ALGORITMI + INCLUSIVI

Intelligenza artificiale e inclusione delle persone LGBTQIA+

Report abstract

EDGE

LGBTI+ LEADERS FOR CHANGE

Premessa 01
Il Report e questo abstract 02
I Bias Algoritmici 03
Gli usi dell'Intelligenza Artificiale e la discriminazione 05
Intelligenza Artificiale e regolazione normativa 06
Per un approccio etico orientato alla soluzione di problemi concreti 07

Premessa

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 o delle donne nelle graduatorie automatizzate o, ancora, degli appartenenti alle minoranze nella valutazione del rischio di recidiva in ambito penale.

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 Report e questo abstract

Il Progetto complessivo si chiama ***"A+I Algoritmi + inclusivi"***, che si articola in diversi filoni di attività e nell'ambito del quale abbiamo creato 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 un **Report** (che verrà pubblicato nel 2026 e di cui questo documento è una anticipazione sintetica dei contenuti) che prende le mosse da quelle interviste ma 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 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.

E non importa quanto accurato sia il risultato giacché l'impatto sulle persone deriva dall'uso dei sistemi.
Ma, appunto, gli aspetti potenzialmente problematici e potenzialmente innovativi sono legati dall'impatto dei sistemi di intelligenza artificiale.

I Bias Algoritmici

Certamente la riflessione non può che partire dal tema dei c.d. ***bias algoritmici***, espressione dalle sfumature diverse ma che può essere usata in senso esteso per indicare non solo i ***bias algoritmici*** in senso stretto ma in generale quei comportamenti dei sistemi di intelligenza artificiale 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.
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**.

Qui vale la pena ricordare che l'emersione amplificata dei ***bias*** veicolati mediante l'uso del linguaggio consente anche una 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)** della Commissione Europea nel 2025 indica i principali sottodomini 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 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 algorimici 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. 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 cd. ***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 all'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 valevoli 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 normativa

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, il **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, invece, 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 antidiscrimi-

natoria 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.
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 abstract ed il Report risultano 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*** dell'advertising algoritmico, 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 proteso 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

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 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+.**

Si tratta di un piano di azioni in costante evoluzione.

Il Report “A+I Algoritmi + Inclusivi” di EDGE

Questo abstract costituisce una sintesi ed una anticipazione del Report ***“A+I Algoritmi + Inclusivi”***, che verrà pubblicato nel corso del 2026.

Il lavoro si propone di fornire in apertura una introduzione ai concetti fondamentali necessari per comprendere l'ambito dell'Intelligenza Artificiale ed una “mappa” di orientamento fra i concetti essenziali sul rischio di discriminazione algoritmica.

Il Report illustra i principali rischi di discriminazione delle minoranze, e in particolare della minoranza LGBTQIA+, nell'ambito dell'IA fornendo esempi concreti e analizzando i principali problemi, a partire dall'effetto ***black box,*** e i principali campi di lavoro per la loro soluzione, a partire dalla ***fairness,*** per poi passare a delineare alcune tecniche di spiegabilità dell'IA e di riduzione dei ***bias*** tramite la selezione e manipolazione dei ***dataset***.
Infine, il tema della discriminazione viene esplorato in alcune applicazioni specifiche e di impatto come l'elaborazione del linguaggio naturale ed il riconoscimento automatizzato di immagini.

Il Report prende poi in esame il mondo dell'impresa, affrontando il tema della consapevolezza dei rischi di discriminazione algoritmica e dell'impegno nel ridurla ed evidenzia la necessità di una cultura diffusa delle questioni legate ai ***bias*** e alla discriminazione a vari livelli, da quello tecnico a quello amministrativo, sia nelle aziende che si occupano di produrre tecnologie di IA, sia in quelle che le utilizzano.
Risultano centrali in questa prospettiva il trattamento dei dati e i rischi legati alla profilazione così come alcuni rischi specifici nell'ambito del ***recruiting*** e della produzione di contenuti per i social media.

Il Report tenta, inoltre, di inquadrare il problema della discriminazione dalla prospettiva delle scienze umane e sociali a partire da una distinzione tra intelligenza artificiale ***debole*** e ***forte***, e tra le preoccupazioni di lunghissimo termine e un approccio più pragmatico che si concentra sui rischi immediati, attuali e concreti di discriminazione e deresponsabilizzazione e sulla datificazione dell'identità sessuale.

Indaga le caratteristiche più significative nel'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 torna sul concetto di ***bias***, sul ***trade off*** tra efficienza e spiegabilità e su una proposta di approccio integrato che interpreti l'apprendimento automatico come socializzazione delle macchine, e il ***bias*** in termini di ***habitus,*** oltre il punto di vista puramente quantitativo.
Il Report esamina poi le principali norme europee in materia di discriminazione e nell'ambito dell'IA, soffermandosi su GDPR, DSA e AI Act.

Le caratteristiche desiderabili per ottenere sistemi di IA che minimizzino il rischio di discriminazione vengono inquadrate nell'ambito dell'AI Act europeo individuando gli spazi di azione per la tutela dei diritti e l'inclusione delle persone LGBTQIA+.

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 ***AI4Good*** 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 Franzecco, 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.