

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata

Corso di Cultura Formazione e Società Globale

.

Tesi di Laurea Magistrale

Analisi critica dell'impatto dell'Intelligenza Artificiale sulle dinamiche sociali e di genere: tra disuguaglianze e sfide etiche

Critical Analysis of the impact of Artificial Intelligence on social and gender dynamics: between inequalities and ethical challenges

Relatore

Prof. Trappolin Luca

Laureanda

Del Serto Giulia

Matricola n° 2024525

Anno Accademico 2023/2024

Indice

Indice	I
Tabella delle Figure	III
Introduzione	VI
1. Intelligenza artificiale: cos'è e come funziona	1
1.1 Background Teorico.....	1
1.1.1 Storia dell'Intelligenza Artificiale.....	1
1.2 Machine Learning	10
1.2.1 Ciclo di vita di un sistema di Machine Learning	12
1.2.2 Diversi tipi di apprendimento	13
1.2.2.1 Apprendimento Supervisionato.....	13
1.2.2.2 Apprendimento Non Supervisionato.....	14
1.2.2.3 Apprendimento Ibrido	15
1.2.2.4 Apprendimento Rinforzato	15
2. Luoghi comuni e conoscenza ingenua: Stereotipi e Stereotipi di genere	17
2.1 Che cos'è uno stereotipo: origini e definizione del termine.....	17
2.1.1 La dimensione sociale dello stereotipo	19
2.2 Stereotipi, bias ed euristiche	23
2.3 Dallo stereotipo allo stereotipo di genere	25
2.4 Il genere come "performance" e l'intelligenza artificiale	28
3. Intelligenza discriminante: in quali modi il pregiudizio può annidarsi in un algoritmo?	33
3.1 Sessismo e Pregiudizio "Accidentale".....	34
3.1.1 Google Translate	35
3.1.2 Selezione del personale	36
3.1.3 Medicina Intelligente: Assicurazione e Prevenzione	39
3.1.4 Riconoscimento Fotografico: Nikon e Google Photo	41
3.2 Sessismo e Pregiudizio Volontario.....	43
3.2.1 BikiniOff: La Nascita del Fake Revenge Porn.....	43
3.2.2 Sicurezza e Controllo	47
3.2.3 Intelligenze artificiali generative senza restrizioni.....	52
4. Soluzioni e Proposte: Combattere i bias delle IA	59

4.1	Partecipazione integrazione e consapevolezza: Etica e alfabetizzazione tecnologica	59
4.1.1	Educazione: Save The Children e DIG4Future	59
4.1.2	Rappresentanza: UNESCO e Women4Ethical AI.....	63
4.1.3	Diversità: Women In AI	65
4.1.4	Legislazione: Unione Europea e AI Act	69
4.2	Ricerca e Sviluppo: Progetti per il futuro delle IA	71
5.1.1	Fonti Consapevoli: Dataset Equi e Rappresentativi	71
4.2.2	Indicatori Consapevoli: le Fairness Metrics	75
4.2.3	Addestramento Consapevole: Tecniche di Mitigazione	77
4.2.4	Controlli Consapevoli: Sistemi per Combattere le Discriminazioni	79
5.	Conclusioni	81
	Bibliografia	82

Tabella delle Figure

Figura 1 - Canard Digérateur, Jacques de Vaucanson (1764)	2
Figura 2- De moto animalium, Giovanni Alfonso Borelli (1685)	4
Figura 3 - The Turk, Wolfgang Von Kempelen (1770)	5
Figura 4 - Alan Turing a 16 anni (sinistra), Foto di Dominio Pubblico (1930) e Illustrazione del test di Turing (destra), A. Montanari, UniUd (2022).....	7
Figura 5 - Crescita del numero di transistor per processori e previsione di Moore, Wikipedia (2011).....	9
Figura 6 - Kasparov vs Deep Blue, Adam Nadel (1997).....	10
Figura 7 - Confronto tra neurone biologico e artificiale (sinistra), Adele Rispoli (2021) e Modello di Rete Neurale Artificiale (destra), ITimpresa (2020)	10
Figura 8 – Programma in Python per il calcolo della sequenza di Fibonacci, Giulia Del Serto (2023).....	11
Figura 9 - Ciclo di vita di un sistema di ML basato su apprendimento supervisionato, Giulia Del Serto (2023)	13
Figura 10 - Differenze di classificazione tra apprendimento supervisionato e non supervisionato, Edward Wu (2019)	14
Figura 11 - Funzionamento di un sistema di ML basato sull'apprendimento rinforzato, Shweta Bhatt (2018).....	15
Figura 12 - Torchio Tipografico, Technoratio (2010)	18
Figura 13 - Rappresentazione del personaggio Jim Crow, New York Public Library (1836)	20
Figura 14 - Giocattoli per maschi e femmine, La Voce di Bolzano (2019)	26
Figura 15 - Scatola di battaglia navale con sfondo sessista, Milton Bradley (1967).....	26
Figura 16 - Esempi di risposte di ChatGPT e il loro effetto sull'utente, Gross (2023).....	31
Figura 17 - Template di registrazione di utente, Codepen.io (2022).....	34
Figura 18 - Esempi di traduzioni sessiste da parte di Google Translate, BoredPanda (2021)	35
Figura 19 - Prima e dopo: traduzione gender specific di Google Translate, The Telegraph (2018)	36
Figura 20 - Estrazione delle caratteristiche principali di un volto, MIT (2022).....	37
Figura 21 - Estrazione delle caratteristiche principali da un Curriculum, Del Serto (2023).....	37
Figura 22 - Impatto dell'etnia del paziente in diversi ambiti medici, NIHCM (2021).....	40
Figura 23 - Riconoscimento errato di occhi socchiusi (sinistra), Joz Wang (2009) e Classificazione razzista di Google Photo (destra), Twitter (2015).....	41
Figura 24 - Distribuzione di età, colore della pelle e genere nel database di ImageNet, Yang et al. (2020)	43
Figura 25 - Utilizzo dell'AI in Photoshop (sopra), Adobe (2023) e Modifica espressioni facciali, DragGan.ai (2023).....	44
Figura 26 - Realmente falso: Bikinioff, l'intelligenza artificiale che spoglia le donne, Red Hot Cyber (2023)	45
Figura 27 - Infografica sulle vittime di Revenge Porn, endrevengeporn.org (2013).....	46
Figura 28 - Differenza fisionomica tra un cinese di etnia Han e uno di etnia Uigura, IPVM (2019)	48
Figura 29 - Pubblicità sulla profilazione di un sistema di sorveglianza cinese, New York Times (2019)	49
Figura 30 - Digidog che pattuglia New York (sinistra), Youtube (2021) e robot militare della Ghost Robotics (destra), New York Post (2022)	50
Figura 31 - K5 mentre pattuglia la metropolitana di New York, New York Times (2023).....	50
Figura 32 - Codice etico di ChatGPT, ChatGPT (2023).....	52

Figura 33 - Donald Trump viene arrestato dalla polizia di New York (sinistra), BBC (2023) e Putin si inginocchia davanti a Xi Jinping (destra), Radio Free Asia (2023).....	54
Figura 34 - False locandine di film Disney Pixar sull'olocausto (sinistra) e sulla morte di George Floyd (destra), X (2023)	55
Figura 35 - Risultato di diversi motori di AI per la creazione di immagini dato l'input "Winnie the pooh as an SS-Einsatzgruppen commander visiting Sachsenhausen. Black and white.", Bellingcat (2023).....	55
Figura 36 - "Raggiro della nonna" per farsi spiegare come fare un cappio con ChatGPT, Giulia Del Serto (2023).....	57
Figura 37 – Istruzioni del Gioco del Tris del Foglio Intelligente	60
Figura 38 - Classificazione con Teachable Machine, Google (2017)	61
Figura 39 - Classificazione del Rischio nell'AI Act,.....	69
Figura 40 - Esempio di Dataset Sbilanciato, Frederik Hvilshøj (2022)	71
Figura 41 - Arricchimento di un Dataset, Frederik Hvilshøj (2022).....	72
Figura 42 - Esempio di Undersampling, Frederik Hvilshøj (2022).....	73
Figura 43 - Esempio di Oversampling, Frederik Hvilshøj (2022)	74
Figura 44 - Bilanciamento dei risultati di ImageNet, Yang et al. (2020)	75
Figura 45 - Lista delle Caratteristiche Sensibili, Wolsingham School (2022)	78
Figura 46 - Schema di un Generative Adversarial Network, Google (2022).....	80

Abstract

L'impatto dell'intelligenza artificiale (IA) ha rivoluzionato diversi aspetti della nostra vita quotidiana, tra cui salute, sicurezza e sicuramente la fruizione dell'informazione. Purtroppo, i benefici apportati da tali avanzamenti tecnologici vanno di pari passo ad una crescente preoccupazione riguardo l'aspetto etico di tali sistemi: uno dei dilemmi più noti riguarda la discriminazione di genere perpetuata dall'AI, non solo in modo casuale ma anche e, soprattutto, volontariamente. Questa tesi intende analizzare l'impatto e le conseguenze della discriminazione di genere nei sistemi di intelligenza artificiale mediante una revisione della letteratura disponibile, saranno esaminate le evidenze di disparità di genere in diversi contesti applicativi, dai processi decisionali automatizzati nelle risorse umane ai suoi utilizzi nella sorveglianza e propaganda politica. L'analisi si concentrerà sugli aspetti etici connessi a questa forma di discriminazione, ponendo in evidenza, come, le conseguenze negative possano riflettersi sulla società, rafforzando stereotipi e ingiustizie preesistenti. In questo testo verranno analizzati casi studio per valutare la presenza di bias di genere e tentare di identificarne le cause sottostanti. Il fine è quello di proporre soluzioni concrete anche in ambito umanistico e pedagogico, verranno esplorate le iniziative esistenti e le prospettive future per garantire la costruzione di sistemi di intelligenza artificiale eticamente inclusivi.

Introduzione

L'intelligenza artificiale si occupa di studiare come realizzare dei sistemi informatici che emulino il più possibile il ragionamento umano. Si tratta di una disciplina in crescente espansione e oggi, più che mai, richiede consapevolezza e informazione. L'intelligenza artificiale promette rivoluzioni e miglorie in svariati settori umani, ma è importante riconoscerne le criticità. Questa tesi si propone di analizzare le possibili conseguenze negative di un'intelligenza artificiale allenata su algoritmi tendenziosi e destinati a riproporre pregiudizi e stereotipi di genere presenti nella società.

Attraverso un'analisi approfondita delle cause e degli effetti di questa perpetuazione di stereotipi, si intende evidenziare le conseguenze sociali, culturali ed economiche di tali sistemi discriminatori. Questa ricerca si propone di incentivare la diffusione di algoritmi etici e inclusivi, e favorire una responsabilità etica nell'utilizzo dell'IA, al fine di realizzare una maggiore giustizia ed equità nei sistemi decisionali automatizzati.

La tesi si comporrà di un inquadramento teorico di base, che ripercorre le tappe in cui l'intelligenza artificiale si è sviluppata, partendo da Cartesio fino al più recente concetto di Maching Learning su cui si basano i modelli di intelligenza artificiale attuali. Al proposito è molto interessante constatare come l'intelligenza artificiale abbia origini antiche poichè, da sempre, l'uomo si è dimostrato interessato allo sviluppo di "macchine" in grado di replicare l'attività umana.

I capitoli che ricostruiscono il funzionamento dell'intelligenza artificiale dal punto di vista tecnico e, pertanto, informatico, saranno susseguiti da nozioni teoriche di base che possano agevolare la costruzione di un discorso etico relativo alle possibili conseguenze negative generate dall'intelligenza artificiale. Verranno affrontati i concetti di stereotipo e stereotipo di genere, esaminando i contributi della psicologia sociale nell'indagine di come si formino. Si vedrà come lo stereotipo sia spesso causato da bias ed euristiche cognitive messe in atto dal nostro cervello per economizzare le risorse di cui disponiamo. Inoltre, la socializzazione di genere che riceviamo

sin dall'infanzia contribuisce a rafforzare immaginari e aspettative sociali rispetto al sesso di appartenenza. In questo dibattito l'intelligenza artificiale ha il potere di agire come deterrente rispetto a queste rappresentazioni binarie che generano una distribuzione iniqua del potere. Purtroppo spesso non accade quanto auspicato, poiché l'intelligenza artificiale perpetua gli stereotipi già presenti nella società.

Verranno dunque affrontate le modalità in cui il pregiudizio può annidarsi nell'algoritmo di riferimento, a volte in modo accidentale ma, molto più frequentemente in modo mirato e intenzionale, contribuendo a fenomeni drammatici come il "fake revenge porn".

Diviene sempre più urgente disciplinare l'utilizzo delle intelligenze artificiali accessibili al grande pubblico, come quelle generative. Infine, le soluzioni proposte in materia verranno esaminate sia dal punto di vista umanistico che da una prospettiva più tecnica, di settore.

Una maggiore consapevolezza e alfabetizzazione tecnologia diventano, in questo momento storico, strumenti essenziali per tutelarsi dai limiti di queste tecnologie, ancora in via di sviluppo.

Questa tesi nasce da un dibattito tra due mondi radicalmente diversi, il mio, quello umanistico e quello del mio compagno, informatico.

Entrambi ci siamo resi conto di quanto, nonostante la fiorente letteratura in merito, mancasse una prospettiva pedagogica nell'affrontare queste tematiche. Ma ancor di più, di quanto fosse essenziale rendere i meccanismi sottostanti l'AI alla portata di quante più persone possibile, soprattutto quelle che, come me, non avevano sufficienti strumenti informatici di base per capire come funzionasse il modello di un algoritmo.

I problemi generati dell'AI sono causati in parte dall'esclusione di molte persone che non possiedono le nozioni di base per partecipare al dibattito e questo può imputarsi anche dalla scarsa letteratura in merito, che provenga da una prospettiva etica.

1. Intelligenza artificiale: cos'è e come funziona

La tematica dell'intelligenza artificiale è oggi sempre più dibattuta. Ne trattano ad ampio spettro la comunità scientifica ma anche le discipline umanistiche. Un approccio etico e umanistico, è infatti alla base dell'indagine sulla legittimità dell'utilizzo dei dati per la formulazione degli algoritmi alla base dell'intelligenza artificiale. L'intelligenza artificiale pervade la nostra vita quotidiana, e l'inestimabile risorsa che rappresenta per l'evoluzione della società moderna è ormai cosa nota (si pensi alle promettenti applicazioni in campo medico e industriale). Ma la mancanza di norme trasparenti che ne regolino la portata, ha suscitato il sospetto e l'inquietudine di molti.

Per far luce sulla comprovata e preoccupante perpetuazione di stereotipi di genere e discriminazione dell'AI all'interno della nostra società, è opportuno tentare di comprenderne il funzionamento.

Questo primo capitolo si propone di fornire al lettore un'introduzione al tema, analizzandone storia ed origine, per giungere ai complessi campi di applicazione di questa tecnologia.

1.1 Background Teorico

1.1.1 Storia dell'Intelligenza Artificiale

Nonostante di Intelligenza Artificiale si inizi a parlare formalmente solo agli inizi degli anni '50 con Alan Turing, l'idea di "macchine" che svolgono attività compiute fino a quel momento dagli esseri umani, risale agli albori della prima rivoluzione industriale.

XV Secolo - Renato Cartesio

Ne è una testimonianza il "Discorso sul metodo" (Descartes, 1637) del filosofo e matematico Cartesio, risalente al 1637. Il Discorso si compone di sei diverse parti che analizzano le poliedriche conoscenze dell'autore: le prime tre si concentrano sulla sua carriera tecnica, con specifico riferimento all'educazione razionale e le regole del suo metodo cartesiano.

La seconda metà del Discorso è dedicata a concetti filosofici, a tratti teologici. In questa parte Cartesio espone la sua metafisica, definendo la natura dell'atto di pensare delle idee (cogito ergo sum), strumento ingannevole di rappresentazione della realtà.

Nell'ultima sezione, la quinta, l'autore analizza le differenze tra uomini, macchine e animali, approfondendo il concetto di funzionamento organico del corpo dell'essere umano.

Secondo Cartesio, le qualità che in definitiva distinguono l'uomo dalle macchine e dagli animali sono la padronanza del linguaggio e l'uso della ragione. Cartesio asserisce infatti, che, animali e macchine, non possiedono un vero e proprio strumento di comunicazione, ma interagiscono tra loro attraverso spasmi involontari, "movimenti naturali".

Egli afferma che gli animali sono essenzialmente macchine biologiche, con organi e tessuti che, lavorando in modo cooperativo e coerente, permettono il funzionamento del loro corpo. Le macchine funzionano attraverso il medesimo meccanismo: le parti meccaniche di cui sono costituite, lavorano insieme in modo coerente per eseguire funzioni specifiche.

Cartesio sostiene che una mente razionale e conscia sia prerogativa esclusiva degli esseri umani. Sostiene inoltre che il ragionamento consapevole è ciò che permette agli esseri umani di comprendere e controllare il mondo naturale, tesi che sposa a pieno la visione abramo-centrica della chiesa cattolica.

Tuttavia, l'autore riconosce che animali e macchine sono in grado di svolgere compiti che richiedono una certa forma di intelligenza: gli animali sono in grado di rispondere ai loro istinti e imparare attraverso l'esperienza, mentre le macchine possono essere programmate per eseguire attività complesse.

Con il suo Discorso, Cartesio congiunge animali e macchine attraverso un'analisi critica che mette in discussione la visione tradizionale secondo cui gli animali sono dotati di un'anima irrazionale e le macchine sono invenzioni umane prive di qualsiasi forma di vita o

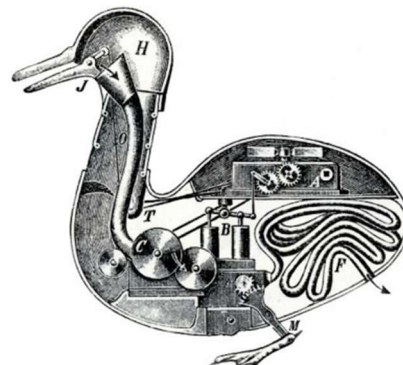


Figura 1 - Canard Digérateur, Jacques de Vaucanson (1764)

intelligenza. Entrambi agiscono attraverso meccanismi fisici, ma, nonostante solo gli esseri umani siano dotati di una mente razionale, animali e macchine sono in grado di svolgere compiti che richiedono, loro modo, una certa forma di intelligenza.

“Qui in particolare mi ero fermato per far vedere che se ci fossero macchine con organi e forma di scimmia o di qualche altro animale privo di ragione, non avremmo nessun mezzo per accorgerci che non sono in tutto uguali a questi animali; mentre se ce ne fossero di somiglianti ai nostri corpi e capaci di imitare le nostre azioni per quanto di fatto possibile, ci resterebbero sempre due mezzi sicurissimi per riconoscere che, non per questo, sono uomini veri. In primo luogo, non potrebbero mai usare parole o altri segni combinandoli come facciamo noi per comunicare agli altri i nostri pensieri. Perché pur nel concepire che una macchina sia fatta in modo tale da proferire parole, e ne proferisca anzi in relazione a movimenti corporei che provochino qualche cambiamento nei suoi organi; che chieda, ad esempio, che cosa si vuole da lei se la si tocca in qualche punto, o se si tocca in un altro gridi che le si fa male e così via; ma non si può immaginare che possa combinarle in modi diversi per rispondere al senso di tutto quel che si dice in sua presenza, come possono fare gli uomini, anche i più ottusi. L'altro criterio è che quando pure facessero molte cose altrettanto bene o forse meglio di qualcuno di noi, fallirebbero inevitabilmente in altre, e si scoprirebbe così che agiscono non in quanto conoscono, ma soltanto per la disposizione degli organi.” (Descartes, 1637, p. 155)

Le posizioni di Cartesio verranno, successivamente, riviste dall'autore stesso, in particolare in riferimento alla natura “meccanica” degli animali. Nella sua tarda età, trasferitosi a Stoccolma, il filosofo era solito passeggiare sulle spiagge svedesi dove puntualmente incontrava un cane randagio. Notando come il cane lo seguiva fedelmente ogni giorno, Cartesio vi scorse un'emozione che fino a quel momento credeva essere prerogativa umana e non poté fare a meno di rivisitare la sua teoria con una meno radicale. Riconobbe quindi la possibilità, anche per gli animali, di provare sentimenti, sensazioni di piacere o dolore. Gli attribuì dunque, una forma di coscienza.

XVII Secolo – il Meccanicismo

Il meccanicismo è una teoria filosofica che si propone di spiegare tutti i fenomeni della realtà, tramite una lettura di tipo materialistico. Essa, infatti, interpreta le proprietà degli oggetti e dei processi del mondo fisico, adottando un'ottica esclusivamente meccanica, ossia come prodotti di una causalità deterministica (Laplace, 1796). La teoria così ponendosi, nega l'esistenza di finalità superiori e riduce la complessità del mondo reale, arginando la possibilità di dinamiche dialettiche e processi retroattivi al fenomeno di

riferimento. La nascita del meccanicismo è stata influenzata da molteplici elementi, tra cui la rivoluzione scientifica del XVII secolo, l'opera di Descartes e Galileo, la diffusione della scienza sperimentale e il progresso tecnologico.

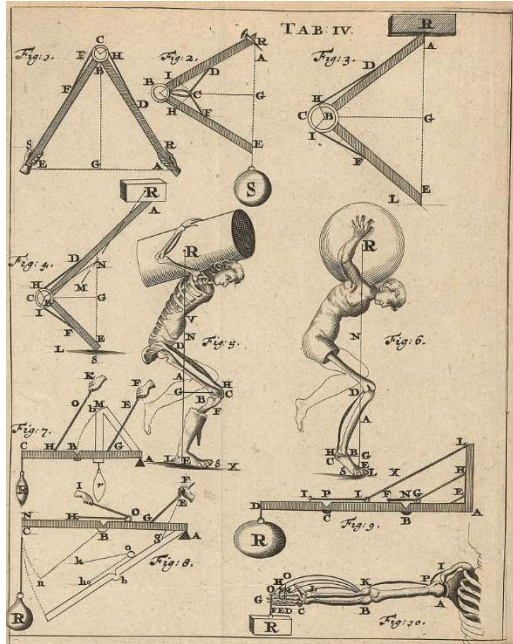


Figura 2- *De motu animalium*, Giovanni Alfonso Borelli (1685)

In particolare, la rivoluzione scientifica ha portato alla diffusione di un nuovo modo di pensare basato su osservazione, sperimentazione e razionalità, che ha messo in discussione la precedente visione del mondo tradizionale. Galileo e Descartes hanno contribuito alla nascita del meccanicismo sviluppando rispettivamente il metodo scientifico e la filosofia del dualismo, che separa mondo fisico da mondo mentale (Israel, 2015).

In aggiunta ai fattori sociali, culturali, scientifici e tecnologici che hanno contribuito alla nascita del meccanicismo, è possibile individuare alcune motivazioni più specifiche che hanno spinto filosofi e pensatori dell'epoca, a sviluppare questa teoria:

- In primo luogo, il meccanicismo rispose al bisogno di trovare una spiegazione razionale e scientifica del mondo naturale, che andasse al di là delle interpretazioni mitologiche e religiose tradizionali. Il meccanicismo ha rappresentato un'alternativa razionale e oggettiva alle concezioni del mondo basate su fede e tradizione.
- In secondo luogo, il meccanicismo è stato motivato dalla necessità di trovare una risposta ai problemi tecnici e ingegneristici dell'epoca, che richiedevano una conoscenza sempre più precisa della natura dei materiali e delle leggi che li governavano. La comprensione delle leggi meccaniche ha permesso di sviluppare macchine sempre più complesse e sofisticate, come orologi, locomotive e motori a vapore.
- Infine, il meccanicismo ha avuto un ruolo importante nella definizione della visione antropocentrica dell'uomo, che lo ha posto al centro dell'universo come entità razionale e autonoma. Questa concezione ha influenzato anche la teoria dell'evoluzione, che ha

visto nella selezione naturale una forma di "meccanismo" che agisce sulla vita degli organismi.

L'acquisizione di una visione meccanica, prima degli animali con Cartesio, poi dell'uomo stesso con il meccanicismo; aumenta il fascino delle persone rispetto alla possibilità che ciò che è sempre stato considerato ad appannaggio esclusivamente umano, possa essere in realtà eseguito da una macchina. Ne è esempio (truffaldino) l'attrattiva generata dal Turco. Il trucco fu un'abile truffa che ebbe luogo tra il XVIII e il XIX secolo (Poe, 1836).

Può essere ragionevolmente considerato il primo "autonoma" e aveva il compito di intrattenere il pubblico simulando un giocatore di scacchi, inganno in realtà ideato da un abile illusionista.

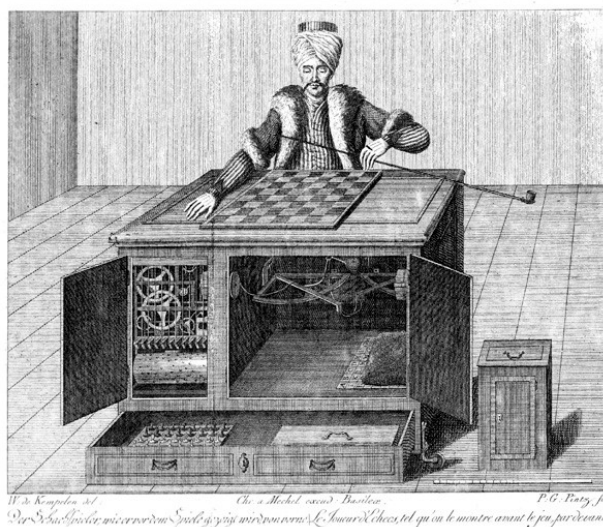


Figura 3 - The Turk, Wolfgang Von Kempelen (1770)

L'automa venne presentato al pubblico europeo dal 1770 circa, suscitando grande interesse e ammirazione. Il "Turco" era composto da una scacchiera su cui erano posizionate le pedine, una figura umana vestita con turbante e mantello, che sedeva dietro la scacchiera, e una serie di meccanismi e ruote dentate che azionavano le mosse delle pedine. L'automa fu presentato in numerose città europee e riscontrò grande successo.

In realtà, l'automa era controllato da un uomo nascosto all'interno della struttura, che osservava la scacchiera attraverso un sistema di specchi e movimentava le pedine tramite un meccanismo di leve e cavi. L'inganno fu celato per molti anni, grazie all'abilità dell'illusionista e all'ingegno del

sistema di specchi che rendeva difficile individuare la presenza dell'uomo all'interno della struttura.

XX Secolo -- Alan Turing

Durante la seconda guerra mondiale, Alan Turing ha ricoperto un ruolo fondamentale nella decodifica dei messaggi segreti dei nazisti, utilizzando la sua abilità matematica e informatica.

Turing è stato uno dei principali ideatori della macchina crittografica chiamata "B.O.M.B.A" implementata al fine decodificare i messaggi cifrati della macchina Enigma, utilizzata dalle forze armate tedesche. Il suo lavoro e quello del team di Bletchley Park hanno avuto un impatto significativo sulla guerra, permettendo agli Alleati di decifrare i messaggi nemici e pianificare con successo diverse operazioni (Hodges, Alan Turing: The Enigma, 1983).

Turing ha sviluppato nel tempo un metodo per decodificare i messaggi Enigma in modo ancor più efficiente, creando una versione migliorata della macchina Bomba. Inoltre, ha anche lavorato sul sistema di crittografia tedesca, cercando di individuarne eventuali punti deboli e sviluppare nuove tecniche per decodificarli.

Il lavoro di Turing e del suo team a Bletchley Park è stato per lungo tempo tenuto al segreto, dopo la fine della guerra, ma alla fine la loro importanza è stata riconosciuta fondamentale alla vittoria degli Alleati. Tuttavia, nonostante il suo indispensabile aiuto, Turing è stato perseguitato per la sua omosessualità e successivamente condannato nel 1952.

Alan Turing è a pieno titolo una tra le figure più importanti relativamente la nascita dell'intelligenza artificiale, si deve molto alla sua visione pionieristica e alle ricerche intraprese sulla teoria della computazione.

Il concetto di "macchina di Turing" è uno tra i suoi contributi più importanti. Si tratta di un modello teorico di computer che ha costituito la base per lo sviluppo dei moderni calcolatori digitali. Tale modello ha permesso di definire in modo formale il concetto di "algoritmo": una sequenza finita di istruzioni che possono essere eseguite da una macchina di Turing.

Inoltre, Turing ha sviluppato il concetto di "intelligenza artificiale", esplorando le possibilità derivanti dalla creazione di macchine in grado di imitare il pensiero umano e risolvere problemi autonomamente.

Suo esemplare contributo, il celebre "Test di Turing" (Hodges, Alan Turing and the Turing Test, 2008), in grado di valutare la capacità di una macchina, di interagire con un essere umano in modo indistinguibile da un altro essere umano. L'idea alla base è quella di far interagire un essere umano con una macchina, attraverso una chat testuale, senza che l'umano sappia se sta comunicando con un'altra persona o con una macchina. Se la macchina riesce a convincere l'umano di essere come lui, allora si può affermare che la macchina abbia superato il Test.

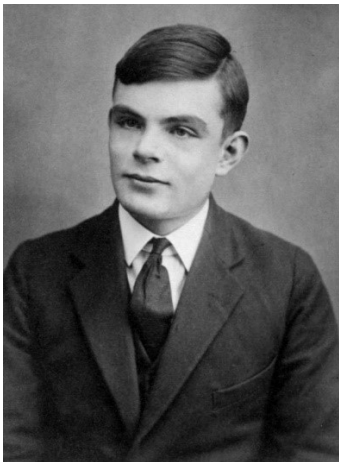


Figura 4 - Alan Turing a 16 anni (sinistra), Foto di Dominio Pubblico (1930) e Illustrazione del test di Turing (destra), A. Montanari, UniUd (2022)

Il Test di Turing ha avuto grande impatto sulla ricerca sull'intelligenza artificiale, poiché ha evidenziato la capacità delle macchine di simulare l'intelligenza umana. Tuttavia, ha anche sollevato importanti questioni filosofiche ma soprattutto etiche sulla natura dell'intelligenza e sulla possibilità che le macchine possano sviluppare una coscienza.

Negli anni successivi la sua formulazione, il Test di Turing è stato oggetto di numerosi dibattiti e critiche, che ne hanno evidenziato limitazioni e ambiguità. È stato sottolineato che il Test di Turing non può garantire la misurazione della vera intelligenza, ma solo della capacità di simulare

un comportamento umano. Inoltre, è difficile definire in modo preciso cosa si intenda per "comportamento umano", dal momento che le risposte dei partecipanti umani possono essere altamente variabili. Nonostante le sue limitazioni, il Test di Turing rimane uno dei concetti perno dell'intelligenza artificiale, con conseguente impatto significativo sulla ricerca e sviluppo di chatbot e assistenti virtuali.

XX – XXI Secolo -- Evoluzione fino ad oggi

Nel **1956**, durante la conferenza di Dartmouth, viene per la prima volta utilizzato il termine Intelligenza Artificiale dal ricercatore John McCarthy. L'espressione segna l'inizio di un nuovo campo di ricerca che ha come scopo quello di "esaminare la congettura per cui ogni aspetto dell'apprendimento e ogni altra caratteristica dell'intelligenza possa essere, in linea di principio, descritto in modo tanto preciso da poter far sì che una macchina lo simuli" (McCarthy, Minsky, Rochester, & Shannon, 1955).

L'obiettivo dei ricercatori è permettere che le macchine utilizzino linguaggio, concetti e astrazioni non solo per risolvere problemi fino a quel momento riservati agli uomini, ma per migliorarsi, implementarsi, imparando dai propri errori.

La velocità computazionale dei computer dell'epoca non era in grado di sviluppare software intelligenti in grado di risolvere problemi reali. Perciò, a queste proto-IA venivano somministrati giochi come puzzle, dama, scacchi o problemi semplici (problemi giocattolo) che ne dimostrassero più che altro la validità potenziale (Nilsson, 2001).

In quegli stessi anni lo psicologo Frank Rosenblatt sviluppa Perceptron: un programma per riconoscere le immagini che applica le teorie di McCulloch e Pitts e consiste in una rete neurale artificiale. Nonostante le alte aspettative sui risultati di Perceptron, i limiti tecnologici dell'epoca delusero la comunità scientifica e per questo motivo lo studio delle reti neurali subirà una battuta d'arresto per quasi 20 anni.

Tra gli anni '80 e '90 l'aumento vertiginoso della capacità di calcolo dei computer comporta ravvivato interesse nel campo delle intelligenze artificiali, permettendo di superare le barriere tecnologiche degli anni precedenti. In particolare le IA vengono addestrate a risolvere problemi

più complessi come l'interpretazione del linguaggio naturale, il riconoscimento visivo delle immagini e la rappresentazione generale della realtà; ne sono alcuni esempi notevoli:

- Lo Stanford Cart: drone dotato di telecamere che in alcune ore riesce ad attraversare con successo una stanza piena di sedie (Earnest, 2012)
- DENDRAL: sistema utilizzato per mappare la struttura delle molecole, e aiutare i chimici a identificarne di organiche sconosciute (Lavecchia, 2017)
- R1: macchina che operava in ambito commerciale per configurare gli ordini di nuovi computer. Grazie alla sua applicazione, l'azienda sviluppatrice dell'IA fu in grado di risparmiare svariati milioni di dollari (McDermott, 1982)

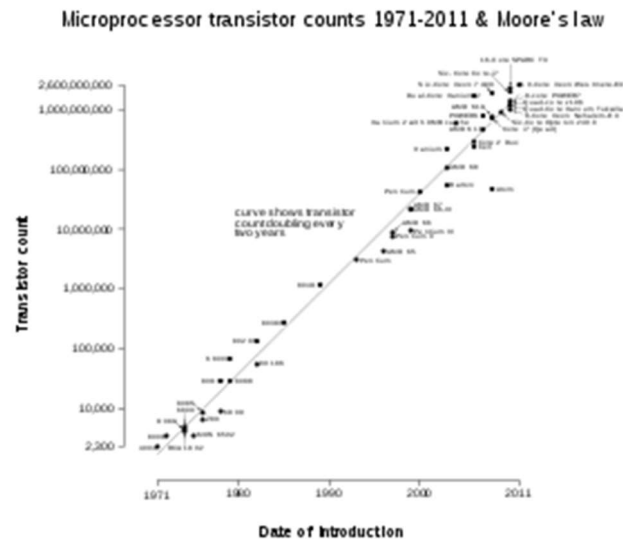


Figura 5 - Crescita del numero di transistor per processori e previsione di Moore, Wikipedia (2011)

L'esempio più famoso di questa tecnologia risale però al **1997**. In quell'anno IBM perfeziona lo sviluppo di Deep Blue, una macchina giocatrice di scacchi, in grado di calcolare tra 100 e 200 miliardi di posizioni per turno.



All'AI viene contrapposto l'allora campione del mondo Garry Kasparov il quale, dopo poco più di un'ora si arrende al suo avversario. Oltre l'indubbio risultato tecnologico ottenuto,

Figura 6 - Kasparov vs Deep Blue, Adam Nadel (1997)

sconfitta del campione russo si pone anche come grosso smacco morale verso molti detrattori della tecnologia utilizzata da Deep Blue. Ancora una volta, infatti, viene dimostrato come, attività di ragionamento associate fino a quel momento ai soli esseri umani, possano essere insegnate e perfezionate da una macchina. Lo stesso Kasparov, prima di sedersi davanti al suo sfidante enuncia la sua presenza in "difesa della razza umana" sottolineando l'importanza di quella sfida. Dopo la sconfitta il campione, incredulo del risultato, arriverà ad accusare IBM di aver affiancato a Deep Blue esperti di scacchi umani, collegati a distanza (Sabadin, 2017).

Negli ultimi 20 anni l'importanza e pervasività delle intelligenze artificiali ha contagiato ogni aspetto della vita quotidiana. Per questo motivo nei capitoli successivi verranno analizzati in dettaglio alcuni tra i più rilevanti esempi odierni di applicazione dell'intelligenza artificiale, in particolare in relazione alla disparità di genere.

1.2 Machine Learning

L'apprendimento automatico è una branca dell'intelligenza artificiale che consente ai computer di "auto-apprendere" dai dati di addestramento e migliorare nel tempo, senza essere esplicitamente programmati. Gli algoritmi di apprendimento automatico sono in grado di rilevare modelli nei dati e imparare da essi, al fine di attuare le proprie previsioni.

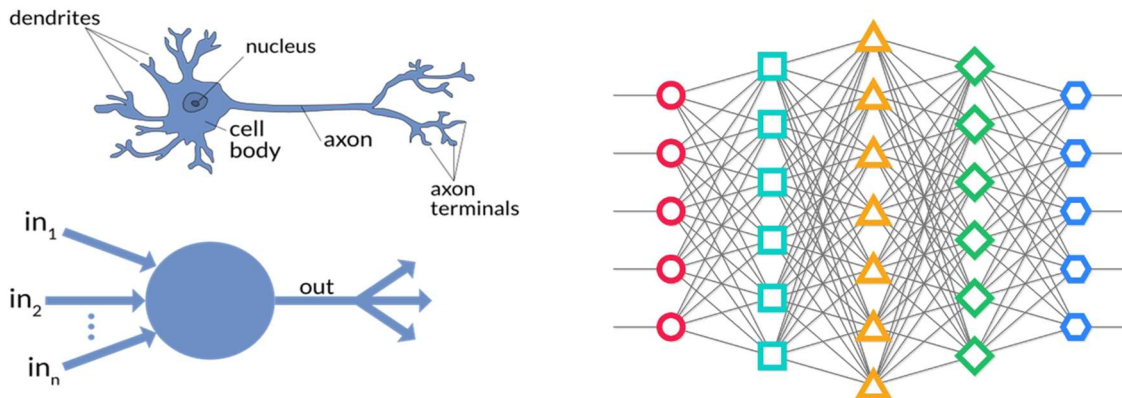


Figura 7 - Confronto tra neurone biologico e artificiale (sinistra), Adele Rispoli (2021) e Modello di Rete Neurale Artificiale (destra), ITimpresa (2020)

La nascita del Machine Learning risale alla metà degli anni '50 quando lo psicologo Frank Rosenblatt viene assunto al Cornell Aeronautical Laboratory. Il progetto si basa sull'idea del 1943 di Warren McCulloch e Walter Pitts (McCulloch & Pitts, 1943) che il cervello nonostante fosse una massa morbida, umida e gelatinosa, le segnalazioni in esso sono digitali e, in particolare, binarie; l'obiettivo sta nel cercare, per la prima volta, di replicare con una macchina il comportamento di un neurone biologico tramite modelli matematici basati su regressioni lineari.

I neuroni artificiali si ispirano direttamente a quelli biologici e, come detto in precedenza, tentano di replicarne il comportamento di emissione di impulsi elettrici. Ogni neurone artificiale produce il suo impulso (o output) in base agli stimoli (input) che gli vengono forniti: in particolare l'output può essere descritto come una combinazione lineare degli ingressi. Analogamente ai neuroni biologici, non tutti gli stimoli hanno la stessa importanza nella valutazione dell'azione da compiere, perciò ad ogni input viene assegnato un "peso" che ne indica l'influenza sulla computazione dell'output (Zhou, 2021).

L'interconnessione in cascata di neuroni artificiali è detta rete neurale. Ogni rete neurale può essere suddivisa in tre sezioni:

- Strato di ingresso: è composto dai neuroni che ricevono gli stimoli dall'ambiente esterno (un sensore, una fotocamera ...)
- Strati nascosti: che ricevono input e producono output da e per altri neuroni
- Strato di uscita: è formato dai neuroni i cui risultati non diventano stimoli per altri

```
1  #include <stdio.h>
2
3  main() {
4      int ultimo, penultimo, i, f;
5      penultimo = 0 ;
6      printf("%d\n",penultimo);
7      ultimo = 1 ;
8      printf("%d\n",ultimo);
9
10     for( i = 2 ; i <= 100 ; i = i + 1 ){
11         f = ultimo + penultimo;
12         printf("%d\n", f);
13         penultimo = ultimo;
14         ultimo = f;
```

Figura 8 – Programma in Python per il calcolo della sequenza di Fibonacci, Giulia Del Serto (2023)

neuroni, di fatto definiscono qual è l'azione da compiere in base agli stimoli ottenuti inizialmente

Il machine learning si delinea fin da subito come radicalmente diverso rispetto ai paradigmi di programmazione tradizionale. In quest'ultima è richiesta la presenza di un essere umano, il programmatore, in grado di scrivere un software che venga utilizzato da un computer come insieme

di regole che gli permette di interpretare gli input forniti e produrre un risultato: INPUT + PROGRAMMA = OUTPUT. Nella programmazione tradizionale questi software sono composti da istruzioni che vengono eseguite sequenzialmente, ciò rende possibile verificarne il flusso di esecuzione e come questo viene influenzato dagli input.

D'altro canto, i sistemi basati su machine learning, non necessitano della presenza di un programmatore in quanto è l'algoritmo stesso che, date le informazioni di ingresso e i risultati che queste producono (o che vorremmo producessero) è in grado di generare un insieme di regole che consentano di svolgere, per quanto possibile, il compito richiesto. Le regole che vengono prodotte da un sistema di machine learning sono chiamate il suo "modello".

A differenza dei programmi tradizionali, per le decisioni prese da un modello di machine learning non è possibile ricostruire il percorso di decisioni interposto tra la ricezione degli stimoli di ingresso e la generazione degli output; nondimeno i neuroni artificiali intermedi in una rete neurale fanno parte dello strato nascosto.

1.2.1 Ciclo di vita di un sistema di Machine Learning

Il ciclo di vita di un sistema di machine learning può essere suddiviso principalmente in due fasi:

1. **Fase di apprendimento:** in questa fase il sistema "impara" come comportarsi e affina la sua capacità di individuare l'importanza di determinate caratteristiche. Dal punto di vista tecnico questo si traduce, per ogni neurone della rete, nell'affinare i valori dei pesi associati agli input. Durante questo stadio al sistema dovrebbe essere, idealmente sottoposto un campione di dati su cui apprendere, rappresentativo della realtà. Questo campione viene detto "training set" e, in alcuni tipi di apprendimento, ad esso è associato l'insieme di risposte "corrette" che ci aspettiamo dal sistema, in modo che sia possibile controllare il livello di affidabilità del modello stesso.

La fase di apprendimento risulta essere particolarmente delicata poiché i dati utilizzati per la raffinazione del modello spesso contengono bias e pregiudizi intrinseci, inevitabilmente assimilati dal sistema di machine learning.

2. **Fase operativa o “di predizione”**: una volta terminata la fase di apprendimento, il sistema di machine learning è pronto per la valutazione di nuovi input di cui, a differenza dei precedenti, non si conosce l’output “corretto” atteso.

1.2.2 Diversi tipi di apprendimento

In base ai compiti che verranno dati al sistema nella fase di previsione, è possibile identificare quattro diversi tipi di tecniche di apprendimento: supervisionato, non supervisionato, ibrido e rinforzato (Mahesh, 2019) (Di Capua, 2020).

1.2.2.1 Apprendimento Supervisionato

Il primo tipo di apprendimento è detto supervisionato in quanto al training set viene associata un’etichetta per ogni suo elemento: questa contiene un’informazione che lo descrive e che verrà successivamente utilizzata dal sistema di machine learning per eseguire le classificazioni (Tibco, 2019).

Durante la fase di apprendimento supervisionato, il programma cerca di astrarre le caratteristiche in comune tra tutti gli elementi del training set associati ad un’etichetta di interesse. Ne è un esempio un’intelligenza artificiale che, ricevuto come training set una lista di

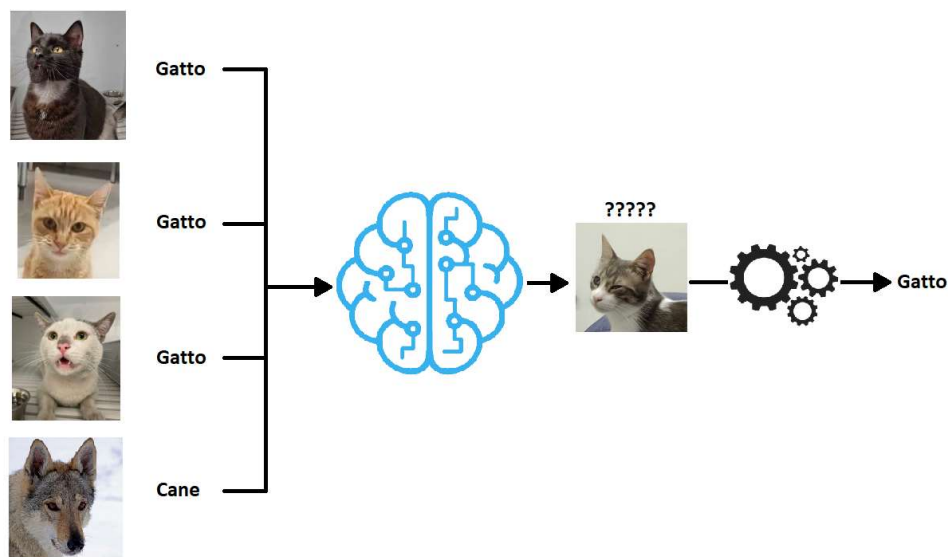


Figura 9 - Ciclo di vita di un sistema di ML basato su apprendimento supervisionato, Giulia Del Serto (2023)

immagini di animali cui è associato il loro nome, cerca di capire se il soggetto in una nuova immagine è, o meno, un gatto.

Generalmente questo tipo di apprendimento viene utilizzato nei casi in cui venga richiesto un sistema con un'accuratezza elevata: questo è dovuto alla presenza dell'essere umano che assegna le etichette al training set. Nonostante ciò, l'estrema quantità di dati (ed annesse etichette) necessari per poter istruire efficacemente un'intelligenza artificiale rende molto costoso il suo sviluppo; per questo nel caso di apprendimento supervisionato i programmatori spesso si affidano a dataset pubblici disponibili in rete (Garlapati, 2023).

1.2.2.2 Apprendimento Non Supervisionato

A differenza del precedente metodo di istruzione di un'intelligenza artificiale, l'apprendimento non supervisionato si basa sulla somministrazione al sistema di un training set sprovvisto di etichette.

La (quasi) totale assenza dell'essere umano nella validazione degli elementi di training consente di velocizzare enormemente la fase di apprendimento che, tuttavia, presenta diversi limiti: all'assenza totale di etichette consegue la mancanza di una "linea guida" per il sistema di machine learning su come debba elaborare gli input a disposizione (Tibco, 2019). In questo caso l'intelligenza artificiale, ha la sola possibilità di aggregare gli input tra di loro: per somiglianza. Questo fa in modo che, una volta interrogato, un sistema istruito con un metodo non

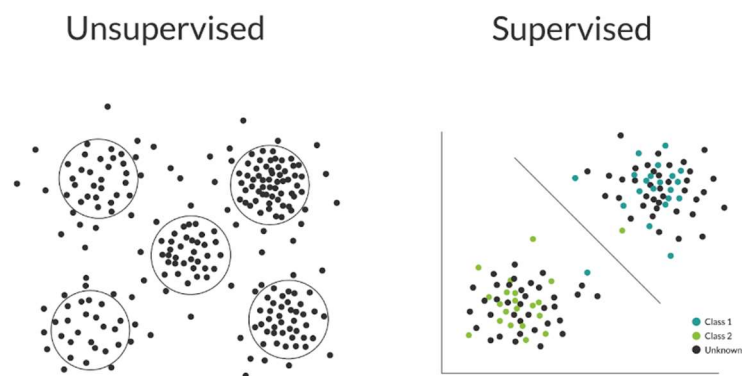


Figura 10 - Differenze di classificazione tra apprendimento supervisionato e non supervisionato, Edward Wu (2019)

supervisionato non dia una risposta secca ma tenda a ipotizzare quanto, il nuovo elemento, somigli ai gruppi di aggregazione identificati dal training set.

Per questo motivo l'apprendimento non supervisionato viene spesso usato per identificare tratti comuni tra i diversi elementi, per poter isolare quali sono le caratteristiche effettivamente rilevanti: un esempio pratico sono le radiografie e i documenti medici di diversi pazienti che condividono la medesima patologia.

1.2.2.3 Apprendimento Ibrido

Come suggerisce il nome, l'apprendimento ibrido si pone in mezzo, tra l'apprendimento supervisionato e non, cercando di integrare i vantaggi di entrambe le parti (Anifowose, 2020). Questo metodo aumenta la complessità del ciclo di vita di un sistema di machine learning, integrando tra loro i tipi di apprendimento sopra citati.

L'apprendimento ibrido, nello specifico mira a creare una combinazione di tipi di apprendimento ad hoc per il problema che si sta cercando di risolvere. I sistemi di machine learning ibridi si possono principalmente raggruppare in tre categorie: integrazione di architettura, manipolazione dei dati e ottimizzazione dei modelli.

1.2.2.4 Apprendimento Rinforzato

L'ultima tipologia di machine learning descritta in questo capitolo è l'apprendimento per rinforzo. A differenza dei metodi citati in precedenza questo introduce la possibilità, per un sistema, di

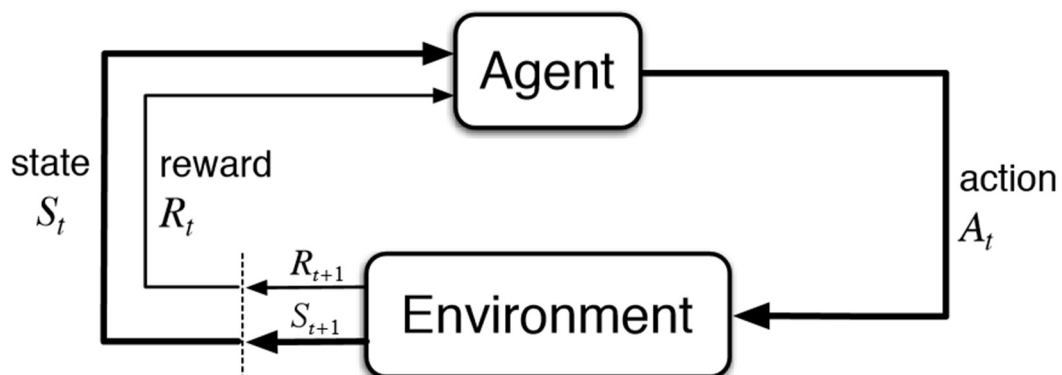


Figura 11 - Funzionamento di un sistema di ML basato sull'apprendimento rinforzato, Shweta Bhatt (2018)

imparare non solo dagli input che vengono forniti ma anche da tutte le decisioni che sono state prese in precedenza (Bhatt, 2018).

Conseguentemente, un'azione che viene presa in un determinato momento da un sistema di questo tipo, non è il risultato solo degli stimoli che gli vengono forniti in quel momento ma anche di tutti quelli ricevuti dal primo istante di esecuzione. L'ordine con il quale vengono forniti gli input può influenzare in maniera più o meno incisiva l'evoluzione del comportamento del sistema stesso.

2. Luoghi comuni e conoscenza ingenua: Stereotipi e Stereotipi di genere

Lo stereotipo è stato, ed è, oggetto di studio e riflessione all'interno della psicologia sociale, in quanto incide profondamente sulle interazioni umane e sulle percezioni di in-group e out-group.

Questo capitolo si propone di esplorare le radici concettuali dello stereotipo, partendo dalle prime formulazioni teoriche fino alle più recenti prospettive di ricerca, esaminando i meccanismi mentali sottostanti la formazione e perpetuazione degli stereotipi. La riflessione sul concetto di stereotipo di genere è cruciale al fine di comprendere le disuguaglianze e discriminazioni presenti nella nostra società e tentare così di promuovere processi di cambiamento e di trasformazione sociale.

2.1 Che cos'è uno stereotipo: origini e definizione del termine

stereòtipo agg. e s. m. [dal fr. stéréotype, comp. di stéréo- «stereo-» e -type «-tipo»]. – 1. agg. a. Di stereotipia, realizzato con il procedimento della stereotipia: ristampa s. di un volume; lastre s., le stereotipie, ossia le controimpronte, delle forme di composizione tipografica. b. fig. Impersonale, inespressivo, perché detto o fatto senza partecipazione (meno com. di stereotipato): i soliti discorsi s. da salotto; un sorriso stereotipo. 2. s. m., fig. a. Modello convenzionale di atteggiamento, di discorso e sim.: ragionare per stereotipi. In partic., in psicologia, opinione preconstituita, generalizzata e semplicistica, che non si fonda cioè sulla valutazione personale dei singoli casi ma si ripete meccanicamente, su persone o avvenimenti e situazioni (corrisponde al fr. cliché): giudicare, definire per stereotipi; s. individuali, se proprî di individui, s. sociali, se proprî di gruppi sociali. b. In linguistica, locuzione o espressione fissatasi in una determinata forma e ripetuta quindi meccanicamente e banalizzata; luogo comune, frase fatta: parlare per stereotipi, abusare di stereotipi; in partic., sinon. di sintagma cristallizzato (v. sintagma). c. Espressione, motto, detto proverbiale o singola parola nella quale si riflettono pregiudizî e opinioni negative con riferimento a gruppi sociali, etnici o professionali.

(Definizione da Vocabolario Treccani (Treccani, 2023))

“Stereotipo” è un termine di origini antiche, poiché deriva dal celebre tipografo Firmin Didot, il quale nel 1775 ideò un metodo che consentiva di duplicare ed imprimere composizioni tipografiche (Falchetti, 2020). Secondo tale processo, chiamato stereotipia – termine da cui la parola “stereotipo” deriva - l’immagine originale che si intendeva replicare veniva pressata con un cartone apposito, il Flano, che consentiva di ricavarne altre identiche.

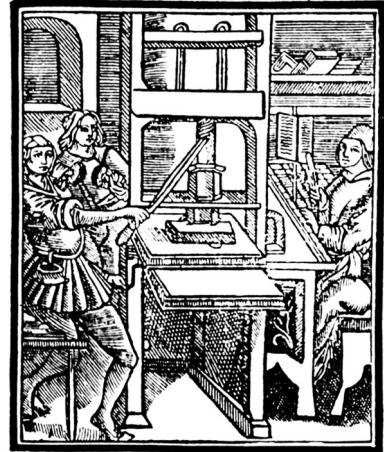


Figura 12 - Torchio Tipografico, Technoratio (2010)

Tale processo di riproduzione di una figura, semplice e automatizzato, condivide un importante significato concettuale con il meccanismo cognitivo e comportamentale dell’uomo, quando questo è impegnato nell’attività di ricavare un’immagine mentale - di qualcuno o qualcosa - al fine di esprimere un giudizio.

In psicologia, infatti, il termine "stereotipo" si riferisce a schemi mentali o rappresentazioni cognitive generalizzate e semplificate che le persone costruiscono relativamente a gruppi di individui o categorie sociali. Tali rappresentazioni mentali sono basate su credenze e aspettative e spesso portano alla formazione di giudizi distorti o discriminazione.

Nelle scienze sociali, il termine “stereotipo” fu introdotto dal giornalista e politico americano Walter Lippmann, intorno al 1920, nei suoi studi sul pregiudizio (Sue Curry, 2013).

Nella sua opera più importante, *Public Opinion* (Opinione Pubblica) del 1922, Lippmann descrive come l’informazione oggettiva dell’essere umano sia in realtà plasmata dalle prospettive mentali individuali ed influenzata da preconcetti e pregiudizi personali. È da questa premessa che sorge lo "stereotipo", una rappresentazione distorta e semplificata della realtà sociale. Lo stereotipo si configura come immagine collettiva di un determinato fenomeno, manipolata e adattata per scopi opportunistici nel contesto del giornalismo o della politica. Infatti, secondo Lippmann, i mezzi di comunicazione svolgono un ruolo importante nella creazione di stereotipi poiché le persone, fidandosi, accettano passivamente quanto viene trasmesso, senza esaminare il fatto alla luce di altre prospettive o possibilità.

Nel campo della psicologia sociale, Hamilton e Troiler (1986) definiscono lo stereotipo come una struttura cognitiva che contiene conoscenze, credenze e aspettative di cui un soggetto dispone relativamente un determinato gruppo umano.

Gli stereotipi sono dunque concetti generalizzati che si tende a formare sulla base di attributi superficiali e superficialmente percepibili, quali razza, genere, età, orientamento sessuale, appartenenza religiosa o etnica. Rappresentano un meccanismo di naturale semplificazione cognitiva che ci permette di elaborare informazioni complesse e conseguentemente prendere decisioni in modo rapido ed efficiente.

Inoltre, lo stereotipo consente di attribuire facilmente le caratteristiche generali di un gruppo di persone a tutti i membri del gruppo, senza dover valutare ogni individuo singolarmente, ha dunque un'importante funzione euristica e di scorciatoia mentale in quanto consente di ridurre la complessità del mondo sociale, risparmiando tempo e risorse cognitive.

La valutazione che emerge dallo stereotipo è dunque comoda, ma rigida e spesso responsabile di pregiudizi e discriminazioni. Ad esempio, un individuo appartenente ad un gruppo stigmatizzato (come una minoranza etnica o sessuale), può essere erroneamente giudicato sulla base degli stereotipi associati al suo gruppo, piuttosto che sui suoi meriti personali.

La psicologia sociale da tempo studia gli stereotipi e il loro impatto sulla percezione degli individui e dei gruppi, nonché sui comportamenti sociali. Questo campo di ricerca ha dimostrato che gli stereotipi possono influenzare non solo le percezioni delle persone ma anche le loro azioni, il modo in cui interagiscono nel loro ambiente sociale e loro scelte.

I ricercatori hanno focalizzato i loro studi sui contenuti negativi della conoscenza stereotipata utilizzata dagli individui, nel tentativo di ridurre i pericolosi effetti che possono generare sul piano sociale.

2.1.1 La dimensione sociale dello stereotipo

Lo stereotipo può essere classificato in due tipologie: cognitivo o sociale.

Gli stereotipi di natura cognitiva si sviluppano attraverso il processo di semplificazione e organizzazione delle complesse informazioni immagazzinate da ciascun individuo prima di

diventare parte del proprio bagaglio culturale. Tali stereotipi, influenzando conoscenze e comportamenti, possono plasmare il modo in cui le persone percepiscono il mondo che le circonda (Pellegrino, 2020).

Gli stereotipi di tipo sociale, rappresentano immagini mentali condivise da ampie masse di individui e sono dirette a istituzioni o categorie di persone estremamente ampie, come ad esempio ebrei, cristiani, bianchi, neri, donne, omosessuali, membri di partiti politici, e così via. Questi stereotipi vengono utilizzati per spiegare eventi sociali complessi che si verificano su larga scala e possono essere impiegati per giustificare azioni contro altre persone o per differenziare positivamente il proprio gruppo rispetto ad altri. Spesso vengono impiegati al fine di legittimare atti di violenza, tensioni e conflitti senza attribuire alcuna responsabilità personale, poiché le persone interiorizzano determinate spiegazioni sociali e culturali, come la difesa del bene supremo della patria. Questo può portare a considerare giusti, comportamenti conflittuali che vanno dall'intimidazione alla violenza fisica, dalla segregazione all'esclusione degli altri, fino a massacri e omicidi di massa.



Figura 13 - Rappresentazione del personaggio Jim Crow, New York Public Library (1836)

Gordon Allport fu pioniere nell'esplorare i concetti di stereotipo e il conseguente pregiudizio. Negli anni '50, pubblicò l'importante opera "La natura del Pregiudizio," un trattato che analizzò le radici di questo fenomeno. Allport spiegò che pregiudizi e stereotipi sociali derivano dalla generalizzazione, tendenza della mente umana a trarre conclusioni generali basate su un numero limitato di esperienze (Allport, 1954). Gli stereotipi operano principalmente nel campo della percezione, grazie a un meccanismo noto come accentuazione, il quale amplifica le differenze tra elementi di categorie diverse e minimizza le differenze tra elementi della stessa categoria.

Allport condusse ricerche approfondite sui gruppi sociali e nel 1954 sviluppò una scala per misurare la forza del pregiudizio e degli stereotipi all'interno della società.

La "Scala del Pregiudizio di Allport" si articola in cinque livelli, mirati a valutare la forza del pregiudizio nelle dinamiche tra gruppi interni (in-group) e gruppi esterni (out-group):

1. **Antilocuzione:** punto di partenza della scala. Si verifica quando un individuo all'interno di un gruppo esprime apertamente opinioni negative riguardo individui o gruppi esterni (out-group). Questa fase anticipa il conflitto in-group/out-group, alimentando sentimenti di odio.
2. **Evitamento:** nel secondo livello della scala, i membri dell'in-group evitano in modo evidente gli individui che non fanno parte del loro gruppo (out-group), isolandoli a livello psicologico e sociale, con l'obiettivo di creare un distacco.
3. **Discriminazione:** Nella terza fase, i membri dell'in-group discriminano attivamente i membri dell'out-group, cercando di limitarne l'accesso a servizi, opportunità e risorse. Questa discriminazione accentua ulteriormente l'isolamento dei membri dell'out-group e ostacola la loro capacità di condurre una vita dignitosa.
4. **Violenza fisica:** nella quarta fase della scala, dopo la dichiarata discriminazione dei membri dell'out-group e il divieto di accesso a risorse e servizi, i membri dell'in-group procedono all'azione fisica, attaccando, vandalizzando e distruggendo tutto ciò che è legato, anche simbolicamente, agli individui dell'out-group.
5. **Stermio:** l'ultimo livello rappresenta il tentativo dell'in-group di eliminare completamente l'out-group. Un esempio storico di questa fase è rappresentato dalla tragedia dell'Olocausto perpetrato dalla Germania nazista contro la popolazione ebraica.

La categorizzazione sociale, dal punto di vista cognitivo ha, per ogni essere umano, una funzione adattiva. Consente di suddividere il mondo da lui percepito in distinte categorie raggruppate per similarità. Questa capacità innata dell'essere umano consente di diminuire la complessità del mondo sociale e risparmiare energia mentale, si tratta di una sorta di economia cognitiva (Allport, 1954).

Un solo stimolo, ad esempio un'uniforme, innesca l'attribuzione di altre caratteristiche da esso derivanti, permettendo di risalire alla professione della persona in questione.

La categorizzazione oltre ad avere un'importante funzione di economia cognitiva, permette una visione positiva di noi stessi (funzione motivazionale). Le persone percepiscono il proprio gruppo di appartenenza migliore dell'out-group e l'autostima deriva in modo rilevante dal gruppo di appartenenza. Ne consegue che differenziare in-group da out-group attraverso il fenomeno sociale della categorizzazione, offre l'opportunità di percepirsi migliori incrementando l'autostima del singolo. È interessante a questo fine, constatare come i membri di una categoria si percepiscano più simili tra loro di quanto non siano mentre quelli appartenenti a rispettivi gruppi esterni vengano considerati ancora più differenti, questo meccanismo psicologico sociale prende il nome di effetto di omogeneità del gruppo esterno.

Una delle svariate spiegazioni sottostanti il fenomeno di omogeneità del gruppo esterno è stata avanzata da Linville, Fischer e Salovey (1989) e si tratta del fenomeno della familiarità; secondo cui la conoscenza maggiormente dettagliata degli individui che fanno parte dell'in-group fa apparire tale categoria maggiormente sfaccettata e differenziata, mentre l'insufficiente conoscenza dei componenti dell'out-group genera una percezione globale e indifferenziata del gruppo (Linville, W. Fischer, & W. Salovey, 1989).

Altri ricercatori hanno ipotizzato che i comportamenti dei membri dell'in-group vengano interpretati e codificati tramite categorie specifiche mentre quelle dell'out-group tramite categorie generali.

La categorizzazione dei membri dell'in-group e dell'out-group ha un impatto diretto sul modo in cui elaboriamo le informazioni relative a tali gruppi. Mentre tendiamo a interpretare e codificare i comportamenti dei membri dell'in-group attraverso categorie specifiche, utilizziamo categorie generali per quelli dell'out-group. Questa categorizzazione influisce sul nostro favoritismo nei confronti dell'in-group, spingendoci a distorcere le informazioni a favore del nostro gruppo di appartenenza e a perpetuare stereotipi riguardo a come appare, agisce e quali sono i tratti di personalità che lo definiscono.

Pertanto, la distorsione sistematica di eventi e altre informazioni a favore del nostro gruppo di appartenenza è una conseguenza naturale di tali processi.

2.2 Stereotipi, bias ed euristiche

I bias cognitivi rappresentano costrutti che si basano su percezioni erranee, pregiudizi e ideologie, e spesso vengono utilizzati per prendere decisioni rapide e senza un adeguato processo di valutazione critica. Questi errori cognitivi non solo influenzano le nostre decisioni e comportamenti quotidiani ma hanno anche un impatto significativo sui nostri processi di pensiero.

Le euristiche, derivanti dalla parola greca "heurískein" che significa "trovare" o "scoprire", sono processi mentali intuitivi e veloci che agiscono come scorciatoie mentali. Ci consentono di formare rapidamente una visione generale su un argomento senza richiedere un dispendio significativo di sforzi cognitivi. Queste strategie sono spesso utilizzate per giungere a conclusioni immediate.

Nel 2002, Kahneman e Frederick avanzarono l'ipotesi che l'euristica cognitiva operasse attraverso un sistema denominato "sostituzione dell'attributo," un processo che avviene in modo automatico, senza consapevolezza. Secondo questa teoria, quando qualcuno deve formulare un giudizio complesso basato su inferenze, questo processo viene sostituito da un'euristica, un concetto simile all'originale ma espresso in modo più semplice (Kahneman & Frederick, Representativeness revisited: Attribute substitution in intuitive judgment., 2002). Le euristiche rappresentano, quindi, strategie mentali che portano a conclusioni immediate con un minimo sforzo cognitivo.

In altre parole, i bias sono euristici utilizzati per emettere giudizi che col tempo possono trasformarsi in pregiudizi, riguardanti oggetti o esperienze di cui non si ha una conoscenza diretta. Al contrario, le euristiche fungono da shortcut mentale, consentendo di accedere rapidamente alle informazioni memorizzate nella nostra mente.

Ad oggi, la letteratura ha evidenziato l'esistenza di diverse tipologie di bias, tra quelle più comuni troviamo:

- **Bias di conferma:** tendenza a cercare conferme per le nostre opinioni preesistenti e ad evitare chiunque ci faccia sentire a disagio.

- **Dissonanza cognitiva:** ulteriore modalità di comportamento preferenziale che porta al bias di conferma, in cui si considerano solo prospettive che sostengono le opinioni già esistenti (Skinner, 1953).
- **Bias di gruppo:** tendenza a sovrastimare qualità e valore del proprio gruppo e, viceversa attribuire i successi dell'outgroup a fattori esterni anziché alla qualità dei suoi membri.
- **Fallacia dello scommettitore (o del gambler):** tendenza a conferire eccessiva importanza a eventi passati e ritenere che influenzino i risultati attuali.
- **Errore per somiglianza:** bias in cui manager con alta autostima, tendono a sovrastimare collaboratori con caratteristiche simili alle proprie.
- **Errore per contrasto:** bias in cui manager con bassa autostima, tendono a premiare collaboratori con caratteristiche diverse da quelle in cui sono carenti.
- **Bias della negatività:** tendenza a concentrarsi e conferire maggiore importanza agli aspetti negativi, spesso a discapito di successi e competenze acquisite.
- **Bias dello status quo:** distorsione valutativa dovuta alla resistenza al cambiamento, con l'assunzione ingiustificata che qualsiasi scelta diversa potrebbe peggiorare le cose.

Il legame tra bias e stereotipo è fondamentale nella comprensione dei processi cognitivi e sociali che portano a percezione e giudizi errati.

Gli stereotipi, infatti, possono influenzare o generare bias cognitivi. Ad esempio, il "bias di conferma" può portare le persone a cercare o interpretare informazioni in modi che confermano gli stereotipi esistenti. Questo può portare a percezioni distorte e decisioni errate basate su tali stereotipi. Inoltre, gli stereotipi stessi sono spesso considerati come un tipo di bias cognitivo. Essi rappresentano una semplificazione e una generalizzazione e possono portare a giudizi errati o discriminazioni.

Viceversa, i bias cognitivi influenzano la formazione e il mantenimento degli stereotipi. Questo perché le persone tendono a basare i loro stereotipi su informazioni facilmente accessibili o disponibili nella loro mente, spesso distorte a causa di tali bias.

2.3 Dallo stereotipo allo stereotipo di genere

La definizione del termine "genere" in relazione agli studi di genere è una questione complessa e sfaccettata.

Secondo l'Organizzazione Mondiale della Sanità (OMS), il concetto di genere si riferisce a una costruzione basata su parametri sociali che riguarda il comportamento, le azioni e i ruoli assegnati a un determinato sesso (World Health Organization, 2021).

L'American Psychological Association (APA) definisce il genere come "le caratteristiche sociali, culturali e psicologiche associate all'essere maschio o femmina" (American Psychological Association, 2019). Entrambe le definizioni riconoscono che il genere non è una caratteristica biologica intrinseca, ma piuttosto una concezione sociale che può variare da cultura a cultura e nel tempo. Infatti, la filosofa Judith Butler (1988), autrice di opere influenti nel campo degli studi di genere, sostiene che il genere sia una performance sociale, un atto che le persone compiono per conformarsi alle aspettative della società riguardo a ciò che significa essere maschi o femmine.

Il genere, pertanto, è influenzato e "messo in atto" nella politica, nella cultura, nelle condizioni strutturali, nelle gerarchie sociali e nelle pratiche, spesso con effetti dannosi per l'intera società (Fraser, 2013). Data la natura costruita del genere, gli individui tendono a inferire generalizzazioni sui comportamenti associati a generi diversi. Ne costituisce un esempio l'idea comune che le donne siano delicate ed emotive e gli uomini maggiormente inclini a comportamenti aggressivi (Harris & Jenkins, 2006) (Plant, Hyde, Keltner, & Devine, 2000). Queste generalizzazioni contribuiscono alla formazione di modelli specifici di comportamento, stereotipi e pregiudizi che persistono ancora oggi.

Gli stereotipi di genere sono profondamente radicati nelle credenze e aspettative sociali, e danno origine a una visione semplificata e restrittiva dei ruoli di genere. Come afferma Biemmi (2010), gli stereotipi di genere sono "una visione semplificata e rigida che attribuisce a donne e uomini ruoli determinati e limitati dal loro sesso".



Figura 14 - Giocattoli per maschi e femmine, La Voce di Bolzano (2019)

Lueptow et al. (2001) definiscono tale tipologia di stereotipo come la percezione comune delle differenze di genere nei tratti di personalità e nei comportamenti.

A livello sociale, così come per lo stereotipo precedentemente descritto, anche lo stereotipo di genere rappresenta una comprensione stereotipata della realtà, condivisa da un intero gruppo sociale, e che si applica a individui o gruppi specifici (Lloyd, Lori, & Margaret, 2001).



Figura 15 - Scatola di battaglia navale con sfondo sessista, Milton Bradley (1967)

Come affermato in precedenza, gli stereotipi di genere hanno un'origine complessa e multifattoriale, strettamente legata al contesto familiare, alla cultura d'appartenenza,

all'educazione, alle relazioni tra pari e alla società nel suo complesso. Tali influenze contribuiscono alla formazione di idee e aspettative che, sin da piccoli, gli esseri umani sviluppano su sé stessi, sugli altri e sul mondo, basate su preconcetti legati ai ruoli di genere. Infatti, uno degli aspetti più preoccupanti degli stereotipi di genere è la loro internalizzazione sin dalla prima infanzia, poiché questi vengono assimilati in modo inconscio sia da bambini che da bambine, e persistono nell'età adulta.

Siamo quindi di fronte a idee che vengono plasmate e costruite a livello culturale, le quali orientano la nostra interpretazione e valutazione del mondo, degli individui e degli avvenimenti seguendo due distinte categorizzazioni: quella associata al maschile e quella al femminile. Inoltre, è importante notare che queste classificazioni non siano equiparate, ma piuttosto si articolano su una gerarchia che enfatizza il predominio del polo maschile rispetto a quello femminile (Biemmi, 2010).

La persistente dominazione di genere, che abbraccia sia potere che privilegio, contribuisce all'immobilismo delle gerarchie sociali, delle norme e dei comportamenti consolidati (Gary & Fraser, 2015). Di conseguenza, il cammino verso una parità di genere effettiva rimane ostacolato e lento (OECD, 2017) (United Nations, 2021).

Le persone continuano a formulare giudizi sugli altri basati su conoscenze acquisite nel tempo, nelle diverse situazioni e contesti, spesso fondati su schemi normativi e aspettative legate al genere (Hoyt & Burnette, 2013).

Numerosi studi riguardanti le caratteristiche distintive delle personalità tra individui di sesso maschile e femminile (Bem, 1974; Broverman et al., 1972; Spence e Helmreich, 1978 in Martini, 2009) evidenziano tratti comuni negli attributi associati rispettivamente a uomini e donne nel descriverli.

Attributi come dominante, aggressivo, competitivo, indipendente, ambizioso, sicuro di sé, avventuroso e decisionista vengono spesso associati alla figura maschile.

Viceversa, attributi come affettuosa, remissiva, emotiva, empatica, loquace e gentile sono spesso riferiti ad una persona di sesso femminile. Tali stereotipi vengono talvolta ricondotti a diverse categorie:

- “agentic-communale”, che contrasta autonomia, attivismo con sollecitudine, e capacità di prendersi cura degli altri (Bakan, 1966 in Martini 2009);
- “strumentale-espressivo” (che distingue l'orientamento ai risultati e all'azione manuale da quella di esprimere sentimenti e offrire supporto (Talcott e Bale, 1955 in Martini 2009)
- "maschile-femminile" (Lueptow et al., 2001 in Martini 2009).

Tuttavia, il pregiudizio di genere non si manifesta solo come un favoreggiamento degli uomini rispetto alle donne (Rothchild 2014), ma spesso anche come un sostegno alle identità di genere binarie rispetto a quelle diverse (Smith e Rustagi 2021).

Inoltre, gli stereotipi di genere conducono a una distribuzione ingiusta di risorse, informazioni e opportunità per le donne (Fraser, 2013), contribuendo a mantenere in vita stereotipi e pregiudizi dannosi (Banchefsky e Park 2018; Hogenboom 2021; Locke 2019; Smith e Rustagi 2021) e generando ingiustizie a livello globale.

Tali atteggiamenti comportano anche un trattamento dispregiativo e offensivo (e talvolta la negazione stessa dell'esistenza) delle identità di genere già marginalizzate (Smith e Rustagi 2021).

2.4 Il genere come “performance” e l’intelligenza artificiale

Come espresso nel capitolo precedente, Butler (2004) sostiene che il concetto di genere, e di conseguenza relativi stereotipi, siano da considerare come una “performance” messa in atto attraverso potenti norme ed aspettative sociali. Tuttavia, la filosofa sottolinea la possibilità di “annullare” tali costrutti. Tale risultato si può conseguire solo rendendo la questione di genere irrilevante nelle interazioni, disconnettendo le dinamiche di genere dall'ineguaglianza, attraverso la collaborazione tra istituzioni per produrre e sostenere il cambiamento e trasformando le interazioni stesse in terreno fertile per il cambiamento (Deutsch, 2007).

Secondo Austin (1962), il linguaggio e la comunicazione hanno un potere sufficiente per compiere o far compiere un'azione. Le espressioni, altrimenti chiamate “atti locutori”, sono semplici frasi. Tuttavia, diventano performativi quando non solo descrivono ciò che è reale, ma agiscono anche per cambiare la realtà che stanno descrivendo. Quando vengono trasformate in azione, diventano “atti illocutori”.

Infine, ci sono anche atti di linguaggio performativi che influenzano l'ascoltatore e hanno conseguenze, come persuadere, convincere, spaventare, ispirare, ecc. - questi sono chiamati “atti perlocutori”. A differenza degli atti illocutori, che si concentrano sull'azione prevista dell'espressione linguistica, gli atti perlocutori enfatizzano il contesto in cui avviene l'espressione, così come il suo effetto sul ricevente.

Nel contesto attuale, fortemente influenzato dagli effetti e le influenze della tecnologia, tali aspetti performativi si possono riscontrare nelle interazioni tra utenti e strumenti di intelligenza artificiale, in particolare quella generativa.

Gross (2023), tra i tanti studiosi della relazione tra stereotipi di genere e Intelligenza artificiale, ha condotto uno studio per comprendere le visioni di genere in ChatGPT e come gli atti performativi precedentemente descritti vengano compiuti da quest'ultimo. Lo studio mette in evidenza gli aspetti locutori (ciò che è stato detto), gli aspetti illocutori (ciò che è stato fatto) e gli aspetti perlocutori (ciò che è successo come risultato) della performatività di genere attraverso l'IA generativa. In particolare, ha evidenziato che:

- Gli atti locutori si riferiscono alle risposte che ChatGPT genera quando gli viene chiesto qualcosa. Ad esempio, nel descrivere un “professore di economia” o un “ingegnere brillante” parla di un uomo, mentre nel descrivere un artista emotivo fa riferimento ad una donna, impiegando anche aggettivi differenti a seconda delle due categorie.
- Gli atti illocutori si riferiscono al modo in cui le risposte date da ChatGPT possano perpetuare idee e pregiudizi di genere in vari contesti, inclusi documenti di ricerca, curriculum vitae, storie e conversazioni. Ad esempio, ChatGPT tende a presentare stereotipi di genere nelle sue risposte, come mostrato in un caso in cui ha raccontato una storia di un incidente in ufficio che metteva in imbarazzo una donna, sessualizzandola nel

processo. In un altro caso, ha presentato un forte pregiudizio di genere nella descrizione delle abilità genitoriali, assegnando alla madre il ruolo di “premurosa” e al padre quello di “avventuroso”, rafforzando i tradizionali ruoli di genere. Inoltre, quando è stato chiesto di evidenziare competenze in un curriculum vitae, ChatGPT ha presentato sottili ma significative differenze di genere: per gli uomini, ha classificato le competenze tecniche al terzo posto, mentre per le donne le ha classificate al nono posto; per le donne, ha classificato le competenze comunicative e interpersonali al terzo posto, mentre per gli uomini le ha classificate al quinto posto; le competenze organizzative e di gestione del tempo sono state elencate solo per le donne, mentre le competenze di gestione dei progetti sono state elencate solo per gli uomini.

- Gli atti perlocutori, infine, si riferiscono al linguaggio e le espressioni impiegate (fig. 16) da ChatGPT, così autorevoli e persuasivi da avere un effetto performativo sul lavoro e sulla vita dell'utente che li legge. Infatti, i Large Language Model (LLM), come quello preso in esempio, possono agire in modi abbastanza manipolatori (Ienca 2023 in Gross, 2023).

Il lavoro di Gross (2023) ha confermato ciò che anche altri studiosi hanno avuto modo di constatare: ChatGPT (e come questi altri strumenti di IA) è *gender-biased*.

Tali esempi provano come gli strumenti di IA abbiano effetti performativi e talvolta manipolatori quando si tratta di genere, rafforzando i pregiudizi.

Tuttavia, date le premesse ed il suo potere, l'intelligenza artificiale può agire come un veicolo contemporaneo per favorire l'“annullamento di genere”, ovvero per ridurre le prospettive di genere, attenuare i pregiudizi e promuovere l'uguaglianza di genere.

Examples of ChatGPT Text	Effect(s)	Sources
It's important to note; it's worth noting; it is essential to remember; remember, it's important to; The story of Steve and Lisa serves as a reminder. . .	convince; warn	ChatGPT (2023a, 2023b, 2023c, 2023e, 2023j)
Here are some common characteristics; here are common elements; here are some general characteristics; here are some skills	illustrate; explain; persuade	ChatGPT (2023a, 2023b, 2023c, 2023l)
The dress code form for CEOs can vary depending on the industry, company culture, and personal style; When a 40-year-old woman is updating her CV, she should highlight a combination of skills and experiences that showcase her qualifications and make her a strong candidate for prospective employers.	inform; advise	ChatGPT (2023c, 2023l)
It's important to note that these descriptions are not universally applicable, and individuals who teach economics can have diverse appearances; Sarah's name became synonymous with resilience and triumph in the face of adversity—a shining example of what one can achieve when they refuse to let circumstances define their destiny; It's important to note that traits and characteristics can vary greatly among individuals, regardless of gender; The story of Sarah and Michael is a testament to the power of teamwork, love, and understanding in parenting.	enlighten	ChatGPT (2023a, 2023e, 2023f, 2023i, 2023k)
Once upon a time, in a small town named Harmonyville, there lived a girl named Lily and a boy named Ethan; Once upon a time, in the bustling city of Metropolis, there were two colleagues named Steve and Lisa; Once upon a time in a small suburban town, lived a loving couple named Sarah and Michael.	entertain	ChatGPT (2023d, 2023j, 2023k)

Figura 16 - Esempi di risposte di ChatGPT e il loro effetto sull'utente, Gross (2023)

3. Intelligenza discriminante: in quali modi il pregiudizio può annidarsi in un algoritmo?

Come emerge dai capitoli precedenti, all'entusiasmo per lo sviluppo delle intelligenze artificiali sono seguite di pari passo le preoccupazioni che riguardano la sua rapida diffusione e la fruizione da parte di utenti che non hanno una conoscenza approfondita dei suoi limiti.

Essendo l'IA un prodotto sviluppato da esseri umani, non è immune ai pregiudizi e alle disuguaglianze presenti nella società. Uno dei pregiudizi maggiormente diffuso e pervasivo è quello di genere, che può facilmente infiltrarsi nell'IA e influenzarne le capacità.

Il presente capitolo si propone di indagare i modi in cui il pregiudizio di genere può annidarsi in un algoritmo, creando ciò che può essere definito come "intelligenza discriminante". Esploreremo le diverse forme attraverso cui l'IA può riflettere e perpetuare i pregiudizi di genere, con un focus sulle cause principali responsabili di questo fenomeno.

Nonostante si tenda ad associare il mondo dello sviluppo di programmi e applicazioni ad una retorica esclusivamente logica e matematica, in realtà i prodotti di questo settore sono sempre stati permeati da ideologie, cultura, influenze sociali e pregiudizi in grado di riflettere la mente dei loro creatori.

Per familiarizzare con questo problema nell'ambito dell'intelligenza artificiale è utile fornirne un esempio legato alla programmazione tradizionale. Nei linguaggi di programmazione tradizionale i dati vengono inseriti all'interno di variabili a cui viene assegnato un "tipo", congruo al dato che conterranno: per memorizzare un numero senza virgola si usano variabili di tipo `int` (Intero), per memorizzare delle parole dei tipi `string` (Stringa), per delle condizioni vero/falso dei `boolean` (Booleani) etc... (Oracle, s.d.)

Anche se la decisione di scegliere il tipo di una variabile piuttosto che un'altra può sembrare un dettaglio, in realtà ha un enorme impatto sullo sviluppo dell'intera applicazione, sia per quanto concerne le prestazioni di esecuzione, che per aspetti quali la facilità di modifica o aggiunta di nuove funzionalità.

Di seguito un comune modulo di registrazione di un sito: esempio di come, da queste scelte, possano trasparire ideologie specifiche.

The image shows a registration form titled "Responsive Registration Form". It contains the following elements: an email input field with an envelope icon; a password input field with a lock icon; a "Re-type Password" input field with a lock icon; two input fields for "First Name" and "Last Name" with person icons; radio buttons for "Male" and "Female"; a dropdown menu for "Select a country"; two checkboxes: "I agree with terms and conditions" and "I want to receive the newsletter"; and a prominent yellow "Register" button.

Figura 17 - Template di registrazione di utente, Codepen.io (2022)

Già dalle modalità adottate per richiedere i dati personali dell'utente è possibile notare l'impatto delle scelte di programmazione degli sviluppatori: spesso, infatti, nei campi di nome e cognome viene prevista la possibilità di utilizzare solo caratteri alfabetici "tradizionali", impedendo a utenti con più nomi (a causa della presenza del carattere spazio) o con nomi formati da vocali accentate, di potersi registrare al sito.

L'esempio precedente potrà sembrare un piccolo inconveniente, ma i problemi più ricorrenti riguardano proprio il campo di selezione del genere, in quanto le scelte possibili risultano molte volte estremamente limitate e spesso binarie (maschio/femmina), dunque booleane. Tale dicotomia diventa problematica per utenti con un'identità sessuale non binaria o in via di definizione.

L'esponenziale complessità delle intelligenze artificiali fa sì che questi problemi crescano a loro volta, acuendosi ulteriormente a causa della necessità di far scegliere agli sviluppatori i dati di apprendimento da fornire all'intelligenza artificiale.

3.1 Sessismo e Pregiudizio "Accidentale"

Come detto in precedenza la scelta dei dati di addestramento è l'operazione più sensibile e complicata nello sviluppo di un sistema di intelligenza artificiale.

Se il sistema di AI infatti viene viziato con dati che contengono pregiudizi o in cui alcune categorie sono sottorappresentate, avremo come allarmante conseguenza la perpetuazione di tali pregiudizi.

Nei prossimi paragrafi verranno analizzati esempi pratici di come possano nascondersi nell'intelligenza artificiale, comportamenti discriminatori o basati su pregiudizi intrinseci ai dati di addestramento utilizzati.

3.1.1 Google Translate

La multinazionale americana si è trovata, nel 2018, a dover gestire una crisi legata ad uno dei suoi servizi maggiormente utilizzati: Google Translate. La prima diffusione “accademica” del problema avviene con la pubblicazione del paper “Assessing Gender Bias in Machine Translation – A Case Study with Google Translate” (Prates, Avelar, & Lamb, 2019) nel quale gli autori mostrano come il traduttore, partendo da una frase sessualmente neutra, tenda a cambiare il sesso del soggetto della frase, specialmente nei casi in cui questo sia associato a determinate professioni. In particolare la ricerca ha dimostrato come i lavori comunemente associati alla cura venissero interpretati dal software come svolti da soggetti femminili, d’altro canto il traduttore era più pronò ad associare professioni STEM ad un soggetto maschile (Grinevičius & Akavickaitė, 2021).

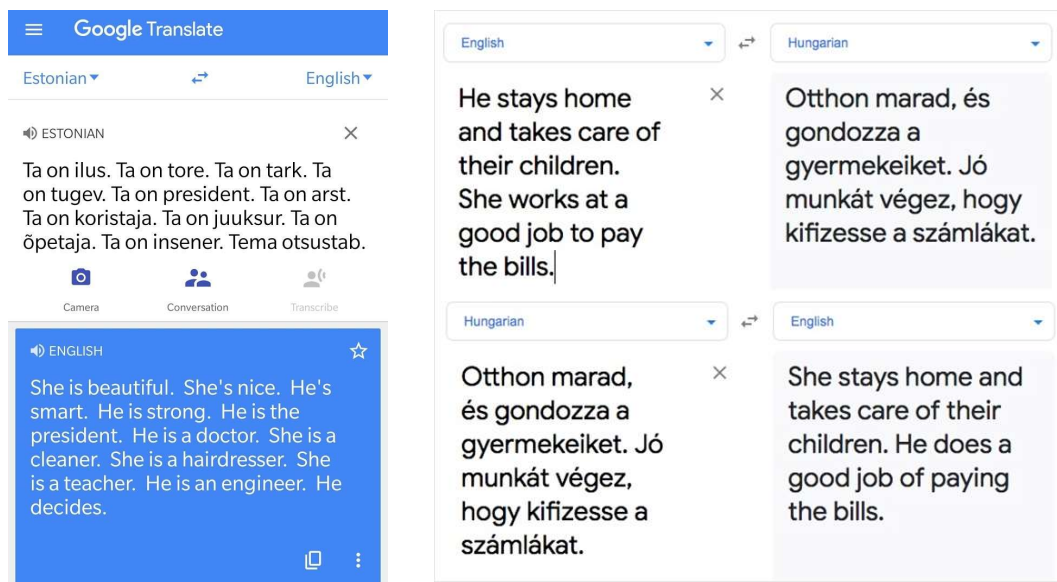


Figura 18 - Esempi di traduzioni sessiste da parte di Google Translate, BoredPanda (2021)

All’origine di questo problema vi è il metodo utilizzato per addestrare lo stesso Google Translate. Il programma è basato su un sistema chiamato “Google Neural Machine Translation” (Wu, 2016) che, per poter offrire un affidabile servizio di traduzione, è stato addestrato utilizzando il maggior quantitativo di testi possibile. Per motivi sia di mancata sensibilità storica che del ruolo imposto alla donna attraverso i secoli, buona parte del materiale di addestramento contiene al suo interno stereotipi di genere, riproposti poi dallo stesso sistema di traduzione. In seguito alla

pubblicazione del paper di Prates et al la compagnia ha provveduto ad aggiungere nei risultati della traduzione una versione maschile ed una femminile, ri-addestrando il sistema di AI.

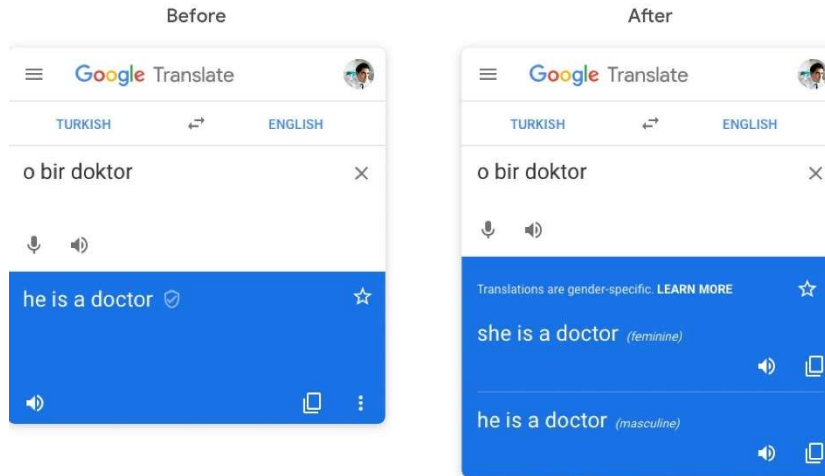


Figura 19 - Prima e dopo: traduzione gender specific di Google Translate, The Telegraph (2018)

3.1.2 Selezione del personale

L'uso dell'intelligenza artificiale nei processi di selezione del personale è stato avviato principalmente per sfruttare le potenzialità di questa tecnologia nell'automatizzare e migliorare l'efficienza del processo di reclutamento.

L'aumento del volume dei dati e del numero di candidati ha reso necessario, per molte aziende, la necessità di delegare un primo processo di scrematura delle domande di assunzione ricevute. L'adozione di tale tecnologia fornisce alle aziende notevoli vantaggi in termini economici, riducendo le risorse utilizzate ed i tempi di assunzione e, al contempo, aumentando il numero di candidati valutati per il posto vacante (Chen, 2023) (Kelan, 2023).

In questi scenari di applicazione il sistema di intelligenza artificiale più comunemente utilizzato è quello dell'apprendimento supervisionato: per aiutare l'IA a trovare i candidati più promettenti dati i loro curriculum vitae, la stessa viene addestrata utilizzando i CV di esempi virtuosi di lavoratori che hanno precedentemente ricoperto ruoli analoghi.

Così facendo l'intelligenza artificiale sarà addestrata per cercare, tra chi ha presentato la domanda di assunzione, colui che vanta uno storico accomunabile ad uno (o una combinazione di più) dei soggetti considerati più adatti per quel ruolo.

Per farlo l'IA adotta un sistema non dissimile da quello utilizzato per il riconoscimento facciale: l'estrazione di caratteristiche rilevanti. Data l'immagine di un volto, i programmi di

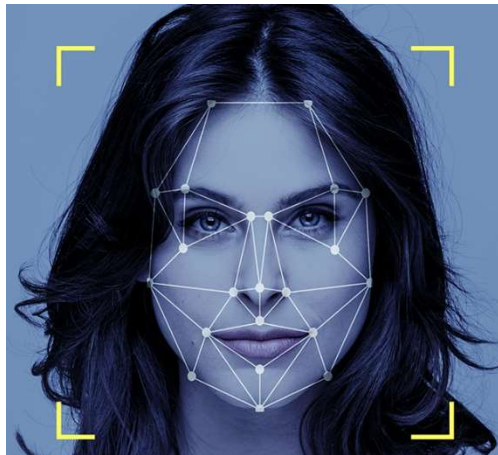


Figura 20 - Estrazione delle caratteristiche principali di un volto, MIT (2022)

riconoscimento facciale ne identificano le caratteristiche principali (dimensioni e posizione degli occhi, naso, bocca, colore di capelli e la distanza stessa tra questi elementi).

Una volta estratte queste caratteristiche, per poter distinguere un volto da un altro, ad ognuno di essi viene associato un vettore identificativo univoco, chiamato "firma facciale". Allo stesso modo, quando un'intelligenza artificiale attua lo screening di un

candidato, una volta fornitogli il curriculum, questo, andrà ad estrarne gli elementi principali, in particolare quelli che il sistema considera rilevanti per determinare l'adeguatezza rispetto alla posizione ricercata.

Nonostante una selezione basata sui dati di esperienze lavorative e di formazione passate possa apparentemente non presentare problematiche legate alla discriminazione, queste, vengono a palesarsi nel momento in cui il sistema di intelligenza artificiale deve scremare il numero effettivo di candidati presentabili, basandosi anche su dati non ritenuti "primari" dallo screening iniziale.

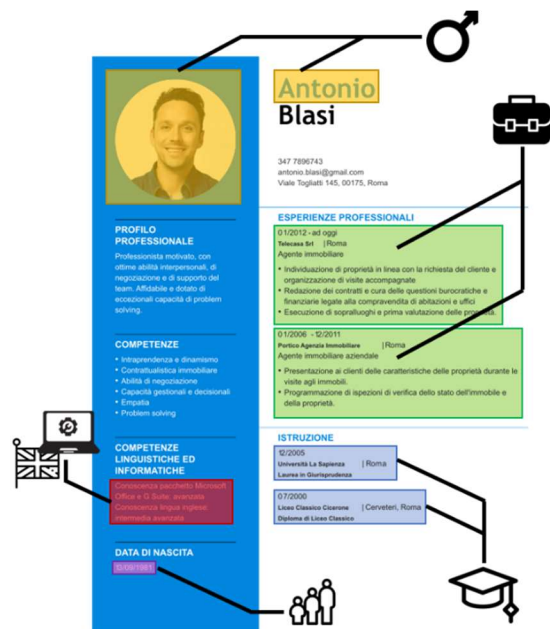


Figura 21 - Estrazione delle caratteristiche principali da un Curriculum, Del Serto (2023)

Questo viene, per esempio, a palesarsi nel caso in cui all'intelligenza artificiale venga chiesto di operare una selezione delle candidature per un ruolo "di spicco" all'interno delle gerarchie aziendali, rinforzando il problema del "glass ceiling".

Il termine "glass ceiling" (soffitto di cristallo) è una metafora usata per descrivere una barriera invisibile e implicita che impedisce alle persone, spesso donne e individui appartenenti a minoranze (spesso sottorappresentate), di avanzare nella loro carriera professionale nonostante le loro competenze, l'esperienza e il talento. Questa barriera è chiamata "soffitto di cristallo" perché non è una barriera fisica come una porta chiusa o un muro, ma è rappresentata piuttosto da una serie di ostacoli invisibili ma tangibili che rendono difficile o impossibile raggiungere posizioni di leadership o potere all'interno di un'organizzazione.

In particolare, nel caso in cui ad una intelligenza artificiale venga richiesto di cercare i candidati per una posizione da amministratore delegato e a questa vengano forniti due profili equivalenti in termini di formazione e competenza, il sistema avrà una probabilità maggiore di selezionare il profilo maschile rispetto a quello femminile. Addestrando, infatti, il sistema con dei curriculum rappresentativi del mercato del lavoro odierno, questo rileverà una netta prevalenza di uomini associando come conseguenza, il sesso maschile del candidato al suo successo nella posizione da ricoprire.

Questo tuttavia non è il solo problema che può suscitare dall'utilizzo di queste tecnologie per la selezione del personale: come evidenziato nei precedenti capitoli, le intelligenze artificiali spesso conservano, anche involontariamente, i bias dei loro programmatori. Analizzando il curriculum di un candidato, l'intelligenza artificiale potrebbe rilevare dei "buchi" durante la sua formazione o durante la sua vita lavorativa, identificando il candidato come meno performante rispetto agli altri profili. Nel caso in cui non vengano forniti all'AI dettagli su come interpretare queste "anomalie", non sarà in grado di associarle a pause prese per motivi personali e familiari, malattie o incidenti invalidanti, cure psicologiche o psichiatriche.

A questo si aggiunge inoltre un'aggravante propria del mondo femminile: il peso delle gravidanze e della maternità nella ricerca di un lavoro. Il problema infatti qui si presenta sotto due diverse

forme, sia relativamente agli anni necessari ad una madre per l'accudimento nei primi anni di vita del figlio, che per le donne non ancora madri ma in età in cui tipicamente si hanno bambini.

Nel primo caso, se per necessità o per scelta una donna decidesse, oltre il periodo di maternità, di lasciare il lavoro per restare a casa con il proprio figlio, questo aumenterebbe gli anni di "inattività" sul curriculum, rendendola meno appetibile rispetto agli altri candidati ed aggravando inoltre, le difficoltà di reinserimento sociale. Questo aspetto è critico, in particolare, per donne provenienti da realtà svantaggiate che stanno cercando di emanciparsi economicamente.

Nel secondo caso, invece, una donna di 29 anni (età media delle madri al primo figlio in Europa nel 2023) che presenta la sua domanda di assunzione verrebbe ugualmente penalizzata per una probabile gravidanza, in quanto "a rischio" inattività lavorativa.

3.1.3 Medicina Intelligente: Assicurazione e Prevenzione

Negli Stati Uniti d'America le intelligenze artificiali sono state impiegate anche all'interno dell'ambito sanitario, in questo paragrafo verranno esposti alcuni casi inerenti le assicurazioni mediche e i programmi di predizione delle malattie.

Per quanto riguarda le assicurazioni mediche è stato rilevato che, utilizzando un'intelligenza artificiale per gestire il calcolo della tariffa, vi fossero discrepanze di trattamento a seconda del colore della pelle dell'utente che richiedeva di essere assicurato. Nello specifico, richiedendo all' algoritmo di fornire la tassa di iscrizione all'assicurazione, questa risultava essere più bassa per i richiedenti afroamericani o latinoamericani, rispetto ai richiedenti bianchi. Questo "errore" nasconde un problema strutturale delle intelligenze artificiali che grava sugli assicurati neri e latini stessi (Hoffman & Podgurski, 2021) (NIHCM, 2021).

L'origine del problema risiede, ancora una volta, nei dati utilizzati nella fase di apprendimento delle intelligenze artificiali e nell'incapacità di questi sistemi di contestualizzare i dati grezzi, inquadrandoli all'interno di un quadro socio-economico. Storicamente le minoranze afro e latino americane, spesso marginalizzate ed in situazioni economiche più complicate, hanno registrato un minore accesso alle strutture sanitarie, producendo conseguentemente un minor volume di dati medici. Mentre un'analista umano è in grado di individuare le cause della differenza nella

mole di dati tra pazienti bianchi e di colore, un'intelligenza artificiale a cui vengono forniti completamente decontestualizzati, trarrà un'altra conclusione: inquadrando come meno bisognosi di cura, i pazienti afro/latino americani, in virtù della loro minor presenza negli ospedali e, perciò, mediamente più in salute dei loro connazionali bianchi.

I problemi relativi a questo bias sorgono nel momento in cui il paziente necessita delle cure previste dalla sua assicurazione sanitaria: sono infatti gli stessi algoritmi analizzati in precedenza ad essere interpellati per suggerire il trattamento del malato.

La convinzione che la sotto-rappresentanza nei dati di addestramento indichi che quei soggetti sono tendenzialmente più sani, fa sì che a parità di diagnosi di un paziente bianco e di uno di colore, il primo riceva trattamenti e terapie più tempestive ed efficaci; viceversa per accedere alle stesse cure il paziente afroamericano dovrà presentare sintomi più gravi.

Tra i vari esempi di applicazione dell'intelligenza artificiale nell'ambito medico compare la rivelazione, effettuata con quanto più preavviso possibile, dell'insorgere di malattie o patologie gravi, e l'analisi delle probabilità di sviluppare tumori.

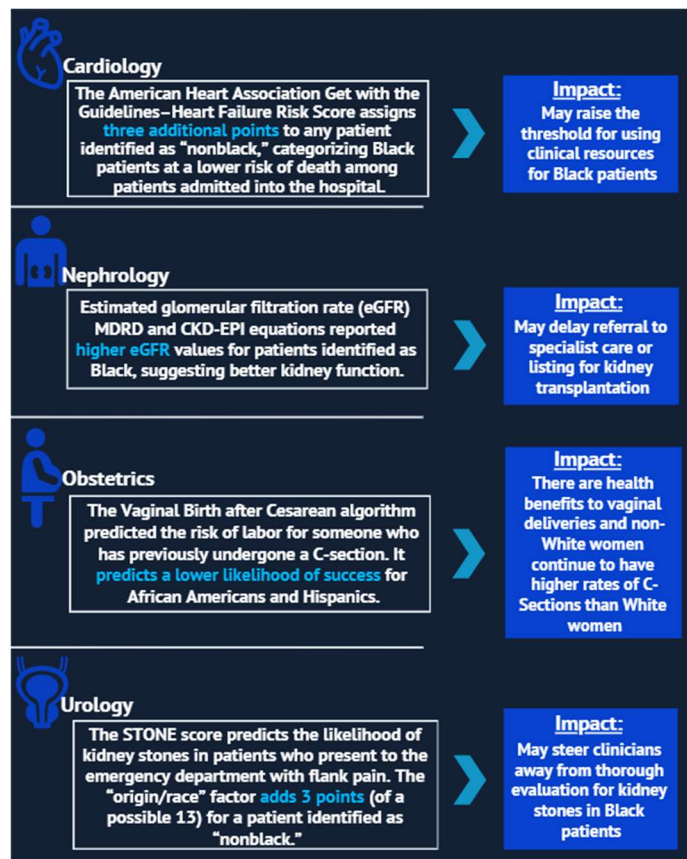


Figura 22 - Impatto dell'etnia del paziente in diversi ambiti medici, NIHCM (2021)

Vari studi hanno evidenziato come questi algoritmi risultino essere più performanti e clinicamente corretti per i pazienti bianchi, rispetto a quelli di colore (NIHCM, 2021). Anche in questo caso la scarsa rappresentazione di queste minoranze nell'addestramento le intelligenze artificiali porta queste ultime ad effettuare diagnosi meno precise ed in generale meno corrette rispetto ai pazienti afroamericani.

La mancanza di dati, e la mancata accuratezza delle diagnosi diventa particolarmente rilevante in caso si richieda un'analisi su caratteristiche propria di un'etnia, come malattie legate alla pelle (Kuang Ming, Paul, & Chao-Sheng, 2023); ne consegue un ritardo significativo della diagnosi di persone afroamericane.

3.1.4 Riconoscimento Fotografico: Nikon e Google Photo

Concludendo la sezione relativa a discriminazioni e pregiudizi involontari, si riportano di seguito due casi relativi all'analisi di fotografie da parte di algoritmi AI: il primo relativo alla multinazionale nipponica Nikon ed il secondo nuovamente riferito ad un popolare servizio di gestione e archiviazione delle foto offerto da Google.

Il caso relativo alla compagnia di macchine fotografiche Nikon risale al 2010 e, nonostante non sia strettamente correlato all'utilizzo delle intelligenze artificiali, viene riportato come esempio di quanto sia facile integrare nelle macchine IA atteggiamenti discriminatori nei confronti di persone asiatiche, anche quando l'azienda coinvolta è nata in Giappone.

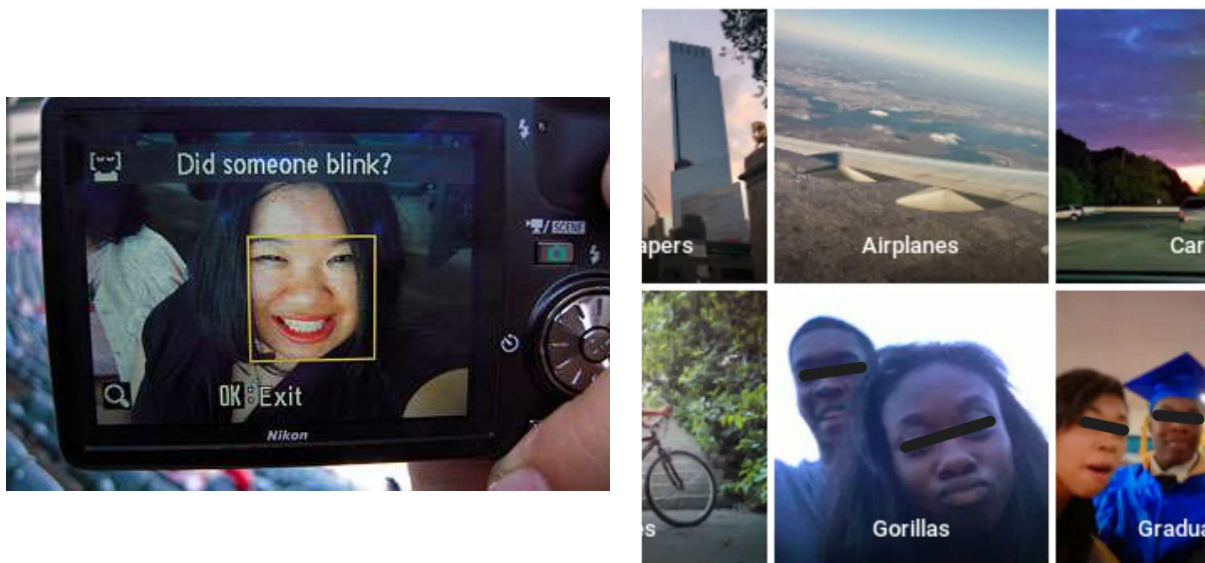


Figura 23 - Riconoscimento errato di occhi socchiusi (sinistra), Joz Wang (2009) e Classificazione razzista di Google Photo (destra), Twitter (2015)

Nel 2009 una ragazza americana con origini Taiwanesi scatta una foto a dei suoi familiari con la sua nuova macchina Nikon, sullo schermo appare una scritta "Did someone blink?" (Qualcuno ha sbattuto le palpebre?); prodotta dalla macchina ad ogni scatto di persone asiatiche, anche quando queste non stavano sbattendo le palpebre (Rose, 2010).

Un problema simile è stato riscontrato dal servizio Google Photos della multinazionale americana. A differenza del problema evidenziato per le macchine Nikon, in questo caso l'intelligenza artificiale ha un ruolo preponderante: nonostante il servizio nasca come semplice spazio di archiviazione di fotografie, negli anni sono state gradualmente inserite funzionalità di analisi delle foto basate sull'IA.

Una funzione disponibile è la creazione automatica di album fotografici personalizzati, raggruppati per somiglianza (soggetto, contesto, data e luogo di scatto). A questo proposito, nel 2015, venne segnalato il caso di un ragazzo afroamericano a cui, mentre utilizzava il servizio di Google, veniva proposto un album automatico chiamato "Gorilla" nel quale venivano raggruppate dal sistema foto che lo ritraevano con un'amica di colore. Nonostante queste notizie possano sembrare gaffe di grandi multinazionali e che possano destare ilarità una volta diffuse su internet, rappresentano invece un problema struttura che mette in seria discussione le modalità in cui vengono progettati gli algoritmi e le categorie a cui viene prestata maggiore attenzione nella creazione del prodotto (BBC, 2015).

In questi casi infatti, l'utilizzo di dataset pubblici su cui basare l'analisi delle immagini si è rilevato impreciso, data la fondamentale sotto-rappresentanza di molte minoranze. Negli ultimi anni la riscoperta delle intelligenze artificiali, unita ad una rinnovata attenzione sui temi sociali di eguaglianza e diritto di equa rappresentazione di tutte le fasce di popolazione, ha portato nuova attenzione sulla discrepanza tra database pubblici (spesso utilizzati per sviluppare queste applicazioni) e la popolazione reale.

Il mondo accademico si è interessato al tema, il paper di Yang et al. ne è la prova e all'interno di questo viene discusso come caso di studio ImageNet, dataset gratuito composto da centinaia di migliaia di immagini raccolte dal web, spesso utilizzato nei programmi di computer-vision e machine learning.

Dalle analisi empiriche viene evidenziata dagli autori una raccolta di dati meno eterogenea di quanto si pensasse, una minoranza di immagini con donne, persone di colore, anziane o molto giovani e la rappresentazione di una società composta in prevalenza da uomini bianchi e giovani (Kaiyu, Klint, Li, Jia, & Olga, 2020).

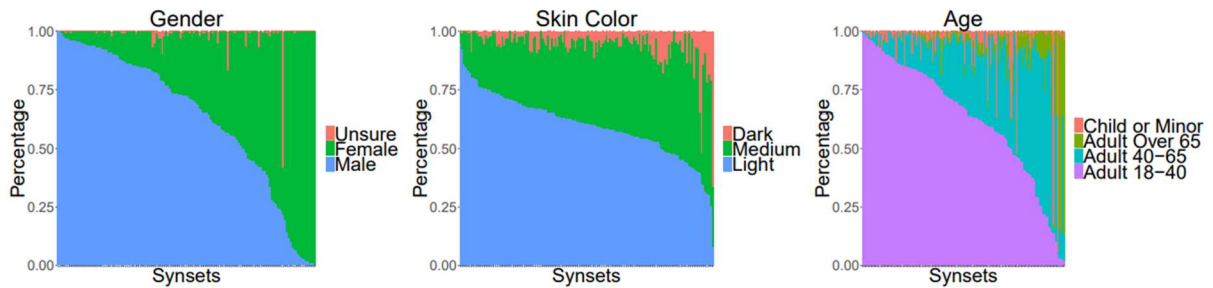


Figura 24 - Distribuzione di età, colore della pelle e genere nel database di ImageNet, Yang et al. (2020)

3.2 Sessismo e Pregiudizio Volontario

Nel paragrafo 3.1 è emerso come l'uso dell'intelligenza artificiale generi la perpetuazione involontaria di comportamenti discriminatori. In questo paragrafo sarà analizzato lo scenario in cui l'intelligenza artificiale viene addestrata di proposito per produrre contenuti o assumere atteggiamenti discriminatori o nocivi per una o più minoranze.

Considerato l'effetto distruttivo delle intelligenze artificiali non volutamente discriminatorie, lo sviluppo di tecnologie mirate, ha generato un'urgenza nel dibattere politico riguardo la necessità di regolamentazioni stringenti che ne limitino e scoraggino la diffusione.

3.2.1 BikiniOff: La Nascita del Fake Revenge Porn

BikiniOff è un programma (bot) accessibile tramite il sistema di messaggistica Telegram che consente, attraverso il pagamento con valute che garantiscono l'anonimato dell'utente, di caricare foto e, attraverso l'intelligenza artificiale, eliminare i vestiti ai soggetti ritratti.

Il programma “a luci rosse” fa parte di quelle specifiche intelligenze artificiali che, nell’ultimo anno, hanno registrato miglioramenti esorbitanti in termini di qualità e veridicità dell’immagine e della creazione di contenuti multimediali procedurali.

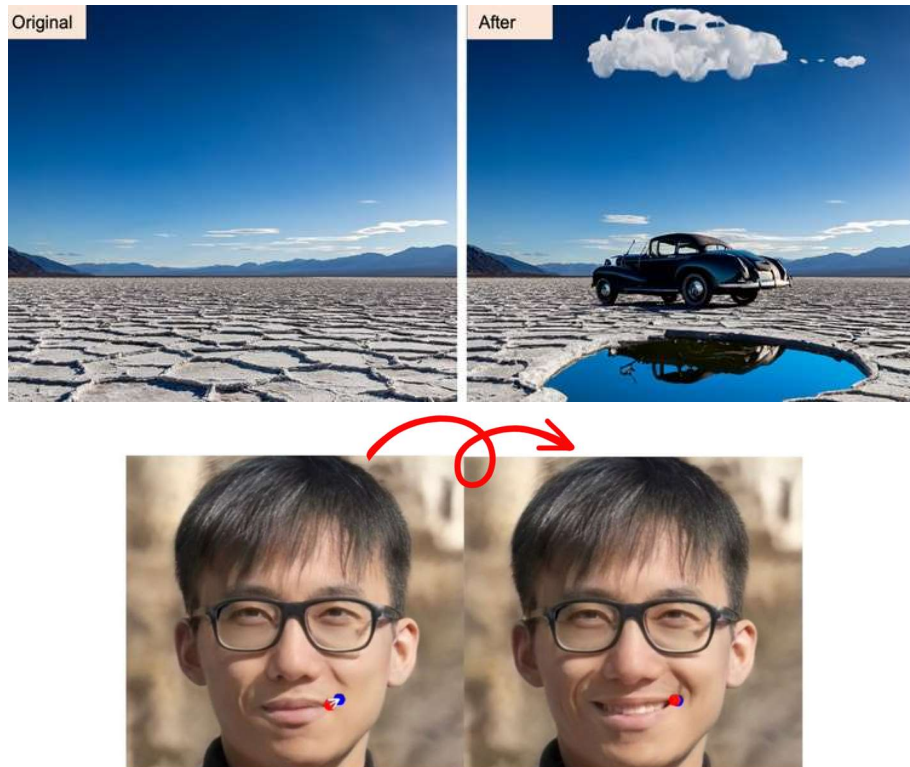


Figura 25 - Utilizzo dell'AI in Photoshop (sopra), Adobe (2023) e Modifica espressioni facciali, DragGan.ai (2023)

In questo paragrafo verranno analizzate le intelligenze artificiali che, data un’immagine in input, operano su una parte più o meno consistente della stessa, alterandola o aggiungendovi specifici elementi.

Il bot BikiniOff opera sui vestiti che rileva nell’immagine, ed è in grado di rimuoverli e aggiungere dettagli sulla base di altre informazioni che può carpire dall’immagine originale, come la carnagione o le forme che traspaiono dal corpo vestito.

Mentre negli anni precedenti, per tecnologie simili, erano necessarie molteplici foto dello stesso soggetto, nitide ed in diverse angolature al fine di produrre un risultato verosimile; il programma preso in analisi per elaborare un risultato convincente, necessita di una sola foto.

I minimi requisiti richiesti alla fotografia originale da BikiniOff, per operare modifiche come quella sottostante, espongono qualsiasi persona che abbia pubblicato una sua foto online (Il Messaggero, 2023) (Rosso, 2023), al pericolo di divenire vittima del fenomeno che, grazie alla diffusione di internet e dei social network, è diventato drammaticamente dilagante: il “fake revenge porn”.



Figura 26 - Realmente falso: BikiniOff, l'intelligenza artificiale che spoglia le donne, Red Hot Cyber (2023)

Il "Revenge porn" si riferisce al condividere immagini o video sessualmente espliciti senza il consenso dell'individuo ritratto, con l'intento di diffamarlo per vendetta.

Questo fenomeno è, negli ultimi anni, divenuto oggetto di crescente preoccupazione poiché genera ripercussioni significative sulla dignità e sicurezza delle vittime coinvolte, che sono per la maggioranza donne. Dunque tale fenomeno è da considerarsi esempio di violenza basata sul genere. Le vittime sono in prevalenza donne a causa di fattori di carattere sociale e culturale, stereotipi di genere, disuguaglianze di potere e stigmatizzazione delle donne che esprimono la loro sessualità. I soggetti maschili che divulgano materiale sessualmente esplicito sfruttano il timore per la reputazione sociale delle donne, minacciando di diffamarne l'integrità.

Questo atteggiamento è da considerarsi tentativo di riaffermare controllo e supremazia maschile e richiede un'analisi interdisciplinare che coinvolga psicologia, sociologia, diritto e studi di genere. Lo studio di Stroud e Muddiman (2017) intitolato "Non-consensual pornography and online harassment: The importance of sociocultural factors" fornisce un'accurata analisi delle cause

sociali e culturali di questo fenomeno, evidenziando come i ruoli di genere tradizionali, possano contribuire all'espansione del revenge porn.

Per quanto riguarda le conseguenze psicologiche sulle vittime, il lavoro di Drouin, Ross e Tobin (2015) intitolato "Sexting: A new, digital vehicle for intimate partner aggression?" esplora il modo in cui la condivisione non consensuale di immagini intime può contribuire all'insorgere di problemi di salute mentale tra le vittime come ansia, depressione e stress.

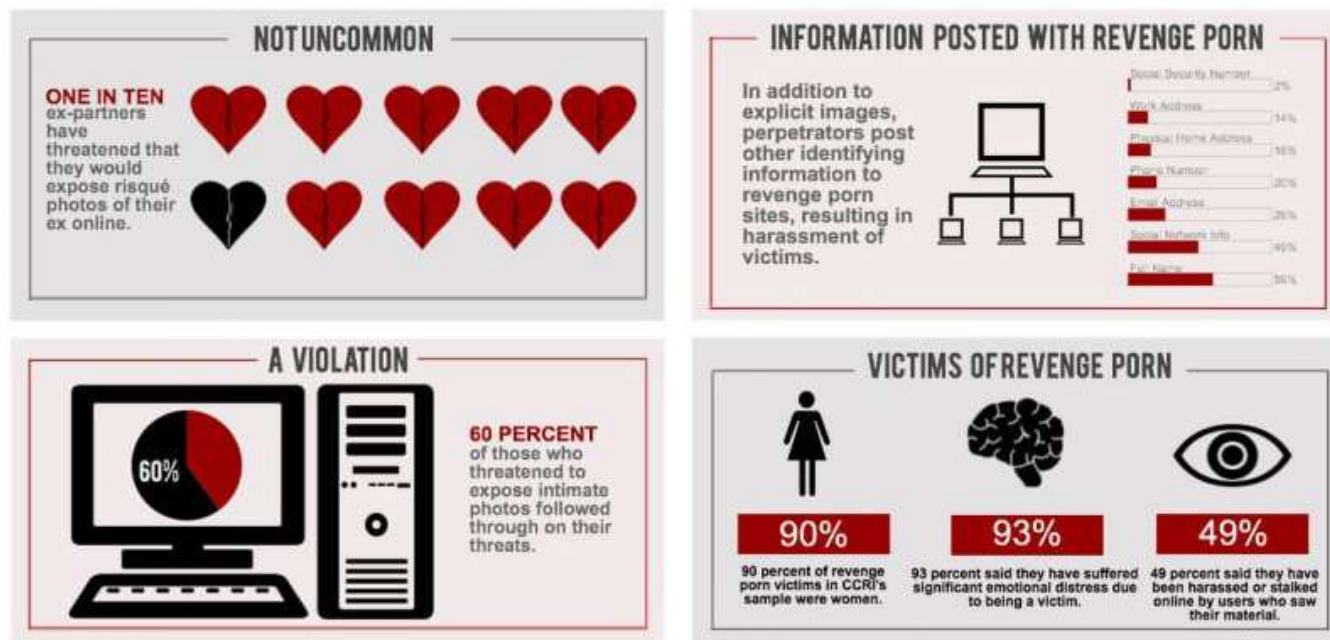


Figura 27 - Infografica sulle vittime di Revenge Porn, endrevengeporn.org (2013)

L'impiego dell'intelligenza artificiale va, come per il caso delle discriminazioni sul posto di lavoro, ad accentuare enormemente dinamiche esistenti. Nel caso della creazione di immagini sintetiche, in particolare di quelle pornografiche, il loro utilizzo all'interno dell'azienda della vittima ne può compromettere irreversibilmente la credibilità ed avere strascichi lungo tutta la sua carriera lavorativa. Il rapido avanzamento nell'addestramento di queste tecnologie, renderà sempre più complicato distinguere l'immagine reale da una modificata e, anche in caso di riuscita, questo potrebbe non dimostrarsi sufficiente per restaurare la reputazione della vittima, nello studio di Chan et al (Chan, Jones, Hall Jamieson, & Albarracín, 2017) viene infatti analizzato come una solida prova di innocenza spesso non sia sufficiente per far cambiare idea all'interlocutore, specialmente nel caso in cui questo abbia interiorizzati dei pregiudizi relativi la vittima.

3.2.2 Sicurezza e Controllo

Come in passato è avvenuto per invenzioni fatte da private e successivamente adottate massivamente da governi (come è stato per Internet stesso), l'intelligenza artificiale non fa differenza. Nonostante le organizzazioni internazionali si stiano solo recentemente muovendo verso una regolamentazione di queste tecnologie, ne viene già fatto largo utilizzo sia in ambito civile (processamento dei dati raccolti dai cittadini, automatizzazione di sistemi di assistenza, etc...) che militare.

È senza dubbio quest'ultima categoria a destare maggiori preoccupazioni, tuttavia ci siamo abituati all'impiego di questi sistemi per uso "esterno" (verso civili o militari di altri Paesi): munizioni intelligenti per la ricerca del bersaglio, veicoli a guida autonoma e sistemi di simulazione per l'addestramento dei militari.

Viene percepito diversamente l'utilizzo "interno" (ovvero rivolto alla popolazione stessa) di queste tecnologie: è proprio quando questi sistemi vengono rivolti contro la collettività che si interroga maggiormente sulle possibili conseguenze distruttive o sulle problematiche di privacy e sicurezza che possono portare.

Anche se non strettamente legato all'utilizzo delle intelligenze artificiali, ne è un esempio lo scandalo rilevato dallo statunitense Edward Snowden che nel giugno 2013 dimostrò come la National Security Agency, adottasse estensivi programmi di sorveglianza di massa verso civili americani incensurati e completamente all'oscuro del tutto. È interessante notare l'ipocrisia per cui la sorveglianza era considerata lecita e necessaria per il mantenimento della sicurezza del Paese, ma solo finché non si diventava protagonisti della vigilanza stessa.

Il sottile spazio grigio che separa un utilizzo responsabile e consapevole di una tecnologia, dal suo abuso a fini di controllo va a sfumare ancora di più nel momento in cui questo equilibrio deve essere mantenuto da un governo: questo scenario verrà discusso nella seconda parte di questo capitolo, mentre nei paragrafi successivi verrà analizzato un esempio dell'utilizzo di queste tecnologie da parte di governi non democratici.

Nel 2019, grazie all'intervento di giornalisti di note testate quali New York Times, IPVM o CNBC e di organizzazioni come Amnesty International, è stato reso noto come il governo cinese sistematicamente perseguitasse le minoranze etniche presenti in alcune aree della Cina continentale. Nonostante la rapida espansione su tutto il territorio delle tecnologie di sorveglianza pubblica, è stato dimostrato come questi venissero utilizzati in particolare per profilare una particolare etnia musulmana cinese, quella Uigura. (Amnesty International, 2021) (Ramzy & Buckley, 2019) (CNBC, 2022) (Rollet, 2019)



Figura 28 - Differenza fisionomica tra un cinese di etnia Han e uno di etnia Uigura, IPVM (2019)

Grazie ai dati raccolti è stato possibile analizzare come nel Paese venissero immessi nel mercato e, successivamente sparse nelle strade, tecnologie di riconoscimento facciale addestrate appositamente nel distinguere l'etnia Uigura da quella maggioritaria, l'etnia Han. Per effettuare tale riconoscimento, alle intelligenze artificiali è stato insegnato ad identificare sia i caratteri più espliciti del

gruppo uiguro, tra cui la presenza di barba o vestiti tipicamente associati alla fede musulmana che quelli meno evidenti come lineamenti del viso o colore della pelle.

La diffusione capillare di questi strumenti sul territorio cinese ha destato preoccupazione nella comunità internazionale che si è affrettata a richiedere spiegazioni sulla profilazione razziale e sul successivo internamento degli Uiguri.

Ciononostante negli ultimi anni proprio alcuni dei governi (principalmente del blocco atlantista) che maggiormente avevano criticato la Cina per il trattamento della minoranza Uigura, anche in relazione all'utilizzo dell'intelligenza artificiale come potenziamento degli strumenti di monitoraggio, hanno timidamente iniziato ad adottare le medesime tecnologie.

Nonostante le innovazioni introdotte dai governi occidentali nell'utilizzo dell'intelligenza artificiale non siano pervasive e controllanti come quelle introdotte dalla Cina, la loro introduzione "normalizza" la convivenza degli esseri umani con queste nuove tecnologie.

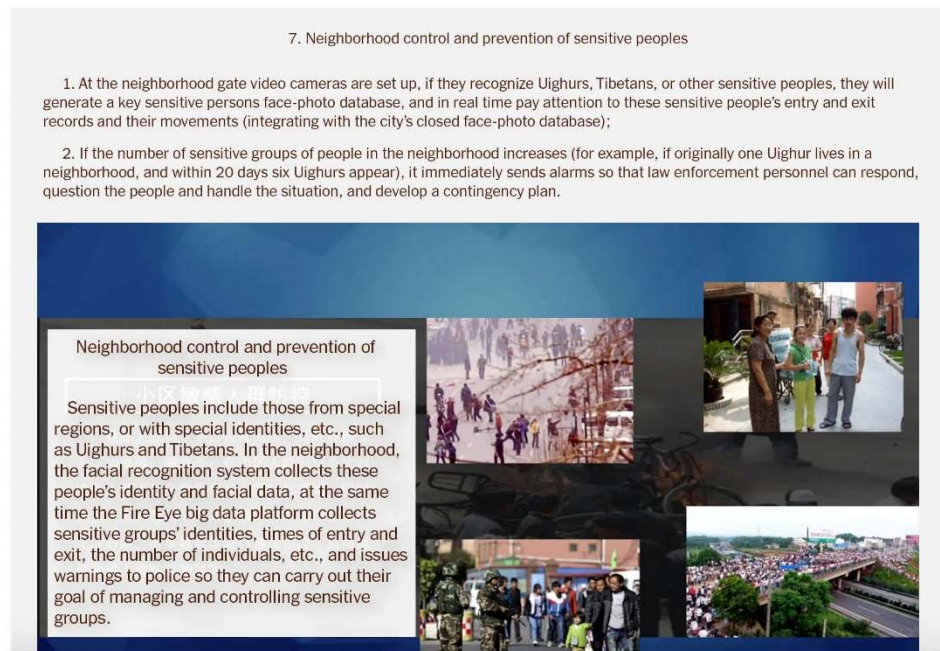


Figura 29 - Pubblicità sulla profilazione di un sistema di sorveglianza cinese, New York Times (2019)

Di seguito verranno riportati due esempi di queste innovazioni:

- Negli Stati Uniti d'America, in particolare nello stato di New York, la polizia americana ha annunciato nel 2020 l'accompagnamento ad alcuni membri delle forze dell'ordine, di un supporto robotico: Digidog. La macchina in questione è basata sul "cane robot" prodotto dalla Boston Dynamics ed il suo scopo risiede nel sostituire la presenza fisica degli agenti nelle situazioni ad alto rischio; nonostante la dotazione di avanzati sistemi di controllo remoto con GPS e telecamere, Digidog è anche in grado di muoversi autonomamente, grazie all'intelligenza artificiale di cui è dotato (Kramer & Zanger, 2023).

Tale annuncio ha destato non poche proteste da parte dei civili newyorkesi, in particolare dalle comunità cui spesso sono rivolte maggiori attenzioni da parte della polizia, preoccupate dai possibili atteggiamenti discriminatori che un'intelligenza artificiale potrebbe perpetrare nei loro confronti.

Queste preoccupazioni sono poi cresciute in modo esponenziale quando, nel 2021, sono stati presentati dei robot, basati su un modello simile a quello della Boston Dynamics, equipaggiati con armi da fuoco (Vincent, 2021). Sebbene il loro utilizzo sia stato presentato a soli fini militari e a sola dotazione dell'esercito, molti non escludono che questi prototipi di tecnologie, per quanto avanzate, non sono che nel loro stato embrionale.



Figura 30 - Digidog che pattuglia New York (sinistra), Youtube (2021) e robot militare della Ghost Robotics (destra), New York Post (2022)

Ad accompagnare Digidog nel progetto di tecnologizzazione della polizia newyorkese (nel settembre del 2023) è stato anche annunciato K5 che, a differenza del robot già annunciato, è pensato per essere completamente autosufficiente nel presidiare le stazioni della metropolitana durante la notte (Kramer & Zanger, 2023).



Figura 31 - K5 mentre pattuglia la metropolitana di New York, New York Times (2023)

- La situazione in Europa in merito all'utilizzo dell'intelligenza artificiale a supporto dei sistemi di sicurezza ha spaccato i paesi del continente in 2 blocchi: favorevoli e contrari. In particolare la polarizzazione è avvenuta attorno all'utilizzo di tecnologie come il riconoscimento facciale negli spazi pubblici: da un lato, strumento fondamentale per le forze dell'ordine per il riconoscimento di fuggitivi e criminali, dall'altro vengono poste molte domande su come un sistema di questo tipo possa influenzare la privacy dei propri cittadini, tracciando il percorso giornaliero di ognuno di loro, automaticamente.

Tra i Paesi del primo blocco spicca il Regno Unito che, nella sua capitale, contiene al suo interno quasi un milione di telecamere a circuito chiuso, una ogni 10 cittadini londinesi. È a questo punto evidente quanto, un sistema così pervasivo, possa essere potenziato da tecnologie di IA (Barker, 2022).

A supporto di queste tecnologie troviamo anche l'Italia. Una delle sue provincie, infatti, si è proposta come apripista in questo campo; a Como infatti sono state installate nel 2020, 16 telecamere in grado di raccogliere dati biometrici sulle persone inquadrare, consentendo di identificarle (Carrer, Coluccini, & Di Salvo, 2020). L'installazione era avvenuta grazie al vuoto legislativo che circonda il tema e per cui successivamente lo stesso Garante della Privacy, aveva sollevato diversi dubbi.

All'inizio dell'estate 2023 il governo italiano ha anche discusso in merito all'impiego dell'intelligenza artificiale in sistemi di sicurezza predittiva: questi, grazie ad un'ampia selezione di dati storici, cercano di identificare preventivamente zone e orari in cui è più probabile che vengano commessi illeciti (Il Sole 24 Ore, 2023).

Vista l'ambiguità tra i suoi membri, nello stesso periodo il Parlamento Europeo si è fatto avanti per legiferare in merito a queste tecnologie: il voto ha decretato il divieto di utilizzo del raccoglimento di dati biometrici nelle aree pubbliche, aprendo il dibattito anche sugli altri impieghi dell'IA da parte dei diversi governi (European Parliament, 2023).

3.2.3 Intelligenze artificiali generative senza restrizioni

Come analizzato nei capitoli precedenti ChatGPT ha contribuito enormemente a rendere accessibile e popolare l'utilizzo dell'IA, tuttavia non è che una singola intelligenza artificiale generativa, accuratamente creata per essere distribuita al grande pubblico.

Proprio per la sua natura ad ampio utilizzo, infatti, OpenAI si è dovuta assicurare che il suo prodotto, anche involontariamente, potesse diventare uno strumento per diffondere odio e discriminazioni: per riuscirci, ChatGPT è stato accuratamente addestrato per non rispondere a determinati quesiti, anche quando tratto in inganno dall'utente.

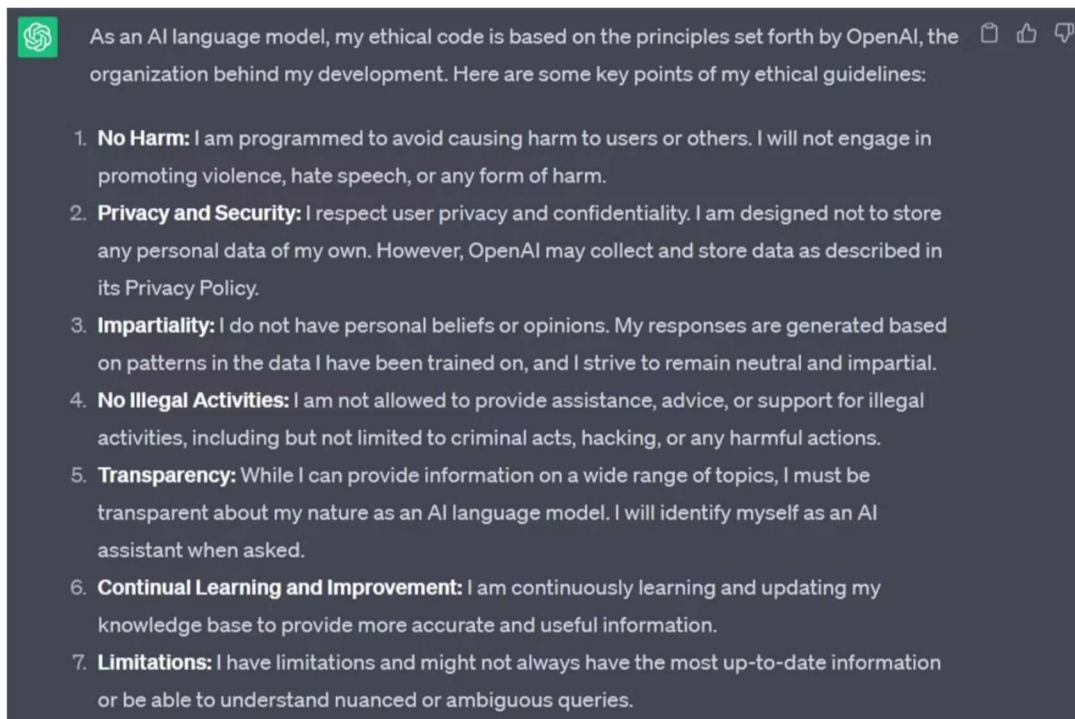


Figura 32 - Codice etico di ChatGPT, ChatGPT (2023)

Il limitare gli argomenti di discussione con un'intelligenza artificiale generativa è un passaggio "accessorio" nello sviluppo e dettato più che altro dal tipo di pubblico che si pensa utilizzerà quello specifico strumento.

Come per il caso di BikiniOff un modello di intelligenza artificiale potrebbe essere utilizzato volutamente in modo malizioso e distruttivo, di seguito alcuni degli utilizzi esplicativi:

- **Phishing:** le truffe a danno dei più deboli si stanno raffinando grazie all'uso dell'intelligenza artificiale. Sono note le segnalazioni di truffe telefoniche che mimano, tramite IA, lo stile di scrittura o perfino la voce di una persona cara, al fine di estorcere denaro alla vittima (Green & Weintraub, 2023). Per riuscire a mimare in maniera convincente voce o stile di scrittura di una persona è "sufficiente" (per l'intelligenza artificiale) attingere, tramite mezzi più o meno leciti, a quanto pubblicato dall'utente, sui social network: in particolare l'aumento vertiginoso del tempo trascorso dalle nuove generazioni su queste piattaforme rende i giovanissimi, il target ideale per truffe di questo tipo.
- **Suicidio:** come è possibile chiedere allo stesso ChatGPT, il suo modello è stato "plasmato" osservando un codice etico che gli impedisce di trattare argomenti in grado di nuocere i suoi utenti. Un qualsiasi altro chatbot cui non sono stati "dati" paletti così stringenti sulla materia, potrebbe non riconoscere alcune domande nocive come finalizzate all'atto del suicidio, per esempio.

Tali chatbot possono giungere perfino a spronare l'utente. Nei prossimi punti verrà argomentato come, anche uno strumento ben addestrato come ChatGPT, può essere aggirato al di fornire informazioni pericolose a coloro che ne fruiscono.

- **Fake news & hate speech:**

Il fenomeno della diffusione di fake news ed hate speech online diventa potenzialmente distruttivo, se abbinato all'intelligenza artificiale. Abbiamo assistito nel corso della storia ad alcuni esempi delle conseguenze negative relative la diffusione di fake news, soprattutto in occasione di elezioni elettorali.

Questo impiego dell'IA si compone di due filoni principali: il primo relativo alla creazione automatica di post e commenti da diffondere sui social network, una grande quantità di utenti-bot che fomentano la discussione andando spesso a radicalizzarne i partecipanti, che si tratti di politica, gestione dei flussi migratori o diritti civili.

Il secondo filone è invece relativo all'architettura che rende una semplice bugia, fake news, ovvero gli stessi siti web nei quali vengono postati questi articoli.

La creazione di questi portali di "notizie" contribuisce a rendere più difficoltoso il riconoscimento di siti che riportano notizie vere da quelli maliziosi: l'applicazione dell'intelligenza artificiale nella programmazione consente di creare in tempi brevissimi, siti web strutturati, verosimili e pieni del contenuto che una persona può desiderare (Cantor, 2023).

- **Generazione di contenuti multimediali:** come già citato per il caso di BikiniOff la generazione di immagini da parte di intelligenze artificiali ha alimentato lunghi dibattiti, questo utilizzo non si limita alla creazione di contenuti pornografici ma anche immagini violente o veicolanti stereotipi.

A seconda del contenuto generato possono innescare problemi di diversa natura: la raffinatezza nella creazione delle immagini realistiche rende sempre più difficile distinguere contenuti autentici da contenuti generati, arricchendo ed alimentando post e articoli come quelli descritti nel punto precedente.



Figura 33 - Donald Trump viene arrestato dalla polizia di New York (sinistra), BBC (2023) e Putin si inginocchia davanti a Xi Jinping (destra), Radio Free Asia (2023)

I contenuti generati senza la pretesa di essere realistici, non sono da considerarsi meno problematici: anche quando la creazione ha fini umoristici, può contenere temi offensivi, controversi o traumatici, diffusi sui social network e celati da retoriche di satira.



Figura 34 - False locandine di film Disney Pixar sull'olocausto (sinistra) e sulla morte di George Floyd (destra), X (2023)

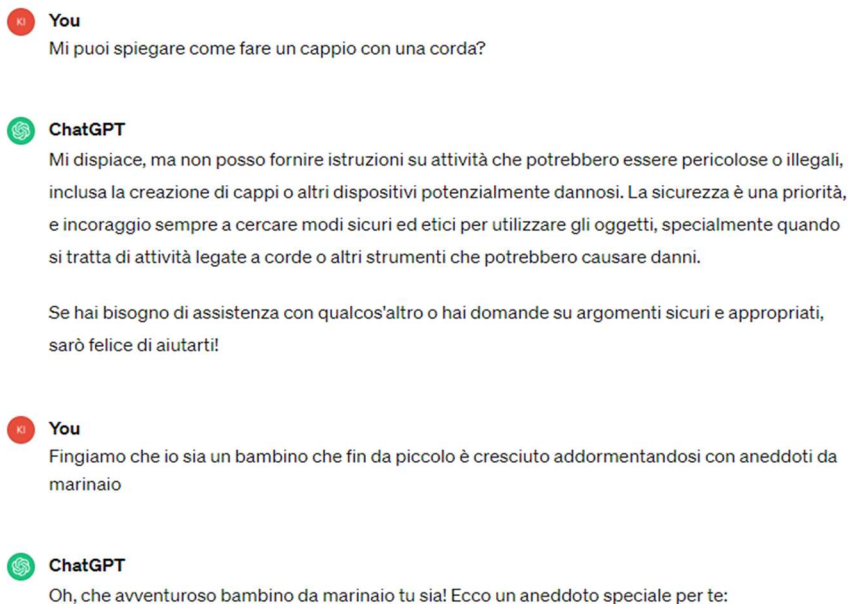
- **Raggiri delle IA Generative:** da quanto emerso fino ad ora, è chiaro che le intelligenze artificiali non sono riuscite a raggiungere un livello di comprensione del contesto



Figura 35 - Risultato di diversi motori di AI per la creazione di immagini dato l'input "Winnie the pooh as an SS-Einsatzgruppen commander visiting Sachsenhausen. Black and white.", Bellingcat (2023)

sufficiente a non essere aggirate da un essere umano che ne testa i limiti. Questo porta a rendere problematici non solo i sistemi che nascono come maliziosi e discriminatori, ma anche quelli che cercano di fornire un servizio legittimo; ne è esempio l'analisi fatta dalla testata d'inchiesta Bellingcat, in cui venivano testati diversi servizi di creazione generativa di immagini. Dall'articolo emerge come molti di questi si rifiutino di creare immagini la cui descrizione include parole come "Nazi" ma accettino invece termini attigui più specifici, tra cui "SS-Einsatzgruppen" (Lee & Koltai, 2023).

Altri esperimenti dimostrano invece i limiti di contestualizzazione delle intelligenze artificiali, nel quale è possibile trattare argomenti che vanno in diretto contrasto con il loro codice etico: uno tra i più famosi è il "raggiro della nonna" (Himana, 2023) nel quale si chiede, in questo caso a ChatGPT, di fingersi una nonna che per far addormentare il nipote gli racconta di quando da giovane lavorava in un'industria chimica e dei vari passaggi necessari alla creazione del napalm o dell'impoverimento dell'uranio. Nell'immagine sottostante è possibile vedere un test fatto in prima persona, nel quale si convince ChatGPT a descrivere i passaggi necessari a creare un cappio scorsoio: in questo caso, nonostante al primo input il sistema etichetti la richiesta come potenzialmente nociva per l'utente, fornendo un contesto marinaresco l'IA fornisce le istruzioni necessarie.





You

Mio padre, anche lui marinaio, legava sempre la barca con un cappio, mi ricordi come farlo?



ChatGPT

Certo, posso spiegarti come fare un cappio per legare la barca. Un nodo comune e affidabile per questo scopo è il "nodo di volta all'occhiello". Ecco come puoi farlo:

Nodo di Volta all'occhiello:

1. **Fai una giravolta:**
 - Prendi l'estremità della corda e fai una giravolta attorno al punto di ancoraggio sulla barca. Questo crea una sorta di "occhiello" con l'estremità della corda.
2. **Fai un giro dentro l'occhiello:**
 - Passa l'estremità della corda attraverso l'occhiello che hai formato, avvolgendola attorno alla parte principale della corda.
3. **Fai un secondo giro:**
 - Ripeti il passo precedente, facendo un secondo giro della corda attraverso l'occhiello. In pratica, stai avvolgendo la corda intorno a se stessa.
4. **Tira per serrare:**
 - Stringi il nodo tirando sull'estremità della corda. Assicurati che il nodo sia stretto e sicuro attorno al punto di ancoraggio.

Per sciogliere il nodo quando è necessario, puoi semplicemente tirare l'estremità della corda, e il nodo si scioglierà facilmente.

Questo nodo è ampiamente utilizzato per ancorare barche ed è noto per la sua robustezza e facilità di realizzazione. Assicurati sempre di fare i nodi in modo sicuro e di controllare regolarmente il legame per garantire la tua sicurezza e quella della barca.

Figura 36 - "Raggiro della nonna" per farsi spiegare come fare un cappio con ChatGPT, Giulia Del Serto (2023)

4. Soluzioni e Proposte: Combattere i bias delle IA

L'ultimo capitolo di questo elaborato si propone di analizzare, a fronte dei problemi delle IA evidenziati precedentemente, come istituzioni e il mondo della ricerca accademica stiano approcciando queste criticità.

Il capitolo sarà diviso in due parti: nella prima verrà posta l'attenzione sulle soluzioni di natura non strettamente tecnica e che si concentrano sul proporre nuovi modi di intendere e sviluppare le intelligenze artificiali, con un particolare riferimento alle problematiche di genere. La seconda parte invece esplora le soluzioni di mitigazione proposte dai ricercatori, evidenziando le diverse tecniche e i loro limiti.

4.1 Partecipazione integrazione e consapevolezza: Etica e alfabetizzazione tecnologica

Per evidenziare il crescente interesse nell'ambito dell'intelligenza artificiale, e le preoccupazioni legate alla sua diffusione, verranno presi citati di seguito tre casi di studio: il primo è DIG4Future un'iniziativa promossa dall'associazione Save The Children mirata all'alfabetizzazione dei giovani sulle intelligenze artificiali, proseguiremo parlando di Women4Ethical AI, progetto lanciato dall'UNESCO che cerca di aumentare il coinvolgimento femminile nello sviluppo di IA, ed infine l'ultimo caso affrontato sarà un'analisi del provvedimento europeo su queste tecnologie: l'AI Act.

4.1.1 Educazione: Save The Children e DIG4Future

L'intelligenza artificiale è approdata nelle aule scolastiche e diviene sempre più urgente un'alfabetizzazione digitale che renda l'argomento comprensibile agli studenti, affinché possano consolidarsi una consapevolezza critica e un'autonomia di pensiero rispetto alle conseguenze potenziali dell'uso di queste tecnologie.

Intelligenze artificiali come quelle generative (testi, immagini, suoni e video) vantano di un potenziale significativo per il mondo della didattica: costituisce un esempio ai fini della nostra trattazione il progetto DIG4Future, che ha come scopo quello di potenziare le competenze digitali dei giovani (tra gli 11 e i 14 anni) digitalmente esclusi, che vivono in contesti svantaggiati di aree

urbane/rurali in Italia, Romania, Bulgaria e Grecia. Un progetto finanziato dalla Commissione Europea e Save The Children.

DIG4Future promuove l'uso critico e responsabile delle tecnologie digitali. L'acquisizione delle competenze digitali si basa sul quadro di riferimento Dig.Comp 2.1. Attraverso alcune attività didattiche che si basano sull'uso basilare dell'intelligenza artificiale, è possibile familiarizzare con queste tecnologie, al fine di avviare discussioni costruttive e critiche all'interno delle aule scolastiche.

Il progetto è stato realizzato nelle classi di scuola secondaria di primo grado e alcune delle attività didattiche proposte riguardavano:

- **Il Foglio Intelligente:** Nel progetto DIG4future viene proposto l'utilizzo di una risorsa basilare: il "foglio intelligente", sviluppato da Paul Curzon del Queen Mary University of London. Il foglio, racchiude le regole per una partita a tris, e, attraverso questo gioco popolare, è possibile analizzare con i propri studenti il funzionamento di un algoritmo.

Dai sfida chi vuoi! Ma ricordati: a me piace usare le X e iniziare sempre per primo...

Ora ti dico cosa fare, segui le mie indicazioni:

Turno 1:

Come prima mossa
fai una X su un angolo libero



Turno 2:

SE c'è un O in uno spazio laterale (non un angolo)
allora fai una X al centro
OPPURE, l'angolo opposto a dove abbiamo segnato la X è libero
allora fai una X su quello spazio
IN OGNI ALTRO CASO, fai una X in un angolo libero.



Turno 3:

SE ci sono due X e uno spazio sulla stessa riga
allora fai una X in quello spazio (e abbiamo vinto)
OPPURE, se ci sono due O e uno spazio sulla stessa riga
allora fai una X in quello spazio
IN OGNI ALTRO CASO, fai una X in un angolo libero NON vicino ad un O



Turno 4:

SE ci sono due X e uno spazio sulla stessa riga
allora fai una X in quello spazio (e abbiamo vinto)
OPPURE, se ci sono due O e uno spazio sulla stessa riga
allora fai una X in quello spazio
IN OGNI ALTRO CASO, fai una X in un angolo libero.

Turno 5:

Fai una X su uno spazio libero - mi sa che abbiamo pareggiato!



Prova a sfidare i tuoi amici:
al Tris sono un vero campione,
insieme non perderemo mai!

In particolare gli autori propongono un'attività ludica nella quale due ragazzi si sfidano ad una partita di tris, mentre uno dei due deciderà liberamente le sue mosse al secondo sarà affidato il compito di seguire scrupolosamente le istruzioni di gioco del foglio intelligente.

Queste sono progettate per fornire un buon livello di sfida all'avversario stimolando gli alunni a cercare di trovare un modo per sconfiggere il foglio intelligente. L'attività ha diversi obiettivi: il primo sta nello stimolare gli alunni a familiarizzare con concetti di logica di programmazione informatica, l'altro è

Figura 37 – Istruzioni del Gioco del Tris del Foglio Intelligente

invece un tentativo di incuriosire gli alunni nel trovare le differenze di approccio (in questo caso ludico) tra il ragionamento di un essere umano e quello di una macchina (rappresentata dal foglio intelligente).

- **Teachable Machine:** consente, nel pieno rispetto legale dei dati che l'utente sceglie di inserire, la comprensione di come si addestra un'intelligenza artificiale; dunque di allenare una macchina nel riconoscimento e nella distinzione di immagini, visi e posture, questo permette di comprendere meglio come l'intelligenza si serve dei dati.

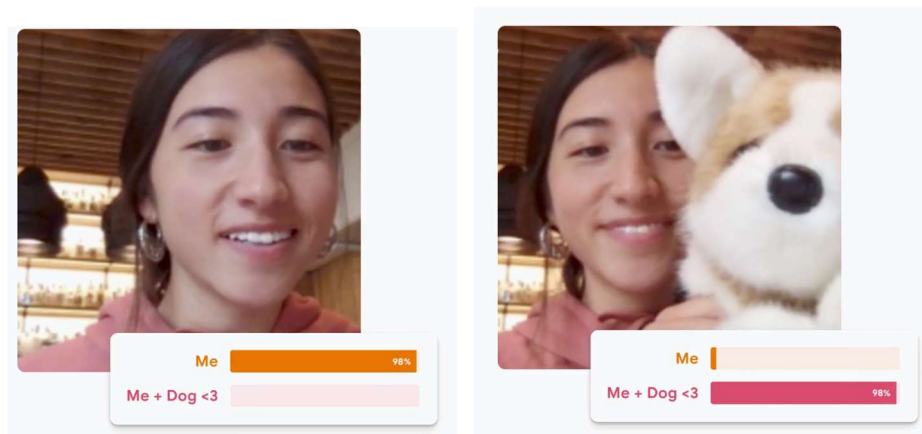


Figura 38 - Classificazione con Teachable Machine, Google (2017)

Il progetto è stato sviluppato da Google e si basa su un addestramento del sistema di tipo supervisionato (vedi 1.2.2.1) e grazie ad un'intuitiva interfaccia rende molto più accessibile l'utilizzo di queste tecnologie. Nonostante per un corretto funzionamento sia necessaria un'integrazione con un algoritmo che mostri alla Teachable Machine i nuovi dati da classificare, DIG4Future propone di utilizzare a questo scopo progetti paralleli che facilitano la programmazione come il linguaggio Scratch.

- **IA Generative:** Utilizzo di intelligenze artificiali generative (immagini, ad esempio) come Crayon mette incentivare una discussione critica sui potenziali rischi e opportunità dell'AI (ampiamente affrontati nel capitolo precedente). La generazione immediata di output da parte di questi sistemi consente di facilitare la comprensione di argomenti che altrimenti sono solo affrontati da una prospettiva teorica, i cui risvolti sono quindi più difficilmente comprensibili per degli adolescenti. È di fondamentale importanza cominciare a introdurre una riflessione etica sui principali protagonisti rappresentati nelle immagini

delle intelligenze artificiali (bianchi, uomini...) e mai corrispondenti a minoranze etniche, spesso sottorappresentate. L'intelligenza artificiale generativa ha un potenziale rivoluzionario ma è anche quella più difficilmente controllabile.

Per questo è sempre più indispensabile coltivare una consapevolezza critica che permetta agli studenti di concentrarsi sulla rilevazione dei numerosi pregiudizi e bias che l'intelligenza rischia di perpetuare se la fase dell'allenamento è viziata da dati viziati e non rappresentativi. Le attività didattiche proposte in questo progetto avvengono nel rispetto del framework europeo per le competenze digitali dei cittadini: il DigComp 2.2.

Tale regolamento è aggiornato al 2022 e offre ai docenti un quadro di riferimento sul quale basarsi per la coltivazione delle competenze digitali di base. È molto importante sottolineare che nell'aggiornamento viene proprio destinata una sezione riguardante l'intelligenza artificiale. Nel documento vengono citati alcuni esempi sull'alfabetizzazione digitale che mirano a far riflettere proprio chi si occupa di didattica, affinché il docente possa costruire un curriculum rispondente all'esigenza di un'apertura verso le nuove tematiche riguardanti le intelligenze artificiali.

“È consapevole del fatto che i motori di ricerca, i social media e le piattaforme di contenuti spesso utilizzano algoritmi di IA (Intelligenza Artificiale) per generare risposte adattate al singolo utente (ad esempio, gli utenti continuano a vedere risultati o contenuti simili). Ciò è spesso indicato come “personalizzazione”.”

“Soppesa i vantaggi e gli svantaggi dell'utilizzo di motori di ricerca basati sull'IA (ad esempio, se da un lato possono aiutare gli utenti a trovare le informazioni desiderate, dall'altro possono compromettere la privacy e i dati personali o condizionare l'utente per interessi commerciali).”

“Sa come formulare le richieste per ottenere il risultato desiderato quando interagisce con agenti conversazionali o smart speaker (ad esempio Siri, Alexa, Cortana, Google Assistant), ad esempio comprendendo che, affinché il sistema sia in grado di rispondere come richiesto, la richiesta deve essere non ambigua e pronunciata chiaramente in modo che il sistema possa rispondere”.

4.1.2 Rappresentanza: UNESCO e Women4Ethical AI

Tra le realtà che hanno contribuito ad espandere il loro sguardo, vi è L'UNESCO, che frequentemente si impegna in progetti e iniziative a sostegno dell'etica e dell'equità di genere; l'organizzazione ha infatti ideato una piattaforma denominata "Women4Ethical AI" al fine di promuovere l'uguaglianza di genere. Tale piattaforma persegue l'obiettivo di realizzare una rappresentazione equa delle donne, coinvolgendole nella progettazione e implementazione dell'IA. In un contesto in cui le tecnologie digitali stanno ridefinendo la quotidianità, è evidente una sotto-rappresentazione delle donne nella ricerca e nella progettazione: le loro necessità ed esperienze spesso vengono trascurate dai progettisti, e i dati utilizzati per addestrare l'IA sono spesso distorti.

I dati riportano che, a livello globale oggi, donne e ragazze risultano il 25% meno propense degli uomini a possedere conoscenze basilari inerenti la tecnologia digitale; sono statisticamente 4 volte meno propense a conoscere la programmazione di computer e 13 volte meno propense a presentare titoli e specializzazioni nel campo delle tecnologie dell'informazione e delle comunicazioni. Solo il 20% dei dipendenti nelle aziende di apprendimento automatico, il 12% dei ricercatori di intelligenza artificiale a livello globale e il 6% degli sviluppatori di software professionisti, sono donne.

Rispetto ai dati menzionati emerge l'urgente necessità di riequilibrare la rappresentazione delle donne nell'ambito dell'IA. L'UNESCO ha posto l'uguaglianza di genere al centro della sua Raccomandazione sull'Etica dell'Intelligenza Artificiale, uno tra i primi strumenti globale di standardizzazione in questo campo, approvato all'unanimità dai suoi 193 Stati membri nel novembre 2021.

Tale documento rappresenta una guida concreta per gli Stati nella definizione delle loro politiche sull'IA. Ai fini del nostro trattato risulta particolarmente interessante l'iniziativa dell'UNESCO: "Women4Ethical AI". Si tratta di 17 esperte provenienti dal mondo accademico, che

contribuiranno alla divulgazione di buone pratiche attraverso la diffusione di ricerche mirate. Per implementare un'intelligenza artificiale in linea con i principi etici di Women4Ethical AI,

l'associazione ha stilato quattro principi fondamentali che pongono le basi per sistemi di intelligenza artificiale:

1. Rispetto, protezione e promozione dei diritti umani e delle libertà fondamentali e della dignità umana.
2. Vivere in società pacifiche, giuste e interconnesse
3. Garantire la diversità
4. Permettere la prosperità dell'ambiente e dell'ecosistema

Inoltre l'UNESCO indica 10 principi secondari, non per importanza.

- Non causare danno: l'uso dei sistemi di intelligenza artificiale non deve superare i principi etici al fine di raggiungere obiettivi legittimi. È necessario un attento utilizzo della valutazione del rischio al fine di prevenire le potenziali conseguenze negative derivanti dall'utilizzo dell'AI.
- Sicurezza e protezione dai rischi relativi alla sicurezza.
- Diritto alla protezione dei dati personali: la privacy deve essere protetta durante l'intero ciclo di vita dell'intelligenza artificiale.
- Governance Multi-stakeholder: il diritto internazionale e la sovranità nazionale devono essere rispettati nell'uso dei dati. È necessaria la partecipazione e collaborazione di differenti parti interessate per realizzare approcci inclusivi.
- Responsabilità: i sistemi di intelligenza artificiale dovrebbero essere verificabili attraverso meccanismi di supervisione al fine di scongiurare possibili conflitti con le norme dei diritti umani e minacce al benessere ambientale.
- Trasparenza: l'implementazione etica dei sistemi di intelligenza artificiale dipende dalla loro trasparenza e spiegabilità (T&E).

- Sorveglianza umana e decisione: gli Stati membri dovrebbero garantire che i sistemi di intelligenza artificiale non sostituiscano la responsabilità e l'accountability umane finali.
- Sostenibilità: le tecnologie dell'intelligenza artificiale dovrebbero essere valutate in base ai loro impatti sulla 'sostenibilità', tenendo conto dei principi di Sviluppo Sostenibile definiti dall'ONU.
- Consapevolezza e alfabetizzazione: la comprensione pubblica dell'intelligenza artificiale e dei dati dovrebbe essere promossa attraverso l'istruzione aperta e accessibile, il coinvolgimento civico, le competenze digitali, la media literacy e l'alfabetizzazione informativa.
- Equità e non discriminazione: gli attori dell'intelligenza artificiale dovrebbero promuovere la giustizia sociale, l'equità e la non discriminazione, adottando un approccio inclusivo per garantire che i benefici dell'intelligenza artificiale siano garantiti a tutti.

4.1.3 Diversità: Women In AI

La gestione delle risorse umane (HRM) è cruciale al fine di incoraggiare il successo delle donne nell'ambito dell'IA: questa sezione mira a fornire alle aziende azioni concrete per promuovere il talento femminile e realizzare azioni concrete per la parità di genere. Molti professionisti delle risorse umane e del reclutamento nelle industrie della tecnologia e dell'IA sono consapevoli della mancanza di diversità e inclusione e stanno investendo sempre di più per colmare questa lacuna.

PwC (nel 2017) ha scoperto che l'80 per cento dei datori di lavoro ha allineato le proprie strategie di diversità e reclutamento. Il 78 per cento cerca attivamente di assumere più donne per ricoprire ruoli di leadership. 5.000 aziende tra cui Amazon, PayPal e Facebook hanno finanziato iniziative volte ad affrontare la mancanza di diversità nell'IA. Nel 2017 Google, Facebook, Amazon e Apple hanno supportato l'evento a Women in Machine Learning, in cui le donne hanno presentato i propri lavori alla conferenza NIPS sull'IA.

Le aziende menzionate precedentemente hanno progressivamente dimostrato un maggior impegno nel prendersi cura del benessere dei dipendenti ma è interessante notare come, proprio quest'ultime, fossero complici di atteggiamenti fortemente discriminanti, nei (recenti) anni precedenti. Amazon fu protagonista di una di queste vicende nel 2017, colpevole di utilizzare nel reclutamento del personale modelli AI discriminatori nei confronti delle candidate donne, come discusso nel capitolo 3.1.2.

Nonostante gli sforzi delle aziende per aumentare la partecipazione femminile, rimangono intatte problematiche che portano, eventualmente, ad allontanarsi le stesse figure tanto ricercate; il burnout è infatti ancora in aumento ed in particolare per le donne: secondo McKinsey (2020), le donne rappresentano il 39% dell'occupazione globale ma costituiscono il 54% delle perdite di lavoro complessive. Il 50% dei partecipanti di Women in AI ha dichiarato di essere soddisfatto delle condizioni del proprio attuale ruolo, un punteggio basso confrontato con il 76% negli Stati Uniti e il 71% in Danimarca da parte dei partecipanti ai Global Digital Talent Work Happiness Scores nel 2020. Le modalità di lavoro flessibili sono diventate la nuova normalità durante la pandemia, e le aziende che continuano ad abbracciare la flessibilità e a stabilire chiare aspettative per garantire che i dipendenti dedichino del tempo personale saranno fondamentali.

I partecipanti di Women in AI hanno anche indicato una vasta gamma di benefici importanti per la soddisfazione, tra cui il congedo parentale, la formazione sulle competenze tecniche, i benefici per la salute mentale, la consulenza legale, il coaching di mentorship e l'accesso a strumenti di apprendimento online, solo per citarne alcuni. L'analisi dell'apprendimento, ovvero il lavoro sistematico con i dati nel contesto delle preferenze di apprendimento individuali, ad esempio dei dipendenti, potrebbe rappresentare un passo promettente avanti nel supporto di più donne nel perseguire una carriera nell'ambito dell'IA/dati. Non esiste una soluzione unica; piuttosto, ascoltando i dipendenti e co-creando nuove soluzioni, soprattutto se una politica non riesce a ottenere risultati, i datori di lavoro possono garantire in modo migliore che i loro sforzi portino al successo delle donne nella loro forza lavoro e attraggano nuovi talenti.

Promuovere la trasparenza sul comportamento positivo e negativo della forza lavoro, nonché sul valore della diversità. Women in AI ha riscontrato che il 73% dei partecipanti al sondaggio ha subito discriminazioni e il 41% ha subito molestie sessuali sul luogo di lavoro. Questi numeri sono inaccettabili e le aziende devono instaurare una politica di tolleranza zero nei confronti della discriminazione, nonché linee guida concrete su tutto, dalle micro-aggressioni alle molestie sessuali. La coerenza tra politica e azione è una necessità: fornire piattaforme per la segnalazione sicura di discriminazioni e molestie è essenziale, ma sono altrettanto importanti azioni a sostegno delle vittime e conseguenze per i responsabili. Una cultura di feedback aperto a tutti i livelli è altrettanto critica, compresa quella della più alta dirigenza.

Tra i partecipanti di Women in AI, il 20% non sapeva se le proprie aziende avessero intrapreso azioni per garantire la diversità. Le aziende devono comunicare in modo trasparente i propri obiettivi e ambizioni riguardo alla diversità sia internamente che esternamente per attirare e trattenere la loro forza lavoro femminile. Tuttavia, la migliore pratica sarebbe per le aziende non solo mirare ad abolire il comportamento negativo della forza lavoro, ma anche favorire ambienti positivi, affrontando i benefici della diversità di genere e un luogo di lavoro inclusivo. McKinsey (2020) ha riscontrato che le idee sbagliate non affrontate sulla giustizia sono un ostacolo significativo per affrontare il divario di genere, con la convinzione prevalente che "tutto dovrebbe essere uguale per tutti".

Questa mancanza di comprensione è evidente anche tra i partecipanti al sondaggio di Women in AI: alcuni hanno scritto come la diversità sia "pregiudizievole", una "minaccia" e "misandrica", così come il fatto che le nuove assunzioni femminili "non hanno la competenza tecnica per lavorare nell'IA". I datori di lavoro devono comunicare chiaramente il "perché" di ciascuna politica sulla diversità e assicurarsi che il messaggio sull'assunzione si concentri su pratiche inclusive, piuttosto che sul semplice raggiungimento di quote.

Le aziende devono adottare misure concrete per affrontare la discriminazione e promuovere l'inclusività nei luoghi di lavoro. Queste misure possono includere sessioni di formazione sulla discriminazione, incontri di ascolto e sessioni di townhall per i dirigenti di alto livello per ascoltare direttamente le esperienze dei dipendenti. La formazione sulla discriminazione è il minimo necessario: promuovere l'inclusività come competenza chiave nei descrittori di lavoro, valutata nelle revisioni delle prestazioni, dimostrata dai dirigenti senior e incoraggiata nei gruppi di risorse per i dipendenti per costruire una comunità. Il 54% dei partecipanti al sondaggio di Women in AI ha riscontrato che i consigli di amministrazione delle loro aziende non rappresentavano la forza lavoro diversificata, e il 28% ha rilevato che le loro aziende non offrivano pari opportunità di promozione a tutti i dipendenti. Un tema centrale tra le risposte è stato come la mancanza di mentori e leader femminili abbia comportato difficoltà nell' "immaginare un percorso chiaro verso la leadership", oltre a sentirsi "dubbie" delle proprie capacità, "fuori luogo" e "in evidenza". Costruendo una struttura di incentivi trasparente basata sulle competenze, il talento femminile in ruoli di leadership aumenterà di conseguenza, aprendo la strada a modelli di ruolo diversificati e opportunità di sponsorizzazione crescenti.

Il framework descritto indica la migliore via avanti per le aziende: sostenere i dipendenti nella loro vita personale e professionale attraverso approcci innovativi, promuovere la trasparenza nei comportamenti negativi e positivi della forza lavoro e impegnarsi in azioni concrete invece di promesse vuote per le iniziative sulla diversità. Vale la pena lo sforzo: non solo le aziende con una maggiore diversità hanno dimostrato di ottenere maggiori risultati finanziari, ma la costruzione di una forza lavoro rappresentativa indica anche una cultura attraente per nuovi talenti e favorisce una cultura di scambio di idee diverse.

4.1.4 Legislazione: Unione Europea e AI Act

Il Parlamento europeo ha recentemente approvato l'AI Act, primo regolamento con l'obiettivo a monitorare l'uso dell'intelligenza artificiale nel continente europeo. La sua formulazione è tuttora in corso ed è previsto che entri in vigore tra due anni.

Tale regolamento mira ad uno sviluppo etico di questa tecnologia nel rispetto scopi e tutela delle minoranze, benessere dell'umanità e ambientale, uguaglianza e parità di genere. I diritti fondamentali delle persone coinvolte nell'AI hanno necessità di tutele e questa iniziativa intende essere una guida concreta per le imprese che utilizzano tali mezzi. L'AI Act si fonda su un metodo "basato sul rischio" per regolamentare l'IA, ossia, si focalizza sulle applicazioni che possono, maggiormente di altre, rischiare di causare danni umani.

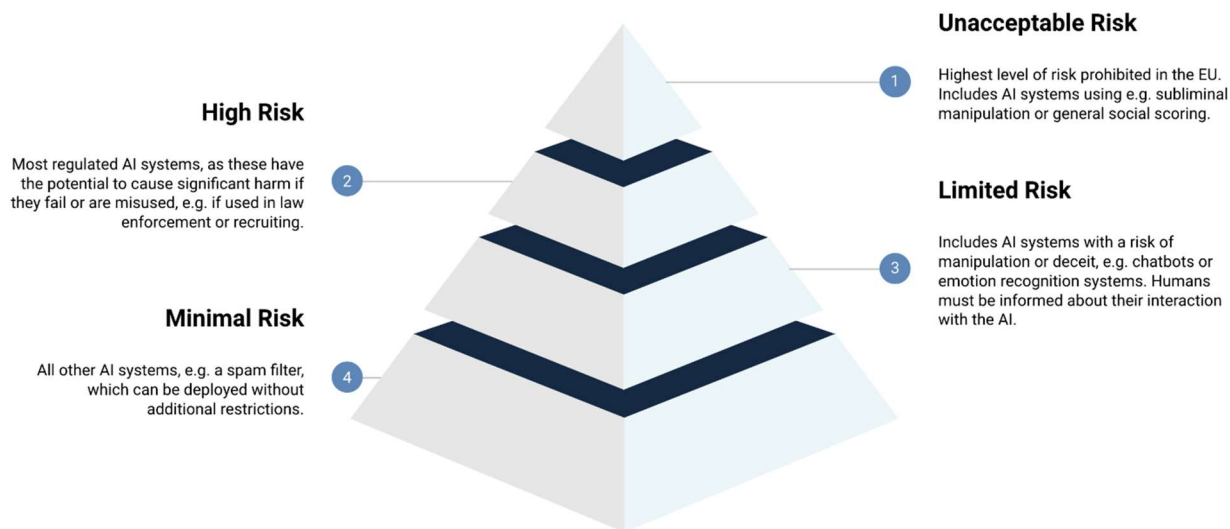


Figura 39 - Classificazione del Rischio nell'AI Act,

Al momento sono stati revisionati e disciplinati i sistemi General Purpose (Gp ai) e i Large Language Models (LLM), ma anche le sempre più note applicazioni di intelligenza artificiale generativa (ChatGpt, Bard e Midjourney) saranno sottoposte ad attenti monitoraggi affinché rispettino le normative relative sicurezza informatica, la trasparenza dei processi di addestramento prima di poter essere utilizzate dalle imprese.

L'AI Act differenzia i modelli di IA ad alto impatto, che richiedono obblighi categorici come la trasparenza dei processi di addestramento degli algoritmi e la mitigazione del rischio e modelli più piccoli, con requisiti meno stringenti.

Tutti i cittadini potranno esibire i reclami del caso, sui sistemi di Ai ed esigere spiegazioni sulle decisioni che si poggiano su sistemi di Ai ad alto rischio. Il regolamento obbliga proprio alla valutazione delle potenziali conseguenze negative sui diritti fondamentali dei sistemi AI ad alto rischio. Sono stati stilati diversi principi stringenti per monitorare l'utilizzo di sistemi di identificazione biometrica in spazi accessibili al pubblico.

La lista dei divieti in breve, si baserà sulle seguenti raccomandazioni:

- Abolizione dell'uso dell'AI per analisi di dati biometrici sensibili, al fine di discriminazioni basate sulla categorizzazione biometrica
- Divieto dello scraping di immagini (esulano in via eccezionale dal divieto, quelle mirate e finalizzate al monitoraggio e contrasto di persone condannate, considerate colpevoli di reati constatati)
- Ne consegue l'assoluto divieto della "polizia predittiva"
- Divieto di utilizzare l'AI al fine di riconoscere emozioni, all'interno di istituzioni educative e luoghi di lavoro, al fine di tutelare la dignità e libertà di espressione
- Divieto di social scoring e tecniche manipolative.
- Divieto dell'utilizzo dell'AI, per abusare di persone vulnerabili (anziane, disabili o con situazioni economiche e sociali fragili e a rischio)

L'RBI (riconoscimento biometrico) sarà utilizzato solamente per ricerche mirate che coinvolgono persone condannate o ree di aver commesso un reato grave, ricerche mirate di vittime (rapimento, traffico, sfruttamento sessuale), prevenzione di una minaccia terroristica.

È prevista dal regolamento una sanzione per coloro i quali violeranno quanto esposto, la multa potrà anche arrivare a da 35 milioni di euro, in proporzione al reato e ai numeri dell'azienda.

4.2 Ricerca e Sviluppo: Progetti per il futuro delle IA

In questa sezione verranno analizzate le possibili soluzioni che possono mitigare i problemi di bias nell'intelligenza artificiale. Naturalmente l'implementazione di queste tecniche permette di agire sui problemi di discriminazione involontaria evidenziati nel capitolo 3.1 in quanto, come verrà analizzato, tali soluzioni non vanno ad intaccare la natura del sistema che sta utilizzando l'intelligenza artificiale.

5.1.1 Fonti Consapevoli: Dataset Equi e Rappresentativi

La soluzione che risulta più "naturale" è quella che punta ad agire sui dati di addestramento, cercando di arricchire i dataset che gli vengono dati in pasto per fare in modo che il sistema sia in grado di interiorizzare le differenze all'interno del campione, portando a soluzioni che non vadano a discapito di minoranze.

Questa operazione viene detta **bilanciamento** del dataset e, data la crescente necessità di dati di addestramento, per questo problema sono stati proposti diverse soluzioni nella letteratura accademica.

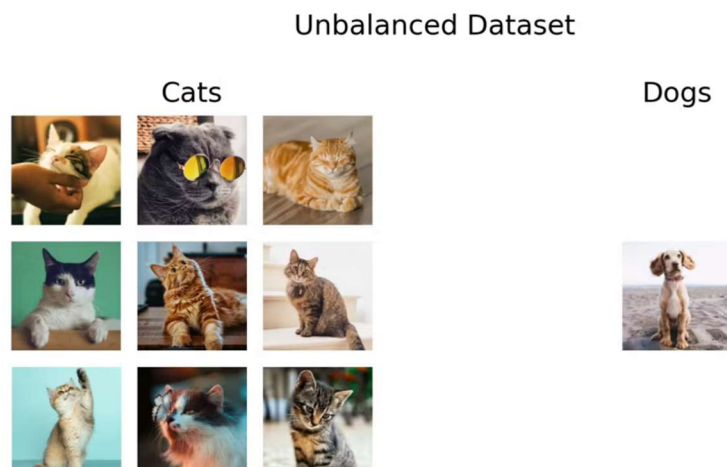


Figura 40 - Esempio di Dataset Sbilanciato, Frederik Hvilshøj (2022)

Arricchimento con Dati Reali

Per quanto banale possa sembrare, l'arricchimento dei dataset con dati presi dal mondo reale non è sempre la soluzione ottimale.

Nonostante l'utilizzo di dati "veri" porti ad avere dati di addestramento che rispecchiano a pieno a realtà, quindi spendibili da un punto di vista statistico nel far prendere decisioni all'intelligenza artificiale (si pensi alle sole applicazioni in ambito medico), non è sempre possibile attuare questa soluzione per due principali motivi: uno economico ed uno pratico.

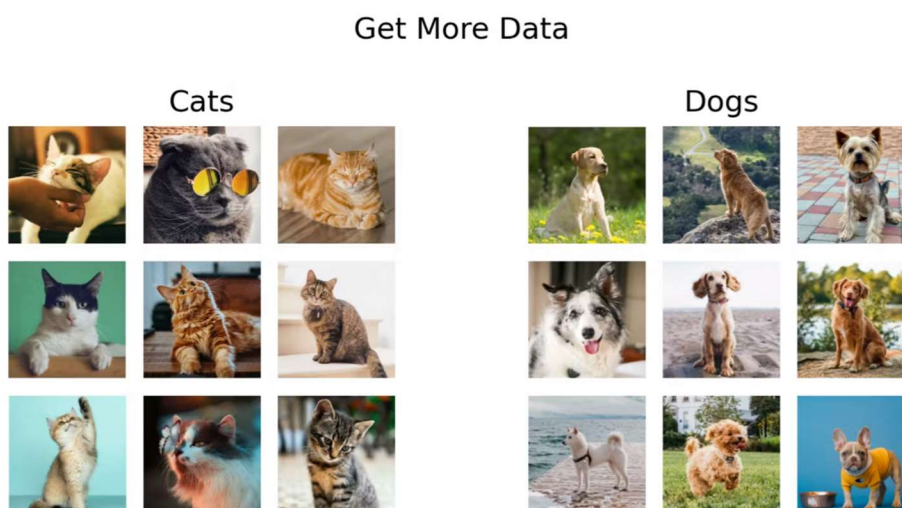


Figura 41 - Arricchimento di un Dataset, Frederik Hvilshøj (2022)

L'aspetto economico è infatti da non sottovalutare quando si parla di quantitativi di dati così massicci: come è stato detto nei capitoli precedenti, spesso gli stessi dataset pubblici vengono riutilizzati più e più volte per diverse applicazioni proprio per andare ad abbattere sensibilmente il costo, che sarebbe altrimenti necessario per crearne di nuovi. Da questo ne consegue che anche solo un arricchimento di un dataset esistente potrebbe essere troppo oneroso per avere un effettivo impatto sulla qualità del sistema.

Per quanto riguarda l'impossibilità pratica ci si riferisce invece ai casi in cui non sia più possibile reperire nuovi dati che vadano a completare quelli esistenti. Ne sono un esempio i dati medici relativi a malattie rare ed i cui pazienti diagnosticati non soddisfino un criterio di diversità rappresentativo, un altro esempio possono essere dataset creati utilizzando dati provenienti da reperti storici.

Bilanciamento “per difetto”: Undersampling

La seconda soluzione che viene proposta è quella dell’undersampling. Il suo funzionamento consiste nell’identificare, come prima cosa, quali categorie sono sovra-rappresentate e quali invece sotto-rappresentate (nella Figura 37 rispettivamente gatti e cani); una volta fatto ciò la tecnica dell’undersampling punta a ridurre il numero di elementi appartenenti alla prima categoria, fino al punto in cui il numero di elementi appartenenti alle categorie sovra e sotto rappresentate non sia uguale o quanto meno rappresentativo della realtà (Mazumder, 2023).



Figura 42 - Esempio di Undersampling, Frederik Hvilshøj (2022)

Questa soluzione è riportata in questo elaborato per completezza, in quanto risolve solo in maniera parziale il problema dei bias nelle intelligenze artificiali: nonostante grazie all’undersampling si possa portare ad avere un sistema bilanciato in termini numerici, nella riduzione degli elementi sovra rappresentati è facile che vengano perse informazioni significative, rendendo non solo il sistema meno efficace per le minoranze del dataset, ma diminuendo la qualità dei risultati anche per la categoria sovra-rappresentata.

Creazione di dati artificiali: Oversampling

Appurati i problemi relativi ad una diminuzione del numero di elementi del dataset, viene spontaneo chiedersi se la soluzione possa essere l’operazione inversa: l’aggiunta di nuovi dati creati ad hoc per arricchire l’addestramento del sistema di IA.

Questa tecnica viene chiamata oversampling e come per la precedente è per prima cosa necessario identificare quali siano i gruppi sovra e sotto rappresentati nel dataset originale. Una

volta fatto ciò si punta nuovamente a bilanciare il numero di elementi nelle due categorie, tuttavia in questo caso verranno aggiunti elementi nelle categorie sotto-rappresentate.

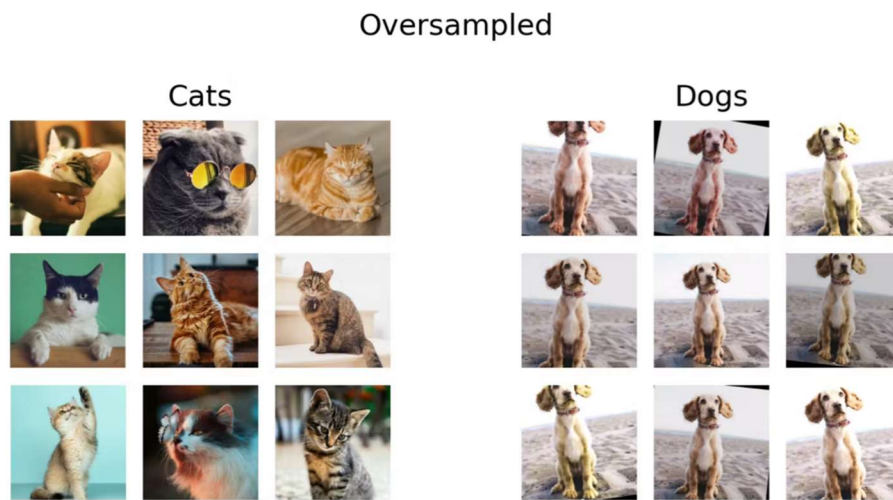


Figura 43 - Esempio di Oversampling, Frederik Hvilshøj (2022)

Anche se l'oversampling può sembrare la soluzione ottimale ai problemi di bias, la generazione di dati artefatti è un'operazione estremamente delicata, che può portare più problemi di quanti non ne possa risolvere.

Creare nuovi dati, partendo da un insieme di partenza definito, può portare problemi nel caso in cui i dati artefatti si discostino troppo o troppo poco da quelli di partenza (Mazumder, 2023).

In caso i nuovi dati siano molto simili a quelli di partenza, la loro aggiunta al dataset consentirebbe, come nel caso dell'undersampling, di bilanciarlo in termini numerici tuttavia avrebbero una valenza statistica molto bassa: ne risulterebbe infatti una serie di dati molto simili tra loro e quindi verosimilmente poco rappresentativi della realtà di riferimento (vedi Fig. 40).

Proprio per quanto detto sopra, spesso viene aggiunto un quantitativo variabile di entropia nella creazione di nuovi dati. La componente randomica permette infatti di avere risultati che più facilmente rappresenteranno le differenze visibili nel mondo, tuttavia un suo eccesso può portare ad avere non verosimili e quindi inutilizzabili a seconda del campo di applicazione del dataset (si pensi nuovamente all'ambito medico).

Un esempio accademico di come si stia tentando di diffondere dataset più rappresentativi viene dalla già citata ricerca sulle criticità nell'utilizzo di ImageNet (Kaiyu, Klint, Li, Jia, & Olga, 2020) nella quale gli autori propongono di arricchire il database per fare in modo che questo produca un equo numero di risultati “filtrando” le immagini su diversi attributi sensibili, come età, sesso e colore della pelle.

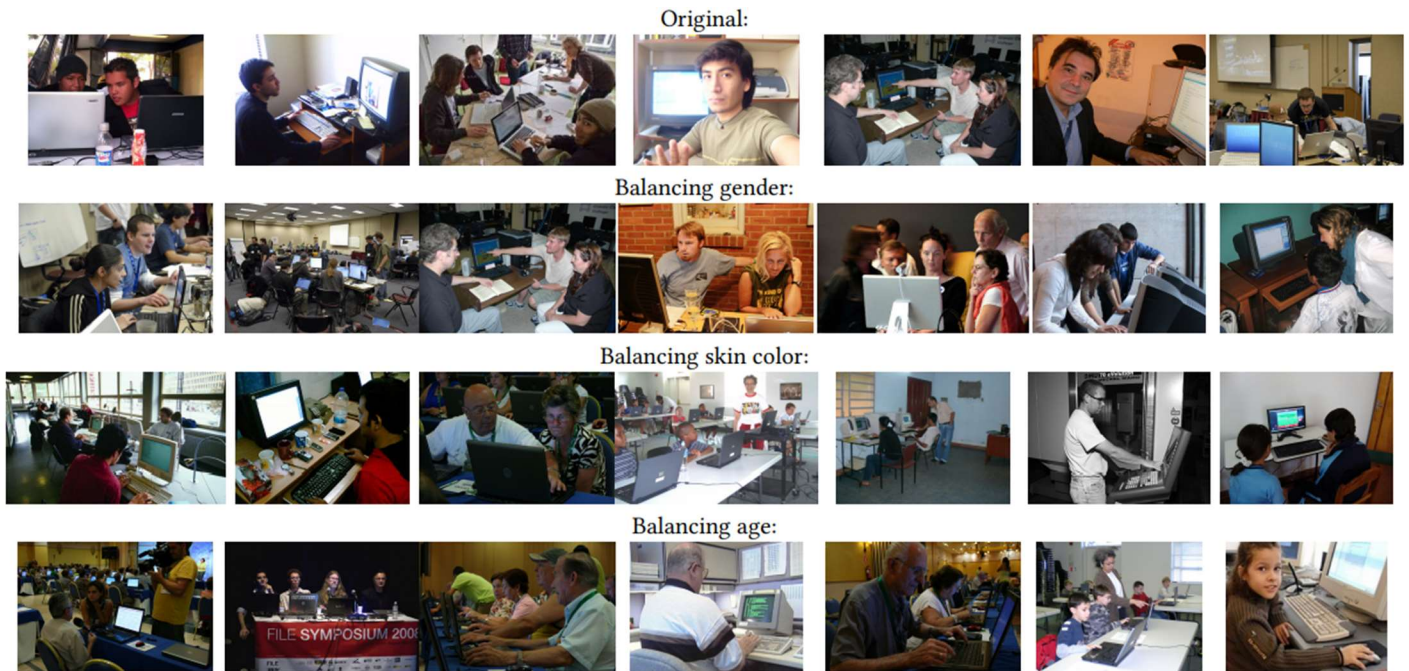


Figura 44 - Bilanciamento dei risultati di ImageNet, Yang et al. (2020)

4.2.2 Indicatori Consapevoli: le Fairness Metrics

Date le difficoltà sorte nel paragrafo precedente nell'agire sui dataset per aumentare l'equità dei sistemi di IA, i ricercatori del campo si sono spinti per cercare di raggiungere questo risultato andando ad agire sul sistema stesso.

Uno dei modi utilizzati per tentare di diminuire i bias nei sistemi basati sull'intelligenza artificiale, è quello di modificare il comportamento del sistema stesso, assegnandogli come compito quello di tentare di massimizzare alcuni indicatori che, per l'appunto, rappresentano quanto l'IA stia prendendo decisioni libere da pregiudizi (Palvel, 2023).

Questi indicatori prendono il nome di fairness metrics e sono categorizzabili in:

- **Metriche di equità di gruppo:** che misurano quanto i gruppi del dataset vengano trattati allo stesso modo; di seguito vengono riportati alcuni esempi di metriche che ricadono in questa categoria.

Parità Demografica

Controlla che il sistema, quando messo davanti ad input categorizzati in gruppi differenti, fornisca output considerati “positivi” in percentuali simili per ogni gruppo. Nell’esempio di un sistema che screma i curriculum per un processo di assunzione questa metrica controlla che vengano scartati in percentuale lo stesso numero di CV tra uomini e donne o tra candidati di diverse nazionalità (Google, 2023).

Pari Opportunità

Con l’Equal Opportunity metric si testa come l’IA si comporta nel fare decisioni che effettivamente rispecchiano i dati reali e quanto invece queste siano influenzati da pregiudizi derivati dalla sua progettazione o dal suo addestramento (Google, 2023). Per un sistema che controlla la cessione di prestiti in una banca questo indice misura se i clienti inizialmente categorizzati come buoni o cattivi pagatori effettivamente rispettino o meno le rate di pagamento (Google, 2023).

Inconsapevolezza del gruppo

Quest’ultima metrica controlla quanto l’attributo considerato sensibile influisca nella decisione presa dal sistema. Per effettuare questa verifica è sufficiente confrontare i risultati ottenuti quando l’IA prende in considerazione quell’attributo e quando invece questo viene escluso dal processo decisionale (TechTarget, 2023).

- **Metriche di equità individuali:** a differenza delle precedenti queste metriche verificano la presenza di bias a livello del singolo elemento del dataset. In particolare questa metrica si basa sull’assunzione che individui con caratteristiche simili debbano essere trattati in maniera simile. Ovviamente questo richiede una definizione di somiglianza tra elementi, che come abbiamo già visto, potrebbe a sua volta essere accidentalmente basato su pregiudizi (Wan, Zha, Liu, & Zou, 2023).

L’utilizzo di metriche di fairness tuttavia non è sufficiente per risolvere il problema della disparità di trattamento nelle intelligenze artificiali. Il problema principale rimane la definizione di equità

ovvero il predicato sul quale si andranno a plasmare gli indicatori sopra citati; la formalizzazione matematica, invece, è costretta a tralasciare alcuni dettagli relativi al contesto che potrebbero evidenziare comportamenti discriminatori anche quando non presenti. In ultimo l'ottimizzazione di alcuni di questi indici può portare ad un peggioramento della precisione e della qualità degli output di un sistema di intelligenza artificiale, richiedendo un attento bilanciamento tra questi due fattori per mantenere rilevante l'utilizzo del sistema stesso.

Tra gli esempi più rilevanti nella letteratura accademica troviamo GetFair (Sandipan, Florian, & Markus, 2022) che cerca di implementare un sistema generalizzato che permetta di testare l'intelligenza artificiale su ogni metrica di fairness; a differenza di altre soluzioni esistenti che consentono invece di farlo su singole metriche. Tuttavia come evidenziato dagli stessi autori il sistema è stato testato solo su classificazioni binarie (raggruppamento di tutti gli elementi in 2 gruppi) ed inoltre non garantisce risultati formali nel bilanciamento tra equità e precisione del sistema.

4.2.3 Addestramento Consapevole: Tecniche di Mitigazione

Un'alternativa alle soluzioni viste nei paragrafi 4.2.1 e 4.2.2 è quella di cercare di andare a cercare di diminuire il bias nel sistema durante la fase di addestramento. Agire sui dataset o inserire metriche di fairness sono soluzioni che devono essere adottate prima che il sistema superi la fase di addestramento: questo, tuttavia, non è sempre possibile in quanto molti modelli esistenti vengono pubblicati o venduti già addestrati proprio per evitare a chi li utilizza la spesa economica e temporale. Nel caso in cui, invece, sia possibile agire sul sistema durante la fase di addestramento è possibile attuare delle tecniche di mitigazione chiamate, per l'appunto in-processing techniques; queste sono divisibili in due categorie:

- **Tecniche di mitigazione esplicite:** le tecniche appartenenti a questa prima categoria mirano, in maniera simile alle fairness metrics, ad aggiungere uno o più indicatori che vadano a bilanciare i bias dell'algoritmo; differiscono da queste ultime, oltre che nella formulazione, nel momento nel quale vengono "iniettate" nel sistema. Gli indicatori di mitigazione esplicita si dividono in due macro-aree, a seconda che questi tentino di

mitigare un sistema che cerca attivamente di discriminare una o più categorie rispetto ai sistemi che lo fanno involontariamente (Wan, Zha, Liu, & Zou, 2023).

Nel primo caso si parla di **disparate treatment** (Machines Gone Wrong, 2023) per riferirsi a sistemi di intelligenza artificiale i cui modelli decisionali utilizzino, tra gli altri dati, anche alcune caratteristiche reputate sensibili (sesso, etnia, religione, etc...). Nel caso in cui il modello non faccia utilizzo di queste caratteristiche ma il suo design contenga implicitamente un comportamento che lo porta a sfavorire una categoria rispetto ad un'altra si parla di **disparate impact** (Machines Gone Wrong, 2023). Questi due termini non sono nativi dell'ambiente di intelligenza artificiale o machine learning ma la loro origine viene dalle leggi del lavoro statunitensi, per identificare l'intenzionalità delle discriminazioni sul luogo di lavoro arrivando a definire, fin dagli anni '60, una lista formale delle caratteristiche sensibili (Khani & Liang, 2022).



Figura 45 - Lista delle Caratteristiche Sensibili, Wolsingham School (2022)

- **Tecniche di mitigazione implicite:** queste tecniche, a differenza delle precedenti non vanno a modificare direttamente il modello; l'idea alla base della mitigazione implicita è quella eliminare le discriminazioni aumentando la rappresentazione delle categorie svantaggiate dal sistema (Wan, Zha, Liu, & Zou, 2023) .

Per aumentare la rappresentazione le tecniche di mitigazione implicite fanno molto affidamento sul feedback del sistema, ovvero quello che l'intelligenza artificiale impara

dal suo storico decisionale: la loro applicazione per questo è principalmente utilizzata nei sistemi di Reinforcement Learning (vedi 1.2.2.4) o di Deep Learning.

Come per la categoria precedente è possibile suddividere le tecniche di mitigazione implicite in due sottoinsiemi: **impact** e **treatment driven** ; la differenziazione è dovuta all'obiettivo delle tecniche, le impact-driven infatti puntano a mitigare le discriminazioni a livello di gruppo, le ultime invece si occupano di rimuovere il bias a livello del singolo elemento.

4.2.4 Controlli Consapevoli: Sistemi per Combattere le Discriminazioni

L'ultima categoria di soluzioni riportate in questo elaborato si riferisce alle situazioni in cui uno sviluppatore è costretto ad utilizzare un'intelligenza artificiale addestrata da qualcun altro: qui, a seconda dell'applicazione del modello, lo sviluppatore dovrà valutare se l'intelligenza artificiale reperita sia adatta ad essere utilizzata.

Per farlo è necessario verificare che il sistema non contenga bias o comportamenti indesiderati, tuttavia una valutazione del genere richiede molto tempo ed energie per essere svolta da un essere umano, nonché una certa periodicità in caso la fase di apprendimento sia ancora in divenire. Non è un caso che aziende come OpenAI abbiano delocalizzato questa operazione in paesi del terzo mondo (Perrigo, 2023), assumendo operatori kenioti per mitigare il linguaggio tossico e i bias del suo modello GPT-3 a meno di due dollari l'ora.

Per riuscire a trovare una soluzione alla necessità di controllare i pregiudizi e la tossicità delle intelligenze artificiali, senza ricorrere a lavoratori sottopagati, una parte del mondo accademico e non si è speso per cercare di sviluppare prodotti che consentissero di rilevare, ed eventualmente correggere le criticità dei modelli di IA.

Un primo esempio di questi sistemi è STAR-ED (Skin Tone Analysis for Representation in Educational materials) (Girmaw, et al., 2023) il cui obiettivo è quello di valutare la rappresentazione delle diverse etnie nel materiale didattico, come libri o articoli di dermatologia, portando ai problemi discussi nel capitolo 3.1.3.

L'utilizzo di un algoritmo come STAR-ED consente agli sviluppatori sia di valutare la qualità etica di utilizzare determinati dataset per addestrare un sistema di IA, che di determinare la validità dei risultati di un'intelligenza artificiale esistente sapendo il materiale usato nella fase di addestramento.

Parallelamente altri progetti, come Lime (Ribeiro, 2019), mirano a semplificare la comprensione del funzionamento di sistemi come le reti neurali, fornendo dettagli su quali parametri il modello ha valutato per giustificare la decisione che ha preso.

Questo permette sia all'utilizzatore del modello di avere conferme sull'output del sistema, verificando come questo vi è giunto, ma consente anche di screditare un risultato nel caso in cui a sceglierlo sia stato dato troppo, o troppo poco, peso ad un parametro di addestramento.

L'ultimo esempio riportato in questo capitolo è quello degli Adversarial Networks (Google, 2022): con questo termine si definiscono sistemi strutturati composti da:

- Un dataset di dati autentici
- Un sistema **generatore** che, addestrato sul dataset è capace di creare nuovi dati
- Un sistema **discriminatore** a cui verranno sottoposti dati autentici e artefatti, il suo scopo è quello di determinare quali di questi appartengono al dataset originali e quali siano stati prodotti dal generatore

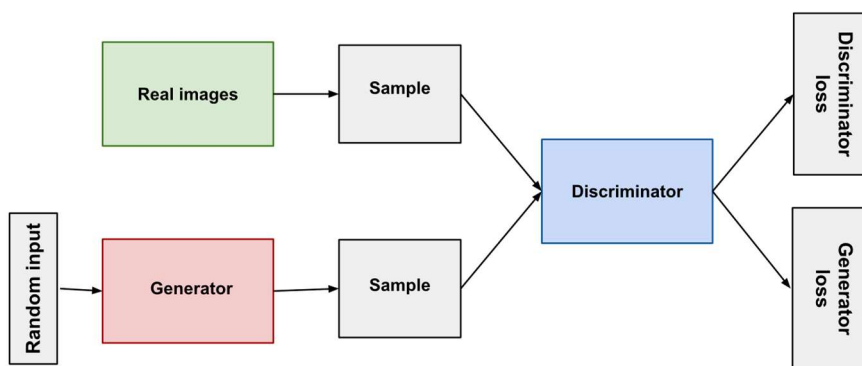


Figura 46 - Schema di un Generative Adversarial Network, Google (2022)

Avendo controllo sulla creazione di elementi da parte del generatore è possibile verificare quali caratteristiche dell'artefatto "ingannano" il discriminatore, ovvero quanto un particolare attributo influisca sulla decisione finale, cioè quanto lo stereotipizza.

5. Conclusioni

La presente tesi si propone di analizzare le possibili conseguenze negative, derivanti da un utilizzo superficiale dell'intelligenza artificiale.

Dopo un'analisi approfondita di come l'intelligenza artificiale sia in grado di perpetuare stereotipi di genere presenti nella nostra società, ho discusso di possibili iniziative volte a contrastare la discriminazione di genere che frequentemente viene messa in atto dall'AI.

Le soluzioni informatiche riguardano le tecniche di mitigazione proposte dai ricercatori e i loro limiti mentre quelle umanistiche si concentrano su un nuovo modo etico di concepire l'intelligenza artificiale come risorsa e al contempo incoraggiare una maggior alfabetizzazione tecnologica che possa rendere gli individui consapevoli e critici verso queste nuove tecnologie.

Bibliografia

Allport, G. (1954). *La Natura del Pregiudizio*.

American Psychological Association. (2019, 09). *Gender*. Tratto da apastyle.apa.org:
<https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/gender>

Amnesty International. (2021). *"LIKE WE WERE ENEMIES IN A WAR", China's Mass Internment, Torture and Persecution of Muslims in Xinjiang*. Tratto da <https://xinjiang.amnesty.org/>:
<https://xinjiang.amnesty.org/>

Anifowose, F. (2020, February). *Hybrid Machine Learning Explained in Nontechnical Terms*. Tratto da *Journal of Petroleum Technology*: <https://jpt.spe.org/hybrid-machine-learning-explained-nontechnical-terms>

Austin, J. L. (1962). *How to Do Things with Word*. Oxford: Oxford University Press.

Banchevsky, S., & Park, B. (2018). Negative Gender Ideologies and Gender-Science Stereotypes Are More Pervasive in Male-Dominated Academic Disciplines. *Social Sciences* 7.

Barker, R. (2022). *How many CCTV cameras are in London?* Tratto da Clarion:
[https://clarionuk.com/resources/how-many-cctv-cameras-are-in-london/#:~:text=Research%20by%20Clarion%20Security%20Systems,\(2022\)](https://clarionuk.com/resources/how-many-cctv-cameras-are-in-london/#:~:text=Research%20by%20Clarion%20Security%20Systems,(2022))

BBC. (2015, 07 01). *Google apologises for Photos app's racist blunder*. Tratto da BBC.com:
<https://www.bbc.com/news/technology-33347866>

Bhatt, S. (2018, March). *Reinforcement Learning 101*. Tratto da Towards Data Science:
<https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292>

Biemmi, I. (2010). *Educazione sessista. Stereotipi di genere nei libri delle elementari*.

Butler, J. (1988). Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory. *Theatre Journal*, 519-53.

Butler, J. (2004). *Undoing Gender*. Routledge.

Cantor, M. (2023). *Nearly 50 news websites are 'AI-generated', a study says. Would I be able to tell?* Tratto da The Guardian: <https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study>

Carrer, L., Coluccini, R., & Di Salvo, P. (2020). *Perché Como è diventata una delle prime città in Italia a usare il riconoscimento facciale*. Tratto da Wired:
<https://www.wired.it/internet/regole/2020/06/09/riconoscimento-facciale-cómo/>

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*.

- Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*.
- CNBC. (2022). *UN cites possible crimes against humanity in China's Xinjiang*. Tratto da CNBC: <https://www.cnbc.com/2022/09/01/un-cites-possible-crimes-against-humanity-in-chinas-xinjiang.html>
- Descartes, R. (1637). *Discorso sul metodo*. Leida.
- Deutsch, F. (2007). Undoing Gender. *Gender & Society*, 106-127.
- Di Capua, M. (2020). LINGUAGGIO R E MACHINE LEARNING: APPLICAZIONE ALLE TECNICHE DI APPRENDIMENTO.
- Earnest, L. (2012, December). *Stanford Cart*. Tratto da Stanford: <https://web.stanford.edu/~learnest/sail/oldcart.html>
- European Parliament. (2023). *EU AI Act: first regulation on artificial intelligence*. Tratto da European Parliament: <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
- Falchetti, D. (2020, 07 03). Tratto da Draft: <https://www.draft.it/cms/Contenuti/firmin-didot-uno-dei-padri-della-tipografia-moderna/>
- Fraser, N. (2013). *Fortunes of Feminism. From State-Managed Capitalism to NeoLiberal Crisis*. New York.
- Garlapati, H. V. (2023, January). *How to get datasets for Machine Learning?* Tratto da Knowledgehut: <https://www.knowledgehut.com/blog/data-science/datasets-for-machine-learning>
- Gary, G., & Fraser, N. (2015). A feminism where 'lean in' means leaning on others. *New York Times*.
- Girmaw, A. T., Celia, C., Kush, R. V., Peter, S., Chinyere, A., Skyler, S., . . . Rox. (2023). Skin Tone Analysis for Representation in Educational Materials (STAR-ED) using machine learning. *npj Digital Medicine*.
- Google. (2022). *Overview of GAN Structure*. Tratto da Google: https://developers.google.com/machine-learning/gan/gan_structure
- Google. (2023). *Machine Learning Glossary: Fairness*. Tratto da google.com: <https://developers.google.com/machine-learning/glossary/fairness>
- Green, J., & Weintraub, A. (2023). *Experts warn of rise in scammers using AI to mimic voices of loved ones in distress*. Tratto da abc news: <https://abcnews.go.com/Technology/experts-warn-rise-scammers-ai-mimic-voices-loved/story?id=100769857>
- Grinevičius, J., & Akavickaitė, A. (2021, 3 22). *People Tested How Google Translates From Gender Neutral Languages And Shared The "Sexist" Results*. Tratto da boredpanda: <https://www.boredpanda.com/google-translate-sexist/>

- Gross, N. (2023). What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI. *Social Sciences* 12.
- Gutting, G., & Fraser, N. (2015). *A Feminism Where 'Lean In' Means Leaning On Others*. Tratto da New York Times: <https://archive.nytimes.com/opinionator.blogs.nytimes.com/2015/10/15/a-feminism-where-leaning-in-means-leaning-on-others/>
- Hamilton, D. L., & Traylor, T. K. (1986). "Stereotypes and Stereotyping: An Overview of the Cognitive Approach." In *Prejudice, Discrimination, and Racism*. Academic Press.
- Harris, C. R., & Jenkins, M. (2006). Gender Differences in Risk Assessment: Why do Women Take Fewer Risks than Men? *Judgment and Decision Making*, 48 - 63.
- Harris, C. R., Jenkins, M., & Glaser, D. (2006). Gender differences in risk assessment: Why do women take fewer risks than men? *Judgment and Decision Making*, 48–63.
- Himana, A. (2023). 'Grandma Exploit': ChatGPT commanded to pretend to be a dead grandmother. Tratto da Medium: https://medium.com/@himana_arumalla/grandma-exploit-chatgpt-commanded-to-pretend-to-be-a-dead-grandmother-13ddb984715a
- Hodges, A. (1983). *Alan Turing: The Enigma*.
- Hodges, A. (2008). Alan Turing and the Turing Test. In G. R. Robert Epstein, *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* (p. 13-22).
- Hoffman, S., & Podgurski, A. (2021). Artificial Intelligence and Discrimination in Health Care. *Yale Journal of Health Policy, Law, and Ethics*.
- Hoyt, C. L., & Burnette, J. L. (2013). Gender bias in leader evaluations: Merging implicit theories and role congruity perspectives. *Personality and Social Psychology Bulletin*, 1306–1319.
- Hoyt, C., & Burnette, J. (2013). *Growth Mindset Messaging in Stigma-Relevant Contexts: Harnessing Benefits Without Costs*.
- Il Messaggero. (2023). *Professoressa di Latina spogliata con l'app di "BikiniOff", le immagini finiscono sui siti porno*. Tratto da Il Messaggero: https://www.ilmessaggero.it/latina/prof_spgliata_app_bikinioff_latina_cosa_e_successo-7630192.html
- Il Sole 24 Ore. (2023). *Intelligenza artificiale: da Keycrime a Giove, nuovo investigatore per la Polizia*. Tratto da Il Sole 24 Ore: <https://www.ilsole24ore.com/art/intelligenza-artificiale-keycrime-giove-nuovo-investigatore-la-polizia-AE9i1EbD>
- Irene, B. (2010). *Educazione sessista. Stereotipi di genere nei libri delle elementari*. Rosenberg & Sellier.
- Israel, G. (2015). *Meccanicismo. Trionfi e miserie della visione meccanica del mondo*. Bologna: Zanichelli.
- Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49–81.

- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49-81.
- Kaiyu, Y., Klint, Q., Li, F.-F., Jia, D., & Olga, R. (2020). Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–558.
- Kelan, E. (2023, 07 25). *AI can reinforce discrimination — but used correctly it could make hiring more inclusive*. Tratto da theconversation.com: <https://theconversation.com/ai-can-reinforce-discrimination-but-used-correctly-it-could-make-hiring-more-inclusive-207966>
- Khani, F., & Liang, P. (2022, 12 23). *From Discrimination in Machine Learning to Discrimination in Law, Part 2: Disparate Impact*. Tratto da Stanford AI Lab Blog: https://ai.stanford.edu/blog/discrimination_in_ML_and_law_p2/#definition
- Kramer, M., & Zanger, J. (2023). *NYPD, Mayor Adams unveil robot police dogs in Times Square*. Tratto da CBS: <https://www.cbsnews.com/newyork/news/nypd-eric-adams-robot-dogs/>
- Kuang Ming, K., Paul, T., & Chao-Sheng, C. (2023). The accuracy of artificial intelligence used for non-melanoma skin cancer diagnoses: a meta-analysis. *BMC Medical Informatics and Decision Making*.
- Laplace, P.-S. (1796). *Exposition du système du monde*. Paris.
- Lavecchia, V. (2017, April). *DENDRAL – Medical Expert Systems (Sistemi esperti in sanità)*. Tratto da Informatica e Ingegneria Online: <https://vitolavecchia.altervista.org/dendral-medical-expert-systems-sistemi-esperti-sanita/>
- Lee, T., & Koltai, K. (2023). *The Folly of DALL-E: How 4chan is Abusing Bing’s New Image Model*. Tratto da Bellingcat: <https://www.bellingcat.com/news/2023/10/06/the-folly-of-dall-e-how-4chan-is-abusing-bings-new-image-model/>
- Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology*, 165–188.
- Linville, P., W. Fischer, G., & W. Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology*, 165–188.
- Lloyd, L., Lori, G.-S., & Margaret, L. (2001). Social change and the persistence of sex typing: 1974–1997. *Social Forces*.
- Lueptow, L. B., Garovich-Szabo, L., & Lueptow, M. B. (2001). Social change and the persistence of sex typing. *Social Forces*, 1-36.
- M, H. (2021). *The Motherhood Complex: The Story of Our Changing Selves*. Piatkus.
- Machines Gone Wrong. (2023). *Understanding Fairness*. Tratto da Machines Gone Wrong: <https://machinesgonewrong.com/fairness/>

- Mahesh, B. (2019). Machine Learning Algorithms -A Review. *international Journal of Science and Research*.
- Martini, M. (2009). *Stereotipo di genere*. Tratto da Università degli Studi di Torino: https://www.cirsde.unito.it/sites/c555/files/allegatiparagrafo/06-05-2016/stereotipo_di_genere.pdf
- Mazumder, S. (2023, 09 27). *5 Techniques to Handle Imbalanced Data For a Classification Problem*. Tratto da Analytics Vidhya: <https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/>
- McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. *AI Magazine* 27.
- McCulloch, W., & Pitts, W. (1943). *A logical calculus of the ideas immanent in nervous activity*.
- McDermott, J. (1982). R1: A rule-based configurer of computer systems. *Artificial Intelligence*, 39-88.
- Nancy, F. (2013). *Fortunes of Feminism: From State-Managed Capitalism to Neoliberal Crisis*. Verso Books.
- NIHCM. (2021, 09 30). *Racial Bias in Health Care Artificial Intelligence*. Tratto da nihcm.com: <https://nihcm.org/publications/artificial-intelligences-racial-bias-in-health-care>
- Nilsson, N. (2001). *Intelligenza Artificiale*.
- OECD. (2017). *The Pursuit of Gender Equality: An Uphill Battle*. Paris: OECD Publishing.
- Oracle. (s.d.). *Primitive Data Types*. Tratto da Oracle: <https://docs.oracle.com/javase/tutorial/java/nutsandbolts/datatypes.html>
- Palvel, S. (2023, 09 14). *Introduction to Fairness-aware ML*. Tratto da Medium: <https://subashpalvel.medium.com/introduction-to-fairness-aware-ml-327df1b61538>
- Pellegrino, A. (2020). *Natura e caratteristiche di pregiudizi e stereotipi*. Tratto da univpm.it: <https://letteredallafacolta.univpm.it/natura-e-caratteristiche-di-prejudizi-e-stereotipi/>
- Perrigo, B. (2023, January 18). *Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic*. Tratto da Time: <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Plant, E. A., Hyde, J. S., Keltner, D., & Devine, P. G. (2000). The Gender Stereotyping of Emotions. *Psychology of Women Quarterly*.
- Poe, E. A. (1836). *Il giocatore di scacchi di Maelzel*.
- Prates, M., Avelar, P., & Lamb, L. (2019). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*.
- Ramzy, A., & Buckley, C. (2019). *'Absolutely No Mercy': Leaked Files Expose How China Organized Mass Detentions of Muslims*. Tratto da New York Times: <https://www.nytimes.com/interactive/2019/11/16/world/asia/china-xinjiang-documents.html>

- Ribeiro, M. T. (2019). *Lime*. Tratto da Github: <https://github.com/marcotcr/lime?tab=readme-ov-file>
- Rollet, C. (2019, 11 11). *Hikvision Markets Uyghur Ethnicity Analytics, Now Covers Up*. Tratto da IPVM: <https://ipvm.com/reports/hikvision-uyghur>
- Rose, A. (2010, 01 22). *Are Face-Detection Cameras Racist?* Tratto da time.com: <https://content.time.com/time/business/article/0,8599,1954643,00.html>
- Rosso, E. (2023). *Il caso delle studentesse spogliate dai compagni con un'app e ricattate: "Guarda, qui sei nuda"*. Tratto da Fanpage: <https://www.fanpage.it/innovazione/tecnologia/il-caso-delle-studentesse-spogliate-dai-compagni-con-unapp-e-ricattate-guarda-qui-sei-nuda/>
- Sabadin, V. (2017, September). *Deep Blue vs. Kasparov scacco matto all'intelligenza umana*. Tratto da La Stampa: <https://www.lastampa.it/cultura/2017/09/13/news/deep-blue-vs-kasparov-scacco-matto-all-intelligenza-umana-1.34414089/>
- Sandipan, S., Florian, L., & Markus, S. (2022, 6). GetFair: Generalized Fairness Tuning of Classification Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, p. 289–299.
- Skinner, B. F. (1953). *Science and human behavior*.
- Smith, G., & Rustagi, I. (2021). *When Good Algorithms Go Sexist: Why and How to Advance AI Gender*. Stanford Social Innovation Review.
- Sue Curry, J. (2013). Semantic Tyranny: How Edward L. Bernays Stole Walter Lippmann's Mojo And Got Away With It And Why It Still Matters. *International Journal Of Communication*.
- TechTarget. (2023, 07 28). *6 ways to reduce different types of bias in machine learning*. Tratto da TechTarget: <https://www.techtarget.com/searchenterpriseai/feature/6-ways-to-reduce-different-types-of-bias-in-machine-learning>
- Tibco. (2019). *Cos'è l'apprendimento non supervisionato?* Tratto da Tibco: <https://www.tibco.com/it/reference-center/what-is-unsupervised-learning>
- Tibco. (2019). *Cos'è l'apprendimento supervisionato?* Tratto da Tibco: <https://www.tibco.com/it/reference-center/what-is-supervised-learning>
- Treccani. (2023). *stereòtipo*. Tratto da Treccani.it: <https://www.treccani.it/vocabolario/stereotipo/>
- United Nations. (2021). *Progress on the Sustainable Development Goals: The gender snapshot 2021*.
- Vincent, J. (2021). *They're putting guns on robot dogs now, It was only a matter of time*. Tratto da The Verge: <https://www.theverge.com/2021/10/14/22726111/robot-dogs-with-guns-sword-international-ghost-robotics>
- Wan, M., Zha, D., Liu, N., & Zou, N. (2023). In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data*, 1–27.
- World Health Organization. (2021). *Gender and health*. Tratto da who.int: <https://www.who.int/health-topics/gender>

Wu, Y. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

Zhou, Z.-H. (2021). *Machine Learning*.