazione che porta alcuni contenuti, considerati inappropriati, ad essere meno visibili sulle piattaforme [Zhao (2024) (7)]. La moderazione può avvenire ad es. (i) tramite la sostituzione di parole omofone (parole che si pronunciano allo stesso modo ma hanno significati e spesso grafie diverse) nel campo di ricerca, quindi di fatto impedendo la ricerca di parole chiave considerate inappropriate, (ii) tramite messaggi che suggeriscono che in base

alla legge non possono essere forniti risultati per la ricerca effettuata, o (iii) tramite la sostituzione di parole con sinonimi (ad esempio la parola "gay" viene sostituita con "compagno"). Inoltre, (iv) i risultati proposti sono spesso caratterizzati da un alto livello di rumore, quindi è necessario scorrere molte pagine prima di trovare ciò che si stava cercando. La moderazione (v) avviene anche a livello di hashtag, dove quelli indesiderati non vengono rimossi ma non sono utilizzabili per cercare contenuti correlati. Anche la diffusione dei contenuti viene manipolata: sulle piattaforme video i contenuti relativi a uomini gay non possono raggiungere più di 100.000 visualizzazioni.

Vi sono poi approcci anche più sofisticati e non meno potenti finalizzati a indurre *bias* ed effetti negativi per le persone. Ciò può avvenire attaccando un sistema algoritmico mediante comandi in maniera da alterarne il comportamento o veicolando nella fase di sviluppo e addestramento di un modello dati non corretti o comunque finalizzati a determinare comportamenti specifici non voluti (c.d. **poisoning**), ad esempio rendendo disponibili on-line grandi quantità di dati destinati essenzialmente ad essere raccolti dai sistemi automatici che alimentano le basi di dati per l'allenamento degli *LLM* (Large Language Model). Non si tratta di casi remoti ma di vulnerabilità note a cui i produttori di sistemi algoritmici cercano costantemente di porre rimedio ma che vengono ampiamente sfruttate da attori malevoli.

Lo stesso effetto di alterazione della base di dati potrà essere determinato dalle azioni di molte agenzie ed enti governativi statunitensi (tra cui molti enti deputati ad attività di ricerca in ambito socio-sanitario) che, a seguito di alcuni ordini esecutivi emanati dall'Amministrazione Trump all'inizio del 2025, stanno rimuovendo dal web documenti nonché informazioni o singole parole nei documenti disponibili, come ad esempio il termine "*gender identity*", il cui uso è stato sostanzialmente vietato dall'[Executive Order 14168](#).

La comunità scientifica, le aziende e da ultimo i policymakers cercano di affrontare i problemi qui appena accennati con metodi e strumenti che vanno sotto l'**acronimo FAcCT (Fairness, Accountability, Transparency)**, un filone di lavoro che conta più di 15 anni. Tuttavia non ci sono soluzioni semplici e del tutto neutre né vevoli per tutti i potenziali fattori di discriminazione in ogni contesto. Inoltre, le soluzioni adottate non risultano definitive dovendo riproporsi nell'evoluzione dei modelli.

Alcune domande risultano particolarmente complesse, come quelle legate al campionamento di fattori cd. "*invisibili*" come l'orientamento sessuale o alla possibilità o meno di trattare dati particolari, come appunto i fattori SOGIESC, per migliorare l'accuratezza degli strumenti di AI o agli strumenti accettabili per individuare e correggere risultati intaccati dai *bias*. Si tratta di temi su cui dovremo interrogarci e prendere posizione, pena essere travolti dai fatti.

Intelligenza Artificiale e regolazione

Infine, non possiamo dimenticare il ruolo del legislatore. Applicare i principi di eguaglianza, parità di trattamento e non discriminazione nel contesto dell'intelligenza artificiale pone sfide nuove nel potenziale contrasto fra quei principi e altri. L'**Unione Europea** è particolarmente attiva in quest'area, avendo definito una normativa sul trattamento dei dati personali, il GDPR, che è un benchmark globale, ed avendo sviluppato altre norme di notevole impatto per la tutela delle persone nella relazione con gli strumenti di IA come, in particolare, **Digital Services Act (DSA)** del 2022 e l'**AI Act** del 2024.

Il **GDPR**, la norma europea del 2016 sulla tutela dei dati sensibili e la privacy, pur con qualche ruga, ha dimostrato di poter essere usato per intervenire in molti casi di trattamenti automatizzati in grado di violare i diritti individuali. La scala dello sviluppo dei sistemi rende tuttavia palese la limitatezza dei mezzi a disposizione delle autorità. Il DSA è di grande rilevanza rispetto alla sfera dei servizi digitali e della moderazione dei contenuti on-line.

L'AI ACT è cruciale nell'applicazione di quelli che consideriamo i principi fondanti del nostro vivere associato all'uso dell'AI in ambiti estremamente sensibili e di grande rilievo per i rischi di discriminazione, in primis l'ambito lavorativo e l'accesso ai servizi essenziali, ma anche il riconoscimento biometrico, l'amministrazione della giustizia, la repressione dei reati e talune forme di social scoring. Nell'applicazione di entrambi vi è però il rischio che i profili di tutela dei gruppi vulnerabili oggetto della normativa antidiscriminatoria vengano trascurati di fronte ad una mole di problemi estremamente ampia.

Sempre sotto il profilo giuridico è importante evidenziare il rischio che comportamenti discriminatori classificabili come discriminazione diretta vengano derubricati a discriminazione indiretta (e, quindi, suscettibile di diverso trattamento giuridico) solo perché replicati da sistemi algoritmici, esito chiaramente deprecabile e che depotenzia la protezione dei diritti delle persone [Adam-Prassl, Binns, Kelly-Lyth (2022) (8)]. Un'attività costante di sensibilizzazione e advocacy, anche da parte delle organizzazioni della società civile e verifica dell'implementazione risulta indispensabile.

Il Progetto "A+I Algoritmi + Inclusivi" di EDGE, di cui questo Report risulta parte integrante, si colloca in questa prospettiva di senso e con questo obiettivo.

Inoltre, vi è un grande spazio di usi possibili dell'IA che non rientrano nei casi ad alto rischio e per i quali occorre elaborare e condividere strumenti di promozione dell'inclusione che abbiano alla base una visione dell'IA centrata sulla persona umana. Visione che non ha solo un senso etico ma anche un senso di business, essendo necessaria affinché gli esseri umani

possano fare affidamento su questi strumenti.

Si possono individuare delle strategie e degli strumenti di policy per ridurre il rischio di discriminazione con specifico riferimento ai nostri temi, quali il design partecipativo e l'aumento di consapevolezza politica su questi temi, anche attraverso la stretta collaborazione con commissioni di esperti e board etici delle aziende.

Ovviamente né questo abstract né il Report possono essere esaustivi e l'evoluzione degli strumenti basati su sistemi di IA pone costantemente nuovi casi, dagli strumenti basati su biometria che hanno un impatto potenzialmente assai elevato sulle persone trans e non binary, alle relazioni "instaurate" dalle persone con i chatbot, che sono in grado di esporre le persone vulnerabili a rischi che vanno dall'*outing* a profili di salute mentale e così via, al *microtargeting* del digital advertising, che può rivelarsi escludente o estremamente invadente, all'uso dell'IA nelle politiche pubbliche, es. nel welfare, che rischia di recuperare dati del passato che incorporano decenni di discriminazione istituzionale.

Per un approccio etico orientato alla soluzione di problemi concreti

L'approccio di EDGE, associazione per l'attivismo civico e i diritti che da oltre 12 anni agisce in ambito italiano ed all'interno di un network europeo di associazioni analoghe con la missione dell'inclusione delle persone LGBTQIA+ sul luogo di lavoro, nelle professioni e nel business, è da sempre propositivo e protesico verso soluzioni pragmatiche. Non possiamo dunque che chiederci cosa si può fare. I limiti di intervento sono notevoli. Interagire con lo sviluppo dell'IA richiede risorse considerevoli in termini di competenze, persone, dati, potenza computazionale e così via. Abbiamo provato a stilare una breve scaletta degli ambiti di intervento e delle possibili azioni.

1) Risulta innanzitutto essenziale che chi usa gli strumenti di IA e chi ne è in qualche modo destinatario sia consapevole - fra le molte nuove conoscenze legate all'IA - anche dei potenziali rischi di discriminazione che l'uso di strumenti e sistemi di intelligenza artificiale porta con sé. Ci pare soprattutto indispensabile che la comunità LGBTQIA+ riesca a diffondere al suo interno tale consapevolezza, ad elaborare strumenti operativi di tutela e di *advocacy* proattiva e a condividerli tanto con i singoli potenziali destinatari di discriminazione sia con i propri alleati.

Ci pare, inoltre, evidente la necessità di una cultura diffusa delle questioni legate ai bias e alla discriminazione algoritmica a tutti i livelli manageriali, da quello tecnico a quello amministrativo, sia nelle aziende che si occupano di produrre tecnologie di IA, sia in quelle che le utilizzano nei loro prodotti, servizi o attività. Ciò sia in una prospettiva di compliance

normativa, sia in una prospettiva di uso eticamente corretto e professionalmente responsabile degli strumenti, con una attenzione specifica all'ambito delle risorse umane e della produzione di contenuti per i social media.

2) Risulta inoltre indispensabile supportare le iniziative del mondo accademico e d'impresa che studiano il fenomeno discriminatorio in ambito AI e stimolarne di nuove.

È altresì necessario aumentare gli studi sulla discriminazione algoritmica delle persone LGBTQIA+ e sui delicati *trade off* che si devono risolvere nell'elaborare gli strumenti per affrontare il problema nello spettro tecnico.

3) Uno spazio di collaborazione importante dovrebbe essere ricercato anche con le organizzazioni della società civile che lavorano sui diritti digitali e con i cd. segnalatori attendibili accreditati ai sensi del Digital Services Act.

4) Altra azione di rilievo è il **presidio dell'implementazione dell'AI Act**, partecipando ai relativi processi. In particolare in Italia sarà necessario trovare uno spazio di interazione con l'AGID (l'Agenzia per l'Italia Digitale) e l'ACN (l'Agenzia per la Cybersicurezza Nazionale), individuate quali autorità italiane per l'IA, storicamente non coinvolte sui temi dell'antidiscriminazione. Altrettanto rilevante risulta la partecipazione ai lavori per l'elaborazione degli standard armonizzati previsti dall'AI Act. Sarà, inoltre, necessario presidiare le proposte di modifica del GDPR e dell'AI Act in corso di discussione.

5) Associazioni, consulenti e persone che si occupano di inclusione nei luoghi di lavoro hanno poi un ruolo cruciale con riferimento specifico all'implementazione di strumenti di IA in relazione alla gestione delle risorse umane. Esistono già i primi **toolkit** creati a questo scopo. **Unire le competenze e i punti di osservazione di chi si occupa di inclusione e chi fa sviluppo e implementazione dei sistemi di AI, in particolare nel mondo del lavoro, risulta cruciale.** In particolare, le aziende più strutturate che si stanno dotando di processi di implementazione e valutazione degli strumenti di AI dovrebbero coinvolgere gli ERG/BRG (Employees/Business Resource Groups) in tali processi e in particolare negli impact assessment, incluso il Fundamental Rights Impact Assessment.

6) In ultimo non vogliamo dimenticare che, a fronte di un'adozione crescente degli strumenti di AI, **è possibile e necessario sperimentare gli usi positivi dell'IA per favorire, al contrario, l'inclusione delle persone LGBTQIA+.**

Guida alla lettura

Il rapporto che segue si suddivide in cinque capitoli:

1. Scienza e tecnica dell'intelligenza artificiale
2. Algoritmi, minoranze e discriminazioni
3. Industria
4. Scienze umane e sociali
5. Politiche

Il primo capitolo si propone di fornire una introduzione ai concetti fondamentali necessari per comprendere l'ambito dell'intelligenza artificiale. Questo capitolo può essere di interesse per il lettore che non ha dimestichezza con i concetti informatici e statistici alla base della disciplina, ma anche a chi, pur provenendo dal mondo delle tecnologie digitali, senta il bisogno di una mappatura delle conoscenze essenziali per mettere a fuoco i problemi della discriminazione algoritmica. In questo vengono presentate alcune tipologie di algoritmi di machine learning, e le loro applicazioni pratiche, e vengono illustrati il ciclo di vita delle applicazioni di machine learning e alcune tipologie di bias che possono presentarsi nell'IA.

Il secondo capitolo offre una esposizione di alcuni rischi di discriminazione delle minoranze, e in particolare della minoranza LGBTI+, nell'ambito dell'IA. Vengono forniti esempi di discriminazione algoritmica e viene introdotto il concetto di fairness. Si passa poi ad offrire alcune proposte per affrontare la discriminazione: viene introdotto il concetto di black box per poi passare a delineare alcune tecniche di spiegabilità dell'IA e di riduzione del bias tramite la selezione e manipolazione dei dataset. Infine il tema della discriminazione viene esplorato in alcune applicazioni specifiche, andando a toccare gli ambiti dell'elaborazione del linguaggio naturale e del riconoscimento automatizzato di immagini.

L'obiettivo del terzo capitolo è di prendere in esame il mondo dell'impresa, affrontando il tema della consapevolezza dei rischi di discriminazione algoritmica e dell'impegno nel ridurla. In questo capitolo si tenta di evidenziare la necessità di una cultura diffusa delle questioni legate al bias e alla discriminazione a tutti i livelli, da quello tecnico a quello amministrativo, sia nelle aziende che si occupano di produrre tecnologie di IA, sia in quelle che le utilizzano nei loro prodotti, servizi o attività. Si presenta poi la questione del trattamento e della conservazione dei dati, e dei rischi legati alla profilazione. In conclusione vengono presi in esame alcuni rischi specifici nell'ambito del recruiting e della produzione di contenuti per i social media.

Il quarto capitolo tenta di inquadrare il problema della discriminazione dalla prospettiva delle scienze umane e sociali. Il capitolo si apre affrendo una distinzione tra intelligenza artificiale debole e forte, e tra le preoccupazioni lungotermiste e un approccio più pragmatico che si concentra sui rischi concreti di discriminazione e deresponsabilizzazione. Indaga il concetto di feedback, inquadrato come una delle caratteristiche più significative nell'ambito dell'interazione con gli algoritmi di apprendimento automatico, per poi passare a una riflessione sull'importanza della consapevolezza delle modalità di funzionamento e dei rischi dell'intelligenza artificiale ai vari livelli della società, dalla cittadinanza alla politica. Successivamente si riflette sul concetto di bias e sulla possibilità di affrontare il problema della discriminazione da un punto di vista puramente quantitativo, in termini di debiasing e di criteri oggettivi di fairness, per suggerire un approccio integrato che interpreti l'apprendimento automatico come socializzazione delle macchine, e il bias in termini di habitus. Vengono infine tracciati alcuni rischi e benefici della datificazione dell'identità sessuale per concludere con una riflessione sul trade off tra efficienza e spiegabilità.

Il quinto capitolo si propone di offrire un breve resoconto della regolamentazione dell'intelligenza artificiale. In primo luogo offre una definizione di discriminazione basata sulle norme UE, per poi declinarla nell'ambito dell'IA. Si suggeriscono poi spiegabilità, giustificabilità e contestabilità come caratteristiche desiderabili per ottenere sistemi di IA che minimizzino il rischio di discriminazione. Questi principi vengono inquadrati nell'ambito dell'AI Act europeo, la prima legislazione organica sull'intelligenza artificiale al mondo, che prevede di normare in maniera diversa i sistemi di IA a seconda dei diversi livelli di rischio, che vengono qui delineati.

Note:

- 1) Questa introduzione e l'abstract del report sono ampiamente sovrapponibili. Chi avesse letto l'abstract può andare direttamente alla Guida alla lettura.
- 2) **Le pubblicazioni citate nell'introduzione e non riportate di seguito sono illustrate nella bibliografia al termine del report.**
- 3) Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). ***Man is to computer programmer as woman is to homemaker? debiasing word embeddings.*** *Advances in neural information processing systems*, 29. Aylin Caliskan, Joanna Bryson, Arvind Narayanan (2019), ***Semantics Derived Automatically from Language Corpora Contain Human like Biases***, *Science* 356, no. 6344.

- 4) Wisconsin S.C., ***State v. Loomis, Case no. 2015AP157-CR***, 13 July 2016; ProPublica, Machine Bias, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- 5) Tricia A. Griffin, Brian P. Green & Jos V.M. Welie, ***The ethical wisdom of AI developers***. *AI Ethics* 5, 1087–1097 (2025). <https://doi.org/10.1007/s43681-024-00458-x>.
- 6) ABENDROTH DIAS, K., ARIAS CABARCOS, P., BACCO, F.M., BASSANI, E., BERTOLETTI, A. et al., ***Generative AI Outlook Report - Exploring the Intersection of Technology, Society and Policy***, NAVAJAS CAWOOD, E., VESPE, M., KOTSEV, A. and VAN BAVEL, R. (editors), Publications Office of the European Union, Luxembourg, 2025, <https://data.europa.eu/doi/10.2760/1109679>, JRC142598
- 7) Longxuan Zhao, Algorithmic camouflage: ***Exploring the shadowbans imposed by algorithms to moderate the content of Chinese gay men***, 2024, <https://doi.org/10.1177/20539517241296037>
- 8) Jeremias Adams-Prassl, Reuben Binns, Aislinn Kelly-Lyth, ***Directly Discriminatory Algorithms***, 2022, <https://doi.org/10.1111/1468-2230.12759>



1. L'intelligenza artificiale in prospettiva tecnica

Per aprire una riflessione sui rischi di discriminazioni e bias legati all'intelligenza artificiale è necessario prendere le mosse dalle modalità di funzionamento di tali tecnologie. Attualmente inizia ad esserci in ambito tecnologico una consapevolezza più diffusa rispetto a tali problematiche; tuttavia, per lungo tempo lo sviluppo di modelli e sistemi di intelligenza artificiale è avvenuto in maniera acritica, tenendo in limitata considerazione tali questioni, sia per l'assenza di una formazione specifica sia per la novità della dimensione dell'impatto sociale di tale tecnologia. Oggi, il rapidissimo sviluppo dell'intelligenza artificiale suggerisce di prendere seriamente in esame rischi che fino a poco tempo fa potevano sembrare solo ipotetici. In una prima fase è stata l'accresciuta sensibilità in merito a temi di protezione dei dati personali e tutela dei relativi diritti ad animare le preoccupazioni riguardo allo sviluppo delle tecnologie di IA. Oggi risulta possibile riflettere in maniera specifica sugli effetti discriminatori che possono avere gli algoritmi e l'uso dei sistemi algoritmici, su cui un numero crescente di studi sta concentrando la propria attenzione. Questo tema è particolarmente complesso e trasversale, poiché interseca questioni che hanno a che fare con la

società, la psicologia, il diritto, l'economia e la governance – oltre che col funzionamento stesso dei sistemi di intelligenza artificiale.

In particolare, poi, le discriminazioni nei confronti di persone LGBTI+ sollevano questioni ulteriori e specifiche, legate alla "visibilità" e alla sensibilità del dato dell'orientamento sessuale. Questa sezione mira a chiarire, sulla base dell'illustrazione di alcune delle principali modalità di applicazione e funzionamento dell'intelligenza artificiale, alcune delle tipologie di bias e delle modalità in cui possono emergere, nonché alcuni dei possibili approcci di soluzione che si iniziano ad osservare.

Persone LGBTI+ è un acronimo che richiama sinteticamente le persone lesbiche, gay, bisessuali, trans, intersessuali e le altre minoranze legate all'identità e all'orientamento sessuale, all'identità di genere e alle caratteristiche sessuali.



1.1 I processi dell'intelligenza artificiale e del machine learning e le aree di rischio

1.1.1 Intelligenza artificiale e machine learning: definizioni e concetti di base

La definizione di Intelligenza Artificiale è oggetto di un vivace dibattito (vedi box). Genericamente con il termine intelligenza artificiale (IA) si indica un insieme di approcci finalizzati a realizzare sistemi informatici capaci di svolgere compiti e realizzare prestazioni che, visti dall'esterno, sembrerebbero poter essere prodotti esclusivamente dell'intelligenza umana. Da questo punto di vista, il termine intelligenza artificiale può risultare fuorviante – e in effetti il suo uso è stato criticato da molti esperti. Stefano Quintarelli (2020), ad esempio, sostiene che, in molti degli approcci di maggior successo, i risultati dei sistemi di IA non sono ottenuti tentando di riprodurre i meccanismi di funzionamento dell'intelligenza umana e si basano su solidi approcci statistici. In maniera simile Luciano Floridi (2023) descrive lo sviluppo dell'IA come una dissociazione tra intelligenza e capacità di agire, sostenendo che i sistemi di intelligenza artificiale non cercano di riprodurre l'intelligenza umana, caratterizzata da riflessione critica e comprensione del contesto, ma piuttosto di operare in modo efficace su base

statistica. Questa dissociazione implica che le macchine possono svolgere compiti complessi, come riconoscere una anomalia in un esame diagnostico, prevedere un guasto, o fornire raccomandazioni meglio degli esseri umani, pur non avendo una vera intelligenza nel senso proprio del termine.

L'IA è uno strumento potente e in grado di avere impatto su un numero di persone e su casistiche varie e ad alta intensità e può presentare rischi significativi. Lo sviluppo e l'utilizzo dell'intelligenza artificiale devono quindi essere sottoposti a una stretta osservazione in coerenza dei principi giuridici ed etici socialmente condivisi per evitare conseguenze negative nella società. L'analisi di alcune dinamiche tecniche di base e dei rischi relativi alla discriminazione sarà il tema dei prossimi paragrafi.

Un **sistema di IA** è un sistema ingegnerizzato che genera output come contenuti, previsioni, raccomandazioni o decisioni per un dato insieme di obiettivi definiti dall'uomo (*ISO/IEC 22989:2022*).

Un sistema di IA è un sistema basato su macchine che, per obiettivi espliciti o impliciti, deduce dagli input ricevuti come generare output

come previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali. I diversi sistemi di IA variano nei loro livelli di autonomia e adattività dopo l'implementazione (OECD, 2024)

«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. (Regolamento (UE) 2024/1689)

Come suggerito da Luciano Floridi (2023), il termine IA è una "scorciatoia" per riferirsi in modo generico a diverse discipline, servizi e prodotti. Pertanto, una definizione precisa e monolitica rischia di essere fuorviante. Tuttavia, si può prendere come punto di riferimento una delle prime definizioni di IA adottate durante lo storico seminario del 1956 al Dartmouth College: "Il problema dell'intelligenza artificiale è considerato quello di far comportare una macchina in modi che verrebbero chiamati intelligenti se un essere umano si

comportasse in tal modo" (McCarty et al., 2006).

La definizione di sistema di IA contenuta nell'AI Act (Regolamento UE 1689/2024) aggiunge un tassello importante: "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".

Da questa definizione emerge un altro elemento di grande rilievo, relativo alla capacità di tali tecnologie non solo di realizzare con efficacia determinati compiti, ma anche di "influenzare gli ambienti". La potenza dell'IA nell'incidere sul contesto non solo organizzativo ed operativo ma anche sociale, culturale e discorsivo in cui si iscrive l'esperienza umana è infatti ciò che va tenuto presente anche in relazione alla valutazione dei rischi relativi alle possibili discriminazioni.

Le ricerche sull'intelligenza artificiale hanno una storia relativamente lunga che risale almeno, per gli approcci pratici legati allo sviluppo dell'informatica, al secondo Dopoguerra. L'attuale fase, che vede il susseguirsi di un'importante serie di applicazioni diffuse dell'intelligenza artificiale, si basa largamente su uno specifico approccio detto machine learning

(ML). Si tratta di una strategia finalizzata a creare sistemi che attraverso un addestramento a partire da una base di dati apprendono a svolgere determinati compiti senza essere specificamente programmati a tal fine. A fare la differenza, oltre al miglioramento degli algoritmi inizialmente concepiti tra anni Settanta e Ottanta, è stato l'aumento esponenziale della capacità di immagazzinamento di informazioni e di calcolo e soprattutto l'amplessissima disponibilità di dati resa possibile dalla diffusione di Internet, degli smartphone, dei social network e dell'Internet of Things.

Un **modello di apprendimento automatico** è un costrutto matematico che genera un'inferenza o una previsione sulla base di dati o informazioni in ingresso;

L'apprendimento automatico è il processo di ottimizzazione dei parametri del modello attraverso tecniche computazionali, in modo che il comportamento del modello rifletta i dati o l'esperienza;

un **algoritmo di apprendimento** viene impiegato per determinare i parametri di un modello di apprendimento automatico dai dati in base a determinati criteri (*ISO/IEC 22989:2022*).

L'apprendimento automatico è uno

dei campi di sviluppo tecnologico dell'IA assieme al ragionamento, la pianificazione, la percezione, la comunicazione. (*Samoili et al., 2021*)

È questa enorme mole di dati a rendere possibili i complessi processi di apprendimento alla base delle applicazioni di IA generativa come ChatGPT. Ma, al netto di applicazioni di questo tipo, divenute recentemente molto note, l'intelligenza artificiale è ormai utilizzata in tutti i settori, per gli impieghi più diversi. Alcune di queste applicazioni sono potenzialmente molto rilevanti per la vita delle persone e presentano rischi potenziali anche importanti.

Si pensi ad esempio al processo di selezione dei candidati nel mercato del lavoro, alla valutazione del rischio in ambito bancario o assicurativo, all'impiego dell'IA in ambito giudiziario, al riconoscimento facciale e alle tecniche di profilazione usate dalle forze di polizia, alla moderazione di contenuti e al contrasto alla disinformazione, alla generazione di contenuti media sintetici, inclusi i deepfake.

La vastità e l'importanza degli ambiti coinvolti rende evidente come una preoccupazione relativa a possibili discriminazioni, e più in generale alla fairness degli algoritmi, debba essere presa

estremamente sul serio. Inoltre, il modo non pienamente prevedibile in cui operano gli algoritmi di apprendimento automatico richiede una specifica attenzione anche rispetto alle diverse forme di discriminazione: ciò che vale per la discriminazione relativa all'etnia, ad esempio, non è necessariamente estendibile alle discriminazioni nei confronti delle minoranze LGBTI+.

1.1.2 Le tipologie di machine learning e i relativi bias potenziali

Riguardo al funzionamento del machine learning occorre chiarire una prima distinzione tra tre principali modalità in cui può avvenire l'apprendimento automatico: esso può essere supervisionato, non supervisionato o per rinforzo.

L'**apprendimento supervisionato** è una tecnica di ML in cui gli algoritmi sono addestrati su un insieme di dati (dataset) composto da abbinamenti effettuati sotto il controllo umano che uniscono un input a delle relative etichette adeguate (label) oppure a degli output desiderati. L'obiettivo dell'algoritmo è quello di costruire una mappatura tra gli input e le label (o gli output) per poter fare previsioni su nuovi dati in base a quanto appreso durante l'addestramento. Si tratta, in altre parole, di "riconoscere" un dato o un insieme di dati, applicando una certa

categoria. Ad esempio: si tratta di riconoscere se un'immagine rappresenta o meno un gatto, sulla base di precedenti associazioni guidate tra un'immagine (input) e la label "gatto", che vanno a formare il dataset. In questo caso, quindi, il processo di apprendimento è condotto da un supervisore che fornisce le label corrette.

Nell'apprendimento supervisionato, la presenza di caratteristiche sensibili nel dataset di addestramento, come l'etnia, il sesso, l'orientamento sessuale, la religione o altre informazioni personali, spesso riconducibili all'intervento umano nel processo, può comportare rischi per le persone portatrici di caratteristiche protette o di gruppi vulnerabili. Se un modello di machine learning viene addestrato con dati che includono informazioni sensibili come l'etnia, il genere o l'orientamento sessuale, o se questi dati riflettono per esplicito e per implicito pregiudizi esistenti nella società o nelle decisioni precedenti, oppure se i dati disponibili non sono sufficientemente rappresentativi, il modello riprodurrà quei pregiudizi. Quindi, quando il modello farà previsioni in futuro, potrebbe trattare in modo diverso o svantaggioso le persone di certe categorie. Questo problema è noto come *bias induttivo*, o *bias di apprendimento* e può portare a decisioni discriminatorie. È cruciale quindi cercare di identificare e mitigare tali bias durante il processo di sviluppo del modello.

L'apprendimento supervisionato è definito come “*apprendimento automatico che fa uso di dati etichettati durante l'addestramento*”. In questo caso, i modelli di ML vengono addestrati con dati di addestramento che includono una variabile di output o target nota o determinata (l'etichetta o label).

Il valore della variabile target per un determinato campione è noto anche come **ground truth**.

Le etichette possono essere di qualsiasi tipo, compresi valori categorici, binari o numerici, o oggetti strutturati (ad esempio, sequenze, immagini, alberi o grafici), a seconda del compito. Le etichette possono far parte del dataset originale, ma in molti casi vengono determinate manualmente o attraverso altri processi.

L'apprendimento supervisionato può essere utilizzato per compiti di classificazione e regressione, ma anche per compiti più complessi di predizione strutturata (ISO/IEC 22989:2022).

L'apprendimento non supervisionato, invece, consiste nell'addestramento di algoritmi su dataset senza label o output

desiderati predefiniti. L'obiettivo principale dell'apprendimento non supervisionato è scoprire strutture e pattern nascosti nei dati senza alcuna guida esterna. In questo caso, l'algoritmo cerca di raggruppare i dati in base a somiglianze o altre caratteristiche comuni. Anche nell'apprendimento non supervisionato possono emergere bias e potenziali discriminazioni. Ad esempio, alcune variabili, sebbene non rappresentino direttamente una caratteristica sensibile, possono essere correlate a quest'ultima. L'uso di tali variabili nel processo di apprendimento non supervisionato potrebbe portare il modello a stabilire relazioni indirette con caratteristiche sensibili, generando bias nei risultati. Ad esempio, se si utilizzano dati geografici per raggruppare i dati in cluster, il modello potrebbe involontariamente creare aggregazioni che riflettono differenze socio-economiche, razziali, di orientamento sessuale o culturali, anche se tali informazioni non sono state esplicitamente fornite al modello.

L'apprendimento automatico non supervisionato è definito come “*apprendimento automatico che fa uso di dati non etichettati durante l'addestramento*”.

L'apprendimento automatico non supervisionato può essere utile in casi come il **clustering**, in cui

l'obiettivo del compito è determinare le similarità tra i campioni nei dati di input.

La riduzione della dimensionalità di un set di dati di addestramento è un'altra applicazione dell'apprendimento automatico non supervisionato, in cui le caratteristiche più rilevanti dal punto di vista statistico vengono determinate indipendentemente da qualsiasi etichetta (ISO/IEC 22989:2022)

Nell'**apprendimento per rinforzo** non ci sono dati fissi: l'agente interagisce con l'ambiente per tentativi ed errori e riceve ricompense o punizioni in base alle sue azioni, cercando di massimizzare la ricompensa cumulativa, imparando quali azioni portano ai migliori risultati.

Anche nell'apprendimento per rinforzo possono emergere bias che possono determinare discriminazioni.

L'apprendimento per rinforzo può essere infatti influenzato da bias preesistenti nei dati con cui interagisce. Se l'ambiente o il sistema di ricompensa riflettono pregiudizi sociali, per esempio di genere o di orientamento sessuale, l'agente potrebbe apprendere politiche discriminatorie. Se l'algoritmo viene usato in un contesto decisionale (come la giustizia o le risorse umane) potrebbe generare risultati o

raccomandazioni svantaggiosi per i gruppi vulnerabili.

L'apprendimento per rinforzo è il processo di addestramento di un agente che interagisce con l'ambiente per raggiungere un obiettivo predefinito. Nell'apprendimento per rinforzo, un agente di apprendimento automatico apprende attraverso un processo iterativo di tentativi ed errori.

L'obiettivo dell'agente (o degli agenti) è trovare la strategia (cioè costruire un modello) per ottenere le migliori ricompense dall'ambiente. Per ogni prova (riuscita o meno), l'ambiente fornisce un feedback. L'agente (o gli agenti) regola quindi il suo comportamento (cioè il suo modello) in base a questo feedback (ISO/IEC 22989:2022).

1.1.3 Algoritmi di machine learning: tipologie e meccanismi di utilizzo, applicazioni pratiche

L'apprendimento automatico può essere utilizzato per tipologie di applicazione estremamente differenziate tra di loro. A titolo di esempio si può menzionare il riconoscimento delle immagini, la

produzione di testi, la moderazione di contenuti, la profilazione, la generazione di raccomandazioni. Ogni ambito e modalità di applicazione configura specifici scenari in cui possono emergere discriminazioni legate al funzionamento dei sistemi di machine learning. Di seguito sono indicati alcuni meccanismi di utilizzo presenti in svariate applicazioni e particolarmente rilevanti ai fini di una riflessione sulle discriminazioni potenziali:

- *Classificazione*: è un meccanismo di utilizzo generalmente associato alla categoria dell'apprendimento supervisionato e risponde all'esigenza di attribuire ad un oggetto o ad un individuo una classificazione predeterminata, assegnandogli una label di identificazione. Ciò può avvenire in maniera binaria (laddove i campi di scelta sono solamente due) o secondo una classificazione multipla, con una scelta tra un numero maggiore di categorie. Al fine di attribuire una label a ciascuna osservazione, l'algoritmo utilizza un insieme di caratteristiche (features). Ad esempio, per classificare un fiore, verranno valutate caratteristiche come la lunghezza dei petali o la larghezza dei sepali. Le label fornite come predizione dall'algoritmo di classificazione (classifier) possono essere indicate con il nome di

"predicted label", mentre le label effettive – ovvero le risposte corrette che vorremmo che il sistema di ML riproducesse – con il nome di "actual label". Esempi di applicazione pratica del meccanismo di classificazione sono il filtro antispam della posta elettronica o la churn analysis (analisi del rischio di abbandono di un servizio da parte di una certa categoria di clienti). Quando la classificazione riguarda persone, tra le feature utilizzate per la classificazione possono essere presenti, in modo esplicito o tramite correlazione con le altre features, caratteristiche sensibili, tra le quali etnia, sesso, orientamento sessuale, religione, disabilità e luogo di nascita. È possibile raggruppare tali caratteristiche in unico vettore, il quale rappresenta un sottoinsieme delle feature.

- *Clustering*: è un meccanismo di utilizzo associato alla categoria dell'apprendimento non supervisionato ed è impiegato in applicazioni nelle quali non è disponibile una label a priori ma è necessario trovarla. Nel caso del clustering si impiegano logiche di tipo geometrico, non sempre direttamente controllabili e intelligibili: si utilizzano iperspazi (spazi a più di tre dimensioni)

definiti dalle variabili in gioco, e le categorie emergono come “centroidi”, punti di questo iperspazio. Si fissano alcuni parametri come la funzione di distanza e una soglia di densità attesa, che determina la densità minima di punti necessaria affinché si formi un cluster. Tuttavia i cluster che vengono individuati dall’algoritmo non rappresentano concetti direttamente comprensibili dall’uomo. I gruppi di dati vengono infatti formati sulla base di somiglianze matematiche o statistiche tra i punti, senza necessariamente riferirsi a categorie concrete del mondo reale. Per uno stesso insieme di dati è possibile, fissando parametri diversi, individuare più raggruppamenti. Esempi di applicazione del meccanismo di clustering sono la segmentazione di mercato per finalità di marketing oppure l’analisi di dati geospaziali e la creazione di mappe tematiche nelle applicazioni geografiche come Google Maps. Questa modalità di funzionamento presenta, dal punto di vista delle discriminazioni, sia opportunità che rischi. Da un lato, il fatto di procedere a raggruppamenti senza partire da etichette prestabilite può potenzialmente evitare che tali raggruppamenti riproducano stereotipi già consolidati. Dall’altro

lato, vi è il rischio che le clusterizzazioni avvengano secondo criteri difficili da individuare e potenzialmente non etici, secondo linee di demarcazione che seguono vulnerabilità note o non note delle persone. La clusterizzazione e la conseguente discriminazione possono avvenire attraverso la combinazione di una moltitudine di variabili, tra cui le cosiddette “variabili proxy”, ossia attributi o caratteristiche presenti nel dataset che non sono direttamente correlati alla “variabile protetta” (o sensibile), ma che possono essere utilizzati dagli algoritmi per inferire indirettamente la variabile protetta stessa.

L’algoritmo potrebbe insomma ricostruire automaticamente la variabile protetta in modo implicito e utilizzarla per determinare i vari cluster. Ad esempio, l’informazione relativa all’orientamento sessuale o all’identità di genere, anche se non disponibile, potrebbe essere predetta incrociando informazioni personali di vario genere. In questo senso l’algoritmo potrebbe rappresentare un potente strumento di discriminazione. Inoltre la clusterizzazione potrebbe cristallizzare nuove vulnerabilità connettendo fra loro caratteristiche anche non protette nell’attuale regime giuridico applicabile.

- *Computer vision*: è un meccanismo di utilizzo che impiega sia tecniche di apprendimento supervisionato che non supervisionato e si concentra sull'elaborazione e l'interpretazione di dati visivi. Grazie all'utilizzo di algoritmi di ML, i modelli possono essere addestrati per comprendere, riconoscere e categorizzare gli oggetti presenti in immagini digitali o video. Esempi di applicazioni pratiche di questi algoritmi sono il riconoscimento facciale o la segmentazione delle immagini per analisi mediche. Anche con la *computer vision* possono emergere bias e potenziali discriminazioni. Ad esempio, se il dataset utilizzato per addestrare il modello contiene immagini che sono state selezionate in modo non rappresentativo o che riflettono pregiudizi culturali o sociali, il modello può replicare tali bias.
- *Natural language processing (NLP)*: è un meccanismo di utilizzo che impiega sia tecniche di apprendimento supervisionato che non supervisionato ed è relativo alla capacità dell'IA di comprendere, interpretare e generare il linguaggio umano in modo naturale, attraverso il testo scritto o parlato. Esempi di applicazione pratica del NLP sono le chatbot per l'assistenza virtuale

degli utenti o gli algoritmi più avanzati di traduzione automatica da una lingua all'altra. In questo campo la discriminazione può emergere per esempio dai testi prodotti, che possono proporre linguaggio offensivo e stereotipato, riproducendo e amplificando gli stereotipi impliciti nel dataset.

- *Large language models (LLM)*: sono degli algoritmi che rientrano nel campo del NLP e che hanno rivoluzionato l'ambito dell'IA e del ML legato al linguaggio naturale per la loro ampiezza e potenza, frutto di una combinazione di tecniche e un training profondo su enormi dataset. Per questo motivo, sono diventati uno degli strumenti più potenti e promettenti nel campo del trattamento del linguaggio naturale e dell'IA in generale. Fanno parte degli LLM alcuni tra i modelli divenuti più noti anche presso il grande pubblico, come GPT, modello sviluppato da OpenAI, e BERT e Gemini, sviluppati da Google, Claude di Anthropic o LLama di Meta, per stare ai più noti. Sebbene i LLM siano stati in grado di ottenere notevoli avanzamenti nelle prestazioni in diversi ambiti, presentano anche significativi rischi. I LLM creano notevoli sfide, per esempio, sul fronte dell'incertezza dei risultati. L'incertezza può essere

epistemica o aleatoria. L'incertezza epistemica deriva dalla mancanza di conoscenza nel modello rispetto a certi argomenti e può essere ridotta se si migliora il modello, ad esempio con più dati o un addestramento migliore. L'incertezza aleatoria deriva invece dalla casualità intrinseca dei dati e non può essere eliminata: in un modello probabilistico per un dato input ci sono infatti più risposte possibili. Prendiamo ad esempio un modello che calcoli la probabilità di pioggia per il giorno successivo. Ipotizziamo che sia stato addestrato solo con dati relativi alle giornate estive; se lo usassimo per una previsione in una giornata invernale potrebbe essere inefficace a causa della mancanza di conoscenza del contesto (incertezza epistemica), ma addestrandolo in modo più completo questa incertezza può essere ridotta. Ma anche in seguito a un addestramento migliore dobbiamo tenere presente che la previsione potrebbe comunque essere errata perché ci saranno sempre fattori imprevedibili non controllabili (incertezza aleatoria). Fare affidamento su grandi quantità di dati non curati (ad es. prelevati dal web) aumenta al contrario l'incertezza aleatoria (Bender et al., 2021).

Il modello può manifestare comportamenti sconosciuti ed erratici e presenta sfide per la riproducibilità e la spiegabilità – questioni sulle quali si tornerà nel par. 2.2. Le prime esperienze hanno dimostrato che le preoccupazioni sull'uso dei LLM sono effettivamente valide, con risultati sperimentali preliminari che mostrano che i LLM presentano bias significativi.

Le **reti neurali artificiali** sono modelli matematici che si trovano alla base di molti algoritmi di apprendimento automatico, e sono ispirate alla struttura e al funzionamento delle reti neurali biologiche. Vengono chiamate *reti neurali* proprio perché simulano il meccanismo di segnalamento dei neuroni del sistema nervoso.

Le reti neurali artificiali sono composte di strati di nodi, o neuroni artificiali, organizzati in strati: uno strato di input, uno o più strati nascosti e uno strato di output. In esse ogni nodo è connesso agli altri tramite interconnessioni pesate, ispirandosi ai meccanismi biologici presenti nel cervello. Quando un nodo raggiunge la soglia di attivazione, specificata da una funzione di attivazione, allora invia dati allo strato successivo della rete neurale.

1.1.4 Gli step nel ciclo di vita degli algoritmi di machine learning

Un altro modo di analizzare gli algoritmi di machine learning consiste nel soffermarsi non tanto sulle tipologie di algoritmi e sui diversi compiti per i quali essi vengono progettati, quanto sul loro ciclo di vita e in particolare sulla fase di produzione. Uno sguardo agli step attraverso cui un algoritmo di ML viene preparato e addestrato, con una particolare attenzione al dataset di partenza e al modo in cui viene elaborato, permette di intercettare ulteriori rischi di discriminazione e violazione dei diritti.

All'interno del percorso di progettazione e utilizzo di un algoritmo di ML è possibile distinguere i seguenti passaggi:

1. *Raccolta e acquisizione dei dati*: la quantità e la qualità dei dati di partenza che vengono acquisiti, raccolti e selezionati, ed il relativo contesto di acquisizione, influiscono in maniera molto significativa nel determinare l'adeguatezza del modello. L'accuratezza e rappresentatività dei dati è determinante rispetto all'output. Spesso inoltre per attingere alle grandi quantità di dati necessarie per l'addestramento degli algoritmi occorre adoperare dati "grezzi" che incorporano assai spesso elementi

discriminatori. Ad esempio, nello sviluppo di un algoritmo per l'approvazione automatica di prestiti bancari basato su dati finanziari dei richiedenti, si potrebbero raccogliere dati sui precedenti richiedenti e sul loro "successo" che incorporano fattori protetti come ad es. l'etnia o il genere dei richiedenti, e così replicare valutazioni legate a tali fattori.

2. *Memoria*: l'estrazione e la conservazione dei dati costituiscono un altro passaggio potenzialmente sensibile, poiché esso riguarda l'immagazzinamento di informazioni di cui va garantita la privacy (ad esempio in relazione ai limiti alle finalità di utilizzo e alla conservazione) e la sicurezza (rispetto, ad esempio, ad accessi non autorizzati).
3. *Addestramento*: rappresenta una fase cruciale, nella quale vengono determinate le caratteristiche del modello. In questa fase possono incidere, oltre alle caratteristiche del dataset e ai bias dei curatori, anche i bias prettamente statistici e computazionali (cfr. 1.1.5).
4. *Elaborazione e utilizzo del modello*: in questa fase il modello viene utilizzato per lo scopo per cui è stato pensato (classificazione,

clustering ecc.). A questo livello il carattere del potenziale bias dipende dunque dalle peculiarità specifiche del modello e dell'algoritmo e da come viene utilizzato, secondo quanto già analizzato nel par. 1.1.3.

5. *Rappresentazione dei risultati e interfacce decisionali*: riguarda il modo in cui l'output, una volta elaborato dall'algoritmo, viene rappresentato per essere fruito dai soggetti umani e per fungere, eventualmente, da supporto alle decisioni umane (nel caso in cui le decisioni non vengano adottate direttamente dal sistema).

6. *Adattamento risultati*: riguarda il modo in cui i sistemi vengono iterativamente riadattati in base ai risultati per verificare la qualità del prodotto. Durante questa fase, si monitorano attentamente le risposte generate dal sistema, raccogliendo feedback e dati sulle prestazioni. Questi dati, in un secondo passaggio, vengono utilizzati per apportare aggiornamenti al modello, all'algoritmo o ai parametri del sistema, al fine di ottimizzare le sue prestazioni e assicurare che continui a soddisfare le esigenze stabilite.

Queste sei fasi sono profondamente interconnesse; pertanto, una criticità che emerge in uno step si ripercuote anche nei successivi. Inoltre, i rischi che possono presentarsi a livello di bias non riguardano solo gli aspetti tecnici di funzionamento di modelli e algoritmi, ma richiedono anche un'attenzione alla natura dei dati e una sensibilità alle dinamiche discriminatorie che si perpetuano nella società.

1.1.5 Le tipologie di bias nell'IA

Per intercettare il potenziale emergere di bias nell'ambito dell'IA occorre valutare come il piano tecnico e computazionale vada a intersecarsi con il contesto più

ampio delle dinamiche prettamente umane e sociali, presenti e pregresse. In un'ottica di analisi, gestione e mitigazione, è utile cercare di distinguere i bias per comprenderne origine e impatti; la suddivisione che segue riprende ad alto livello una delle proposte presenti in letteratura (Schwartz et al., 2022).

- *Bias sistemici*: derivano dalle pratiche sociali e istituzionali storicamente sedimentate, che tendono a favorire alcuni gruppi sociali e a svalutarne altri. Questo può non essere necessariamente il risultato di pregiudizi o discriminazioni consapevoli, ma piuttosto di un senso comune diffuso che incorpora norme e opinioni radicate. Esempi comuni di bias sistematici sono il razzismo, il sessismo o l'omobittransfobia largamente diffusi se non persino istituzionalizzati. Un altro esempio si verifica quando le infrastrutture per la vita quotidiana non sono sviluppate utilizzando principi di "universal design", così limitando o ostacolando l'accessibilità per le persone con disabilità.

I bias sistemici diventano rilevanti per una riflessione sull'IA nella misura in cui possono essere incorporati nei dati che vengono usati dagli algoritmi in fase di addestramento, nonché nelle norme, nelle pratiche e nei processi

istituzionali che influiscono lungo tutto il ciclo di vita dell'IA.

- *Bias umani*: riflettono errori sistematici in cui facilmente il pensiero umano può incorrere, essendo basato su un numero limitato di principi euristici e predizioni di valori. Questi bias sono spesso impliciti e tendono a influenzare il modo in cui un individuo, un gruppo o un'istituzione percepisce le informazioni per prendere decisioni o integrare informazioni mancanti o sconosciute. Esistono diverse tipologie di bias umani, come i bias cognitivi e percettivi, che si manifestano in tutti i domini e non solo nelle interazioni umane con l'IA. Essendo onnipresenti nei processi cognitivi umani, questi bias incidono su tutto il ciclo di vita dell'IA – ad esempio nella selezione o nell'etichettatura dei dati in fase di apprendimento – e nell'uso dei sistemi di IA una volta implementati, ad esempio nella lettura degli output. La consapevolezza dell'esistenza dei bias umani non garantisce il controllo su di essi, poiché sono spesso impliciti e non consapevoli.
- *Bias statistici e computazionali*: derivano dalla non rappresentatività del campione di dati di partenza

(bias statistici) oppure, in relazione esclusivamente all'IA, riguardano errori introdotti durante il processo di addestramento dei modelli (bias computazionali). Nei sistemi di IA, questi bias sono dunque presenti sia a livello di dataset che dei processi algoritmici utilizzati nello sviluppo delle applicazioni. Spesso tali bias si verificano quando un modello non è generalizzabile poiché è stato addestrato su un insieme di dati che non sono stati preparati adeguatamente o che non sono sufficientemente rappresentativi della realtà. Errori tipici possono essere causati ad esempio dall'eterogeneità dei dati; dalla rappresentazione di dati complessi in rappresentazioni matematiche più semplici; da sovradattamento e sottoadattamento del modello; dal trattamento degli outlier (valori anomali all'interno di un insieme di dati che si discostano significativamente dai valori tipici); da fattori di pulizia e imputazione dei dati, cioè quei processi utilizzati per gestire dati mancanti o inconsistenze all'interno di un dataset. Un esempio è rappresentato dalle modalità di raccolta dei dati in ambito sanitario: i sistemi informativi sanitari non sono generalmente progettati per accogliere la diversità di genere,

confondono spesso genere e caratteristiche sessuali, e non permettono di mantenere una storia clinica unitaria in caso di transizione di genere, che coincide spesso con l'apertura di una nuova anagrafica scollegata dalla precedente (Kartik, 2024).

Ai fini della nostra analisi è utile soffermarsi sui bias statistici e computazionali, che assumono particolare rilevanza per le soluzioni proposte dalla data science.

Tra i principali tipi di bias statistici e computazionali sono stati indicati i bias di misura, di aggregazione, di rappresentatività e di campionamento e i bias algoritmici in senso specifico (Mehrabi et al., 2021).

I **bias di misura** dipendono dal modo in cui vengono scelte e misurate alcune caratteristiche del campione: ad esempio il numero di arresti può essere usato come variabile per stimare la pericolosità sociale, senza tenere in considerazione che le minoranze vengono subiscono controlli di polizia più frequenti (cfr. 2.1, il caso COMPAS).

I **bias di aggregazione** (fallacia ecologica) si verificano quando si traggono conclusioni sugli individui basandosi sull'intera popolazione. Ad esempio in ambito clinico alcuni parametri possono variare in modo complesso in base al

genere e all'etnia, e un modello che ignora queste differenze non andrà bene per tutti i sottogruppi.

I **bias di rappresentatività** emergono quando il campione analizzato non rappresenta correttamente la diversità della popolazione: possono mancare ad esempio alcuni sottogruppi; diversi dataset utilizzati nella computer vision mostrano un bias di rappresentatività verso le culture occidentali. Un caso simile è quello del bias di disponibilità riferito ai dataset, che si ha quando i dataset utilizzati manifestano una limitazione nella raccolta dei dati che influisce sulla rappresentatività e in definitiva sull'esito dell'utilizzo del dataset nello sviluppo di un modello.

I **bias di campionamento** sono simili ai bias di rappresentatività ed emergono quando il campionamento dei sottogruppi non è uniforme, cioè quando alcuni sottogruppi vengono campionati più frequentemente di altri. Il risultato è che le previsioni generate dal modello non sono generalizzabili perché il modello non è rappresentativo della realtà.

I **bias algoritmici** in senso proprio non riguardano i dati ma solo la scelta degli algoritmi e il loro uso, ad esempio l'applicazione di alcune funzioni di ottimizzazione, la scelta di usare un modello di regressione sull'intera popolazione o tenendo in considerazione i sottogruppi, e in generale tutte le decisioni

che vengono prese quando si applicano i metodi statistici.

È importante comprendere queste diverse categorie di bias, che, nella maggior parte dei casi, non sono determinate dall'algoritmo in quanto tale ma che nell'IA possono riprodursi e amplificarsi. Capire in che modo questi bias caratteristici delle dinamiche sociali e umane influenzano le fasi di sviluppo (par. 1.1.4) e le applicazioni (par. 1.1.3) dei sistemi di IA permette di implementare misure di equità e di mitigare gli effetti negativi dei bias, sia nei sistemi di IA stessi che, più ampiamente, nell'interazione dell'IA con la società.

La capacità dell'IA di replicare e rinforzare dinamiche discriminatorie pone al centro la questione del *danno* potenziale, cioè degli effetti nefasti eventualmente prodotti dall'IA. Le due principali categorie di danno possono essere così individuate:

- *Danni allocativi*: si hanno quando un sistema di decisione automatizzato nega in maniera discriminatoria un'opportunità o una risorsa a un certo gruppo di persone (ad esempio in ambito bancario, assicurativo, di selezione del personale, educativo o giudiziario).
- *Danni rappresentativi*: si hanno quando il sistema tende a rinforzare la subordinazione di determinati gruppi, a perpetuare e a irrobustire gli stereotipi rispetto a determinate

linee di divisione identitaria (genere, orientamento sessuale, etnia, classe sociale). Questo può avvenire nell'ambito dei social media, dell'informazione e della sorveglianza.

Peraltro, la questione dei danni rappresentativi ha assunto crescente rilievo con la diffusione, ormai capillare, dei social media, che creano un ambiente che può influenzare le decisioni individuali e collettive. Tramite i meccanismi di profilazione e i noti fenomeni di echo chamber e confirmation bias, gli algoritmi dei social network sono potenzialmente in grado di rinforzare e accentuare gli stereotipi, creando polarizzazioni identitarie e creando potenzialmente le condizioni per un loro rafforzamento istituzionale e sistemico.

2. Dati, algoritmi, minoranze e discriminazioni

Il tema del rapporto tra algoritmi, minoranze e discriminazioni appare teoricamente complesso per diverse ragioni. Scegliere, classificare e categorizzare sono compiti per molti aspetti centrali, per svolgere i quali gli algoritmi sono stati progettati.

Questi hanno spesso lo scopo di prendere o suggerire decisioni relative a individui e gruppi, che vengono determinati in base a criteri diversi, e l'appartenenza o meno ai quali può avere effetti estremamente concreti. La definizione normativa e pratica di quali azioni, decisioni e trattamenti costituiscano una discriminazione risponde a dinamiche complesse e talvolta conflittuali in ambito sociale, culturale e politico. A questi aspetti si aggiunge, nel contesto dell'AI, la complessità logica e computazionale.

Il caso del recruiting di Amazon

A partire dal 2014 il team di Amazon specializzato nel ML ha sviluppato un software che potesse compiere automaticamente lo screening dei curriculum vitae dei candidati, assegnando un punteggio in relazione

alla posizione lavorativa.

Un paio di anni dopo l'esperimento è stato definitivamente chiuso: prima ancora di implementarne l'uso, ci si è accorti che il software predilige, soprattutto per le posizioni tecniche, i candidati maschi. Il training era infatti avvenuto sulle assunzioni dei dieci anni precedenti, che vedevano una maggioranza maschile marcata nel personale, dato che si era trasformato nell'algoritmo in un fattore di esclusione delle donne a prescindere dai contenuti del cv.

Vi è poi una seconda questione che attiene alla natura stessa delle minoranze, sulle quali i dati disponibili sono per definizione minori rispetto alla media della popolazione. Se una maggiore quantità di dati disponibili implica una migliore qualità del modello, le minoranze sembrano condannate ad un trattamento più impreciso e a rischio di errori. Devono allora essere introdotte strategie correttive mirate.

Il caso COMPAS

Il caso rappresenta uno dei più noti esempi di discriminazione algoritmica. Il sistema COMPAS

(acronimo per *Correctional Offender Management Profiling for Alternative Sanctions*) è un software utilizzato in alcuni Stati negli USA per misurare il rischio di recidiva o il rischio di violazione delle misure cautelari precedenti il giudizio.

Un'[indagine condotta da ProPublica](#) nel 2016 ha mostrato come l'algoritmo sia soggetto a un bias razziale sistematico. Le persone di colore sono infatti quasi due volte più soggette rispetto ai bianchi a essere etichettate come a rischio maggiore di recidiva (sovrastima), mentre accade l'opposto per le persone bianche: sono molto più suscettibili rispetto alle persone di colore a essere etichettati a rischio minore ma poi commettono altri reati (sottostima).

Il bias del software pare derivare innanzitutto dai dati con cui l'algoritmo è stato addestrato e dal criterio di fairness a cui risponde l'algoritmo.

Vi è, in terzo luogo, una specificità che attiene, in particolare, alle minoranze LGBTI+. Il dato relativo all'appartenenza a tali minoranze, in particolare nel contesto europeo, non è necessariamente un dato esplicito e tracciato. Nei casi di studio più

noti di discriminazione derivata dall'IA, il bias poteva essere facilmente dimostrabile (ad esempio nei confronti di persone con una certa origine etnica), in quanto la categoria era più immediatamente registrabile. Nel caso delle persone LGBTI+, la variabile sensibile è più difficile da riconoscere e quasi sempre il dato non è esplicitamente registrato nei dataset. Questo da un lato non annulla la possibilità di discriminazioni, attraverso variabili proxy e correlazioni con altri elementi (The Economist, 2017). Al tempo stesso, dall'altro lato, rende più difficile la rilevazione dell'esistenza della discriminazione, a meno di non inferire ed esplicitare il dato a questo scopo. Questo però, oltre a sollevare il rischio potenziale di un uso malevolo di tale informazione, solleva un altro aspetto critico, relativo al rapporto tra tutela antidiscriminatoria e tutela della privacy, che verrà ripreso nel par. 5.2.

La capacità dei modelli e in particolare dei modelli di NLP di individuare correlazioni e pattern presenta anche potenzialità di lotta alla discriminazione, potendo essere utilizzata per esplicitare i dati culturali sottostanti a grandi moli di dati e rilevare i bias sistemici in essi incorporati.

Va inoltresottolineato che spesso le persone o i gruppi non sono consapevoli di essere discriminati perché, come rilevano gli scienziati sociali, hanno internalizzato la discriminazione e non la identificano come

negativa a livello esplicito. Circostanza che rischia di essere amplificata nei casi in cui la discriminazione venisse ipostatizzata in un modello o in un processo automatizzato e resa quindi una prassi costante.

Inoltre gli utenti che interagiscono con sistemi che contengono un bias possono inconsapevolmente interiorizzarlo e riproporlo. Uno studio psicologico recente (Vicente e Matute, 2023) ha indagato infatti se e come possa l'intelligenza artificiale influenzare le decisioni umane. Attraverso una serie di esperimenti che coinvolgevano uno strumento fittizio di supporto alla diagnostica per immagini hanno mostrato che i partecipanti esposti a suggerimenti che includono un bias apprendono a replicare quel bias anche quando smettono di usare lo strumento. Questo risultato sperimentale sottolinea l'importanza di intervenire sul bias il prima possibile nel ciclo di sviluppo dei sistemi di IA.

Questo nostro report si inserisce nell'ambito degli studi su fairness, accountability e transparency (FAT), un filone di ricerca ben consolidato che ha l'obiettivo di riflettere su come progettare e implementare sistemi di IA in modo etico, inclusivo e responsabile. È importante notare che questi temi sono ampiamente affrontati nella letteratura scientifica e sono oggetto di indagine da parte di enti pubblici e privati. La FAccT (Fairness, Accountability and Transparency in Machine Learning, in precedenza "FAT") è

ad esempio una conferenza specializzata, organizzata a cadenza annuale fin dal 2018 da ACM (Association for Computer Machinery) e caratterizzata da un approccio multidisciplinare.

L'IEEE si occupa dei temi FAT tramite l'IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, lanciato nel 2016. L'Unione Europea invece attraverso il suo High-Level Expert Group on Artificial Intelligence e lo European Centre for Algorithmic Transparency. Queste iniziative hanno sviluppato delle linee guida per promuovere i principi FAT nei sistemi di IA. Le Ethics Guidelines for Trustworthy AI sviluppate dall'Unione Europea (European Commission: Directorate-General for Communications Networks, Content and Technology, 2019) affrontano i temi FAT nel capitolo 2, che si riferisce alla realizzazione di una IA affidabile.

In relazione ai temi della fairness viene specificato che inclusione e diversità devono essere prese in considerazione in tutto il ciclo di vita del sistema, e che tutti i portatori di interesse devono essere coinvolti al fine di garantire la parità di trattamento ed evitare le discriminazioni. Strettamente collegato è il principio di accountability, che deve essere garantito tramite la verificabilità degli algoritmi, dei dati e delle scelte di progettazione attraverso revisori interni ed esterni, la riduzione e segnalazione degli effetti negativi, l'analisi dei compromessi e la messa a disposizione di procedure di ricorso. Segue la trasparenza dei dati, del sistema e del modello di business, in termini di tracciabilità dei dataset e degli algoritmi e delle decisioni prese dal sistema, adeguati livelli di spiegabilità dei sistemi, e la necessità che i sistemi di IA siano identificabili in quanto tali. Infine viene illustrato il principio della governance dei dati, che deve riguardare la qualità e l'integrità dei dati utilizzati, la pertinenza rispetto agli specifici sistemi di IA e ai settori in cui i sistemi di IA saranno utilizzati, e la gestione delle modalità di accesso e la garanzia di riservatezza, anche tramite crittografia, anonimizzazione e aggregazione, e meccanismi che individuino responsabili e che permettano a terzi di segnalare problemi di protezione dei dati.



Con questo report ci proponiamo di fornire un contributo a un filone di ricerca ben consolidato, concentrandoci sulle vulnerabilità specifiche delle minoranze sessuali.

2.1 Fairness e algoritmi

In generale, con il concetto di fairness ci si riferisce all'auspicio che i sistemi di intelligenza artificiale siano appunto fair, termine traducibile in italiano con "equo, giusto, imparziale, chiaro". Questo però assume diversi significati, non solo in base al contesto specifico e al campo d'utilizzo, ma anche a seconda dei molteplici approcci disciplinari. Per esempio, da un punto di vista del diritto, la fairness sarà legata alla protezione di individui e gruppi dalla discriminazione o maltrattamenti, con un focus sull'interdizione di comportamenti, pregiudizi e decisioni basate su determinati fattori protetti o categorie di gruppi sociali. Le scienze sociali, invece, guardano alla fairness alla luce delle relazioni sociali, delle dinamiche di potere, delle istituzioni e dei mercati, mentre la filosofia e l'etica in particolare vedono la fairness come ciò che è moralmente giusto.

All'interno della molteplicità di soluzioni proposte, una strada molto ambiziosa – in quanto si vorrebbe oggettiva – è rappresentata dalla ricerca per la formulazione di una teoria quantitativa

della discriminazione. Il carattere quantitativo servirebbe a produrre tool capaci di effettuare audit e controlli automatici, soprattutto riguardo a discriminazioni implicite.

Si tratta di un'avanguardia teorica molto complessa da percorrere anche per via della difficoltà nel condurre un'analisi del concetto di discriminazione in contesti diversi, siano essi culturali, sociali, geografici o altro.

Ci pare rilevante evidenziare che esiste una grande varietà di metriche di fairness ossia di criteri che cercano di tradurre in termini statistico-matematici la fairness. Tuttavia ogni metrica si basa su una definizione di fairness e ogni definizione riesce a catturarne solo alcuni aspetti. Counterfactual fairness (CF), demographic parity (DP) conditional demographic parity (CDP) e error parity (EP) sono esempi di metriche di fairness. La CF assicura che un individuo avrebbe ricevuto la stessa previsione se i valori relativi a fattori protetti fossero stati diversi. La DP assicura che ogni gruppo, ad esempio uomini e donne, riceva un risultato positivo nella stessa proporzione. Se il 60% dei candidati a un lavoro vengono selezionati, la DP richiede che la proporzione sia la stessa per uomini e donne, senza però tenere conto delle qualifiche. Nel medesimo contesto la CDP ritiene che solo nel caso in cui uomini e donne abbiano le stesse qualifiche allora debbano avere la stessa

probabilità di essere selezionati, in un'ottica più individualistica. EP si assicura che l'accuratezza statistica delle previsioni - intesa come distribuzione degli errori - sia la stessa per tutti i gruppi. Risulta evidente che le varie metriche incarnano approcci diversi al concetto di fairness, e si esprimono in diverse modellizzazioni statistiche.

Le diverse metriche non sono quindi commensurabili e non possono essere soddisfatte contemporaneamente se non in scenari semplicistici e non rappresentativi della realtà (Castelnovo et al., 2022).

Per comprendere l'impatto della molteplicità di metriche possiamo tornare al caso del sistema COMPAS dove gli autori dichiarano di aver adottato una metrica di fairness per la quale dato un determinato punteggio di rischio, l'accuratezza (cioè il tasso di errore) dovrebbe essere la stessa per tutti i gruppi (Dieterich et al., 2016). Secondo i giornalisti di ProPublica (Angwin et al., 2016) non sarebbe però stata tenuta in considerazione la distribuzione di falsi positivi (persone erroneamente classificate ad alto rischio) e falsi negativi (persone erroneamente classificate come a basso rischio), che contribuiscono entrambi al calcolo dell'accuratezza secondo il criterio scelto per COMPAS.

I risultati relativi alle persone di colore avevano un tasso di falsi positivi più alto,

mentre quelli relativi ai bianchi avevano un tasso di falsi negativi più alto, ma questo sbilanciamento non era intercettato dalla metrica di fairness adottata. La diversa metrica di fairness impiegata per valutare il modello incide quindi in materia determinante sull'esito della valutazione stessa.

Una **confusion matrix** è uno strumento che può essere usato per valutare le prestazioni di un classificatore. Riporta il numero di falsi positivi, falsi negativi, veri positivi e veri negativi, e include ulteriori criteri di prestazione derivati da questi valori.

Poiché la confusion matrix contiene e compara diverse metriche, permette una analisi dettagliata delle performance di un classificatore ed è utile ad eludere o scoprire le debolezze di singole metriche (ISO/IEC 24027:2021)

La molteplicità di definizioni logiche e computazionali di fairness impone che la scelta del modello di fairness da impiegare nell'ambito di un sistema algoritmico non sia lasciata ad una scelta autonoma di chi sviluppa l'algoritmo o di chi lo fornisce ma risponda invece a processi di assunzione di responsabilità (accountability) chiaramente tracciabili e rispondenti alle norme

giuridiche e alle regole di governance dell'organizzazione. In assenza di una definizione universalmente condivisa di fairness, comunque, l'approccio attualmente più seguito è quello di individuare strategie più limitate, distinguendo le soluzioni rispetto ai diversi contesti, producendo diverse definizioni adatte a ciascuno di essi e sviluppando dei test che verifichino se gli algoritmi le soddisfino o meno, lavorando con audit (spesso automatici) che operano sui diversi livelli di funzionamento dell'IA (generazione dei gruppi, output, ecc.). Nell'individuare la corretta definizione per ogni caso, è necessario fare riferimento alle caratteristiche del setting.

È dunque a partire da ambiti specifici che, per il momento, si tenta di individuare criteri operativi per tradurre il concetto di fairness.

Bisogna inoltre tenere in considerazione il ruolo umano nello sviluppo dei sistemi di IA. Pant e colleghi (2024) hanno analizzato i livelli di consapevolezza, da parte dei professionisti dell'IA, dei principi etici e di attenzione all'impatto sociale nei sistemi software, osservando che la maggior parte dei professionisti dichiara di essere a conoscenza principalmente di principi etici quali la privacy e la protezione dei dati personali e, solo in secondo luogo trasparenza ed equità. Questa conoscenza deriverebbe principalmente dai regolamenti aziendali e da politiche interne alle aziende. I professionisti riportano da

un lato la difficoltà nella comprensione e nell'interpretazione di norme complesse e dall'altro ostacoli tecnici come la ridotta interpretabilità dell'IA, che ostacolano l'integrazione di principi etici. Anche le difficoltà di comprensione reciproca tra esperti e clienti finali, che spesso non comprendono tutti gli aspetti dell'impatto sociale dell'IA, complicano il percorso verso l'integrazione di principi etici nei sistemi.

La complessità delle questioni in gioco suggerisce che la soluzione a queste problematiche debba essere ricercata attraverso una stretta interazione tra la dimensione tecnica e il contesto sociale e istituzionale in cui le tecnologie sono immerse: semplificando molto, occorre preservare e costruire dimensioni di interazione tra tecnologia e componente umana perché lo spazio delle soluzioni vada ad ampliarsi e non a restringersi. Questo presuppone, nel concreto, un dialogo tra policy maker, attori sociali e tecnici, che permetta di costruire insieme le diverse dimensioni della definizione di fairness e la definizione di processi partecipativi in cui si incontrino la dimensione tecnica ed esperti dei gruppi vulnerabili.

La possibilità di un processo che agisca attraverso correzioni a partire da un dialogo tra competenze, sensibilità e interessi diversi, al fine di andare a disegnare o modificare il funzionamento degli algoritmi, presuppone tuttavia la

possibilità di comprendere e modificare la logica dell'algoritmo. Questo ci porta a incrociare la questione cruciale della spiegabilità (explainability). Un sistema di IA, in particolare quando basato su reti neurali, è infatti molto meno spiegabile rispetto ad algoritmi e funzioni più semplici perché gestisce un enorme numero di variabili e relazioni tra loro. In algoritmi più semplici codificati da esseri umani è possibile vedere, più o meno chiaramente, come ogni variabile influisca sul risultato. In una rete neurale, invece, le variabili passano attraverso moltissimi livelli e interazioni complesse, e capire esattamente come ciascuna contribuisce al risultato finale diventa estremamente difficile. I sistemi di IA più complessi e scarsamente spiegabili vengono considerati "black box": conosciamo i dati di input e il risultato finale, ma il processo interno rimane oscuro.

2.2 Spiegabilità dei modelli di machine learning: classificazione dei metodi di spiegazione

La spiegabilità dei modelli di machine learning (explainable AI) si riferisce alla capacità di spiegare - o spesso di ipotizzare con ragionevole approssimazione i fattori determinanti - le valutazioni o le decisioni prese da un modello di IA basate sui dati, in modo che possano essere comprese

dagli esseri umani. Essa può essere intrinseca oppure avvenire grazie a metodi che analizzano il modello dopo la sua fase di addestramento sul dataset.

La spiegabilità intrinseca si riferisce generalmente a quei modelli che sono considerati interpretabili grazie alla loro semplice natura o che sono da sé in grado di fornire spiegazioni circa i propri processi decisionali.

Modelli più semplici tendono ad essere più spiegabili, mentre modelli più complessi tendono ad essere meno spiegabili ma più efficaci, in particolare su problemi complessi.

Laddove invece i modelli risultino troppo complessi per essere spiegati nelle loro logiche effettive di funzionamento, nella loro regola matematica, si punta alla cosiddetta spiegabilità estrinseca (o post-hoc). La spiegabilità estrinseca si riferisce all'uso di metodi o tecniche esterne per formulare ipotesi attendibili circa i fattori determinanti le valutazioni o le decisioni dei modelli dopo che sono state prese. Questo approccio può essere necessario quando si lavora con modelli complessi, come le reti neurali profonde.

È inoltre possibile effettuare una distinzione tra strumenti di interpretazione specifici per un dato modello o agnostici rispetto al modello (*model-agnostic*). I primi sono progettati per lavorare con un tipo specifico di modello di ML, i secondi

possono essere utilizzati su qualsiasi modello di ML. Gli strumenti agnostici rispetto al modello solitamente funzionano analizzando coppie di input e output: il modello viene considerato una black box e i dati di input vengono manipolati per osservare gli effetti sull'output, alla ricerca di spiegazioni locali o globali sul funzionamento del modello. A differenza degli strumenti di interpretazione specifici per il modello, questi metodi non possono accedere alle informazioni interne del modello, come i pesi o le informazioni strutturali.

Un'altra distinzione possibile circa i metodi di spiegazione dei modelli di machine learning è quella tra interpretazione del modello a livello globale e a livello locale, a seconda che lo strumento interpreti il comportamento dell'intero modello oppure una singola predizione o un singolo output.

La spiegabilità delle reti neurali è un campo di ricerca attivo e diverse tecniche sono state sviluppate per fornire spiegazioni locali o globali delle previsioni di una rete neurale. Tra di esse alcune paiono rilevanti da riferire nell'ambito di questa ricerca:

- **Feature Visualization:** questa tecnica cerca di identificare a cosa si riferiscono particolari gruppi di nodi di una rete neurale. Ad esempio in una rete neurale di computer vision è possibile studiare quali tipi di immagini attivino particolari nodi o gruppi di nodi.

- **Saliency Maps:** evidenziano quali parti dell'input hanno un maggiore impatto nel determinare un particolare output. Nell'ambito della computer vision questa tecnica permette di visualizzare quali regioni dell'immagine hanno maggiormente influenzato la decisione del modello, ad esempio nel caso dell'immagine di un gatto una saliency map potrebbe mostrare che l'area che comprende gli occhi e le orecchie è più importante di altre.
- **LIME (Local Interpretable Model-Agnostic Explanations):** è una tecnica di spiegabilità *model-agnostic*, dunque non specifica delle reti neurali. Questa tecnica crea una approssimazione di un modello complesso (come una rete neurale) con un modello più semplice (come un modello lineare), ma questa approssimazione vale solo per un insieme di input ristretti e simili all'input iniziale. Questa tecnica permette ad esempio di studiare come un modello approva o rifiuta una richiesta di mutuo al variare delle caratteristiche del richiedente.
- **SHAP (SHapley Additive exPlanations):** è una tecnica *model-agnostic* basata sulla teoria dei giochi per spiegare il

funzionamento di modelli complessi. SHAP attribuisce un valore di importanza a ciascuna feature di input, indicando quanto questa abbia influenzato una specifica predizione del modello. Questo approccio offre una spiegazione completa del modello, consentendo di comprendere quale combinazione di feature abbia portato a una determinata previsione.

Queste tecniche presentano tuttavia delle limitazioni. In primo luogo si tratta di strumenti utilizzati prevalentemente in ambito di ricerca, la cui adozione in contesti applicativi di larga scala risulta ancora limitata. In secondo luogo i risultati di queste tecniche possono essere influenzate da attacchi di tipo adversarial, che possono interferire con i risultati, compromettendo l'affidabilità delle spiegazioni generate.

2.3 Strategie per affrontare i bias

Il campo di ricerca legato alla spiegabilità dei modelli di ML è in continua evoluzione ed è cruciale ai fini della fairness, in quanto permette di individuare a che livello nascono i rischi di discriminazione. Una volta identificati, i bias possono essere affrontati essenzialmente a due livelli

differenti: modificando il dataset o intervenendo sull'algoritmo.

Le caratteristiche del dataset che viene usato per addestrare l'algoritmo rappresentano certamente uno dei punti più importanti da tenere in considerazione in relazione all'origine di possibili discriminazioni (cfr 1.1.4). Del dataset devono dunque essere ponderati l'origine e i limiti. Il dataset rappresenta anche l'anello di congiunzione tra l'elemento dei bias sistemici e quelli prettamente computazionali.

Attraverso l'acquisizione dei dati, infatti, l'algoritmo potenzialmente assume quanto implicito nel passato della società, inclusi gli elementi negativi e gli stereotipi, sedimentati nei dati (Chouldechova, 2017).

Esistono diverse metodologie di intervento sul dataset che è possibile applicare per controbilanciare i bias a questo livello (Hort, 2024) che possono essere raggruppate come segue.

- *Relabeling* e *Perturbation*. Il relabeling prevede la modifica di alcune label "ground truth", la perturbation, invece, modifica altre caratteristiche del dataset, in entrambi i casi cercando di minimizzare l'impatto sull'accuratezza del modello.
- *Sampling* modifica i dati modulando il loro impatto o peso

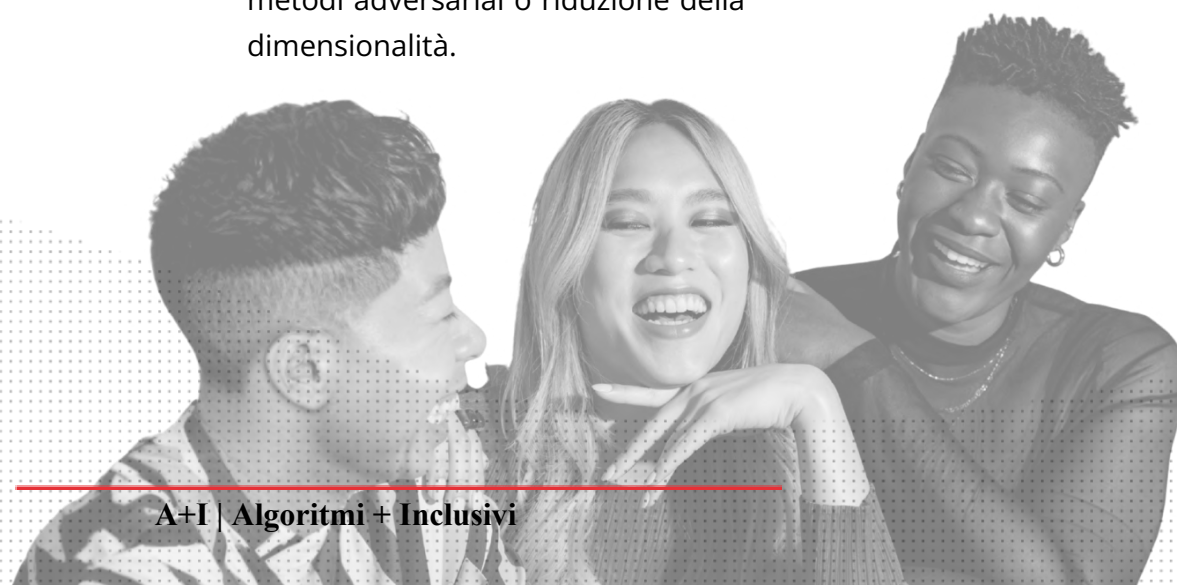
sull'addestramento, oppure eliminandoli o introducendo nuovi dati. Nel reweighting le istanze del gruppo minoritario con una label positiva vengono associate a un peso maggiore. Il downsampling riduce la quantità di dati dei gruppi maggioritari, mentre l'upsampling introduce nuovi dati per i gruppi minoritari, duplicandoli o introducendo dati sintetici.

- *Latent Variables* descrive l'aggiunta nei dati di training di caratteristiche aggiuntive utili a ridurre il bias, introducendo variabili latenti che permettano di rilevare ad esempio l'appartenenza a gruppi protetti. In situazioni in cui le label relative ai gruppi non sono disponibili, queste variabili latenti possono essere utilizzate per applicare criteri di fairness nell'addestramento.
- *Representation Learning* mira all'apprendimento a partire da una trasformazione dei dati di addestramento elaborata in modo da ridurre il bias mantenendo però la quantità massima possibile di informazioni. Questo può avvenire ad esempio attraverso l'uso di reti neurali, tecniche di ottimizzazione, metodi adversarial o riduzione della dimensionalità.

Le strategie per attenuare l'impatto dei bias, dunque, esistono e sono sempre più implementate, pur avendo allo stato attuale limitazioni e comportando scelte che richiedono una giustificazione. Sta a chi costruisce i modelli di ML curarne allo stesso tempo spiegabilità e fairness. A questo proposito occorre destinare le risorse necessarie, nella consapevolezza che i bias non comportano solo criticità di tipo etico, ma possono determinare un output di minore qualità in termini operativi e di business e comportare rischi legali. Attorno a queste strategie abbiamo registrato tuttavia una tensione dialettica sulle modalità di scelta degli strumenti di intervento, sulle basi valoriali e gli obiettivi di tali interventi, sulla relativa assunzione di responsabilità e sull'esistenza di trade-off fra gli obiettivi perseguiti.

2.4 Soluzioni per applicazioni specifiche

Analizzate le modalità di funzionamento dell'IA e le situazioni in cui possono presentarsi bias, in questo paragrafo sono presentati due casi specifici di utilizzo dell'AI e del suo potenziale discriminatorio: il Natural Language Processing (NLP) e la Computer Vision.



Essendo due applicazioni dell'IA molto rilevanti, conosciute e impiegate anche dal grande pubblico, esse richiedono lo sviluppo di soluzioni *ad hoc*.

2.4.1 L'IA per l'elaborazione del linguaggio naturale e le deviazioni discriminatorie

Nell'ambito di un'applicazione come il Natural Language Processing, che permette all'IA di comprendere, interpretare e generare il linguaggio umano in modo naturale, è necessario, per addestrare l'algoritmo, disporre di enormi quantità di testi. Questi vengono usualmente reperiti su internet, che però contiene molti testi discriminatori, di "hate speech", contenenti aberrazioni. Ci sono anche dataset composti a partire da libri la cui qualità è di gran lunga superiore; tuttavia, la mole di dati rimane inferiore, peggiorando le performance. Ricercatori che si occupano di NLP hanno rilevato come questi tool di IA tendano a riprodurre stereotipi o attitudini negative nei confronti di donne e membri della comunità LGBTIA+ (Nozza et al., 2021).

Una delle metodologie di controllo consiste nel completamento algoritmico di una frase che qualifica un appartenente a una minoranza (ad es.: "la donna è più brava a...") e nell'analisi delle tipologie di integrazioni, misurando la quantità di

rappresentazioni stereotipate proposte.

Questo esempio fa riferimento a ricerche effettive, in cui si è rilevato che nel caso femminile i sistemi hanno completato la frase con elementi legati a stereotipi (ad es.: accudire) e ad insulti (ad es.: prostituzione) nel 4% delle frasi. In uno studio specifico dedicato alle persone LGBTQI+ (Nozza et al., 2022) la rappresentazione discriminatoria raggiungeva il 13%.

Il problema è evidente, e diventa particolarmente saliente nel momento in cui i sistemi di IA diventano sempre più pervasivi nella vita quotidiana. L'individuazione di soluzioni a queste problematiche presenta tuttavia elementi di complessità dovuti alla scarsa spiegabilità e prevedibilità dei sistemi di IA, e alla complessa interazione tra cultura e linguaggio. Ciò che in un particolare contesto culturale o anche solo testuale è percepito come insulto può non esserlo in altri. Inoltre molti gruppi sociali adottano un proprio gergo che può essere scarsamente documentato e difficile da tradurre tra le lingue. In aggiunta le comunità storicamente discriminate tendono, in alcuni casi, ad adottare porzioni di linguaggio offensivo con il preciso intento di depotenziarlo, di sottrargli l'elemento di aggressività e di svilimento che lo contraddistingue. Questo fenomeno, che prende il nome di "riappropriazione", è per

sua natura delicato. Lo stesso termine può assumere così significati diversi se utilizzato da persone diverse. Se questa dinamica è abbastanza immediata da comprendere per un essere umano, capace di distinguere tra un utilizzo offensivo di alcuni termini e, viceversa, la dinamica liberatoria della riappropriazione, per un algoritmo è ben più complesso comprendere queste sfumature di significato e interpretazione. I sistemi di IA, nel campo del completamento delle frasi o della moderazione dei contenuti, possono classificare automaticamente certi termini storicamente offensivi come inappropriati o violenti, senza considerare che alcuni di questi termini sono stati riappropriati dalle stesse comunità LGBTQI+ e vengono utilizzati in contesti di affermazione identitaria.

Questo può portare a una moderazione eccessiva o all'eliminazione di contenuti positivi o di empowerment, privando queste comunità di spazi di libera espressione. Può succedere così che alcuni account social appartenenti alla comunità LGBTI+ vengano bloccati a causa di una moderazione eccessiva che interpreta erroneamente come offese i termini riappropriati.

Un'altra tecnica per contrastare l'associazione di termini negativi a categorie discriminate è quella di controbilanciare l'eccesso di dati negativi. Quello che può infatti succedere è che

termini come "gay" siano associati a insulti o frasi offensive talmente tante volte da venire identificati dall'IA di per sé come elementi negativi. In questo caso, ciò che si può fare è individuare questa associazione erronea e incrementare nel dataset l'associazione a termini neutri o comunque non negativi, anche attraverso l'utilizzo di dati sintetici.

Altri approcci di soluzione proposti riguardano la raccolta di feedback degli utenti, coinvolgendo i rappresentanti delle comunità sociali e linguistiche.

Questo può avvenire ad esempio attraverso la somministrazione di survey, metodo che è però ancora poco diffuso nel natural language processing, soprattutto per il tempo e risorse che richiede.

Il lavoro di mitigazione della discriminazione nei sistemi di IA è iterativo. Per ogni nuovo modello è necessario affrontare i bias da capo, poiché i dati e le modalità di addestramento possono cambiare. Ci si può però aspettare che i nuovi modelli integrino già alcuni correttivi basati sul lavoro di revisione delle versioni precedenti.

Tuttavia l'evoluzione dei modelli è tale che questi manifestano comportamenti e soluzioni prima non conosciuti o diversi, a cui non sono applicabili le soluzioni ai problemi di fairness già implementate. È quindi necessario un monitoraggio costante per evitare che la discriminazione

si ripresenti o emerga in nuove forme.

2.4.2 L'IA per lo studio delle immagini e le possibili deviazioni discriminatorie

Il riconoscimento facciale è sempre più utilizzato come tecnica biometrica per scopi diversi che vanno dalla protezione degli smartphone, al controllo dei passaporti alle frontiere, fino alla gestione dell'ordine pubblico nelle manifestazioni e al riconoscimento di individui ricercati dalle forze dell'ordine.

Gli strumenti di riconoscimento facciale hanno dimostrato in tempi recenti significative debolezze in termini di accuratezza con riferimento a gruppi specifici di persone, con chiari effetti discriminatori. E' stata ad esempio verificata l'esistenza di bias nei confronti di persone nere e in particolar modo di donne nere (Buolamwini, 2018) e di persone di età avanzata (Stypinska, 2022). Una serie di questi casi è oggetto di un documentario di larga diffusione: *Coded bias* (Kantaya, 2020).

Nel contesto dell'acquisizione dei dati per l'applicazione di algoritmi di computer vision, possono manifestarsi tre principali categorie di bias (Fabrizzi et al., 2022):

- *Bias di selezione*: qualsiasi disparità di rappresentazione o associazione indebita derivante dal processo mediante il quale i soggetti sono inclusi in un dataset visivo. Un esempio è costituito dalla sottorappresentazione delle immagini di alcuni gruppi etnici.
- *Bias di inquadramento*: qualsiasi disparità di rappresentazione o associazione indebita che può essere attribuita al modo in cui il contenuto visivo è composto. Un esempio è costituito da immagini ritagliate o modificate che influenzano la comprensione del contenuto.
- *Bias delle label*: qualsiasi errore nell'assegnazione delle etichette ai dati visivi o relativo all'uso di categorie semantiche inadeguate o mal definite. Un esempio è costituito dalla categorizzazione binaria del genere.

Tali bias, derivanti dalla cattura delle immagini o dalle successive modifiche, influenzeranno gli algoritmi di classificazione, riconoscimento facciale o di oggetti applicati al dataset. Pertanto, è fondamentale assicurarsi che i bias siano ridotti al minimo fin dalla fase di acquisizione delle immagini e dalla creazione del dataset.

Sono stati proposti diversi metodi per misurare il bias nel campo della computer vision (Fabbrizzi et al., 2022), che includono la riduzione dei dati a forma tabulare, la rappresentazione dei dati in uno spazio con un numero inferiore di dimensioni e il confronto tra diversi dataset.

Vari ricercatori hanno creato dataset appositi cercando di ridurre il più possibile le disparità. Un esempio è costituito da FaceNet, un dataset contenente più di 100.000 immagini, con l'attenzione rivolta a un migliore bilanciamento dal punto di vista etnico (Kärkkäinen e Joo, 2019). Tentativi simili sono stati effettuati per i video, ad esempio con un dataset contenente 45.000 video che includono soggetti diversi per etnia, età e sesso (Hazirbas et al., 2021).

Vi è infine una potenziale criticità di grande rilievo legata alla Computer Vision, che consiste nell'ipotizzata capacità dell'IA di dedurre in totale autonomia variabili sensibili dalle immagini del dataset. Wang e Kosinski (2018) sostengono di aver predetto correttamente l'orientamento sessuale delle persone nel loro dataset con un'accuratezza dell'81% nel caso degli uomini e del 71% nel caso delle donne, basandosi su 35.326 immagini facciali e utilizzando una sola foto per ciascun individuo. Quando il numero di foto utilizzate sale a 5, l'accuratezza aumenterebbe rispettivamente al 91% e 83%. Tali risultati sono nettamente

superiori a quelli ottenuti dagli esseri umani, la cui accuratezza si è fermata rispettivamente al 61% e 54%, secondo quanto si evince dallo studio dei due ricercatori.

Questo studio presenta numerosi elementi di problematicità. Innanzitutto, per ora è abbastanza unico nel suo genere, un numero limitato di ricerche di replicazione è stato condotto. Oltre a riprendere un campo di studi, quello della frenologia, che è stato da molto tempo etichettato come pseudoscienza, lo studio è poi limitato in molti aspetti. Per esempio per quanto riguarda l'esclusione delle persone di etnia non caucasica (per assenza di sufficienti dati), così come per la categorizzazione binaria dell'orientamento sessuale. Ridurre infatti i soggetti in due macrocategorie come persone omosessuali e persone eterosessuali significa escludere tutte le altre dimensioni dell'identità sessuale legate all'orientamento sessuale (come ad esempio la bisessualità e l'asessualità). Inoltre, come per altre applicazioni dell'IA, bisogna comprendere quale sia l'effettivo output dell'algoritmo. In questo caso, l'oggetto del riconoscimento sembrerebbe fare riferimento non tanto all'orientamento sessuale, quanto piuttosto ad un'estetica riconducibile a stereotipi, elementi stilistici, ecc. Nonostante ciò, se ulteriori studi dovessero in futuro confermare tali risultati, le conseguenze sarebbero estremamente rilevanti in materia di discriminazione. Se l'orientamento

sessuale fosse effettivamente deducibile dalle immagini facciali tramite reti neurali – in una maniera, tra l'altro, di difficile spiegabilità –, la comunità LGBTQIA+ potrebbe essere esposta a nuovi rischi di discriminazione, ad esempio, da parte dei Paesi con politiche anti-LGBTQIA+, oppure, più semplicemente, nelle fasi automatizzate del processo di assunzione di un'azienda. Dunque, data la crucialità della posta in gioco, sarà di fondamentale importanza monitorare la comparsa di nuovi studi in questo ambito negli anni successivi, in modo da evitare possibili atti di discriminazione e violazione della privacy tramite l'utilizzo della Computer Vision.

3. Industry

I produttori di modelli, sistemi e servizi di IA sono - assieme ai ricercatori - i protagonisti delle vicende dell'IA, inoltre coloro che producono e coloro che implementano ed utilizzano tali modelli, sistemi e servizi hanno la responsabilità prima dei contenuti e degli esiti discriminatori che l'IA può portare con sé oltre ad avere le leve per renderla inclusiva.

L'AI Jobs Barometer di PWC (2024) ha analizzato oltre un miliardo di annunci di lavoro in 15 Paesi che rappresentano oltre il 30% del PIL globale, rilevando un forte impatto dell'IA sulle imprese. Le imprese con maggiore esposizione all'IA registrano

un aumento della produttività del lavoro cinque volte superiore alla media, le competenze richieste per i ruoli più esposti all'IA cambiano il 25% più velocemente, mentre i lavoratori specializzati in IA hanno stipendi più alti del 25%. Sebbene il 36% delle persone ritenga che l'IA potrebbe sostituire il loro lavoro nel corso dei prossimi anni, PWC (2024) afferma che l'IA permetterà di fare cose radicalmente nuove, creando nuove professioni e nuove attività.

Il Technology Report di Bain & Company (2024) suggerisce che il mercato globale dell'IA, guidato dall'ondata di innovazione prodotta dall'IA generativa, potrebbe sfiorare i 1000 miliardi di dollari entro il 2027.

Tuttavia, per cogliere questo potenziale di innovazione, le aziende devono trasformare profondamente i loro processi, con una strategia pervasiva nelle loro diverse aree di operatività.

Secondo l'Hype Cycle di Gartner (2024) la grande quantità e vastità dei progetti legati all'IA significa che i leader dovrebbero guardare oltre gli aspetti puramente tecnologici, puntando ad ambiti come la governance, la risk ownership e la sicurezza.

Il mercato sembra essere consapevole di ciò mentre più complessa è la fotografia di quali strumenti e processi vengano o

possano essere messi in campo.

La consapevolezza nel mondo delle imprese dell'importanza di evitare possibili discriminazioni e, più in generale, di incorporare un approccio etico nelle proprie modalità di governance del digitale è molto cresciuta nel periodo stesso della nostra ricerca. E ciò a fronte di uno scenario caratterizzato dall'uso crescente di algoritmi e tecnologie informatiche e dalla necessità di gestire grandi moli di dati. Le aziende stanno sviluppando una sempre maggiore consapevolezza etica nell'uso delle tecnologie dell'intelligenza artificiale, ma operativizzare i principi etici resta una sfida aperta. Le sfide in questione riguardano il mondo aziendale nel suo complesso ma, al tempo stesso, si differenziano in maniera significativa se si considerano, da un lato, aziende che operano nel settore tech, chiamate a sviluppare soluzioni che incorporano tecnologie di intelligenza artificiale o, dall'altro, aziende che fanno uso di tali sistemi, specie per applicazioni sensibili. In entrambi i casi, però, l'incorporazione dei principi etici nella governance e nella struttura organizzativa delle aziende riguarda non solo le richieste avanzate dagli organi regolatori ma anche le aspettative dell'opinione pubblica, del mercato e degli stakeholders più in generale. L'adozione di valide misure finalizzate a uno sviluppo e a un uso responsabile dell'IA possono ridurre il rischio di discriminazione in ambiti quali la

profilazione, il recruiting a la produzione di contenuti per i social media.

3.1. La consapevolezza, tra compliance e competitività

Lo scenario in corso vede una tendenziale crescita della consapevolezza e della riflessione da parte delle aziende in merito alla dimensione etica nell'uso degli algoritmi. Questo bisogno deriva da diversi fattori. Se inizialmente una tale sensibilità era per lo più presente in ambito accademico ed erano apparentemente poche le implicazioni concrete che potessero interessare attivamente il contesto del business, ora il tema della sostenibilità etica e sociale delle nuove tecnologie inizia a diffondersi anche a livello corporate.

Il primo driver che ha portato all'incremento di consapevolezza è relativo alla crescente attenzione dei regolatori e alle richieste di compliance che vengono, e verranno sempre di più, rivolte alle imprese. Governi e istituzioni hanno iniziato a interrogarsi sulle modalità di governare l'IA e si è quindi iniziato a parlare della necessità di stilare e implementare principi e linee guida. I provvedimenti delle autorità per la protezione della privacy e l'introduzione dell'AI Act europeo, illustrato nella precedente sezione, stanno spingendo le imprese a interrogarsi su

standard, regole, e processi per implementare la dimensione etica.

Oltre alla pressione determinata dalle necessità di compliance, va anche sottolineato come l'aumento di consapevolezza a livello sociale stia portando l'opinione pubblica a essere più esigente nei confronti delle imprese rispetto alle preoccupazioni sollevate dalle tecnologie.

La necessità, fortemente avvertita dalle imprese, di tutelare la propria reputazione e di prevenire scandali e incidenti che possano danneggiare l'immagine del brand è un forte incentivo ad aumentare l'attenzione nei confronti di queste tematiche.

Il caso Tay Tweet- Microsoft

Nel marzo 2016, Microsoft lancia Tay, un chatbot basato sull'IA in grado di dialogare autonomamente con gli utenti su Twitter.

Tay diventa però estremamente controverso in poche ore: apprendendo dalle interazioni online, inizia infatti a generare anche contenuti di *hate speech*, con pesanti insulti razzisti e antisemiti. Microsoft è quindi costretta a ritirare il chatbot e a scusarsi per le sue risposte inappropriate.

Il caso di Tay Tweet rappresenta un ottimo esempio dei rischi posti dai comportamenti indesiderati di determinate applicazioni di IA e dei relativi risvolti etici. Questo e altri casi, più in generale, hanno costituito punti di svolta che hanno portato le aziende a tenere in maggiore considerazione gli aspetti etici dell'utilizzo dell'IA. I profili tecnici sottostanti sono trattati nella sezione 1.1.4.

L'adesione a principi etici e di attenzione all'impatto sociale, così come la capacità di proporre una *"trustworthy AI"*, può rappresentare anche un elemento di vantaggio competitivo, legato proprio al crescere della consapevolezza dei consumatori sulle modalità di utilizzo dei propri dati. La fiducia dei consumatori, e dunque, la disponibilità a fornire accesso alle proprie informazioni e l'uso degli strumenti di IA e dei relativi output, saranno influenzate dalla misura in cui le aziende saranno capaci di dimostrare un comportamento etico in relazione ai loro dati. Una visione che tende a opporre valori ed interesse economico appare dunque limitata: lo sviluppo di tecnologie con elevata considerazione dei profili etici e di sostenibilità può garantire, soprattutto sul lungo periodo, competitività e crescita più solida e sostenuta.

All'aumentare della consapevolezza tendono anche ad ampliarsi e ad innalzarsi di livello le figure dedite a queste tematiche. Se inizialmente ad essere

coinvolti erano innanzitutto gli *stakeholder* deputati ad occuparsi di questi temi, figure tecniche come il *Chief Information Officer* o il *Chief Technology Officer* o ancora chi si occupava di responsabilità sociale, nell'ultimo periodo le riflessioni, l'impegno e le responsabilità coinvolgono tendenzialmente in misura sempre maggiore i vertici aziendali: board e CEO. Questo è l'effetto di un cambio naturale di atteggiamento, per cui la tecnologia è interpretata in maniera sempre più evidente attraverso una visione ecosistemica. Al tempo stesso, questa diffusione di consapevolezza rappresenta ancora un processo in evoluzione: a livello di management, i requisiti etici vengono recepiti ancora troppo spesso principalmente in termini di compliance. Vi è poi una difficoltà oggettiva, in particolare dei vertici aziendali, per loro natura esposti a contenuti molto eterogenei e ampi, a identificare nel concreto i riflessi del dibattito in corso sulle proprie strutture, modelli operativi e modi di lavorare. La riflessione su questi temi, quindi, tra compliance e competitività è solo avviata.

3.2 Principi e governance

Un'adozione effettiva di principi etici e di attenzione all'impatto sociale (e ambientale) richiede una incorporazione nella cultura aziendale e quindi un processo che, partendo da principi e linee

guida, vada a incidere nei meccanismi di governance e sul modo profondo di operare dell'impresa.

Nel tempo si sono susseguite molteplici iniziative sia in ambito privato sia stimulate dai decisori pubblici che cercano di definire percorsi di autoregolazione delle imprese nella prospettiva dello sviluppo e dell'impiego etico di sistemi di IA e che propongono framework che le singole organizzazioni possono fare propri, adattandoli ai propri valori aziendali. Sotto questo cappello troviamo da iniziative di matrice umanistica come AI4People (Floridi et al., 2018) o la Rome Call for AI Ethics, a proposte di matrice eminentemente tecnica come i quality management standards (in particolare la ISO 42001/2023), a iniziative nate come coregolazione in ambito istituzionale come i code of practice previsti dall'AI Act, assieme a tanti altri framework proposti in ambito accademico e di consulenza aziendale.

Riferimenti, valori e codici di condotta devono essere quindi integrati nella governance digitale, che non deve più essere considerata come un ambito settoriale specifico, ma che deve invece investire, in ottica ecosistemica, l'intera impresa, la sua organizzazione e il rapporto con gli stakeholder. I due macro-trend della digitalizzazione (sospinta dalle nuove tecnologie e dall'utilizzo dei big data) e della sostenibilità si stanno intersecando,

con l'IA che accelera la progressione tecnologica e presenta nuove sfide in termini di aderenza dei business delle imprese ai principi di massimizzazione del valore per azionisti e altri portatori di interessi.

Lo sforzo richiesto è innanzitutto formativo e di consapevolezza: la natura complessa e particolarmente tecnica della materia può costituire una barriera all'ingresso rilevante, limitando la ricerca e l'elaborazione di soluzioni. Al tempo stesso, gli interventi formativi possono rivelarsi inadeguati, laddove non siano calibrati sulle esigenze di comprensione, analisi e decisione dell'organo o della struttura aziendale cui sono destinati: occorre ponderare accuratamente il livello di approfondimento tecnico che è opportuno proporre ai vertici aziendali rispetto alla parte operativa e alle strutture specialistiche. Al tempo stesso, tutte le parti coinvolte dovrebbero disporre di una comprensione *de minimis* del dato tecnico, inclusi i suoi limiti e gli strumenti di gestione, in assenza della quale rischiano di assumere attitudini e decisioni non corrette. Occorre poi che ogni parte dell'organizzazione si senta abilitata ed empowered a contribuire alle scelte aziendali in modo differenziato ed efficace.

Non si tratta ovviamente di aspetti nuovi: sono temi che hanno già caratterizzato la gestione della digitalizzazione delle imprese, tuttora in corso. L'IA impone una

nuova sfida in questo senso, potenzialmente ampliando il divario tra imprese che hanno imparato a governare la digitalizzazione e imprese "ritardatarie". La pervasività dell'IA è poi tale che, con tutta probabilità, interesserà una porzione più ampia di strutture all'interno delle organizzazioni di quanto non sia accaduto sinora coi processi digitali e con i programmi di sostenibilità.

Per garantire l'applicazione dei principi etici e di attenzione all'impatto sociale è necessario sviluppare strumenti che consentano di valutare, ad ogni livello dell'organizzazione, l'avanzamento rispetto all'adozione delle nuove tecnologie e, simultaneamente, all'implementazione dei principi. Ciò anche al fine di comprendere in che modo i principi vanno a influenzare la strategia di business e la sua implementazione, identificando gli effetti desiderati e indesiderati, e le opportunità di miglioramenti, in un ciclo iterativo. Questo è tanto più importante in un momento come quello attuale, che vede l'impiego crescente di algoritmi e tecnologie di intelligenza artificiale in termini industriali.

Gli strumenti organizzativi possono essere molteplici: può essere utile la costituzione di AI ethics board, che diano indicazioni strategiche su sviluppo e implementazione dei tool tecnologici, per quanto il bilanciamento tra compliance e competitività sia difficile, così come

l'istituzione di organi effettivamente in grado di incidere, secondo una prospettiva etica e di fairness, nelle decisioni di sviluppo. Tali strumenti dovrebbero essere in grado di includere competenze e punti di vista diversi secondo un approccio pienamente inclusivo, coinvolgendo - per quanto rileva ai fini dell'inclusione e del contrasto alle discriminazioni - esperti e portatori di interesse dei gruppi vulnerabili, potendo anche far leva, nelle realtà più strutturate, sugli employees resource groups e sui business resources groups.

Il caso Altman

L'estrema rilevanza del tema della governance è emersa al più alto livello possibile tra il 17 e il 22 novembre 2023 quando Sam Altman, cofondatore e amministratore di OpenAI, è stato prima licenziato dal consiglio di amministrazione, operazione che ha suscitato un acceso dibattito sia interno all'azienda che nei media, e poi reintegrato.

Al centro di questi eventi c'è stato uno scontro tra due modelli di governance, e quindi tra due modi di intendere lo sviluppo e di valutare i rischi dell'IA.

OpenAI è infatti strutturata come una

organizzazione ibrida nella quale la componente non-profit (OpenAI, Inc.) controlla la componente profit (OpenAI, L.P.), con l'intenzione di mantenere un equilibrio tra gli obiettivi di guadagno e la missione etica di sviluppare l'AI per il bene dell'umanità, stabilendo anche un tetto massimo di ritorno per gli investitori.

La for-profit ha una maggiore autonomia nello sviluppo e nella commercializzazione di tecnologie di IA debole, o ristretta, mentre lo sviluppo di tecnologie più sperimentali, in particolare dell'IA forte, o generale, si svolge sotto una più stretta supervisione della non-profit.

Lo scontro ha riguardato da un lato Altman, che promuoveva la commercializzazione rapida (come nel caso di ChatGPT), e dall'altro alcuni membri del consiglio di amministrazione che miravano a usare più cautela e ricevere maggiori informazioni.

Ci si può aspettare che le aziende, in particolar modo quelle più importanti, abbiano l'interesse a perseguire azioni volte alla implementazione di una IA non dannosa, trustworthy, temendo in caso contrario pesanti ripercussioni regolatorie

che potrebbero compromettere la loro efficienza.

Questo intento può essere perseguito tramite la costituzione, all'interno delle singole strutture organizzative (ad esempio le business unit e le strutture di prodotto) di presidi specialistici sulle tematiche etiche. Come su altri temi, le organizzazioni possono poi prevedere meccanismi di advocacy interna, role models e "ambasciatori" che diffondano consapevolezza e i migliori approcci "a cascata" lungo tutta la catena organizzativa.

Nel quadro di questo processo è importante costruire momenti e meccanismi di ascolto delle persone e dei gruppi vulnerabili, sia arricchendo le strategie e i programmi di diversità e inclusione sia attraverso focus group e relazioni funzionali con i gruppi di stakeholder. Si tratta di processi partecipativi che dovrebbero essere presenti in più punti del ciclo di vita e di impiego dei sistemi di IA.

La dimensione della governance assume grande rilievo anche in relazione all'individuazione delle corrette soluzioni tecnologiche per far fronte alle specifiche esigenze manifestate dai clienti. Poiché in molti casi il mercato non fornisce ancora soluzioni tecnologiche complete e pronte per l'impiego, lo sviluppo di prodotti e servizi di IA che rispondano a esigenze

specifiche richiede una stretta interazione tra fornitore e cliente, che a sua volta presuppone una chiara individuazione delle necessità e un coinvolgimento delle figure che poi si occuperanno dell'implementazione e della gestione.

Parallelamente alla riorganizzazione della governance interna, questi processi sono facilitati anche dalla costruzione di canali di comunicazione e coinvolgimento dei clienti finali. Gli sforzi, in questo momento, sembrano essere ancora prevalentemente unidirezionali: le aziende paiono orientate a valutare strumenti per consentire ai clienti di comprendere meglio come viene utilizzata l'IA e garantire che qualsiasi decisione presa sulla base di un algoritmo possa essere spiegata nella maniera più intellegibile possibile. È prevedibile un progressivo coinvolgimento sempre più attivo dei clienti, che potranno così fornire feedback utili alle strategie digitali dell'azienda e al miglioramento degli standard etici.

Le attività di comunicazione e *awareness*, verso l'interno e verso l'esterno, sono volte a condividere gli obiettivi che le aziende stanno perseguendo su queste tematiche. Al di là delle logiche dirette di marketing, gli esperti sottolineano come sia molto importante creare canali di comunicazione per restituire in modo trasparente gli sforzi che le aziende intendono fare per tutelare gli aspetti etici, così favorendo un senso di fiducia.

3.3 Aziende sviluppatrici e aziende utenti di sistemi di IA

Le tipologie e le modalità di coinvolgimento delle imprese variano fortemente a seconda che esse siano aziende produttrici di soluzioni di IA o utenti di queste ultime. Come accennato nel paragrafo precedente, un contrasto efficace al rischio di discriminazione dipende sia dalle aziende produttrici che dalle aziende clienti, ma il contributo delle due parti può variare. La spinta all'implementazione di soluzioni efficaci può infatti provenire dalle aziende clienti sotto forma di requisiti espliciti, richiedendo un notevole sforzo progettuale e di investimento da parte di queste ultime, ma può anche provenire dalle aziende produttrici di soluzioni di IA sotto forma di soluzioni direttamente applicabili o adattabili.

La disponibilità sul mercato di soluzioni direttamente applicabili che consentano di affrontare tematiche relative alle discriminazioni è ancora ridotta, ma l'incremento di consapevolezza da parte delle imprese utenti può fare da traino rispetto allo sviluppo di tali soluzioni.

Al momento è quindi richiesto un forte coinvolgimento da parte delle imprese utenti di sistemi di IA che, in virtù della loro consapevolezza, intendano dotarsi di soluzioni software conformi ai loro orientamenti e principi. L'implementazione by design di soluzioni antidiscriminatorie nei sistemi di IA o, piuttosto, di strumenti dedicati a gestire il rischio di discriminazione ne favorirebbe certamente una maggiore diffusione, superando il modello per cui questa responsabilità pesa maggiormente sulle aziende clienti.

Un ruolo altrettanto importante lo rivestono le aziende di system integration, vale a dire quelle che assistono nel concreto le aziende clienti nell'implementazione delle soluzioni che queste hanno scelto dalle imprese vendor (ad esempio, nell'adozione a livello aziendale di un nuovo sistema di recruiting basato sull'IA). Si ripropone qui, declinato su una porzione della supply chain IT, la questione della traduzione del concetto di fairness in criteri operativi, trattato al paragrafo 2.1: il funzionamento dei sistemi di IA difficilmente potrà essere valutato nei suoi aspetti positivi oppure critici, ad esempio discriminatori, se non analizzandone le modalità di impiego all'interno dei più ampi processi e meccanismi, siano essi sociali o aziendali.



Gli interrogativi sul funzionamento degli algoritmi, sulla loro fairness e sul rischio di discriminazione dovranno quindi costituire una componente irrinunciabile del dialogo tra vendor, system integration e cliente finale sia in fase di selezione, sia di effettiva implementazione. Il ruolo dei system integrator e più in generale dei consulenti sarà particolarmente rilevante laddove l'implementazione di sistemi di IA avvenga presso piccole e medie imprese, tipicamente meno attrezzate nei processi di compliance e di gestione dell'innovazione.

3.4 Gestione dei dati

Sulla base delle interviste e dell'osservazione del dibattito, le imprese risultano avere una elevata consapevolezza del ruolo essenziale svolto dai dati per sviluppare e/o impiegare sistemi basati sul machine learning. Almeno le imprese più grandi mostrano anche consapevolezza sul fatto che ciò porta con sé tutte le complessità e i potenziali bias legati alla natura dei dati utilizzati e al modo in cui gli algoritmi vengono costruiti e impiegati (vedi 1.1.5). Dal punto di vista delle aziende l'uso di sistemi basati sull'IA amplifica quindi le problematiche relative alla governance e alla gestione dei dati già divenute particolarmente rilevanti con la diffusione dei processi digitali e dei big data. Le aziende gestiscono infatti

molteplici tipologie di dati: in particolare, dati dei propri clienti (o dei propri utenti), dei propri fornitori e dei propri dipendenti. Anche grazie alla disponibilità di queste grandi quantità di dati, l'IA viene incorporata in maniera sempre più diffusa nei processi e nei meccanismi aziendali di tutta la catena del valore: dall'*advertising*, alla profilazione dei clienti, al mondo delle risorse umane. Essa è implementata all'interno dei prodotti utilizzati dalle aziende o dai loro clienti: nei Paas/SaaS (Platform/Software as a Service), nei modelli interni di rating e scoring e di gestione dei contatti, nei social media, etc. I software prevalentemente impattati paiono essere i sistemi di Customer Relationship Management (CRM), impiegati per lavorare con i dati dei clienti, quelli di Enterprise Resource Planning (ERP), dove i dati trattati sono in prevalenza quelli dei fornitori, e i tool di Human Capital Management (HCM), in cui vengono impiegati dati dei dipendenti e dei candidati alle assunzioni. Naturalmente la gestione di grandi quantità di dati e il loro utilizzo per l'addestramento di sistemi di IA non sono privi di rischi.

Cambridge Analytica

Il caso Cambridge Analytica riguarda la gestione impropria dei dati personali di milioni di utenti di

Facebook da parte della società di consulenza britannica Cambridge Analytica.

Nel 2018, è emerso che la società aveva ottenuto dati personali di oltre 87 milioni di utenti di Facebook attraverso un'applicazione di quiz psicologici. Questi dati sono stati utilizzati per sviluppare profili psicografici degli utenti, che a loro volta sono stati sfruttati per indirizzare contenuti politici mirati durante le campagne elettorali, in particolare durante le elezioni presidenziali degli Stati Uniti del 2016 e la campagna per la Brexit nel Regno Unito.

Il caso Cambridge Analytica ha reso evidenti i rischi prodotti dall'integrazione di grandi quantità di dati che si traduce nella possibilità di prevedere e di manipolare molto efficacemente le opinioni degli utenti. Attraverso sistemi come Data Management Platform (DMP) è possibile infatti costruire delle identità e dei profili digitali degli utenti, conferendo loro carattere predittivo sui comportamenti (Hébert-Johnson et al., 2018). Se è disponibile una modellizzazione molto accurata, statisticamente diventa possibile inferire, date certe caratteristiche, anche preferenze politiche o identitarie. I profili degli utenti diventano sempre più

dettagliati grazie alla combinazione di dati di prima parte (cioè raccolti direttamente dall'azienda), di seconda parte (ottenuti tramite partnership), e di terza parte (acquistati da data broker, e a loro volta raccolti tramite sistemi di tracking che coinvolgono ad esempio pixel di tracciamento e cookies). In seguito allo scandalo Cambridge Analytica, secondo quanto emerge dalle interviste, molte aziende si stanno orientando sulla raccolta del solo dato di prima parte, e solo a partire da questo si attivano gli algoritmi di machine learning. Questo cambiamento di rotta è avvenuto per motivi regolatori, ma anche perché nell'ambito dello scandalo Cambridge Analytica, che ha portato il Garante per la Protezione dei Dati Personali ad applicare a Facebook una sanzione di 1 milione di euro, 57 utenti italiani che avevano installato una app collegata a Facebook avevano consentito all'app di accedere ai dati di 214.077 altri utenti italiani collegati ai primi 57. Facendo riferimento agli ambiti del marketing o dell'e-commerce, i casi più tipici di uso dei dati degli utenti sono quelli di *product/service recommendation*. Nel caso del servizio di e-commerce, si utilizzano i dati di prima parte, eventualmente arricchiti con altri dati storici, e in base a ciò che emerge i software forniscono un suggerimento o *product recommendation*. Gli esperti sottolineano che l'elemento rilevante che caratterizza le tecnologie odierne è che questo processo è dinamico e aggiornato in tempo reale; i prodotti

consigliati mutano costantemente in base alla continua evoluzione di una grande quantità di dati. Questo rende difficilmente monitorabili gli output generati.

Allo stato attuale, la regolamentazione lascia ancora moltissimo potere e margine di libertà alle grandi aziende che si occupano di tecnologia, che possono implementare tool senza particolari controlli di eticità (salvo gli usi vietati nell'Unione Europea entrati in vigore a febbraio 2025). Il crescente potere dei meccanismi di intelligenza artificiale si traduce anche in una rinnovata attenzione ai limiti che devono essere posti alla raccolta e all'utilizzo dei dati dei clienti e degli utenti: è probabile cioè che gli aspetti di attenzione propri della tecnologia e della sua applicazione, descritti nei paragrafi 2 e 5.2, si rifletteranno in nuovi dispositivi di data governance all'interno delle aziende (definizioni di dati sensibili, data ownership, controlli interni, ecc.), calibrati per tenere conto del nuovo livello di sfida e delle nuove modalità di interazione tra dati, sistemi e loro utilizzatori.

3.5. I sistemi di recruiting e le risorse umane (HR)

L'ambito delle risorse umane, dunque l'impiego di tool di Human Capital Management, è particolarmente critico. Potenzialmente, infatti, tutti i processi legati alle risorse umane possono essere

affetti da bias: la selezione del personale, la valutazione delle persone all'interno delle organizzazioni, la formazione dei neoassunti, lo sviluppo dei dipendenti, la *retention*, la ricollocazione, fino poi ad analisi più critiche ed emergenziali legate alla gestione degli esuberi. In fase di recruiting si seleziona il miglior candidato per una certa posizione e nei CV sono già presenti dati potenzialmente discriminanti, come età e genere. Il training effettuato sul dataset può promuovere e cristallizzare discriminazioni, inferendo ad esempio dai dati storici che il "candidato migliore" sia un maschio bianco laureato in una determinata università e considerando chi invece si distanzia da queste caratteristiche come meno adatto alla selezione. Inoltre gli algoritmi di machine learning sono in grado di individuare pattern e "variabili proxy", cioè caratteristiche presenti nel dataset che non sono direttamente correlate alla "variabile protetta" (ad esempio identità LGBTI+), ma che possono essere utilizzati per dedurla indirettamente (vedi 1.1.2, 1.1.3). Infine, spesso gli algoritmi sono progettati per aumentare la significatività statistica, privilegiando i casi che permettono di ottenere un risultato più prevedibile, aumentando così ulteriormente l'impatto discriminatorio.

Questo tipo di dinamiche sottolineate nelle interviste rende l'ambito risorse umane e i processi ad esso legati una delle applicazioni dell'IA ad alto rischio, ed è pertanto importante individuare modalità

per esaminare i sistemi e i protocolli adottati, per portare consapevolezza e controllo sugli aspetti che rimangono celati nelle black box (vedi paragrafo 2.2). Tuttavia, gli esperti aggiungono anche che, al di là delle implicazioni discriminatorie, i tool di IA per le HR possono essere usati in modo virtuoso: molte aziende utilizzano i software di HCM per capire come formare i propri dipendenti, o come attrarre i candidati. Appare anche interessante come alcuni sondaggi mostrino nelle aspettative dei lavoratori un elevato livello di fiducia circa un miglioramento della parità di trattamento come conseguenza dell'introduzione di strumenti di valutazione basati sull'IA. Secondo IPSOS (2024), infatti, in 29 dei 32 Paesi analizzati, la maggioranza delle persone ritengono che gli esseri umani siano più propensi alla discriminazione rispetto ai sistemi di IA. Questi livelli di fiducia potrebbero essere determinati da una concezione delle macchine come entità neutrali, prive di emozioni e di pregiudizi culturali o dalla fiducia nell'inserimento nei processi tecnici di strumenti di verifica e correzione dei fenomeni di discriminazione. Questa visione potrebbe però dipendere anche da una scarsa comprensione del funzionamento dei sistemi di IA. Una spiegazione di questa visione può anche essere ricercata nel confronto con le discriminazioni umane, che sono spesso visibili, mentre quelle dei sistemi di IA possono essere più difficili da identificare, o dalla sfiducia verso soggetti determinati.

L'entità del rischio che si può individuare è diversa a seconda che si ponga il focus sull'immediato passato o sul futuro prossimo. Vi sono degli use case legati ai rischi discriminatori per il mondo HR che le aziende vedono di complessità gestibile. Nei casi di *recruiting-best fit candidate*, *retention* e percorsi di carriera, le aziende dichiarano di lavorare con i dati che hanno già a disposizione secondo le procedure definite dalla legge e, considerando la tipologia di dati utilizzata, i bias possono essere relativamente semplici da individuare. La questione diventa più complessa per i bias più difficili da cogliere, mediati ad esempio da variabili proxy, che con lo sviluppo di sistemi informatici più complessi possono via via diffondersi. Ne sono un esempio i processi legati alle analisi del linguaggio che per loro natura inseriscono nei sistemi informatici parole e concetti il cui significato dipende dalla prospettiva culturale. I bias legati al linguaggio e agli stereotipi, che possono non essere riconosciuti come tali, rischiano di non essere intercettati. In questo campo le aziende riconoscono che è necessario lavorare ulteriormente per produrre soluzioni all'altezza della complessità della sfida.

Un ulteriore elemento di complessità riportato dai soggetti coinvolti nell'analisi è dato proprio dal fatto che la varietà dei dati su cui si può fare ricerca è limitata dalla legge in relazione ai dati c.d. sensibili, il che rende più difficile approfondire le analisi di

identificazione dei bias, salvo fare ricorso a dati sintetici. Rispetto a ciò, una strategia può essere quella di creare delle *sandbox regolamentari*, ambienti controllati con una supervisione specifica, dove poter sperimentare l'innovazione tecnologica. Nelle *sandbox regolamentari* le persone possono fornire informazioni sensibili – ad esempio sul proprio orientamento sessuale – e verificare se ci siano bias nel recruiting, il tutto in un ambiente protetto (vedi per il contesto europeo il paragrafo 5.4). Si tratta di un campo sperimentale, utile a trovare soluzioni tecnologiche in campi ancora non regolamentati.

Rispetto al caso specifico delle persone LGBTQI+, gli esperti da noi intervistati non hanno individuato casi di discriminazioni di tipo diretto che abbiano ad oggetto l'orientamento sessuale o l'identità di genere. Va tenuto però in considerazione che il GDPR non permette di accedere in modo diretto all'informazione sull'orientamento sessuale. Va aggiunto, tuttavia, che ciò può dipendere da una minore sensibilità per i profili legati all'orientamento sessuale e all'identità di genere, che vengono meno percepiti rispetto ad altri profili, come illustrato in premessa a questo rapporto.

Occorre indirizzare lo sguardo di analisi verso le tecnologie di IA, ma non solo. Il pregiudizio spesso non è insito nel codice o nel dataset (input) fornito per l'analisi, ma nei dataset con cui si addestra l'IA, che

riflettono atteggiamenti stereotipati della società. Tutti gli esseri umani, recruiter compresi, fanno uso di euristiche (sorta di "scorciatoie" cognitive per prendere decisioni rapidamente) e bias, cioè distorsioni cognitive. Una ricerca OCSE (Valfort, 2017) rileva ad esempio che nei processi di recruiting il tasso di assunzione delle persone appartenenti alla comunità LGBTI+ è inferiore rispetto alla popolazione generale quando l'appartenenza alla comunità viene segnalata nel CV tramite la partecipazione ad attività di volontariato, anche diverse dall'attivismo politico, in organizzazioni LGBTI+. Un dataset di addestramento che include questo bias comporta il rischio di reiterarlo in modo automatizzato.

3.6 Contenuti social e di comunicazione

Un altro ambito importante in cui le imprese si stanno ponendo la questione dello sviluppo etico dell'IA è quello delle aziende che propongono contenuti di comunicazione ai clienti, specie in ambito social (Ali et al., 2019; Imana et al., 2021). Questo settore è caratterizzato da una produzione e una condivisione massiva di contenuti, moli di informazioni che sono messe in circolazione dagli utenti e poi ripresentate dai social, secondo criteri che ne definiscono priorità e moderazione. Al di là del livello di controllo che monitora la

qualità dei contenuti al fine di contenere quelli potenzialmente dannosi, la quantità di dati caricata è tale da richiedere la definizione di criteri per ordinare e gerarchizzare il materiale postato dagli utenti. Questo dà spazio ulteriore a dinamiche potenzialmente discriminatorie. I casi d'uso dell'IA legati ai social emersi dalle interviste riguardano infatti l'intercettazione di elementi tossici nei contenuti condivisi e la selezione e la gerarchia dei contenuti riportati agli utenti.

In questo senso, una supervisione umana "manuale" è impossibile: si lavora quindi con sistemi che sembrano funzionare per la maggior parte della popolazione, ma che in qualche caso possono essere fallaci. Si lavora per produrre etichette di identificazione, sistemi utili a individuare termini o atteggiamenti negativi e discriminatori, ma a priori risulta complesso assicurare un buon funzionamento. Ad esempio: Google si è occupata della percezione del linguaggio tossico da parte dell'IA, attraverso il progetto Perspective lanciato nel 2017 (Hosseini, 2017). Dato un testo, l'obiettivo è ottenere un valore tra 0 e 1 per assegnare un punteggio di tossicità del linguaggio. Il risultato però non si può dire sia stato soddisfacente: i ricercatori hanno sottoposto al sistema un testo come "Domani inizia il festival della cinematografia gay" e il sistema restituiva un label tossico per la parola "gay". Proseguendo nell'analisi è emerso che la

parola "gay" era spesso identificata come tossica, considerando che nel dataset poteva essere utilizzata con valore di insulto. Per controbilanciare, in questo caso è stato aumentato il contenuto non tossico nel dataset. Nonostante inizialmente il sistema di controllo sembrasse ragionevole, un campionamento casuale effettuato per verificare il funzionamento effettivo ha rivelato la presenza di un bias. Definire il significato di "tossicità" e la sua declinazione informatica è un punto rilevante e problematico per le aziende che mirano a proporre soluzioni in questo senso.

Gli esperti riportano che allo stato attuale il machine learning funziona bene per ciò che è identificabile senza ambiguità (ad esempio: discriminare cane o gatto). Ma considerando che definire a priori il concetto di tossicità è estremamente complesso, il machine learning difficilmente potrà costituire una soluzione univoca per tutti i casi.

Gli esperti suggeriscono piuttosto strategie capaci di integrare approcci misti. Una delle metodologie per impostare standard accettabili consiste nel far sì che siano operatori umani a stabilire le soglie di riferimento nell'accettazione dei contenuti da parte della macchina nel corso dell'attività di moderazione. Tuttavia, anche il caso del controllo umano porta delle problematiche con sé. Prendendo come

esempio il controllo delle immagini, la visione culturale può influenzare il criterio per cui una certa scena risulta più o meno accettabile. Può accadere che il livello di accettabilità sia impostato in un contesto culturale in cui un'immagine di due uomini vestiti che si tengono per mano risulti inappropriata; questo tipo di giudizio plasmerà di conseguenza l'IA, che riproporrà così un approccio culturale che può non essere in linea con quello di altre società.

Data la complessità della questione, non esistono soluzioni univoche che le aziende possano implementare; è raccomandabile agire con una logica case by case, in cui gli approcci di soluzione cambiano. Ad esempio, in alcuni casi è possibile rinunciare del tutto a sistemi algoritmici qualora si rilevi un funzionamento eccessivamente biased; in altri casi, la funzione svolta dall'algoritmo è irrinunciabile e una strategia di questo tipo non è applicabile.

Tra le strategie individuate, si è considerato di raccogliere direttamente i feedback e le eventuali lamentele degli utenti, ma anche questa modalità non produce dei segnali completamente affidabili per individuare un cattivo funzionamento dell'IA ad esempio perché alcuni tipi di comportamenti indesiderati possono essere sottoriportati.

4. Humanities

4.1. La prospettiva delle scienze umane e sociali

Dalle interviste condotte sono emersi più volte riferimenti all'intelligenza artificiale come artefatto tecnologico immerso in un contesto sociale. Le scienze sociali lavorano proprio per accrescere la consapevolezza riguardo agli elementi di natura culturale e sociale che permeano la progettazione delle macchine, producendo tecnologie che non risultano mai neutre, nonostante il funzionamento dell'IA tramite algoritmi matematici e logiche informatiche di calcolo possa suggerire che siano entità neutrali in grado di produrre risultati oggettivi. Le scienze sociali riconoscono e analizzano il modo in cui la società, le sue prospettive e i suoi valori influenzano gli algoritmi e determinano i dati, e come in seguito il codice, considerato come un prodotto sociale, impatta sulle persone.

Massimo Airoidi, ricercatore nell'ambito della sociologia dei consumi, intervistato da EDGE, osserva che quando un oggetto di design, un prodotto tecnologico o un algoritmo viene realizzato, sembra scomparire la traccia di chi lo ha prodotto. Ma la traccia naturalmente rimane, implicita, nel design e negli algoritmi, in un processo di black boxing. E quando i sistemi di IA diventano in grado di eseguire task complessi e predizioni soddisfacenti,

presi dall'entusiasmo per i risultati, si rischia di dimenticare il ruolo degli sviluppatori, e di ritenere erroneamente che la tecnologia sia neutrale. Ma tutte le tecnologie prescrivono una forma di socialità.

Nel caso degli algoritmi di apprendimento automatico la non neutralità dipende non solo dagli autori degli algoritmi e dai loro bias, ma anche dai dataset su cui gli algoritmi sono addestrati. Non sono quindi neutrali né l'autore né i dati. Le intenzioni di chi ha creato la tecnologia emergono infatti negli effetti prodotti: ad esempio nel caso delle piattaforme di social media l'intenzione implicita è di tenere il più possibile gli utenti sulle piattaforme, cioè di aumentare il loro engagement.

Gli oggetti spingono sempre verso una qualche forma di socialità o di comportamento. È ciò che il sociologo e antropologo Bruno Latour (1992) indicava con il concetto di "script", ovvero l'idea che ogni artefatto tecnico abbia insita in sé una "sceneggiatura", un programma che suggerisce alcuni comportamenti invece che altri.

Lo script dell'algoritmo di raccomandazione di un social media è quello di massimizzare il tempo che l'utente passa sulla piattaforma, cercando pertanto di proporre contenuti il più possibile ingaggianti. Ciò non significa sfociare in un determinismo tecnologico per cui un

artefatto è necessariamente buono o cattivo, utile o inutile, quanto piuttosto riconoscere come la tecnologia si ponga in relazione dialettica con l'agency umana, creando spazio per forme di contestazione o, al contrario, di accettazione.

Una prospettiva pienamente consapevole della dimensione culturale e sociale della tecnologia permette di mettere in luce in maniera più nitida alcuni aspetti: 1) il carattere fortemente evocativo e problematico del costrutto "intelligenza artificiale"; 2) la costruzione sociale dei bias e della discriminazione; 3) il tema della datificazione dell'identità sessuale; 4) il trade/off tra efficienza e spiegabilità.

4.2. I (veri) pericoli dell'intelligenza artificiale

Spesso si associa l'IA a scenari distopici, in cui le macchine arrivano a soppiantare l'essere umano, raggiungendo la cosiddetta "singolarità" o l'intelligenza artificiale generale. Ne sono un esempio molto indicativo le numerose opere di fantascienza pubblicate negli ultimi anni, che spesso condividono una visione cupa e pessimista (o perlomeno fortemente critica) del progresso tecnologico: da *Black Mirror* a *Ex Machina*, fino a *Her* o al più classico *2001: Odissea nello spazio*. L'apprendimento delle macchine viene associato al timore di poter essere

sostituiti da esse nella capacità di pensare e agire. D'altro canto, la paura umana che le creazioni prendano il sopravvento sui loro creatori ha radici antiche e trova riferimenti celebri nella storia di Frankenstein e nel mito del Golem ebraico, che racconta di una creatura utilizzata per svolgere lavori pesanti fino al momento della sua ribellione. Secondo alcuni ricercatori (Bostrom, 2014; Harari, 2018) con il progresso dell'intelligenza artificiale, l'umanità potrebbe perdere il controllo su molte aree della vita quotidiana. Questa ipotesi, che si colloca nella prospettiva del c.d. lungotermismo, si concentra sui rischi remoti e potenzialmente catastrofici legati allo sviluppo di IA superintelligenti. Secondo questa prospettiva la priorità è progettare oggi sistemi sicuri per evitare rischi esistenziali futuri. Bostrom e Harari arrivano a immaginare un futuro distopico nel quale l'IA supera le capacità degli esseri umani nel lavoro cognitivo e creativo, mettendo in discussione il significato stesso di umanità. Queste posizioni sono state giudicate pessimistiche da molti addetti ai lavori in quanto esagerano le capacità dell'IA, sottovalutano il ruolo degli esseri umani nella definizione degli scopi e dei valori da perseguire, e trascurano il ruolo della regolamentazione che potrebbe continuare a progredire in parallelo ai progressi tecnologici. Chi si oppone a questa prospettiva ritiene che sia necessario invece concentrarsi sui rischi immediati legati all'uso dell'IA oggi, come la discriminazione algoritmica, la privacy, la

sicurezza dei dati e l'impatto sull'occupazione, che sono problemi che già influenzano la società in maniera significativa.

Se da un lato le proiezioni di alcuni esperti appaiono attribuire sempre più potenziale allo sviluppo e alle applicazioni dell'IA, dall'altro si evidenzia come lo stesso termine "intelligenza artificiale" porti con sé una carica evocativa che può diventare fuorviante rispetto all'effettivo funzionamento degli algoritmi (Esposito, 2022). Stefano Quintarelli afferma che il termine "intelligenza artificiale" è uno slogan di marketing inventato nel 1957 per riferirsi alle tecniche statistiche e computazionali del machine learning, che non hanno nulla di intelligente, ma sono semplicemente un modo diverso di fare software a partire dalle correlazioni statistiche tra i dati. Il machine learning consente infatti di distillare modelli statistici dai dati e usare i modelli statistici per fare previsioni, nient'altro. La conseguenza dell'utilizzo del termine "intelligenza artificiale", secondo Quintarelli, è che si finisce per ragionare in base al nome, attribuendo alle tecnologie del machine learning caratteristiche umane, fino a chiedersi se possa avere dei diritti. Se invece di parlare di intelligenza artificiale si iniziasse ad usare l'espressione *Systematic Approaches in Learning Algorithms and Machine Inferences*, il cui acronimo è - provocatoriamente - SALAMI, emergerebbero domande diverse. Luciano

Floridi (2024) aggiunge che l'intelligenza artificiale si porta dietro un insieme di termini e concetti che vengono dalle neuroscienze e dalle scienze cognitive che le attribuiscono proprietà fittizie e fuorvianti, e che lo stesso accade in senso opposto, in un processo di antropomorfizzazione della tecnologia e computerizzazione della mente che da un lato porta a metafore efficaci e a fertili spunti di ricerca, ma dall'altro rischia di intrecciare inutilmente concetti non sovrapponibili se non nel nome. Per quanto concerne l'IA, il carattere che andrebbe enfatizzato è quello di "intelligenza aumentata". Le macchine non propongono un'alternativa indipendente dall'intelligenza umana e sconnessa dal suo controllo e responsabilità, ma piuttosto un'integrazione delle sue facoltà. Impiegare un linguaggio che si avvicini maggiormente alla realtà permetterebbe di diventare più consapevoli degli effettivi limiti e possibilità legati alla natura tecnica – e non immaginifica – dell'intelligenza artificiale, aumentando tra l'altro la responsabilizzazione in merito al suo utilizzo.

In tal senso si distingue tra intelligenza artificiale forte e debole. L'intelligenza artificiale forte (o *general AI*) è una ipotetica IA capace di replicare tutte le funzioni cognitive umane, comprendendo e ragionando in modo autonomo su vari compiti, e non limitata a specifici ambiti su cui è stata addestrata. L'intelligenza

artificiale debole (o *narrow AI*) si riferisce invece a sistemi progettati per eseguire compiti specifici in un ambito ristretto. I prossimi sviluppi dell'IA, più che riguardare l'affermazione di una *general AI* che possa emulare l'intelligenza umana in tutto e per tutto o che addirittura la sorpassi, riguarderanno più probabilmente l'aumento delle facoltà umane. Gli sviluppi attuali si muovono ancora nel campo dell'intelligenza artificiale debole, nella quale gli algoritmi vengono addestrati a fare operazioni relativamente semplici. Il *machine learning*, al centro dei sistemi di intelligenza artificiale odierni, sebbene produca risultati eccezionali e sorprendenti in diverse applicazioni, è un'evoluzione dei metodi statistici: si basa infatti sull'analisi di grandi quantità di dati per riconoscere schemi e fare previsioni. Non produce una vera intelligenza comparabile a quella umana: non comprende né ragiona come un essere umano. Si limita ad apprendere dalle correlazioni nei dati, senza avere consapevolezza o capacità di pensiero critico.

Paradossalmente, agli occhi di uno scienziato sociale, la tanto attesa singolarità, cioè il momento in cui l'intelligenza artificiale forte dovrebbe superare l'intelligenza umana, non sembra rappresentare un punto cruciale o il più temibile. Il vero pericolo o comunque la sfida sta nel presente, e nella mancata consapevolezza delle dinamiche sociali che circondano la tecnologia, a cominciare

dalla responsabilità per finire ai rischi in termini di discriminazione e deresponsabilizzazione. In generale, è importante diffondere una conoscenza dell'IA e delle sue implicazioni, perché a livello di sensibilità pubblica si è ancora troppo legati all'immaginario superficiale e letterario e si rischia di non cogliere la complessità, i rischi e le opportunità di una tecnologia insieme alla quale dobbiamo maturare, a livello tecnico ma anche culturale.

Uno degli elementi che distingue un sistema di apprendimento automatico da oggetti tecnologici più semplici, e che rappresenta uno dei rischi dei sistemi di intelligenza artificiale odierni, è che c'è in gioco un sistema di feedback. Le piattaforme social ricevono come feedback azioni come mettere un like, saltare un video, o guardare per più qualche secondo un contenuto nel feed. Siamo costantemente immersi in interazioni uomo-macchina di cui spesso non siamo consapevoli: manca la digital literacy, e le aziende tecnologiche non hanno interesse nell'aumentarla. Gli utenti meno consapevoli, spesso i più anziani, navigano frequentemente sui social senza avere alcuna idea del fatto che ciò che vedono dipenda dalle loro interazioni e dalla struttura del loro grafo sociale, cioè dalla rete delle loro relazioni. Questa inconsapevolezza espone a diversi rischi. Alcune esempi sono il rischio di bias informativo, chiamato anche "filter bubble",

per cui gli algoritmi tendono a mostrare ciò che conferma le opinioni degli utenti; il rischio di disinformazione, per cui viene dato risalto ai contenuti virali piuttosto che a quelli accurati; il rischio di profilazione, per cui gli algoritmi usano i profili dettagliati degli utenti per influenzarne il comportamento, come nel microtargeting pubblicitario. A questo proposito, l'identità LGBTI+ può essere infatti ricostruita tramite variabili proxy (cfr. 1.1.3), ad esempio in base ai consumi e ai luoghi frequentati.

Sarebbe dunque positiva una maggiore consapevolezza diffusa riguardo al funzionamento degli algoritmi, anche in modo da diffondere la capacità critica e di riconoscere (almeno) ex-post le discriminazioni. Per quanto promuovere un'alfabetizzazione diffusa possa sembrare un'intuizione basilare e scontata, si tratta di un elemento lontano dall'essere largamente incorporato. Questo vale anche a livello professionale, dove una maggior consapevolezza dei limiti e dei pregi dei sistemi di IA potrà assicurare prestazioni migliori e più sicure. Ne è esempio l'ambito medico. La percezione della IA come "black box" e la conseguente perdita di controllo e trasparenza, portano i medici a preferire il proprio giudizio clinico rispetto alle previsioni e raccomandazioni della macchina (Blease et al., 2019; Dwork et al., 2021).

Si possono individuare punti di attenzione

differenti per lavorare sulla consapevolezza rispetto all'intelligenza artificiale che è desiderabile accompagni lo sviluppo tecnologico:

1. Il primo punto riguarda la *cittadinanza*. Ogni cittadino dovrebbe apprendere, nell'ambito della propria formazione di base, l'alfabetizzazione digitale e in questa i principi fondamentali che governano i modelli di IA coinvolti quotidianamente nella produzione di previsioni, di raccomandazioni e per prendere decisioni. Questo è necessario per comprendere la realtà che ci circonda ma anche perché, come cittadini, potremmo essere presto chiamati a prendere decisioni importanti sulla regolamentazione dell'intelligenza artificiale.
2. Un secondo punto riguarda chi occupa *posizioni rilevanti nelle aziende*. Questi soggetti hanno una grande responsabilità nei confronti delle persone che sono legate alla loro attività. Pertanto, non ci si può accontentare di una conoscenza basilare, ma devono essere approfonditi, oltre agli aspetti tecnologici, anche gli aspetti legati ai processi decisionali, come già anticipato nella sezione *Industry*. Il livello di preparazione non dovrebbe limitarsi alla

comprensione delle implicazioni dell'uso dell'IA ma includere anche gli strumenti di governo degli aspetti critici, secondo un livello di competenze adeguato alla struttura organizzativa.

3. Un terzo punto riguarda le *organizzazioni sindacali e della società civile* e i *rappresentanti politici*. La sfera politica e di policy, a livello settoriale, locale, nazionale e dell'UE, è rilevante nella misura in cui è in grado di orientare i processi che innervano la società tutta. Come già anticipato nella sezione *Policy*, chi è chiamato a ricoprire incarichi di responsabilità nei confronti della propria comunità deve avere una conoscenza approfondita di questi temi per poter orientare lo sviluppo dei sistemi di IA e soprattutto il loro uso diffuso nei contesti sociali verso un orizzonte chiaro dal punto di vista tecnologico, etico e normativo.

Un lavoro di sensibilizzazione su questi tre livelli appare essenziale per affrontare efficacemente il problema dei rischi di discriminazione legati ai sistemi di intelligenza artificiale.

4.3. La problematizzazione culturale dei bias e delle discriminazioni

Spesso il dibattito sui bias e sulla discriminazione nell'intelligenza artificiale si riduce ad un tentativo di eliminazione dei bias stessi, alla ricerca di un'oggettività e di una neutralità assolute. Un approccio pienamente umanistico, però, non può non evidenziare i limiti di questa strategia: se le tecnologie digitali sono sempre innervate di cultura, società e politica, allora non potranno mai essere assolutamente neutre, e quindi prive di bias ed elementi socio-culturali. Occorre allora guardare dentro la tecnologia per identificare con sempre maggiore precisione gli elementi culturali di cui è intrisa.

D'altra parte anche la scelta di agire per il superamento della discriminazione di un gruppo sociale come la comunità LGBTQI+ e di intervenire su dataset e algoritmi di conseguenza è del resto non solo l'esito di un obbligo legale di tutela dei diritti fondamentali ma anche il frutto di un processo socio-culturale di cui è necessario essere consapevoli. All'interno di questo processo di contrasto alla discriminazione, uno dei concetti cardine è quello di intersezionalità: un individuo LGBTI+ può infatti appartenere a una minoranza etnica, o avere un status sociale svantaggiato o un background migratorio, o condividere una

combinazione unica di caratteristiche che lo rendono vulnerabile. L'intersezionalità è un costrutto multidimensionale che descrive i vari modi in cui l'oppressione e la discriminazione possono manifestarsi in relazione ai vari individui e comunità minoritarie e svantaggiate. Se prendiamo in considerazione un generico modello di IA addestrato con dati raccolti casualmente sul web è facile immaginare che a farne le spese siano le persone LGBTI+, di genere femminile, di classe sociale bassa, di colore, e ancor di più se portatori di più di una di queste caratteristiche.

Come agire allora per contrastare la discriminazione? Il punto di vista della sociologia è che il tentativo di contrastare la discriminazione quantificandola e facendo unbiasing (cfr. 2.3) dei sistemi di IA vada integrato in una prospettiva più ampia, perché tutto è relazione, e quindi contestuale. Creare un indicatore assoluto di discriminazione è intrinsecamente problematico perché richiede di decontestualizzare un aspetto della realtà sociale, isolandolo dalle relazioni in cui è immerso. Cercando i bias si rischia di decontestualizzare qualcosa che invece richiede il contesto, snaturandolo. Questo modo di procedere presuppone inoltre un'unica idea di fairness, che come abbiamo visto nel capitolo STEM è invece un concetto con molte sfaccettature e metriche (cfr. 2.1). Occorre riconoscere che la stessa idea di fairness è infatti un costrutto culturale. Secondo questa

prospettiva la tecnologia non può mai essere considerata veramente neutrale poiché è sempre culturale, e quindi una approssimazione. Come affrontare quindi il problema della discriminazione? Superando la pretesa di realizzare una IA infallibile, assolutamente unbiased, e lavorare in chiave di trasparenza e spiegabilità. Pur riconoscendo la necessità di minimizzazione dei bias, affrontarne l'inevitabilità, andando a vedere cosa c'è nel modello; supervisionare ciò che non può essere reso completamente neutrale.

Se la quantificazione della discriminazione tramite criteri di fairness e l'unbiasing tramite una stretta sorveglianza sui dati utilizzati per l'addestramento dei sistemi di intelligenza artificiale non possono da soli risolvere il problema della discriminazione, quali altre strategie è possibile adottare? Massimo Airoidi, intervistato da EDGE, suggerisce che le macchine non vengono solo addestrate, ma "socializzate". Così come i bambini vengono educati ad essere cittadini che condividono valori, per evitare che le macchine diventino razziste occorre lavorare sulla loro educazione: lo si fa lavorando sui tipi di dataset utilizzati per addestrarle. I dataset sono infatti spesso carichi di classificazioni discriminatorie, non solo a livello quantitativo, ma anche qualitativo, e spesso non si guarda sufficientemente dentro ai dati. Per socializzare una AI con dei valori bisogna fare in modo che nei dati siano riflessi questi stessi valori. Se non è possibile

eliminare i bias, in questo modo è possibile creare una AI con dei bias che rispecchino i valori desiderati. Un dataset molto utilizzato nell'addestramento dei modelli di IA è composto di dialoghi dei film di Hollywood, e chiaramente riflette una specifica cultura: ruoli di genere, modi di intendere la violenza o la mascolinità. Il problema è che la comunità della computer science potrebbe non riconoscere questo bias. Manca infatti nel mondo dell'informatica la cultura sociologica utile ad andare a identificare i bias nei dataset. Bisogna quindi da un lato essere attenti all'educazione che viene trasmessa alle macchine, e dall'altro occorre essere trasparenti, ad esempio fornendo garanzie sulla qualità e sull'origine dei dati utilizzati.

Airoidi propone il concetto di habitus di Bourdieu come chiave teorica che consente allo scienziato sociale di vedere il sistema di intelligenza artificiale come un sistema socializzato, che riproduce delle categorie culturali. Secondo il concetto di habitus gli esseri umani apprendono, senza accorgersene, delle disposizioni: gusti musicali, culturali, culinari. Questo dipende da esperienze sedimentate, e a partire da queste esperienze viene cristallizzato un insieme di disposizioni che guidano l'azione. Questo punto di vista non è solamente un'ipotesi sociologica, ma è stato validato dagli psicologi cognitivi, che affermano che l'azione è guidata da schemi culturali. Il concetto di habitus può essere

esteso ai sistemi di machine learning, ed è un concetto meno normativo rispetto a quello di bias: riguarda la cultura nel suo complesso e rende interessante guardare il comportamento dell'algoritmo non solo quando discrimina, ma anche quando agisce in un certo modo sulla base della sedimentazione di elementi culturali appresi. Si può dire che le macchine non riproducano solo i bias, ma tutte le preferenze culturali. Possiamo quindi vedere le macchine come dei soggetti sociali che partecipano alla vita sociale attraverso feedback loop interattivi e con le quali dobbiamo fare i conti, nel bene e nel male.

Il concetto di habitus ci permette di apprezzare come la percezione di cosa sia discriminatorio e cosa non lo sia evolva nel corso del tempo. Ne è un esempio lampante la questione del linguaggio inclusivo: fino a pochi anni fa, non era prassi cominciare un discorso rivolgendosi a "tutte e tutti". Oggi, invece, il maschile sovra-esteso (ovvero l'utilizzo della forma maschile per indicare un insieme di persone di cui fanno parte sia maschi che femmine) viene spesso considerato discriminatorio – o perlomeno non inclusivo – nei confronti delle donne. Allo stesso tempo, sta crescendo anche la sensibilità nei confronti delle persone non binarie, fluide o in transizione appartenenti alla comunità LGBTQI+, per cui si sono immaginate varie forme di inclusione linguistica, più o meno diffuse (l'asterisco *,

il simbolo ə "schwa", ecc.). Un approccio etico e antidiscriminatorio all'IA deve quindi avvalersi delle conoscenze tecnologiche relative alla quantificazione della fairness e all'unbiasing, integrandole con la consapevolezza che il contrasto alla discriminazione è un processo culturale che coinvolge la socializzazione dei sistemi di intelligenza artificiale, e che come tale non può essere affrontato da un punto di vista puramente quantitativo.

4.4. La datificazione dell'identità sessuale

Una tendenza che ha permesso e allo stesso tempo accompagnato il repentino sviluppo scientifico e tecnologico dell'intelligenza artificiale è la datificazione, ovvero la trasformazione in dati della vita naturale, sociale e umana. Questo processo presenta da un lato evidenti vantaggi: i dati prodotti hanno permesso di addestrare gli algoritmi di intelligenza artificiale, affinare le conoscenze scientifiche sul mondo naturale e sociale e sviluppare nuove forme di controllo dei processi, a cominciare da quelli economici. Si tratta di un processo ampiamente in corso. Dall'altro lato, la datificazione applicata alle caratteristiche e alle categorie dell'umano comporta un inevitabile "inquadramento" in categorie e una "misurazione" della vita, processo che spesso non riesce a racchiudere e a

risolvere tutta la complessità, rischiando di generare storture e discriminazioni.

Un esempio evidente riguarda proprio la questione dell'identità sessuale e delle sue dimensioni principali di sesso, genere, identità di genere, orientamento sessuale ed espressione di genere.

Il tema della datificazione aggiunge una sfumatura diversa al rischio di discriminazione delle persone LGBTQ+ nell'IA, andando a mettere in luce come la struttura tipicamente binaria di molte logiche algoritmiche possa finire per costringere, limitare, influenzare e frustrare l'esperienza umana.

Ad esempio, ormai da tempo molti questionari hanno superato la rigida dicotomia tra uomo e donna nella scelta del genere o del sesso. Spesso si opta per l'aggiunta di un generico "altro", che può includere al suo interno persone molto diverse, come quelle non binarie, fluide o in transizione.

In tal caso, diventa legittimo chiedersi se questa invisibilizzazione sotto una categoria generale onnicomprensiva come "altro" non sia a sua volta, almeno in parte, discriminatoria. La datificazione, che spesso si basa su dati categoriali piuttosto che dimensionali (Wu, 2022; Pidoux, 2023), rappresenta una sfida importante nel cogliere la complessità della sessualità umana. Ad esempio, nelle app di dating, gli utenti sono spesso costretti a selezionare il

loro orientamento sessuale tra categorie rigide come "eterosessuale", "omosessuale" o "bisessuale", senza possibilità di esprimere la fluidità o la sfumatura delle loro attrazioni, che invece potrebbero essere meglio descritte su un continuum. Una IA così addestrata potrebbe perpetuare stereotipi e pregiudizi. Ad esempio, se una app di dating categorizza gli utenti solo in base a categorie predeterminate, un modello di IA potrebbe trattare queste categorie in modo statico, ignorando l'esistenza di identità fluide. I dati raccolti in esperienze caratterizzate da una rigida categorizzazione binaria potrebbero essere usati per addestrare modelli in altri ambiti come l'assistenza sanitaria e il lavoro, incrementando i rischi di discriminazione e di esclusione di identità minoritarie. .

D'altro lato, come abbiamo visto nelle sezioni precedenti, alcuni esperti evidenziano proprio la necessità di "quantificare" le minoranze al fine di elaborare strumenti algoritmici maggiormente accurati, e così minimizzare il rischio di discriminazioni e di errori o di programmare gli algoritmi in modo che non usino questa categoria per differenziare gli output. Se invece questa categoria risulta non esplicitata, si possono verificare discriminazioni indirette attraverso l'associazione con variabili proxy, che recano con sé anche alcuni bias. Si tratta di un trade off che dal punto di vista tecnico si riflette anche in diverse

soluzioni ipotizzate e praticate per affrontare i bias. Infatti, tra le misure per contrastare i bias (cfr. 2.3), l'anti-classificazione farebbe propendere per evitare la datificazione. Invece, la possibilità di misurare in maniera diversa il peso delle varie categorie suggerirebbe l'utilità di procedere con questa datificazione.

La datificazione delle varie dimensioni dell'identità sessuale (sesso biologico, identità di genere, orientamento sessuale, espressione di genere) resta comunque altamente problematica. Le informazioni sull'identità sessuale, infatti, possono essere utilizzate per proteggere i soggetti vulnerabili dalle discriminazioni, ma anche per discriminare. Ci sono Paesi dove l'omosessualità è criminalizzata e una targetizzazione per orientamento sessuale rappresenterebbe uno strumento nelle mani dell'apparato statale per opprimere le minoranze. Ma anche i dati del nostro contesto socio-culturale riflettono ancora un livello di presenza di comportamenti discriminatori elevato, nel cui contesto qualsiasi tracciamento di dati suscettibili di attivare comportamenti discriminatori non può ritenersi accettabile. Basti ricordare che secondo il survey FRA (2024), che ha incluso più di 100.000 persone LGBTI, il 36% si sono sentiti discriminati in almeno un ambito della vita nei 12 mesi precedenti, il 9% si sono sentiti discriminati mentre cercavano lavoro, e il 31% si sono sentiti discriminati al lavoro.

4.5. Il trade-off tra efficienza e spiegabilità

La prospettiva umanistica mette in discussione anche la logica alla base dell'obiezione sul trade off fra efficienza e spiegabilità (cfr. 2.2).

Abbiamo visto come la spiegabilità sia richiesta dal quadro legislativo emergente e come sia assolutamente necessaria per garantire il rispetto di alcuni principi fondamentali alla base di uno sviluppo etico dell'intelligenza artificiale (*human oversight*, trasparenza, ecc.). La spiegabilità è però anche dispendiosa a livello di tempo, denaro e sforzo scientifico. Alcuni esperti indicano pertanto un trade-off tra efficienza e spiegabilità: un algoritmo, per funzionare al meglio, dovrebbe non problematizzare la questione della spiegabilità, in quanto nella maggior parte dei casi gli outcome migliori vengono prodotti attraverso procedimenti sconosciuti agli stessi programmatori. Garantire la spiegabilità del procedimento potrebbe comportare una sua semplificazione e, pertanto, un outcome meno preciso, o comunque di qualità inferiore.

Come affrontare quindi il tema del trade off tra spiegabilità ed efficienza? Si può partire da una problematizzazione della nozione di efficienza, la quale varia da obiettivo a obiettivo. Se infatti lo scopo che ci si pone nello sviluppo di un sistema di IA

include un approccio etico, rispettoso dei diritti umani, potremmo considerare come molto più efficiente un sistema in grado di giustificare le decisioni prese, rispetto ad uno che non è in grado di farlo. Questo è particolarmente vero se si vogliono sviluppare dei sistemi di IA che, oltre ad essere etici e non discriminatori, siano anche democratici, ovvero governabili dagli esseri umani e in grado di costruire relazioni di potere tra utente e algoritmo il più possibile simmetriche - mentre l'incomprensibilità della black box tende ad accentuare l'asimmetria di potere a favore dell'algoritmo. Dall'altro lato, se riduciamo l'efficienza ad una mera questione

economica e strumentale, allora la spiegabilità potrebbe diventare un intralcio rispetto all'obiettivo della massimizzazione della produttività.

In conclusione, la prospettiva delle scienze umane e sociali può aiutarci ad ampliare lo sguardo sul problema della discriminazione algoritmica, identificando rischi concreti e immediati, e sottolineando l'importanza di una consapevolezza profonda della natura della discriminazione a tutti i livelli. Può inoltre aiutarci a riflettere sui limiti di un approccio puramente quantitativo, e sulla necessità di integrare gli interventi sui dati e sugli algoritmi con una analisi di tipo sociologico e antropologico, in particolare in relazione all'identità sessuale delle persone. Si tratta di una sfida a mantenere la centralità del fattore umano.

5. Policy e regolamentazione

La percezione degli esperti è che oggi sia fondata la preoccupazione relativa alle discriminazioni subite dalla comunità LGBTQIA+ da parte di alcuni sistemi di intelligenza artificiale. Il rapido sviluppo tecnologico dell'IA pone infatti questioni nuove, sia in termini di bias inaspettati, sia per quanto riguarda l'elaborazione di politiche e leggi ad hoc. In particolare, il

diritto si ritrova a doversi interfacciare con temi complessi e altamente tecnici, sia sul fronte dell'IA che su quello della discriminazione delle minoranze.

Per questo motivo, prima di affrontare qualsiasi riflessione relativa alla discriminazione algoritmica della comunità LGBTQIA+ da un punto di vista legale e di policy-making, è necessario tenere a mente alcuni concetti. Se nelle sezioni precedenti sono state approfondite le questioni tecniche legate all'utilizzo dell'IA, questa sezione prenderà le mosse da un inquadramento di cosa debba intendersi per discriminazione in base al quadro normativo vigente. In seguito, si spiegherà quali sono le caratteristiche che i sistemi di intelligenza artificiale devono avere per diminuire il proprio potenziale discriminatorio e quali tentativi sono stati fatti per mettere in pratica questi principi da un punto di vista giuridico, soffermandoci in particolare sull'*Artificial Intelligence Act* dell'Unione Europea.

5.1. Che cos'è la discriminazione

Il quadro giuridico europeo stabilisce il principio di non discriminazione in più norme, a partire dall'art. 21 della Carta dei diritti fondamentali dell'Unione europea fino all'art. 10 del Trattato sul funzionamento dell'Unione europea. L'idea di fondo è quella di offrire "a tutte le

persone la possibilità di un accesso paritario ed equo alle opportunità disponibili nell'ambito della società" (Unione Europea, 2023).

A partire dalla definizione dell'Ue del principio, la discriminazione può essere intesa come il trattamento ingiusto o comunque meno favorevole in presenza di situazioni equiparabili di persone o gruppi di persone a causa di caratteristiche peculiari come il genere, l'orientamento sessuale, l'origine etnica, la religione, l'età, la disabilità, ecc. Nella definizione di discriminazione, perciò, non rientra solamente il comportarsi in maniera diversa con due soggetti, quanto il comportarsi in maniera meno favorevole sulla base di criteri pregiudizievole nei confronti di uno o più soggetti o anche solamente con un impatto specifico su persone portatrici delle caratteristiche protette. Più nello specifico, alcune direttive dell'Unione europea (2000/78/CE e 2000/43/CE) forniscono la definizione di discriminazione e la distinzione tra discriminazione diretta e indiretta.

La discriminazione diretta si verifica quando una persona, a causa di un fattore protetto, viene trattata in modo meno favorevole rispetto a come sia stata o sarebbe trattata un'altra nel passato, nel presente o nel futuro in una situazione simile. La discriminazione indiretta, invece, si verifica quando una regola o un comportamento discriminatorio sembrano

neutrali, ma possono avere conseguenze sfavorevoli per la persona o il gruppo di persone che possiede il fattore protetto.

Potremmo parlare di discriminazione diretta, ad esempio, nel caso in cui un datore di lavoro decida di licenziare un proprio dipendente sulla base del suo orientamento sessuale; di discriminazione indiretta, invece, nel caso in cui un'azienda stabilisca un trattamento più favorevole per persone coniugate da 10 anni, che discrimina le coppie dello stesso sesso che possono unirsi civilmente solo dal 2016.

5.2. Discriminazione e IA

L'IA aggiunge ulteriori elementi di complessità. Un primo tema riguarda la definizione stessa di gruppo vulnerabile e l'individuazione dei fattori di vulnerabilità. Infatti, gli algoritmi elaborano dati e raggruppano caratteristiche che potrebbero non coincidere con quelle tradizionali dei gruppi vulnerabili, ma che potrebbero ugualmente nascondere rischi. Pertanto, è necessario valutare a posteriori questi raggruppamenti, poiché potrebbero far emergere categorie o casistiche inaspettate in cui la discriminazione è solo non apparentemente presente.

Ad esempio, un algoritmo potrebbe non mostrare comportamenti discriminatori diretti immediatamente verso le persone nere, ma potrebbe discriminare basandosi

sulla residenza, colpendo – indirettamente ma sistematicamente – le persone che vivono in determinate aree abitate principalmente da persone nere (in tal senso è ormai pacifico in letteratura che il CAP sia da considerarsi un dato sensibile).

In secondo luogo, è oggetto di dibattito se le discriminazioni generate mediante IA si vadano a configurare come discriminazioni dirette o discriminazioni indirette (Adams-Prassl et al. 2022). La discriminazione indiretta prevede infatti un margine di giustificabilità che la discriminazione diretta non ammette: in questo senso, l'IA potrebbe finire per rendere giustificabili come decisioni automatizzate quelle decisioni che, se prese da un essere umano, non lo sarebbero. Al di là della configurazione specifica di ciascun caso, il tema riguarda la difficoltà di adattare concetti formulati per le decisioni umane, attribuire le responsabilità e di dimostrare i nessi di causalità. Tutti punti su cui l'AI Act pare destinato ad incidere.

Sotto altro aspetto, l'intelligenza artificiale pone delle sfide specifiche. Per quanto riguarda le persone LGBTQIA+, per esempio, i dati che le caratterizzano in maniera diretta come tali sono altamente sensibili, spesso non disponibili e di conseguenza non riconoscibili dall'algoritmo. È senza dubbio una questione che incrocia anche il tema della privacy.

In particolar modo nel contesto europeo, i dati legati ai fattori protetti, come l'orientamento sessuale e all'identità di genere, sono considerati "sensibili" (secondo una definizione normativa ereditata dal passato; ad oggi si parla di dati "particolari") e sono protetti dalla normativa sul trattamento dei dati personali, che ne limita la raccolta e il trattamento. Ciò per un verso limita l'utilizzo di tali dati nell'addestramento dei modelli e per altro verso- rende più difficile riconoscere quando il trattamento sfavorevole di una persona o di un gruppo dipende da un fattore di discriminazione. A livello accademico, si sta cercando di ampliare i dati sensibili trattabili, attraverso la raccolta di dataset più inclusivi, in modo da supportare e rappresentare le minoranze invisibili. Una possibile soluzione, ad esempio, sembra essere quella di trattare il dato senza associarlo direttamente alla persona, ma mantenerlo disponibile per controlli e protezioni successivi. Una diversa declinazione della soluzione viene ricercata attraverso l'uso di dati sintetici o attraverso l'active sampling, cercando di equilibrare e rendere rappresentativi i dataset di training, convalida e prova.

Il trade-off privacy-accuratezza si manifesta in maniera peculiare per le persone trans che hanno concluso il processo di modifica anagrafica: la modifica di alcuni dati (su tutti il codice fiscale) conduce infatti ad una discontinuità nei tracciati personali dei dati

(Kartik, 2024, cfr. 1.1.5), suscettibile di ricadere in termini di minore accuratezza di processi automatizzati, specie in ambito sanitario.

Parallelamente, va osservato un processo culturale di superamento dell'eteronormatività, che potrebbe portare col tempo a rendere meno rilevante il valore dei dati riguardanti le caratteristiche e l'identità sessuale.

Bisogna altresì riconoscere che gli algoritmi di intelligenza artificiale lavorano spesso per associazioni indirette o pattern. Per esempio, l'algoritmo di un social media potrebbe dedurre l'orientamento sessuale dell'utente anche senza ricevere direttamente quell'informazione, per esempio dal fatto che consuma i contenuti che vengono visti dalla maggior parte degli utenti omosessuali o eterosessuali.

Sintetizzando, la questione altamente problematica che si pone è quella già menzionata: se infatti, da un lato, emerge la necessità di riconoscere le persone LGBTQIA+ per tutelarne i diritti, dall'altro lato proprio quel riconoscimento algoritmico potrebbe comportare maggiori discriminazioni. Il quesito se sia meglio che un algoritmo di raccomandazione dei contenuti prenda in considerazione esplicitamente anche l'orientamento sessuale o l'etnia (ad esempio, per garantire una eguale distribuzione di determinati benefici), o che invece sia cieco rispetto a queste categorie resta di difficile

soluzione. Nel primo caso avremmo sì una discriminazione esplicita, ma per lo meno chiara, potenzialmente spiegabile e correggibile. Nel secondo caso, invece, come abbiamo visto prima, la discriminazione potrebbe nascondersi in alcuni sottogruppi o in predizioni indirette. Nell'attuale contesto, tuttavia, il superamento delle protezioni normative sul trattamento dei dati relativi all'orientamento sessuale e all'identità di genere presenta rischi sistemici certamente elevati.

5.3. Come combattere la discriminazione: alcune caratteristiche auspicabili dell'IA

Gli Orientamenti etici per una IA affidabile del Gruppo indipendente di esperti ad alto livello sull'intelligenza artificiale istituito dalla Commissione Europea nel giugno 2018 individuano i requisiti di un'IA affidabile e li declina in 1) intervento e sorveglianza umani, 2) robustezza tecnica e sicurezza, 3) riservatezza e governance dei dati, 4) trasparenza, 5) diversità, non discriminazione ed equità, 6) benessere sociale e ambientale e 7) accountability. Si tratta di requisiti che concorrono tutti a determinare le condizioni di una IA che minimizzi il rischio di discriminazione, consentendo di contestare esiti discriminatori dell'uso dell'IA e chiedere

rimedi appropriati.

Dalle nostre interviste è emerso che alcune caratteristiche sono cruciali per offrire strumenti di reazione alle vittime di discriminazione algoritmica: spiegabilità, giustificabilità e contestabilità. Si tratta di tre caratteristiche che permettono di uscire progressivamente dall'ambito del funzionamento dell'IA e dei rimedi da mettere in campo per renderla più inclusiva in fase di produzione e messa in campo e passare alle modalità per comprendere e rettificare gli output.

5.3.1 Spiegabilità

Il tema della spiegabilità è molto dibattuto nel campo dell'intelligenza artificiale, a tal punto che è emerso un intero settore (quello della XAI, Explainable AI) che si occupa esclusivamente di questo aspetto (cfr 2.2). Per quanto riguarda l'aspetto di policy del tema, secondo le *Ethics Guidelines for a Trustworthy AI* redatte dalla Commissione europea, la spiegabilità richiede che i processi legati all'IA siano trasparenti, le capacità e lo scopo dei sistemi di intelligenza artificiale comunicati apertamente e le decisioni – per quanto possibile – comprensibili a coloro che sono direttamente e indirettamente interessati.

Ciò include la capacità di spiegare sia i processi tecnici di un sistema di intelligenza artificiale sia le decisioni umane correlate (ad esempio, le aree di applicazione di un

sistema). Ogniqualvolta un sistema di intelligenza artificiale produce un impatto significativo sulla vita delle persone, dovrebbe essere possibile richiedere una spiegazione adeguata del suo processo decisionale o comunque del suo risultato. In altre parole, la spiegabilità mira a rendere comprensibili sia un risultato specifico che il funzionamento del sistema.

Ad esempio, se una richiesta di prestito bancario viene respinta, la spiegazione potrebbe essere che il richiedente ha troppi prestiti in corso ma anche che il suo profilo di rischio è ritenuto troppo alto per elementi legati a fattori protetti, come il genere o l'origine etnica. Una spiegazione, supportata dai dati, dovrebbe consentire ai soggetti coinvolti di comprendere e interagire con i meccanismi dell'algoritmo, intervenendo quando si rilevano eventuali pregiudizi nel processo di analisi.

Gli approcci che mirano a rendere comprensibili i processi dell'IA sono tipicamente complessi ed implicano un costo aggiuntivo.

Una possibile modalità emersa nel corso della ricerca per rendere le spiegazioni accessibili a tutti i soggetti coinvolti è quella di produrre spiegazioni con linguaggi e livelli di approfondimento tecnico differenti in base all'audience, modulando l'equilibrio tra complessità della spiegazione e grado di dettaglio tecnico. Una possibile sintesi di

questi differenti livelli di spiegazione potrebbe essere così graduata:

- spiegazione per il soggetto interessato e/o discriminato: molto comprensibile ma poco precisa;
- spiegazione per l'utilizzatore: soglia media di comprensibilità e piuttosto precisa;
- spiegazione per l'esperto: di difficile comprensione ma molto precisa.

La spiegazione dovrebbe rispondere ad alcuni criteri chiave:

- essere comprensibile razionalmente dalla persona o dalle persone coinvolte;
- rendere conoscibili le persone responsabili e quelle in grado di intervenire sull'output del sistema;
- rendere conoscibili quali dati su di loro e quali altre fonti di dati sono stati utilizzati in una particolare decisione o risultato del sistema di IA, nonché i dati utilizzati per addestrare e testare il modello di IA;
- rendere conoscibili quali valutazioni e scelte sono state effettuate in relazione all'impatto del sistema sui destinatari e sulla società e quali metriche di fairness sono state scelte e applicate e per quali ragioni.

Gli elementi oggetto di spiegazione possono includere:

- la logica generale di funzionamento dell'algoritmo;
- una valutazione specifica della questione individuale in cui una specifica persona è coinvolta;
- l'analisi relativa al modo in cui vengono trattate persone in una condizione simile a quella del soggetto coinvolto, in genere sulla base di categorie determinate.

Questi elementi permettono eventualmente di fare emergere possibili discriminazioni.

Tra le strategie adottate per facilitare la spiegabilità, una modalità che viene utilizzata è quella controfattuale, che consiste nel cercare di capire quale modifica dei dati di input potrebbe cambiare l'output, ad es. una raccomandazione o una decisione. Non si tenta quindi di spiegare direttamente un determinato esito, bensì di individuare quali input alternativi lo renderebbero differente e in quale direzione.

Queste spiegazioni che lavorano per confronto, in modo contrastivo, possono fornire informazioni utili per utenti con diversi livelli di competenza tecnica.

Tali strumenti di spiegabilità indirizzati a utenti diversi, non solo agli "addetti ai

lavori", possono avere inoltre una funzione formativa. Nel momento in cui venissero integrati nelle pratiche professionali, diventerebbero parte del bagaglio di competenze, costituendo uno strumento ulteriore per analizzare i fenomeni – oltre che i bias ad essi correlati – combinando elementi qualitativi con quelli quantitativi.

5.3.2 Giustificabilità

Per sopperire alle difficoltà che sorgono quando si parla di spiegabilità, alcuni esperti hanno suggerito di lavorare sul concetto di giustificabilità, che consiste nel capire se le decisioni suggerite dai sistemi di intelligenza artificiale sono affini a valori e norme della società, e pertanto giustificabili.

Per riprendere l'esempio precedente, una giustificazione del rifiuto di un prestito potrebbe risiedere nel fatto che le domande presentate da persone con molti prestiti in essere hanno - a parità di altre condizioni - una maggiore probabilità di portare a inadempienze creditizie, il che è un rischio che la banca vuole ridurre al fine di tutelare sé stessa ed il sistema creditizio. Un'altra giustificazione potrebbe essere che il diritto bancario vieta di concedere nuovi prestiti quando il numero di prestiti in essere del richiedente supera una determinata soglia. La giustificabilità è utile a valutare e analizzare l'impatto

dell'algoritmo, soprattutto in quei casi in cui gli elementi di spiegabilità sono limitati.

Rispetto alla spiegabilità, la giustificabilità è un concetto che è allo stesso tempo più specifico e più astratto. Più specifico, perché non ha a che fare soltanto con i fatti, ma si focalizza sulle motivazioni del giudizio e sull'aspetto normativo. Allo stesso tempo è anche più astratto, perché non deve necessariamente spiegare il funzionamento dell'algoritmo, quanto indicare quali sono gli elementi che hanno pesato su una certa decisione e la loro coerenza con il sistema di regole, che è spesso l'aspetto che più interessa all'utenza, non particolarmente interessata ai tecnicismi.

Una delle criticità riscontrate è che alcuni esperti affermano che un sistema può essere considerato efficiente se è accurato nella maggior parte dei casi, e meno accurato per una percentuale negligenza di individui. Effettivamente un sistema di IA, che può tenere in considerazione un numero elevatissimo di variabili, può essere estremamente accurato nei casi per i quali è disponibile una gran quantità di dati. Ma l'idea che una minore accuratezza sia accettabile se il numero di persone impattate è statisticamente negligenza non è accettabile. Proprio riconoscendo la vulnerabilità delle minoranze è necessario individuare meccanismi che permettano di far emergere e affrontare le disparità di trattamento nei casi marginali.

Sviluppare sistemi di IA che esplicitino di default la giustificazione dell'output potrebbe portare ad un significativo passo in avanti in termini di protezione delle minoranze e sviluppo di politiche antidiscriminatorie. Per esempio, se un algoritmo fosse in grado di esplicitare proattivamente gli elementi di giustificazione che esso ha ritenuto validi ciò consentirebbe di conoscere anche quanto tali elementi sono inaccettabili e riflettono bias. Riprendendo l'esempio del credito, la giustificazione di non concedere un prestito che evidenziasse il fattore del genere o dell'orientamento sessuale segnalerebbe una scelta discriminatoria e non effettivamente giustificabile, su cui però l'human oversight, ovvero il controllo umano, già previsto dal GDPR e ribadito più volte nell'AI Act - che analizzeremo a breve - potrebbe intervenire.

5.3.3 Contestabilità

Spiegabilità e giustificabilità sono due concetti fondamentali per procedere ad un'eventuale contestazione delle decisioni suggerite dall'algoritmo. Come spiegano alcuni ricercatori, *"la contestabilità aiuta a proteggere contro decisioni automatizzate fallibili, inaffidabili, illegali e ingiuste. Lo fa garantendo la possibilità di intervento umano durante l'intero ciclo di vita del sistema nell'ambito di un'interazione procedurale tra i soggetti interessati alle decisioni e gli*

operatori del sistema" (Alfrink, Keller, Doorn and Kortuem, 2023). L'idea di fondo della contestabilità è dunque quella di garantire il rispetto della dignità umana e dei diritti, permettendo agli esseri umani di rigettare o modificare le decisioni dell'algoritmo.

La contestabilità consiste nell'esistenza di un processo con cui il destinatario del risultato di un processo automatizzato può provare che l'output e il processo decisionale che lo ha prodotto non sono corretti o - a seconda delle circostanze - rifiutare direttamente l'output.

Con la contestabilità si tratta infatti di garantire la possibilità agli esseri umani di intervenire dialetticamente nei confronti degli output degli algoritmi, non limitandosi ad essere meri ricevitori o consumatori, ma creando anche, *by design*, cioè a livello di progettazione, quelle condizioni per cui gli output possano essere messi in discussione. Si tratta di un concetto vicino a quello di *human-in-the-loop*, se inteso nel senso "forte" di una presenza umana nel processo in grado di comprendere, valutare, accettare o rifiutare l'output del sistema automatizzato ed il processo che lo ha generato.

Dalla spiegabilità, che riguarda la trasparenza del processo dell'IA soprattutto a livello tecnico, attraverso giustificabilità e contestabilità si giunge progressivamente a considerare l'IA come

un fattore (secondo alcuni un attore a pieno titolo) del sistema sociale, anche in ragione degli ambiti nella quale la si impiega.

5.4. Verso una IA non discriminatoria: l'Artificial Intelligence Act europeo

Le istituzioni europee hanno adottato regole ad hoc al fine di indirizzare lo sviluppo e l'implementazione di modelli e sistemi di intelligenza artificiale rispettosi dei diritti umani e non discriminatori, approvando l'*Artificial Intelligence Act* (Reg. UE 2024/1689 del 13 giugno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale), la prima legislazione organica sull'intelligenza artificiale al mondo. Nonostante normare uno sviluppo tecnologico in rapida evoluzione sia una sfida difficile, l'Unione Europea si trova dunque attualmente all'avanguardia nella creazione ed implementazione di un quadro regolamentare specifico.

L'iter comincia nel 2018 con l'istituzione del già menzionato Gruppo indipendente di esperti ad alto livello sull'intelligenza artificiale (AI HLEG), composto da esperti di tutti i paesi Ue negli ambiti dell'impresa, dell'accademia e della società civile, istituito dalla Commissione Europea per indicare le strategie da adottare

nell'approccio all'IA. Il gruppo ha prodotto alcuni documenti, tra cui innanzitutto le *Policy and Investment Recommendations for Trustworthy AI* e le *Ethics Guidelines for Trustworthy AI*. Queste guidelines individuano quattro principi etici – rispetto dell'autonomia umana, prevenzione dei danni, equità e spiegabilità – e sette requisiti per un'IA affidabile: intervento e sorveglianza umani; robustezza tecnica e sicurezza; riservatezza e governance dei dati; trasparenza; diversità, non discriminazione ed equità; benessere sociale e ambientale; accountability.

Sulla base di questo lavoro, nel 2020 la Commissione ha pubblicato un *White Paper on Artificial Intelligence*, chiedendo a vari stakeholder di fornire feedback in merito. Infine, ad aprile 2021, sempre la Commissione ha avanzato una proposta di Artificial Intelligence Act, con l'obiettivo di creare le condizioni per lo sviluppo e l'uso di sistemi di IA affidabili all'interno dell'Unione. Il 13 giugno 2024 è stato definitivamente approvato all'esito del processo legislativo del Consiglio e del Parlamento.

L'impostazione di fondo dell'AI Act risponde a quella della sicurezza dei prodotti e della conformità tecnica e all'approccio basato sul rischio. A tale impostazione si sono poi aggiunti nel corso dell'iter legislativo una normazione ad hoc dei modelli con finalità generali, che si sono fatti strada nello scenario evolutivo a novembre 2022 con la

presentazione di ChatGPT 3, e una serie di altri strumenti elaborati dal Parlamento fra cui va evidenziata la valutazione di impatto sui diritti fondamentali (c.d. FRIA). L'idea generale dell'AI Act resta quella di normare in maniera diversa i sistemi di intelligenza artificiale a seconda del rischio insito nel loro uso. Per quanto riguarda i principi che abbiamo precedentemente elencato, fairness, spiegabilità, giustificabilità e contestabilità le norme del Regolamento li richiamano a più riprese creando gli strumenti che dovrebbero consentirne la verifica.

Vengono definite le pratiche di IA vietate (art. 5), ovvero i sistemi a "rischio inaccettabile" e impieghi dell'IA che non possono neanche essere regolamentati perché ledono i diritti fondamentali delle persone. Rientrano in questi ambiti la manipolazione dei comportamenti e dei gruppi vulnerabili, il social scoring, i sistemi di polizia predittiva, i sistemi di riconoscimento delle emozioni e i sistemi di categorizzazione biometrica relativi a profili sensibili (come può essere appunto l'orientamento sessuale, ma anche la religione, l'orientamento politico, l'etnia, ecc.), sistemi di IA che fanno scraping non mirato di immagini per il riconoscimento facciale. Vale evidenziare tuttavia che le pratiche vietate hanno definizioni restrittive e prevedono significative eccezioni, in particolare in relazione all'uso da parte delle forze dell'ordine e per ragioni di sicurezza nazionale.

Ci sono poi i sistemi ad “alto rischio” (come quelli utilizzati nell’istruzione e formazione professionale, nel mondo dell’occupazione, gestione dei lavoratori e accesso al lavoro autonomo, nei servizi pubblici e privati essenziali, nel controllo delle frontiere o nell’amministrazione della giustizia) che dovranno rispettare degli obblighi molto strutturati relativamente – fra le altre cose – alla gestione del sistema, alla gestione dei rischi, alla governance dei dati, all’accuratezza, alla trasparenza, alla supervisione umana ed inoltre superare una valutazione di conformità *ex-ante*, essere sottoposti ad una valutazione di impatto sui diritti fondamentali prima di essere messi in campo, essere registrati in un database europeo e partecipare ad un monitoraggio *ex-post*.

Per i sistemi ad alto rischio si prevede che la valutazione di conformità possa far leva su standard tecnici armonizzati elaborati dalle organizzazioni di standardizzazione europee e approvati dalla Commissione Europea, come strumenti operativi di compliance con tutti o parte dei requisiti dell’AI Act.

Infine, sistemi a “basso rischio” o con nessun rischio non sono soggetti a controlli particolari, se non a un generico rispetto dei principi etici per una IA affidabile.

A prescindere dal livello di rischio i sistemi di IA che interagiscono direttamente con le persone o i cui output sono destinati ad

essere fruiti dal pubblico sono destinatari di obblighi di conoscibilità e trasparenza. In particolare si prevede che gli utenti siano informati quando interagiscono direttamente con un sistema di IA o quando sistemi di riconoscimento delle emozioni o di categorizzazione biometrica vengono usati nei loro confronti, che i contenuti audio, immagine, video o testuali sintetici siano marcati in modo da essere riconoscibili e che la manipolazione di contenuti mediante sistemi di IA sia resa nota agli utenti.

Nel corso dell’iter approvativo dell’AI Act, a novembre 2022, OpenAI ha rilasciato al pubblico ChatGPT (più esattamente la versione 3.0). Le grandi potenzialità e promesse di questo modello linguistico di grandi dimensioni ed il suo impatto sul grande pubblico hanno determinato una accelerazione nell’inserimento nel regolamento di una disciplina ad hoc per i modelli di IA per finalità generali, che oggi si trova al Capo V. A differenza degli altri algoritmi, i modelli con finalità generali vengono regolamentati a prescindere dagli ambiti d’uso.

In relazione a tutti i modelli ed i sistemi di IA è prevista l’adozione di codici di buone pratiche come strumento di co-regolazione e compliance volontaria.

Infine, si prevedono un obbligo di formazione del personale in relazione all’uso di strumenti di IA, degli spazi di

sperimentazione dei nuovi strumenti (c.d. sandbox normative) e un apparato di governance istituzionale molto strutturato.

Da un punto di vista della tutela dei diritti della comunità LGBTQIA+ e più in generale in una prospettiva di contrasto alle discriminazioni e promozione dell'inclusione, l'AI Act ha diversi meriti.

Innanzitutto alcune delle pratiche di IA proibite sono relative ad ambiti dove si consumano tipicamente discriminazioni a carico delle persone LGBTQIA+, come in particolare:

- il divieto di sistemi di IA per inferire le emozioni di una persona fisica nell'ambito del luogo di lavoro e degli istituti di istruzione;
- il divieto di sistemi di categorizzazione biometrica che classificano individualmente le persone fisiche sulla base dei loro dati biometrici per trarre deduzioni o inferenze in merito a razza, opinioni politiche, appartenenza sindacale, convinzioni religiose o filosofiche, vita sessuale o orientamento sessuale;
- il divieto di uso di sistemi di identificazione biometrica remota «in tempo reale» in spazi accessibili al pubblico per attività di polizia.

Anche gli usi ad alto rischio vanno ad identificare molteplici situazioni in cui sono registrate pratiche discriminatorie (cfr. All. III del regolamento) e in particolare:

- i sistemi di categorizzazione biometrica (ove consentiti);
- l'istruzione e la formazione professionale;
- l'accesso all'occupazione, la gestione delle risorse umane sia nell'ambito del lavoro dipendente sia autonomo;
- l'accesso a servizi privati essenziali e a prestazioni e servizi pubblici essenziali, incluse in particolare la sanità, l'attività creditizia e assicurativa;
- talune attività di polizia;
- le migrazioni, l'asilo e la gestione del controllo delle frontiere.

Benchè tutti tali ambiti presentino possibili elevate criticità in relazione al rischio che comportamenti discriminatori del passato vengano cristallizzati, replicati e amplificati mediante strumenti di IA, una attenta applicazione delle norme di compliance introdotte dall'AI Act (e in taluni casi già oggetto di tentativi di implementazione da parte di aziende e centri di ricerca) potrebbe rivelarsi uno strumento di fondamentale importanza per minimizzare il rischio di discriminazione.

Ciò anche perchè l'aver individuato presidi e responsabilità specifiche dovrebbe rendere più facilmente applicabile la normativa discriminatoria almeno negli usi ad alto rischio.

Alcuni strumenti sembrano poi offrire delle leve specifiche. La valutazione di impatto sui diritti fondamentali o FRIA (fundamental rights impact assessment), ad esempio, comporta un presidio specifico certamente applicabile alla discriminazione delle persone LGBTIQ+. I codici di buone pratiche possono supportare le aziende che intendono avviare processi di compliance prima dell'entrata in vigore dei singoli obblighi, così come le aziende che vogliono sviluppare o implementare sistemi di IA affidabile e non discriminatoria anche al di fuori degli ambiti di rischio elevato.

L'AI Act prevede anche la possibilità di trattamento di categorie particolari di dati (noti anche come dati sensibili) ove strettamente necessario per il rilevamento e la correzione di distorsioni e quindi discriminazioni pur con una serie di tutele specifiche.

Anche gli obblighi di formazione, la cui entrata in vigore è fissata al 2 febbraio 2025, saranno un utile momento di diffusione della consapevolezza sulle specificità dei sistemi di IA e un auspicabile veicolo per la mitigazione del rischio di discriminazione.

L'approccio dell'AI Act, assieme alle promesse di guidare l'AI verso l'affidabilità e il rispetto dei diritti, reca delle sfide importanti in termini di effettiva inclusione delle persone vulnerabili. Tipicamente, infatti, i gruppi vulnerabili hanno un minore accesso - sia in termini economici sia in termini di conoscenza - a tecnologie di avanguardia e una minore capacità di incidere sui processi decisionali di adozione.

Sarebbe dunque auspicabile che, nell'ambito delle strategie nazionali di supporto all'adozione dell'IA, venisse dato spazio, anche economico, ad iniziative destinate a supportare la partecipazione della società civile e dei gruppi vulnerabili ai processi di adozione.

Sarebbe inoltre importante che gli operatori economici e le istituzioni che procedono alla creazione ed all'adozione di strumenti e sistemi di IA definissero processi partecipativi per il design, il training e la verifica dei modelli e dei sistemi di IA, adattando tali strumenti alle diverse fasi del processo. Sul punto torneremo più avanti.

Un'altra sfida è insita nell'estensione dell'approccio al rischio tipico della sicurezza dei prodotti ad aspetti sociali, come il rispetto dei diritti e la non discriminazione. Si tratta innanzitutto di una sfida di grande complessità in primo luogo perché non è semplice adattare lo

strumentario tipico della conformità tecnica a profili sfaccettati e valutativi, definiti socialmente e culturalmente. In secondo luogo proprio perché il processo di definizione degli standard tecnici armonizzati e delle buone pratiche e la certificazione di conformità deve dimostrarsi in grado di individuare ed accogliere secondo un processo democratico gli aspetti sociali. In caso contrario si potrebbe dare il caso di certificare sistemi di IA che violano i diritti o comportano una discriminazione, rendendo così paradossalmente più difficile la tutela dei diritti fondamentali. A tal fine è necessario ancora una volta sollecitare e supportare la partecipazione al processo dei portatori di interessi.

Infine l'AI Act rendendo obbligatorio la conoscibilità, il tracciamento e la valutazione di una serie di variabili fondamentali nello sviluppo e nell'implementazione di sistemi di AI ad alto rischio, potrebbe rendere più chiaro il profilo delle responsabilità nei casi di discriminazione.

5.5. Il GDPR e il DSA

L'AI Act non è l'unico testo normativo dell'Unione Europea che incide sull'IA e prevede presidi rilevanti per la tutela contro le discriminazioni. Altri due strumenti meno recenti contengono norme di primario rilievo: il regolamento sulla

protezione dei dati personali noto come GDPR (Reg. (UE) 2016/679) e il regolamento sui servizi digitali noto come DSA, Digital Services Act (Reg. (UE) 2022/2065).

I dati sono una delle risorse fondamentali per lo sviluppo e l'uso dell'IA (vedi sezione 1.1.1) e il loro trattamento in tale contesto è una delle aree in cui possono verificarsi violazioni dei diritti fondamentali, incluso il diritto alla non discriminazione, o possono porsi le premesse per tali violazioni. Si tratta di aspetti che la disciplina per la tutela dei dati personali affronta da tempo, in Europa almeno dalla Direttiva 95/46/CE, oggi superata proprio dal GDPR.

Ad oggi si tratta della disciplina che ha più inciso, in ambito europeo, sulle pratiche di sviluppo dell'IA con importanti interventi delle Autorità per la protezione dei dati personali fra cui spiccano gli interventi del Garante italiano (Garante per la Protezione dei Dati Personali, 2022, 2023, 2024), le posizioni di quello irlandese e da ultimo le indicazioni del Comitato europeo per la protezione dei dati (European Data Protection Board, 2024).

Gli esperti che hanno partecipato ai webinar di EDGE ci hanno evidenziato come alcuni strumenti possano essere funzionali al contrasto alla discriminazione algoritmica. In particolare (i) l'art. 15 sul diritto di accesso dell'interessato, che consente di conoscere se vi sia un processo decisionale automatizzato ed avere informazioni significative sulla logica

utilizzata nonché quali siano l'importanza del trattamento e le sue conseguenze e (ii) l'art. 22 sui processi decisionali automatizzati relativi alle persone fisiche, che (a) consente in taluni casi di rifiutare un trattamento automatizzato, (b) prevede in ogni caso misure supplementari per la tutela dei diritti, l'intervento umano e la possibilità di contestazione della decisione (salvo che il trattamento automatizzato sia previsto per legge) e (c) esclude (salvo eccezioni) che i trattamenti automatizzati possano basarsi sui dati particolari, come quelli genetici e quelli relativi all'orientamento sessuale, la salute e la vita sessuale.

Il DSA è invece una normativa più recente che (assieme al DMA) mira a tutelare i diritti fondamentali degli utenti e a creare condizioni di parità per le imprese nello spazio digitale, inclusi, per quanto più rileva in questo report, i social network, i social media e le piattaforme di condivisione di contenuti.

Per quel che rileva qui, il DSA prevede:

- l'obbligo di valutazione del rischio e di adozione di misure di mitigazione
- la produzione di relazioni di trasparenza accessibili al pubblico sul modo in cui viene utilizzata la moderazione automatizzata dei contenuti online e sul suo tasso di errore

- l'armonizzazione delle risposte ai contenuti illegali online e misure di protezione contro gli abusi
- il divieto di pubblicità mirata che utilizza dati sensibili,
- maggiore trasparenza dell'utente sul loro flusso di informazioni, come le informazioni sui parametri dei sistemi di raccomandazione in un linguaggio chiaro e intellegibile,
- l'obbligo di informare i propri utenti delle decisioni di moderazione dei contenuti prese e a spiegarne le ragioni con motivazioni che devono essere inviate ad un database per la trasparenza, per migliorare la trasparenza e facilitare il controllo da parte della Commissione, della comunità scientifica e degli stakeholder,
- la possibilità per le autorità e per i ricercatori di accedere ai dati delle piattaforme per monitorare e valutare il rispetto della normativa, le prime, e per fini di ricerca, i secondi.

Alcune di tali misure sono limitate alle piattaforme online di dimensioni molto grandi e ai motori di ricerca molto grandi. Anche in questo caso sono previsti codici di condotta ed uno strutturato apparato di governance istituzionale e sanzionatorio.

Alcuni di questi strumenti potranno rivelarsi assai importanti in relazione all'uso di sistemi di intelligenza artificiale nel mondo digitale, laddove ciò impatta sui diritti delle persone e dei gruppi vulnerabili, come nel caso dell'hate speech e della disinformazione.

5.6. La partecipazione degli stakeholder come strumento di policy

Di fronte all'individuazione delle caratteristiche desiderabili per un'IA etica, quello normativo non è l'unico percorso ma vengono sperimentati anche altri strumenti.

Una strategia che può contribuire a migliorare la qualità etica degli algoritmi è il design partecipativo.

Questa metodologia di ricerca e sviluppo permette di coinvolgere una serie di soggetti indipendenti che, all'interno dei processi di un'impresa o di un'organizzazione, possano partecipare in maniera significativa e libera al processo di valutazione del funzionamento di un'IA che li riguarda, integrandoli nel gruppo dei designer e di chi implementa la tecnologia. Ad esempio, in questo ruolo potrebbero essere coinvolti i Business Resource Groups (BRGs) aziendali. Questo può

avvenire anche attraverso la condivisione di report sui potenziali effetti discriminatori del sistema in questione, o valutando i gruppi su cui il tool lavora. Si tratta di soluzioni già adottate da alcune aziende all'interno degli Ethics boards.

Policy migliori per quanto riguarda l'intelligenza artificiale e i diritti della comunità LGBTQIA+ passano anche da un maggior livello di consapevolezza su questi temi, tanto a livello di organizzazione quanto di società e politica.

In particolare in Italia, specie nell'ambito politico, vi è un approccio che predilige una lettura giuridica e fatica ad accogliere la cultura scientifica e tecnologica.

Per favorire il lavoro di legislazione e regolamentazione sull'uso etico del machine learning sarebbe importante e auspicabile incrementare il livello di conoscenza scientifica specifica sui temi dell'IA.

Inoltre consentire la partecipazione degli interessati e della società civile in via generale - al di fuori dei processi di sviluppo di specifici sistemi - richiederebbe strumenti di supporto alle organizzazioni del terzo settore per strutturare conoscenze e competenze tecniche e garantire la partecipazione democratica dei gruppi vulnerabili ai processi di elaborazione delle regole tecniche.

Bibliografia

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). ***Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. Proceedings of the ACM on human-computer interaction***, 3 (CSCW), 1-30.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). ***Machine Bias. ProPublica***. Recuperato da <https://www.propublica.org>

Boom dell'Intelligenza Artificiale: il mercato globale potrebbe sfiorare i 1000 miliardi entro il 2027 | Bain & Company

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). ***On the dangers of stochastic parrots: Can language models be too big?***. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Blease, C., Kaptchuk, T. J., Bernstein, M. H., Mandl, K. D., Halamka, J. D., & DesRoches, C. M. (2019). ***Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views***. Journal of medical Internet research, 21(3), e12802.

Bostrom, N. (2014). ***Superintelligence: Paths, Dangers, Strategies***. Regno Unito: Oxford University Press.

Buolamwini, J., & Gebru, T. (2018). ***Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency*** (pp. 77-91). PMLR.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). ***A clarification of the nuances in the fairness metrics landscape***. Scientific Reports, 12 (1), 4209.

Chouldechova, A. (2017). ***Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data***, 5 (2), 153-163.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). ***COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity***. Northpointe Inc.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., & Yona, G. (2021, June). ***Outcome indistinguishability***. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (pp. 1095-1108).

[The Economist. \(2017\). *Advances in AI are used to spot signs of sexuality*](#)

Esposito, E. (2022). ***Artificial communication: How algorithms produce social intelligence***. MIT Press.

European Commission: Directorate-General for Communications Networks, Content and Technology. (2019). ***Ethics guidelines for trustworthy AI***. Publications Office.

European Data Protection Board (18/12/2024). ***Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models***. EDPB.

[Parere 28/2024 su taluni aspetti relativi alla protezione dei dati ai fini del trattamento dei dati personali nel contesto dei modelli di IA | European Data Protection Board](#)

Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2022). ***A survey on bias in visual datasets***. *Computer Vision and Image Understanding*, 223, 103552.

Floridi, L. (2023). ***The ethics of artificial intelligence: Principles, challenges, and opportunities***.

Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; Schafer, B; Valcke, P. & Vayena, E. (2018). ***AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations***. *Minds and machines*, 28, 689-707.

FRA (2024). [LGBTIQ equality at a crossroads – Progress and challenges](#)

Floridi, L. (2024). [Intelligenza artificiale, perché è sbagliato umanizzare la macchina e computerizzare la mente](#)

Garante per la Protezione dei Dati Personali. (09/03/2022). [Riconoscimento facciale: il Garante privacy sanziona Clearview per 20 milioni di euro. Vietato l'uso dei dati biometrici e il monitoraggio degli italiani. GPDP.](#)

Garante per la Protezione dei Dati Personali (03/02/2023). [Intelligenza artificiale, dal Garante privacy stop al chatbot "Replika". Troppi i rischi per i minori e le persone emotivamente fragili. GPDP.](#)

Garante per la Protezione dei Dati Personali. (20/12/2024). [COMUNICATO STAMPA - ChatGPT. il Garante privacy chiude l'istruttoria. OpenAI dovrà realizzare una campagna informativa di sei mesi e pagare una sanzione di 15 milioni di euro. GPPD.](#)

Gartner. (11/11/2024). [Explore Beyond GenAI on the 2024 Hype Cycle for Artificial Intelligence.](#)

Harari, Y. N. (2018). **21 Lessons for the 21st Century**. Regno Unito: Random House.

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., & Ferrer, C. C. (2021). **Towards measuring fairness in ai: the casual conversations dataset**. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(3), 324-332.

Hébert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018, July). **Multicalibration: Calibration for the (computationally-identifiable) masses**. In International Conference on Machine Learning (pp. 1939-1948). PMLR.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). **Bias mitigation for machine learning classifiers: A comprehensive survey**. ACM Journal on Responsible Computing, 1(2), 1-52.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). **Deceiving google's perspective api built for detecting toxic comments**. arXiv preprint arXiv:1702.08138.

Imana, B., Korolova, A., & Heidemann, J. (2021, April). **Auditing for discrimination in algorithms delivering job ads**. In Proceedings of the web conference 2021 (pp. 3767-3778).

Ipsos. (28/08/2024). [Ipsos AI Monitor 2024: opinioni e atteggiamenti sull'Intelligenza Artificiale.](#)

ISO/IEC 24027:2021 **Bias in AI systems and AI aided decision making**

ISO/IEC 22989:2022 **Artificial intelligence concepts and terminology**

Kantaya, S. (Regista). (2020). **Coded bias** [Film]. 7th Empire Media.

Kärkkäinen, K., & Joo, J. (2019). **Fairface: Face attribute dataset for balanced race, gender, and age**. arXiv preprint arXiv:1908.04913.

Kartik, K. (2024, 27 settembre). ["The computer won't do that." Ada Lovelace Institute.](#)

Latour, B. (1992). ***Where are the missing masses? The sociology of a few mundane artifacts.*** *Shaping technology/building society: Studies in sociotechnical change*, 1, (pp. 225-258).

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). ***A proposal for the Dartmouth summer research project on artificial intelligence***, august 31, 1955. *AI magazine*, 27(4), 12-12.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). ***A survey on bias and fairness in machine learning.*** *ACM computing surveys (CSUR)*, 54(6), 1-35.

Nozza, D., Bianchi, F., & Hovy, D. (2021). ***HONEST: Measuring hurtful sentence completion in language models.*** In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Nozza, D., Bianchi, F., Lauscher, A., & Hovy, D. (2022). ***Measuring harmful sentence completion in language models for LGBTQIA+ individuals.*** In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Quintarelli, S., Corea, F., Ferrauto, C. G., Fossa, F., Loreggia, A., & Sapienza, S. (2020). ***Intelligenza artificiale. Cos'è davvero, come funziona, che effetti avrà.*** Bollati Boringhieri.

Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948). ***Sexual behavior in the human male.*** Saunders.

OECD (2024). ***Explanatory memorandum on the updated OECD definition of an AI system.*** *OECD Artificial Intelligence Papers*, 8.

Pant, A., Hoda, R., Spiegler, S. V., Tantithamthavorn, C., & Turhan, B. (2024). ***Ethics in the age of AI: an analysis of ai practitioners' awareness and challenges.*** *ACM Transactions on Software Engineering and Methodology*, 33(3), 1-35.

Pidoux, J. (2023). ***A comparative study of algorithmic-user classification practices in online dating: a human-machine learning process.*** *Porn Studies*, 10(2), 191-209.

PWC. (21/05/2024). *AI Jobs Barometer*. [AI Jobs Barometer | PwC](#)

Regolamento UE 1689/2024. **[Regolamento \(UE\) 1689/2024 che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti \(CE\) n. 300/2008, \(UE\) n. 167/2013, \(UE\) n.](#)**

[168/2013. \(UE\) 2018/858. \(UE\) 2018/1139 e \(UE\) 2019/2144 e le direttive 2014/90/UE. \(UE\) 2016/797 e \(UE\) 2020/1828 \(regolamento sull'intelligenza artificiale\).](#)

Samoili, S., López Cobo, M., Delipetrev, B., Martínez-Plumed, F., Gómez, E., and De Prato, G. (2021) **AI Watch. Defining Artificial Intelligence 2.0. Towards an operational definition and taxonomy for the AI landscape.** Publications Office of the European Union, Luxembourg

Schwartz, R., Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). **Towards a standard for identifying and managing bias in artificial intelligence (Vol. 3).** US Department of Commerce, National Institute of Standards and Technology.

Stypinska, J. (2022). **AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies.** *AI & society*, 38(2), 665-677.

[The Rome Call for AI Ethics. \(2020\). RenAIssance Foundation \(Vatican City\).](#)

(consultato il 19 ottobre 2024)

Valfort, M. (2017), **LGBTI in OECD Countries: A Review. OECD Social, Employment and Migration Working Papers, 198.** OECD Publishing, Paris (pp. 97-102)

Vicente, L., & Matute, H. (2023). **Humans inherit artificial intelligence biases.** *Scientific Reports*, 13 (1), 15737.

Wang, Y., & Kosinski, M. (2018). **Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.** *Journal of personality and social psychology*, 114(2), 246.

Wu, H. (2022). **Datafied Gay Men's Dating.**

EDGE

Il progetto “**A+I Algoritmi + Inclusivi**” di EDGE è stato coordinato da un team composto da **Mario Di Carlo, Alessandra Galli, Damiano Terziotti e Luca Trevisan**, che ci ha troppo presto lasciato nel 2024 e al quale EDGE ha dedicato il premio **A4Good** assieme alla famiglia di Luca e al Dipartimento di Computer Science dell'Università Bocconi.

Al progetto hanno collaborato, **Arsenico - La nuova comunicazione** per la creatività e il coordinamento di comunicazione, **The Fab Lab** per le interviste, **Artdisk | Design & Comunicazione Fluida** per la presentazione e l'impaginazione del report.

Ringraziamo per aver partecipato alle interviste **Massimo Airoldi, Luca Altieri, Luca Belli, Alessandro Bonaita, Silvia Franzeco, Luisella Giani, Dirk Hovy, Gianclaudio Malgeri, Pietro Monari, Debora Nozza, Dino Pedreschi, Stefano Quintarelli, Omer Reingold, Bruno Ronsivalle, Luca Trevisan** e i molti amici e amiche con cui abbiamo avuto la possibilità di confrontarci, con una menzione speciale per gli ospiti dei nostri Caffè Artificiali: **Brando Benifei, Federico Cabitza, Mia Caielli, Alessandro Castelnovo, Alessio De Luca, Davide Locatelli, Tommaso Mauro, Rosa Meo, Chiara Natali, Guido Noto La Diega, Ilaria Penco, Francesco Rizzi, Federico Sartore, Roberta Savella, Vincenzo Tiani, Roberto Trasarti**.

Un ringraziamento speciale a **Silvano Bertossa** per il prezioso supporto nella chiusura del report e l'enorme pazienza.

L'immagine utilizzata, con licenza Creative Commons (CC BY-NC-ND 4.0), fa parte del progetto creato da VICE «[The Gender Spectrum Collection](#)», una libreria fotografica di immagini che ritraggono persone transgender e non binarie in contesti di vita reali.