Discriminazione e bias nei sistemi di IA

Comitato Pari Opportunità
Consiglio dell'Ordine degli Avvocati di Torino
Avv. Manuela Monti
21. 06. 2023

Normativa antidiscriminatoria

Il diritto antidiscriminatorio ha una composizione variegata, ed è il risultato dell'incontro di norme di diritto nazionale, di norme di recepimento di direttive comunitarie, di norme primarie UE, tra cui

Carta dei diritti fondamentali dell'Unione europea

- **Articolo 21 - Non discriminazione**
- 1. È vietata qualsiasi forma di discriminazione fondata, in particolare, sul sesso, la razza, il colore della pelle o l'origine etnica o sociale, le caratteristiche genetiche, la lingua, la religione o le convinzioni personali, le opinioni politiche o di qualsiasi altra natura, l'appartenenza ad una minoranza nazionale, il patrimonio, la nascita, la disabilità, l'età o l'orientamento sessuale.
- 2. Nell'ambito d'applicazione dei trattati e fatte salve disposizioni specifiche in essi contenute, è vietata qualsiasi discriminazione in base alla nazionalità.
- **Articolo 23 co. 1** secondo cui «la parità tra uomini e donne deve essere assicurata in tutti i campi, compreso in materia di occupazione, di lavoro e di retribuzione»

Articolo 14 CEDU il godimento dei diritti e delle libertà riconosciuti nella Convenzione «deve essere assicurato senza nessuna discriminazione, in particolare quelle fondate sul sesso, la razza, il colore, la lingua, la religione, le opinioni politiche o quelle di altro genere, l'origine nazionale o sociale, l'appartenenza a una minoranza nazionale, la ricchezza, la nascita od ogni altra condizione».

Normativa antidiscrimnatoria connessa al genere in diverse direttiva (quali la dir. 2006/54 rimangono oggi la dir. 79/7 sulla parità di trattamento tra gli uomini e le donne in materia di sicurezza sociale, la dir. 2010/41)

Il **d.lgs. 11 aprile 2006, n.198 (Codice delle Pari Opportunità), novellato dal d.lgs. 5/2010** Nel nostro ordinamento, la normativa antidiscriminatoria connessa al genere si trova, in gran parte, all'interno del d. lgs. 198/2006, c.d. codice delle pari opportunità tra uomo e donna

La **l. 162/2021** ha istituito, per i datori di lavoro che attuano politiche «per ridurre il divario di genere in relazione alle opportunità di crescita in azienda, alla parità salariale a parità di mansioni, alle politiche di gestione delle differenze di genere e alla tutela della maternità», una certificazione della parità di genere (art. 46 bis d. lgs. 198/2006)

GDPR _ Regolamento UE 2016/679 **Articolo 9 Trattamento di categorie particolari di dati personali**

Discriminazione di genere

Il Rapporto sullo Sviluppo Umano dell'UNDP del 1995 affermava:

"Per troppo tempo si è dato per scontato che lo sviluppo fosse un processo ... neutro nel suo impatto di genere. L'esperienza ci dice qualcosa di diverso»

La discriminazione di genere è la disparità di trattamento delle persone in base al loro genere. Include la concessione di privilegi a un certo genere o l'emarginazione di qualcuno a causa della sua identità di genere.

Disparità di retribuzione, molestie sessuali, accesso limitato o impossibile a diritti come l'istruzione e l'assistenza sanitaria sono forme di discriminazione di genere.

Discriminazione di genere: un concetto in evoluzione

Il **genere** è una tipizzazione sociale, culturale e psicologica . E' un costrutto sociale, a differenza del sesso che definisce l'aspetto biologico-anatomico di una persona.

Certi concetti evolvono nel tempo, e cosi l'identità di genere, con il superamento di una concezione non più pienamente binaria.

L'**identità di genere** riguarda il modo in cui ci si identifica con il proprio sesso e come si vuole essere percepiti da altri, può corrispondere al **sesso biologico** o discostarsene ed è un concetto diverso dall'**orientamento sessuale**.

Partita come differenze tra maschi e femmine, si è evoluta. Esiste una molteplicità di identità di genere che va oltre la dualità, come ad esempio: transgender, agender, non binario (persone LGBTIQ+)

- Cosa significa **discriminazione in base all'identità di genere**?

Si parla di discriminazione sulla base dell'identità di genere, quando persone di qualsiasi identità di genere, vengono svantaggiate o svalutate a causa di esso.

Discriminazione intersezionale

- Per discriminazione intersezionale si intende una situazione nella quale intervengono diversi motivi di discriminazione che interagiscono tra loro, per esempio il genere con altri motivi di discriminazione quali razza, età, orientamento sessuale, disabilità, tra gli altri, in modo tale da essere indissociabili e da produrre tipologie specifiche di discriminazione. I motivi di discriminazione si intrecciano, creando una tipologia unica di discriminazione, senza che un motivo prevalga sull'altro

- La discriminazione multipla ha luogo quando ciascun tipo di discriminazione può essere dimostrata e trattata in maniera indipendente.

Intelligenza artificiale, algoritmi e machine learning

- Il termine **Intelligenza artificiale** (IA o AI) si riferisce a sistemi che utilizzano algoritmi di apprendimento automatico in grado di analizzare grandi volumi di dati di addestramento per identificare correlazioni, schemi e altri metadati che possono essere utilizzati per sviluppare un modello in grado di fare previsioni o raccomandazioni basate su dati futuri.
- Gli **algoritmi** sono istruzioni matematiche, procedure e insieme di istruzioni passo-passo utilizzate per eseguire un'attività.
- Alcuni algoritmi informatici sono progettati per consentire ai computer di imparare da soli (cioè, facilitare l'apprendimento automatico).
- Gli usi dell'apprendimento automatico – o **machine learning** - includono l'estrazione dei dati e il riconoscimento di modelli. Nell'apprendimento automatico, gli algoritmi si basano su più set di dati, o dati di addestramento per apprendere un modello, fare predizioni.
- **Deep learning** è un'evoluzione del machine learning con tecniche basate su reti neurali artificiali organizzate in diversi strati, che elaborano le informazioni in modo più evoluto, senza necessità di dati strutturati, elimina una parte dell'intervento umano
- Gli algoritmi sfruttano enormi volumi di macro e micro-dati per influenzare le persone e le decisioni che riguardano le persone in una serie di ambiti molto diversi tra loro.

Bias e pregiudizio di genere

- **Bias:** costrutti derivanti da percezioni errate, automatismi mentali che generano credenze, distorsioni cognitive che conducono a veloci valutazioni e decisioni fondati
- Il **pregiudizio o bias di genere** è un comportamento che mostra un favoritismo verso un genere rispetto a un altro. Si tratta di una forma di pregiudizio più o meno inconscio o implicito, che si verifica quando un individuo attribuisce determinati atteggiamenti e stereotipi a un certo genere di persona o gruppo di persone.
- **AI bias** si riferisce a sistemi di intelligenza artificiale che producono sistematicamente e ingiustificatamente risultati meno favorevoli, ingiusti o dannosi per i membri di specifici gruppi.
- I pregiudizi dell'IA possono manifestarsi in sistemi che prendono decisioni, producono risultati o forniscono prestazioni meno accurate o trattano le persone in modo meno favorevole sulla base di una caratteristica tra cui, ad esempio, razza, identità di genere, orientamento sessuale, età, religione o disabilità.
- I sistemi «biased» sono quelli che **discriminano** sistematicamente e ingiustamente gli individui o alcuni gruppi sociali: così i pregiudizi sociali vengono perpetuati con gravi conseguenze.
- I bias (umani, statistici, computazionali) possono essere presenti in tutte le fasi di vita dell'algoritmo, introdotti a partire dalla selezione dei dati e delle sorgenti, nell'elaborazione o nel raggruppamento dei dati, mutuati da altri sistemi, possono non essere presenti nel dato di partenza ma generati nel processo di apprendimento tramite connessioni complesse e comunque gravare l'output finale.
- Un sistema può amplificare simultaneamente più stereotipi (es. genere e razza)

Dati rappresentativi

Uso di dati rappresentativi e di alta qualità per addestrare il sistema: fondamentale per garantire un'intelligenza artificiale giusta, equa e democratica.

I sistemi di intelligenza artificiale dovrebbero rispecchiare la nostra società in tutta la sua diversità, altrimenti di default si perpetuano pregiudizi e discriminazioni .

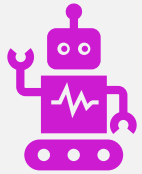
Representation gap.

I sistemi ad apprendimento automatico imparano estraendo informazioni da grandi quantità di dati, se i dati sono espressione di pregiudizi, non sono correttamente rappresentativi della realtà e della varietà umana, ma solo di una sua parte, oppure se riflettono modelli storici discriminatori, **si perpetuano e amplificano gli stereotipi e i pregiudizi**, inclusi quelli di genere, e gli **effetti discriminatori**.

Il comportamento della macchina può risultare iniquo, come se la macchina, avendo conosciuto solo una parte di un certo contesto, ne avesse elaborato una **conoscenza ristretta e limitata**.

A volte i bias sono chiari e percepibili, altre volte più sfumati e può esserne difficile misurarne l'espressione.

Rischio di un’unica prospettiva sul mondo, invece di una visione che rappresenti diversi tipi di culture e identità



Intelligenza artificiale:

- ha il potenziale di generare nuove conoscenze e di rendere i processi più efficienti, presenta opportunità

- può promuovere la diversità, l ’equità, l'inclusione e l'accessibilità

- presenta rischi: difetti di progettazione, di sviluppo e/o nell'implementazione dei sistemi di IA, possono generare inaffidabilità, mancanza di interpretabilità, patterns discriminatori, decisioni inique, difficoltà nell’individuare e dimostrare i profili di responsabilità

In che modo big data e algoritmi si intersecano con la discriminazione?

- Molti sistemi automatizzati si basano su grandi quantità di dati per trovare schemi o correlazioni, per poi applicare questi schemi a nuovi dati per eseguire attività o fare raccomandazioni e previsioni.
- Esiti possono avere effetti discriminatori deflagranti
- La reiterazione della discriminazione amplifica effetto discriminatorio
- La potenziale discriminazione nei sistemi automatizzati può derivare da diverse fonti, tra cui problemi con:
- - Dati e set di dati
- - Opacità del modello e dell'accesso («*black box*»),
- - Progettazione e utilizzo

Artificial Intelligence Act e discriminazione

- ***Artificial Intelligence Act*** approvato **in data 14 giugno 2023** dal Parlamento Europeo - in attesa di avviare i colloqui con i Governi per il testo definitivo - affronta il problema del possibile utilizzo dell'intelligenza artificiale con **finalità o modalità discriminatorie**.
- L'obiettivo di minimizzare «il rischio di discriminazione algoritmica» è un fine specifico del Regolamento, che è formulato tenendo conto dei principi affermati dalla Carta dei diritti fondamentali dell'Unione europea (art. 21) e tutte le fonti del diritto antidiscriminatorio prodotto dall'Ue

- Emendamento 213 Proposta di regolamento introduce nuovo art. 4 bis
- Principi generali applicabili a tutti i sistemi di IA
- 1. *Tutti gli operatori che rientrano nel presente regolamento si adoperano al massimo per sviluppare e utilizzare sistemi di IA o modelli di base conformemente ai seguenti principi generali che istituiscono un quadro di alto livello che promuova un approccio europeo antropocentrico coerente a un'intelligenza artificiale etica e affidabile, che sia pienamente in linea con la Carta e con i valori su cui si fonda l'Unione*: *…(omissis)*
- e) *"diversità, non discriminazione ed equità": i sistemi di IA sono sviluppati e utilizzati in modo da includere soggetti diversi e promuovere la parità di accesso, l'uguaglianza di genere e la diversità culturale, evitando nel contempo effetti discriminatori e pregiudizi ingiusti vietati dal diritto dell'Unione o nazionale*

- Emendamento 27 Proposta di regolamento Considerando 9 bis (nuovo)
- *È importante osservare che i sistemi di IA dovrebbero fare il possibile per rispettare i principi generali che istituiscono un quadro di alto livello che promuova un approccio coerente e antropocentrico a un'IA etica e affidabile, in linea con la Carta dei diritti fondamentali dell'Unione europea e i valori su cui si fonda l'Unione, tra cui la protezione dei diritti fondamentali, l'intervento e la sorveglianza umani, la robustezza tecnica e la sicurezza, la riservatezza e la governance dei dati, la trasparenza, la non discriminazione e l'equità nonché il benessere sociale e ambientale*.

- Emendamento 228 Proposta di regolamento Articolo 5 – par. 1 bis (nuovo)
- L' art. 5 prevede una serie di divieti (ad es. uso di sistemi di identificazione biometrica remota "in tempo reale" in spazi accessibili al pubblico)
- Viene introdotto il par. 1 bis *Il presente articolo lascia impregiudicati i divieti che si applicano qualora una pratica di intelligenza artificiale violi un altro atto legislativo dell'Unione, tra cui l'acquis dell'UE in materia di protezione dei dati, non discriminazione, tutela dei consumatori o concorrenza»*

Artificial Intelligence Act e discriminazione

- Emendamento 40 Proposta di regolamento Considerando 17
  - Il Considerando parla degli strumenti di «punteggio sociale» assegnato ai cittadini , con riferimento ai quali si rileva che *possono portare a risultati discriminatori e all'esclusione di determinati gruppi. Ledono inoltre il diritto alla dignità e alla non discriminazione e i valori di uguaglianza e giustizia….. Il punteggio sociale ottenuto da tali sistemi di IA può determinare un trattamento pregiudizievole o sfavorevole di persone fisiche o di interi gruppi in contesti sociali che non sono collegati ai contesti in cui i dati sono stati originariamente generati o raccolti, o a un trattamento pregiudizievole che risulta ingiustificato o sproporzionato rispetto alla gravità del loro comportamento sociale*.». Per questo se ne ritiene opportuno il divieto.
  - Nel prosieguo, l'AI Act considera invece delle possibili applicazioni dell'intelligenza artificiale da ritenersi «ad alto rischio».
- Emendamento 54 Proposta di regolamento Considerando 27
  - *E' opportuno che i sistemi di IA ad alto rischio siano immessi sul mercato dell'Unione, messi in servizio o utilizzati solo se soddisfano determinati requisiti obbligatori. Tali requisiti dovrebbero garantire che i sistemi di IA ad alto rischio disponibili nell'Unione o i cui output sono altrimenti utilizzati nell'Unione non presentino rischi inaccettabili per interessi pubblici importanti dell'Unione, come riconosciuti e tutelati dal diritto dell'Unione, inclusi i diritti fondamentali…*
- Emendamento 74 Proposta di regolamento Considerando 41
  - *Il fatto che un sistema di IA sia classificato come un sistema di IA ad alto rischio a norma del regolamento non dovrebbe essere interpretato come un'indicazione del fatto che l'utilizzo del sistema sia necessariamente lecito o illecito a norma di altri atti giuridici dell'Unione o del diritto nazionale compatibile con il diritto dell'Unione, ad esempio in materia di protezione dei dati personali. Qualsiasi siffatto utilizzo dovrebbe continuare a verificarsi solo in conformità ai requisiti applicabili risultanti dalla Carta (e dagli atti applicabili di diritto derivato dell'Unione e di diritto nazionale*
- Emendamento 80 Proposta di regolamento Considerando 50
  - *(Omissis)… Gli utenti del sistema di IA dovrebbero adottare misure per garantire che il possibile compromesso tra robustezza e accuratezza non produca risultati discriminatori o negativi per sottogruppi minoritari.*

Artificial Intelligence Act e discriminazione

- Emendamento 78 Proposta di regolamento Considerando 44

Un accesso ai dati di alta qualità svolge un ruolo essenziale nel fornire una struttura e garantire le prestazioni di molti sistemi di IA, in particolare quando si utilizzano tecniche che prevedono l'addestramento di modelli, al fine di garantire che il sistema di IA ad alto rischio funzioni come previsto e in maniera sicura e che non diventi una fonte di discriminazione vietata dal diritto dell'Unione. Per disporre di set di dati di addestramento, convalida e prova di elevata qualità è necessaria l'attuazione di adeguate pratiche di governance e gestione dei dati. I set di dati (omissis)….incluse le etichette, dovrebbero essere sufficientemente pertinenti, rappresentativi, adeguatamente verificati in termini di errori e il più possibile completi alla luce della finalità prevista del sistema. Dovrebbero inoltre possedere le proprietà statistiche appropriate, anche per quanto riguarda le persone o i gruppi di persone in relazione ai quali il sistema di IA ad alto rischio è destinato a essere usato, prestando particolare attenzione all'attenuazione di possibili distorsioni nei set di dati, che potrebbero comportare rischi per i diritti fondamentali o risultati discriminatori per le persone interessate dal sistema di IA ad alto rischio. Le distorsioni possono ad esempio essere intrinseche agli insiemi di dati di base, specie se si utilizzano dati storici, inseriti dagli sviluppatori degli algoritmi o generati quando i sistemi sono attuati in contesti reali. I risultati forniti dai sistemi di IA sono influenzati da tali distorsioni intrinseche, che sono destinate ad aumentare gradualmente e quindi a perpetuare e amplificare le discriminazioni esistenti, in particolare nei confronti delle persone che appartengono a determinate minoranze vulnerabili o etniche o comunità razziali. In particolare, i set di dati di addestramento, convalida e prova dovrebbero tenere conto, nella misura necessaria alla luce della finalità prevista, delle caratteristiche o degli elementi peculiari dello specifico contesto o ambito geografico, comportamentale o funzionale all'interno del quale il sistema di IA ad alto rischio è destinato a essere usato. Al fine di proteggere i diritti di terzi dalla discriminazione che potrebbe derivare dalla distorsione nei sistemi di IA, è opportuno…..che i fornitori siano in grado di trattare anche categorie particolari di dati personali, come questione di rilevante interesse pubblico, al fine di garantire il rilevamento e la correzione delle distorsioni in relazione ai sistemi di IA ad alto rischio. Le distorsioni negative dovrebbero essere intese come distorsioni che creano un effetto discriminatorio diretto o indiretto nei confronti di una persona fisica. I requisiti relativi alla governance dei dati possono essere soddisfatti ricorrendo a terzi che offrono servizi di conformità certificati, compresa la verifica della governance dei dati, dell'integrità dei set di dati e delle pratiche di addestramento, convalida e prova dei dati.

IA, qualità dei dati e data governance

- I sistemi ad alto rischio dovranno essere formati e testati con un set di dati **pertinenti, rappresentativi, verificati, inclusivi e completi** per ridurre al minimo il rischio di pregiudizi e garantire che questi possano essere affrontati attraverso individuazione, correzione e altre misure di mitigazione. I set di dati di addestramento, convalida e prova sono soggetti a misure di governance dei dati adeguate al contesto di utilizzo come pure alla finalità prevista del sistema di IA.

- Tali misure riguardano in particolare:.
- f) *un esame atto a valutare le possibili distorsioni che possono incidere sulla salute e sulla sicurezza delle persone, avere un impatto negativo sui diritti fondamentali o comportare discriminazioni vietate dal diritto dell'Unione, in particolare quando i dati di output influenzano gli input per operazioni future ("circuiti di feedback", feedback loops), nonché misure adeguate per individuare, prevenire e attenuare le possibili distorsioni*
- (Emendamento 285 Proposta di regolamento Art. 10 – par. 2 – lett. f)

- I sistemi devono inoltre essere **tracciabili e verificabili, spiegabili**, in modo che eventuali violazioni degli obblighi in materia di uguaglianza di genere obblighi di parità di genere possano essere indagate e affrontate dalle autorità competenti e dai tribunali.
- Più un sistema è in grado di spiegare decisioni e processi, più è affidabile.

Sistemi di IA, rischi e misure di mitigazione

- Quando i sistemi di intelligenza artificiale vengono impiegati in contesti ad alto rischio, dove hanno un impatto diretto sui diritti delle persone e sull'accesso alle opportunità, senza una progettazione e applicazione responsabile e senza misure adeguate di gestione dei rischi, possono danneggiare anche gravemente la vita delle persone, ma anche le aziende o gli enti pubblici che possono subire conseguenze legali, finanziarie e reputazionali
- Danni individuali + danni sistemici.
- Persone con posizione di estremo svantaggio informativo: maggiore trasparenza su quando vengono utilizzati i sistemi di IA e su come sono stati progettati, sviluppo di tecniche per maggiore comprensione e test per identificare e affrontare conseguenze e danni.
- Strumenti di IA progettati da un'azienda e utilizzati da altre in contesti diversi: test necessari a garantire che i sistemi di IA funzionino come previsto. Gli strumenti di IA imparano e si adattano all'uso in tempo reale, devono essere verificati negli ambienti in cui vengono utilizzati, su base ricorrente.
- Occorre educare sviluppatori, proprietari e utenti dell'IA sui rischi potenziali e sulla necessità di identificarli, misurarli e mitigarli.
- Necessità di formulare standard di misurazione del rischio e affidabilità condivisi e di utilizzare certificazioni

Rischio discriminatorio e misure di mitigazione

Complesso gestire i bias e i pregiudizi discriminatori nei sistemi di IA:

- l'individuazione dei bias include tecniche che esaminano il sistema per rilevare qualsiasi tipo di bias sistematico
- l'equità è l'assenza di pregiudizi o favoritismi nei confronti di un individuo o di un gruppo in base alle loro caratteristiche intrinseche o acquisite. Pertanto un algoritmo non equo è un algoritmo le cui decisioni sono orientate verso un particolare gruppo di persone. Ma non esiste un metodo universalmente condiviso per valutare se un sistema opera in modo equo *(fair)* – sviluppare approcci che esaminano la percezione delle diverse parti interessate rispetto all'esito decisionale dell'algoritmo (ad es. con l'uso di questionari e test statistici)
- gli sviluppatori devono valutare la natura del sistema che stanno creando per determinare quale sia la metrica più appropriata per valutare la parzialità e mitigare i rischi che potrebbe comportare
- il processo del sistema di IA dovrebbe essere non influenzato o influenzabile da preferenze, giudizi, pregiudizi personali, o altre limitazioni introdotte dal contesto, che deve sempre essere valutato
- due potenziali fonti di iniquità nei risultati dell'apprendimento automatico: quelle che derivano da pregiudizi nei dati e quelle che derivano dagli algoritmi
- correlazioni non specificabili potrebbero rendere difficile o impossibile l'accesso ai diritti: Inspiegabilmente, e per ragioni apparentemente innocue e non connesse, certi soggetti potrebbero essere classificati come a rischio dal sistema di previsione algoritimica
- E' importante la gestione della spiegabilità
- .- Valutazione periodica della validità dei modelli sviluppati in modo da individuare e risolvere tempestivamente i potenziali bias che introducono ulteriori vulnerabilità
- - Sviluppo di approcci per eliminare i bias esistenti, gli squilibri, ecc. che possono degradare le prestazioni del modello
- .- Sviluppo di set di dati standardizzati secondo determinati requisiti per riprodurre e confrontare in modo affidabile le soluzioni esistenti basate sull'intelligenza artificiale
- - Creazione di team che progettano e gestiscono i sistemi di IA dovrebbero riflettere la diversità degli utenti e della società (White Paper su IA - "Un approccio europeo all'eccellenza e alla fiducia" del 2020)

Bias nel ciclo di vita del sistema di IA

- **BIAS RISK MANAGEMENT FRAMEWORK**
- I pregiudizi possono insinuarsi in tutte le fasi del ciclo di vita di un Sistema di IA
- pregiudizi derivanti da difetti nei dati utilizzati per addestrare gli algoritmi, pregiudizi umani storici, dati incompleti o non rappresentativi.
- **FASE DI PROGETTAZIONE**
- l'obiettivo del Bias Risk Management Framework è identificare se l'uso dell'IA è appropriato per il progetto in questione. I rischi potenziali includono i bias di formulazione del problema.
- **ACQUISIZIONE DEI DATI**
- raccolta di set di dati da utilizzare per addestrare il modello e identificare i modelli che gli consentiranno di fare previsioni sui dati futuri, che possono introdurre pregiudizi (pregiudizio storico; dati non rappresentativi; bias di etichettatura)
- **FASE DI SVILUPPO**
- preparazione dei dati e definizione del modello. Rischi potenziali : proxy bias, bias di aggregazione.
- convalida, test e revisione del modello. Eventuale revisione del modello per ridurre i rischi di distorsione ritenuti inaccettabili.
- **FASE DI IMPIEGO**
- distribuzione e utilizzo. Prima dell'utilizzo, valutare efficacia misure mitigazione applicate . Gestione apriori di eventuali usi impropri. Monitoraggio delle prestazioni del sistema di AI I rischi potenziali includono:
- bias di distribuzione, derivanti da impiego nel mondo reale
- bias derivanti da uso improprio. L'impiego di un sistema di IA in un ambiente che differisce in modo significativo dalle condizioni per cui è stato progettato o per scopi non coerenti con i casi d'uso previsti può esacerbare i rischi di distorsione.

NIST e AI Risk Management Framework

NIST (National Institute of Standards and Technology) ha da poco pubblicato un documento denominato «framework per la gestione dei rischi connessi allo sviluppo dell'Intelligenza Artificiale» - AI Risk Management Framework (AI RMF) al fine di promuovere uno sviluppo dell'IA affidabile e responsabile.

Il NIST Framework fornisce raccomandazioni dettagliate su come le aziende possono mappare, misurare e gestire i rischi presentati dai diversi usi dell'IA, definendo anche le caratteristiche dell'IA affidabile per le quali le aziende dovrebbero valutare i loro sistemi, e indicazioni sui soggetti (es. i proprietari degli algoritmi, designers, sviluppatori, utenti finali) che dovrebbero essere coinvolti in questo processo.

Utilizzo dell'intelligenza artificiale: una questione critica aperta

- Necessità di informazioni quando i contenuti sono prodotti dall'intelligenza artificiale.
- Obbligo di documentare e rendere pubblici i dati usati per addestrare i modelli e le architetture sottostanti.
- Il problema del legame capitale privato/ricerca finanziata dai produttori/ politiche irresponsabili
- Vampirizzazione delle informazioni
- Problemi di attribuzione di responsabilità e dimostrabilità

- Costruire macchine che lavorano per noi e non adattare la società a essere leggibile e scrivibile dalle macchine
- Scelte politiche precise evitando il rischio di allucinazioni

- Due diversi approcci:
- - uno più moderato e attento alla questioni tecniche
- - un altro più radicale che vede nella stretta regolamentazione dell'uso dell'intelligenza artificiale l'unico modo per contenere i rischi dell'uso dell'IA che sono strutturali e in sostanza non superabili

Grazie per l’attenzione!