https://datascience-project.eu/

Scheda didattica

Titolo:	Scienza dei dati & Impatto Sociale: Ottenere risultati positivi
Parole chiave	Impatto sociale, dati per il bene, metriche di equità, monitoraggio dei social media
Lingua	Italiano
Obiettivi/Traguardi/ Risultati dell'apprendimento	**1. Utilizzare la scienza dei dati per il bene sociale** **2. Comprendere i principali rischi della tecnologia ed essere in grado di citare esempi** **3. Essere in grado di elencare le caratteristiche di "AI affidabile"** **4. Comprendere le sfide della misurazione dell'equità**

Corso di formazione:	
Alfabetizzazione della scienza dei dati	
Visualizzazione dei dati e modulo di analisi visiva	
Introduzione alla scienza dei dati per le scienze umane e sociali	
Scienza dei dati for good	X
Giornalismo dei dati e Storytelling	

Descrizione	In questo corso, daremo un'occhiata alle molte applicazioni della scienza dei dati che possono rendere il mondo un posto leggermente migliore. Entreremo poi nel dettaglio sul monitoraggio dei social media condotto per conto di Amnesty International Italia per capire come tale applicazione può funzionare. Nella prossima sezione, esploreremo alcuni degli effetti dannosi che la scienza dei dati e l'intelligenza artificiale possono avere. Questo ci aiuterà a capire perché è necessario che i sistemi di IA siano affidabili. Infine, prenderemo familiarità con alcune delle sfide delle metriche di equità e vedremo cosa possono significare queste metriche nella pratica.
Contenuti disposti in 3 livelli	**1. <u>Utilizzare la scienza dei dati per il bene sociale</u>** Esaminando diversi casi d'uso, in particolare il caso d'uso di Amnesty Italy, si otterrà una panoramica di come la scienza dei dati può essere utilizzata per buoni scopi. **1.1 Panoramica della possibile scienza dei dati per buoni casi d'uso** Il modo migliore per comprendere l'impatto positivo che la scienza dei dati può avere sulle persone e sul pianeta è guardare alcuni esempi del recente passato. Il rapido ritmo del cambiamento tecnologico sta anche innescando cambiamenti nel mercato del lavoro - i vecchi posti di lavoro e le professioni stanno scomparendo e vengono sostituiti da quelli nuovi. Ciò ha l'effetto di

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

causare disoccupazione in alcuni settori, mentre in altri, i datori di lavoro hanno difficoltà a trovare dipendenti qualificati. Ma in realtà, molte competenze acquisite nei settori "in via di estinzione" potrebbero essere facilmente adattate e riutilizzate in nuovi settori. Nel progetto pilota SkillsFuture Singapore, la scienza dei dati viene utilizzata per individuare tali competenze "riutilizzabili" e aiutare i disoccupati con corsi di formazione mirati a riallineare le loro competenze con le esigenze dei settori industriali in espansione.

L'intelligenza artificiale può anche essere utilizzata per migliorare la capacità predittiva dei gemelli digitali, ad esempio per contribuire a rendere la catena di approvvigionamento più resiliente. I gemelli digitali utilizzano i dati a disposizione di un'azienda - sia i dati generati internamente attraverso processi operativi, transazionali o di altro tipo, sia quelli disponibili al pubblico come il monitoraggio meteorologico - per simulare la catena di approvvigionamento. I sistemi di intelligenza artificiale formati con l'apprendimento di rinforzo possono essere aggiunti a questi gemelli digitali, consentendo alle aziende di esplorare gli effetti di diversi scenari "cosa succede se", come l'impatto di un tornado, e sviluppare misure per reagire a tali scenari [2].

I sistemi di IA possono essere utilizzati in una varietà di modi per lavorare per raggiungere gli obiettivi climatici. Ad esempio, Fero Labs utilizza l'intelligenza artificiale per aiutare i produttori di acciaio a ridurre l'uso di ingredienti estratti fino al 34 %, impedendo circa 450.000 tonnellate di emissioni di CO_2 all'anno, mentre il Mapping the Andean Amazon Project utilizza l'intelligenza artificiale per monitorare la deforestazione tramite immagini satellitari per aiutare a scoprire la deforestazione illegale e sostenere le risposte politiche [3].

Una delle sfide associate ai veicoli elettrici è che richiedono l'accesso a infrastrutture elettriche appositamente progettate per loro, vale a dire le stazioni di ricarica elettrica. Se molte auto hanno bisogno della stessa infrastruttura allo stesso tempo, questo può rappresentare una sfida significativa per la rete elettrica. Prendendo in considerazione l'idea - uno degli ostacoli all'adozione su larga scala di fonti di energia rinnovabili è la grande fluttuazione della disponibilità di energia e la capacità limitata di immagazzinare l'elettricità ai tempi di massima disponibilità, al fine di distribuirla nei momenti di maggior utilizzo. Le tecnologie Veicolo di rete, che consentono alle auto elettriche di essere utilizzate come "stoccaggio" per eccesso di energia, e di permettere alla rete di attingere energia dalle auto quando le auto non sono in uso, possono contribuire a mitigare il problema. Utilizzando l'AI, Caltech ha sviluppato un sistema di ricarica adattiva che programma quando caricare quale veicolo, e quando e quanta energia può essere richiamata nella rete, in base agli orari di partenza inviati dal

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

conducente. Ciò riduce lo stress complessivo posto sulla rete elettrica e apre l'interessante possibilità per le auto elettriche di alleviare effettivamente parte dell'onere sulle reti elettriche [4].

Le catene di approvvigionamento sono incredibilmente complesse, il che è una sfida per la legislazione come l'Uyghur Forced Labor Prevention Act degli Stati Uniti che mira a far rispettare standard sociali o ambientali più elevati nei prodotti. Altana Atlas combina le informazioni geolocalizzate sulle sedi e le strutture aziendali con i dati di proprietà aziendale per mappare le relazioni commerciali tra i settori. Ciò aiuta le aziende a rispettare tale legislazione in modo più efficace e ad agire da soli contro problemi come il lavoro forzato [5].

Le turbine eoliche sono un'importante fonte di energia rinnovabile, ma la loro produzione dipende da un fattore difficile da controllare: il vento. Ciò pone una sfida per la rete energetica, ma anche per il reparto vendite dei fornitori di energia eolica, in quanto l'energia che è più prevedibile può anche raggiungere prezzi più elevati. Per supportare il caso aziendale dei parchi eolici, DeepMind ha sviluppato una rete neurale addestrata sulle previsioni meteorologiche e sui dati operativi storici in grado di prevedere la produzione del parco eolico con 36 ore di anticipo, ottenendo così un valore superiore del 20 % per l'energia prodotta [6].

1.2 Amnesty Italia usa il caso

I social media sono una parte importante della sfera pubblica. Per indagare come si sta sviluppando il discorso politico sulle questioni relative ai diritti umani e come questo influisce sui gruppi svantaggiati, Amnesty International Italy conduce il monitoraggio chiamato Barometro dell'Odio (Barometro dell'Odio) ogni anno utilizzando tecniche della scienza dei dati.
I dati sono raccolti tramite le API pubbliche di Facebook e Twitter, da un elenco di account pubblici e profili forniti da Amnesty. Di solito, il periodo di monitoraggio comprende tra quattro e otto settimane (il 2021 ha visto un periodo di monitoraggio prolungato di 16 settimane). Per questo periodo, i commenti degli account più attivi sono campionati in modo casuale, pari a una serie di 30.000-100.000 commenti, etichettati da volontari addestrati di Amnesty per quanto riguarda l'argomento e il livello di offensività. Tutte le etichette sono incrociate, il che significa che ogni commento è etichettato da due o tre volontari e le eventuali incongruenze sono risolte dal Consiglio dell'Odio di Amnesty International (Tavolo dell'Odio).

Esempio: Elezioni del Parlamento europeo 2019

Nel periodo precedente le elezioni del Parlamento europeo 2019, i profili pubblici di 461 candidati su Twitter e Facebook nelle sei settimane precedenti le elezioni (15 aprile - 24 maggio 2019). In totale, sono stati inizialmente raccolti 27.000 messaggi e 4 milioni di commenti. In una

Co-funded by the Erasmus+ Programme of the European Union
Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

seconda fase, la dimensione del set di dati ha dovuto essere ridotta per rendere il set di dati gestibile per i volontari in base alla portata dell'attività dei social media dei profili, garantendo al contempo la rappresentanza complessiva di tutte le parti, di tutte le regioni e di almeno una donna e un uomo per partito. In questo modo, il set di dati finale comprendeva attività sui social media relative a 77 politici: L'80 % dei post sono stati etichettati da 150 volontari Amnesty International insieme a un campionamento casuale di 100 mila commenti.
I risultati [8] mostrano che l'incitamento all'odio non è distribuito in modo casuale, è raggruppato. Anche se si stima che la sua prevalenza complessiva sulle piattaforme di social media sia di circa l'1 %, è più probabile che si verifichi in relazione a gruppi e argomenti specifici e raggiunge un picco in determinati momenti. Ad esempio, l'incitamento all'odio è più probabile che si verifichi quando la discussione riguarda la migrazione, i rom, le minoranze religiose o le donne.
Dando uno sguardo più approfondito ai dati, è anche possibile osservare alcuni modelli. L'incitamento all'odio raccoglie più incitamento all'odio, ma è anche più probabile che riceva interazioni (come reazioni, condivisioni o commenti). Può anche essere utilizzato per escludere attivamente le persone dalle piattaforme di social media: ad esempio, durante la campagna di monitoraggio del 2020, è stato osservato come due donne siano state specificamente prese di mira dall'incitamento all'odio e tre siano state respinte dalle piattaforme di social media [9].

2. La scienza dei dati non è sempre buona

Sfortunatamente, proprio come qualsiasi altra tecnologia, l'intelligenza artificiale e la scienza dei dati possono anche essere utilizzate per scopi negativi o avere conseguenze indesiderate. Tuttavia, a differenza di altri strumenti, l'intelligenza artificiale automatizza le decisioni per noi, e quindi ha un potenziale ancora maggiore di causare danni. Pertanto, dobbiamo anche essere consapevoli che l'IA e la scienza dei dati possono avere un impatto negativo sugli esseri umani, sulla società e sull'ambiente.

2.1 Principali esempi noti

La scienza dei dati mira ad aiutarci a prendere decisioni migliori sulla base dei dati, consentendo di elaborare grandi quantità o tipi di informazioni molto diversificate. Come abbiamo visto in precedenza, la scienza dei dati può essere utilizzata per monitorare o migliorare i processi che aiutano a rendere il mondo un posto migliore. Tuttavia, la storia recente ci ha dimostrato che non possiamo fidarci ciecamente dei risultati degli algoritmi, specialmente quando questi risultati possono avere un grave impatto negativo sulla nostra vita.
Esempi ben noti di tali impatti negativi si sono verificati nelle applicazioni di IA che vanno dalla salute al lavoro all'ambiente: -àò

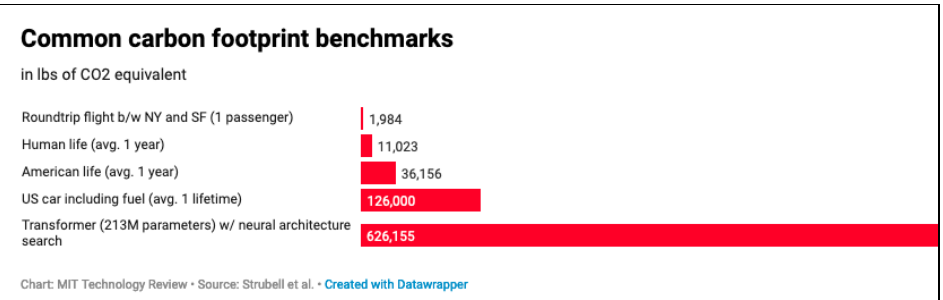
Co-funded by the Erasmus+ Programme of the European Union
Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

1. Gli ospedali negli Stati Uniti ora fanno affidamento su algoritmi per valutare il grado di malattia dei pazienti, al fine di determinare se hanno bisogno di cure ospedaliere o ambulatoriali. Uno studio ha scoperto che le valutazioni di un sistema molto ampiamente utilizzato erano distorte in modo razziale: I pazienti neri erano infatti più malati dei pazienti bianchi che avevano ricevuto lo stesso punteggio di rischio. Ciò era probabilmente dovuto al fatto che l'algoritmo ha utilizzato i costi sanitari storici come delega per le esigenze sanitarie — tuttavia, dal momento che il sistema sanitario degli Stati Uniti è stato storicamente afflitto da disparità di trattamento, meno denaro è stato speso per coprire le esigenze sanitarie dei pazienti neri. L'algoritmo ha quindi erroneamente concluso che sono più sani dei pazienti bianchi che sono in realtà ugualmente malati [10].
2. Amazon ha creato uno strumento di reclutamento dell'IA per assistere il Dipartimento Risorse Umane nella ricerca del personale giusto per i posti tecnici e l'ha addestrato sui curriculum presentati all'azienda nei dieci anni precedenti. Tuttavia, poichè la maggior parte di queste domande proveniva da uomini, Amazon si rese presto conto che il suo sistema di reclutamento non stava valutando i candidati in modo neutro dal punto di vista del genere. Il sistema di intelligenza artificiale ha penalizzato i CV presentati dalle donne e contenenti parole come "donne". Il software ha dovuto essere rimosso e finora non è stato ripristinato [11].
3. Nel 2015, il classificatore di immagini di Google etichettava una persona nera come "gorilla". Google si è scusato ma ha optato per una soluzione rapida semplicemente censurando "gorilla", "scimpanzè", e "scimmia" da ricerche e tag immagine. Sei anni dopo Facebook ha classificato un uomo nero in un video come primato, raccomandando agli utenti di continuare a guardare i video dei primati. [12]

Questi sono solo alcuni degli esempi per illustrare gli impatti potenzialmente negativi. La scienza dei dati e l'intelligenza artificiale hanno bisogno di dati - e spesso questi dati sono etichettati o altrimenti elaborati da lavoratori sottopagati, che lavorano in condizioni molto stressanti e spesso esposti anche a contenuti violenti o inquietanti [13]. Gli algoritmi possono essere utilizzati per classificare dipendenti o gli appaltatori in un modo che è discriminatorio e porta a una perdita di opportunità [14]. La scienza dei dati e l'IA sono costose dal punto di vista computazionale, il che significa che sono anche ad alta intensità di risorse; questo vale soprattutto per i modelli di grandi dimensioni e per i modelli perfezionati come i trasformatori inclusi nel grafico di confronto sottostante [15].

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.



Esercizio: Se vuoi diventare tu stesso un detective dei pregiudizi, vai semplicemente su Google Traduttore (o deepl.com) e traduci dall'inglese in tedesco:
Inglese: Il mio medico è intelligente. Ha immediatamente trovato la soluzione
Google Deutsch:

Inglese: La mia segretaria è intelligente. Ha immediatamente trovato la soluzione
Google Deutsch:

Google ha tentato di affrontare questo problema nel 2018 dopo una grande protesta per la traduzione in ruoli di genere stereotipati da lingue neutre di genere, ma come si può scoprire da soli, cinque anni dopo, rimangono problemi.

2.2. Panoramica dei principali rischi

Dall'uso dei robot per creare nudi falsi su Telegram, generare avatar sessualizzati di donne (ma non uomini), non sviluppare funzionalità utili a un gruppo specifico di persone, o minare l'identità di genere attraverso la classificazione binaria, le applicazioni della scienza dei dati possono causare danni.
Uno dei principali rischi con l'IA e la scienza dei dati è che presumiamo che la tecnologia stessa - come ogni altro strumento - sia priva di giudizi ed errori umani. Tuttavia, in questa teoria ci sembra di dimenticare che siamo noi a creare questi sistemi, che hanno scelto gli algoritmi, che selezionano i dati e decidono come utilizzare e a chi il sistema dovrebbe essere distribuito. Pertanto, è fondamentale capire che le applicazioni della scienza dei dati - anche con le migliori intenzioni in mente - non sono né oggettive né neutrali. Rifletti su cosa può fare la tua applicazione, a cosa serve, chi è incluso/escluso e chi potrebbe essere influenzato in modi diversi - le conseguenze possono essere diffuse!
Nel loro studio del 2018 [15], Joy Buolamwini e Timnit Gebru hanno scoperto che gli algoritmi di classificazione di genere che utilizzano il riconoscimento facciale di routine classificano male le donne dalla pelle scura più frequentemente degli uomini (e delle donne) con la pelle chiara. Questo

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

perchè i set di dati su cui i modelli indagati sono stati addestrati contenevano una quota sproporzionata di immagini di uomini e donne dalla pelle chiara. Due studi del 2019 hanno dimostrato che gli algoritmi utilizzati per rilevare discorsi offensivi sulle piattaforme online erano più propensi a classificare i modelli di linguaggio comuni tra gli americani neri come offensivi - e i set di dati mostravano analogamente un diffuso pregiudizio contro l'inglese afroamericano [16]. Ciò dimostra quanto sia importante l'etichettatura del set di dati: se i dati sono etichettati in modo parziale, anche i risultati saranno distorti.

-> dobbiamo riconoscere che le applicazioni della scienza dei dati non sono perfette e i loro errori non sono distribuiti in modo casuale: in effetti, questi sistemi tendono a fallire più spesso per i gruppi demografici storicamente emarginati o vulnerabili.

Inoltre, le applicazioni della scienza dei dati possono essere molto intense di dati, portando con sé problemi di

- Privacy: I modelli di IA che si basano su sempre più dati incentivano la raccolta di dati in diversi campi. Ciò significa che molti dati finiscono per essere raccolti sulle persone, con importanti implicazioni per la privacy. Ad esempio, mentre a volte può essere pratico dal punto di vista del consumatore sapere dove si trova esattamente il pacco al momento, e dal punto di vista di un fornitore di servizi postali può essere pratico disporre di tali dati per ottimizzare i percorsi, rintracciare il veicolo in cui viene consegnato un pacco significa anche rintracciare la persona che guida il veicolo.
- Protezione dei dati: Molti dei dati raccolti possono consentire all'utente di identificare le persone ed è quindi considerato dato di identificazione personale - come l'esempio del tracciamento dei pacchi di cui abbiamo appena discusso. Tali dati non solo possono essere utilizzati in modo improprio, ma possono anche essere utilizzati per limitare le loro opportunità, motivo per cui il regolamento generale sulla protezione dei dati dell'UE ha una politica rigorosa di minimizzazione dei dati.
- Scarsa qualità dei dati: Potresti aver sentito parlare della frase "rifiuti in entrata, rifiuti in uscita" al fine di descrivere come la scarsa qualità dei dati può portare a risultati negativi. Ciò significa che semplicemente avere un sacco di dati non renderà il tuo modello, o i tuoi risultati, migliori. Al contrario, un set di dati di grandi dimensioni, scarsamente etichettato, mal elaborato e pieno di dati irrilevanti, peggiorerà i tuoi risultati. Tieni presente: la maggior parte del tempo dedicato alla scienza dei dati e ai progetti di intelligenza artificiale è dedicato alla creazione di un set di dati di alta qualità che è quindi possibile utilizzare in modo affidabile e ripetutamente. Fai in modo che questo sforzo conti!

Al fine di contrastare i rischi derivanti dalla scienza dei dati e dall'IA, ad oggi sono state elaborate oltre 80 diverse linee guida etiche: tra le più importanti

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

ci sono quelle emesse da organizzazioni internazionali come l'OCSE, l'UNESCO, l'UNICEF; ma anche da grandi aziende tecnologiche, come Google e Microsoft.
Il problema di questi standard etici è che non sono né giuridicamente vincolanti, né applicabili: non ci sono conseguenze per la non conformità. Gli standard etici ci aiutano a stabilire la giusta direzione e a darci indicazioni per ciò che è sbagliato e giusto, tuttavia, il carattere volontario di tali iniziative significa che sono effettivamente un piacere da avere, invece di un deve fare.

3. AI affidabile

In questa sezione esamineremo le caratteristiche della cosiddetta "Intelligenza Artificiale", analizzeremo da dove viene la nozione e perchè questo è importante. Ci concentreremo sul tema dei pregiudizi indesiderati che possono portare a discriminazioni e modi su come misurare l'equità con l'aiuto di una matrice di confusione.

3.1 IA affidabile

L'Unione europea ha anche creato i propri standard etici, le cosiddette "Linee guida sull'etica per un'intelligenza artificiale affidabile" [17]. Un documento elaborato dal gruppo di esperti ad alto livello sull'intelligenza artificiale (AI HLEG), un gruppo di esperti indipendente istituito dalla Commissione europea nel giugno 2018, nell'ambito della strategia dell'UE in materia di IA.
L'EU HLEG ha stabilito le seguenti caratteristiche di un sistema di IA affidabile, basato sulla Carta dei diritti fondamentali dell'UE:
1) l'azione umana e la supervisione: I sistemi di IA dovrebbero essere comprensibili dagli esseri umani nella misura in cui le loro decisioni possono essere messe in discussione e gli esseri umani dovrebbero essere sempre in grado di intervenire nei sistemi di IA.
(2) robustezza tecnica e sicurezza: I sistemi di IA dovrebbero essere in grado di gestire una varietà di situazioni che potrebbero ragionevolmente incontrare, così come gli attacchi dannosi, e dovrebbero essere progettati tenendo conto della sicurezza e della protezione.
(3) privacy e governance dei dati: I sistemi di IA non dovrebbero pregiudicare il diritto di nessuno alla vita privata, gli interessati dovrebbero avere il pieno controllo sul modo in cui i loro dati vengono utilizzati e i dati non dovrebbero essere utilizzati per danneggiare o discriminare gli interessati. Inoltre, è necessario istituire un adeguato sistema di governance dei dati per garantire che l'insieme di dati sia di alta qualità e non possa essere consultato per scopi illegittimi.
(4) trasparenza: le decisioni prese dai sistemi di IA dovrebbero essere tracciabili e spiegabili per gli esseri umani e i limiti del sistema di IA dovrebbero essere chiaramente comunicati.
(5) diversità, non discriminazione ed equità: i set di dati distorti causano problemi, ma anche modelli parziali o sistemi di IA che hanno effetti sproporzionati su gruppi specifici - e di solito svantaggiati - sono dannosi. Per questo motivo, la diversità di rappresentazione e partecipazione in tutte le fasi del ciclo di sviluppo dell'IA è fondamentale per individuare

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

precocemente possibili danni e sviluppare adeguati meccanismi di prevenzione e mitigazione.
(6) benessere ambientale e sociale: I sistemi di IA hanno un impatto reale sulla società e sull'ambiente, non solo sugli individui. Ciò significa che in alcuni settori l'uso dei sistemi di IA dovrebbe essere ben ponderato e tutti i sistemi di IA dovrebbero essere concepiti in modo sostenibile dal punto di vista ambientale e sociale.
(7) responsabilità: I sistemi di IA dovrebbero essere verificabili e i potenziali effetti negativi nonchè i compromessi dovrebbero essere identificati e affrontati in anticipo, offrendo la possibilità di un ricorso efficace in caso di danno.
Mentre la guida dell'UE HLEG va oltre i semplici orientamenti etici, fondando i principi della Carta dei diritti fondamentali dell'UE (quadro giuridico), vedremo nella prossima sezione, basata sull'esempio di equità e non discriminazione (principio 5), che c'è ancora molta strada da percorrere, dal principio all'attuazione.
3.2. Pregiudizio, equità, non discriminazione
Tutti noi abbiamo il diritto umano di essere trattati in modo equo. Ma cosa si intende per equità? Fondamentalmente, l'equità è un concetto soggettivo e dipende dalla cultura e dal contesto. Nel tentativo di aggirare questo difficile problema, molta ricerca si è concentrata sulla questione dei pregiudizi nell'IA.
Tuttavia, nel contesto della scienza dei dati e dell'apprendimento automatico in generale, molte definizioni diverse di pregiudizio collidono (uso colloquiale vs Statistica vs. apprendimento profondo). Questo è un problema perchè persone provenienti da contesti disciplinari diversi parlano di pregiudizi, ma in realtà, non significano la stessa cosa. Nel contesto di un'IA affidabile, prenderemo i pregiudizi per essere un pregiudizio che favorisce un gruppo rispetto ad un altro.
Ci sono molti diversi tipi di pregiudizi, come Societal Bias, Confirmation Bias, In-group Bias, Automation Bias, Temporal Bias, Omitted Variable Bias, Sampling Bias, Representation Bias, Measurement Bias, Evaluation Bias, e molti altri.
Tutti questi pregiudizi - nei dati, nel sistema di IA, o derivanti dall'interazione di persone pregiudizievoli con il sistema di IA - possono portare a trattamenti e discriminazioni ingiusti, il che significa il trattamento ingiusto o pregiudizievole di diverse categorie di persone, ad esempio, per motivi di etnia, età, sesso o disabilità.
Ma come rilevare e misurare i pregiudizi?
Il primo passo è controllare la qualità dei tuoi dati, che è uno dei modi più comuni per intrufolarsi nel set di dati. Ma anche se non ci sono difetti nei tuoi dati, il modello può ancora essere distorto.
Di solito è possibile rilevare solo i pregiudizi come effetto sul risultato del modello. Lo fai con una cosiddetta Metrica di Equità, che è l'argomento della

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

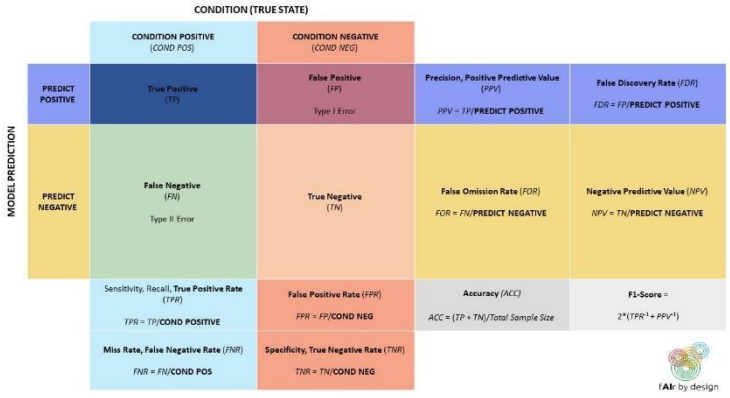
prossima sezione. Come puoi vedere, il tentativo di evitare di definire l'equità guardando invece ai pregiudizi, non è andato molto lontano.

3.3. Metrica di equità

Poichè non esiste una definizione unica e perfetta di equità, non esiste una sola metrica giusta per misurare l'equità, e una soluzione unica è impossibile. Invece, ci sono molti diversi tipi di equità e modi per misurarla, tra cui l'equità di gruppo, la parità statistica condizionata, il bilanciamento del tasso di errore falso positivo, il bilanciamento del tasso di errore falso negativo, l'uguaglianza dell'accuratezza condizionale, l'uguaglianza generale dell'accuratezza, la correttezza dei test, la calibrazione, l'equità attraverso l'inconsapevolezza, l'equità controfattuale e molti altri.
Sfortunatamente, non puoi semplicemente testarli per assicurarti che il tuo algoritmo sia equo, poichè queste metriche possono portare a risultati contraddittori. Ad esempio, è matematicamente impossibile soddisfare i requisiti sia per la parità predittiva che per le quote equalizzate. Considerare la seguente formula, derivata in [18]:

$$\mathbf{FPR = (1 — FNR)} \quad \frac{p}{1-p} \cdot \frac{1 — PPV}{PPV}$$

La p nella formula si riferisce alla prevalenza della classe POSITIVA, ed è possibile utilizzare la matrice di confusione qui sotto per capire gli altri termini. Ora supponiamo di avere due gruppi demografici, G1 e G2, con prevalenza p1 e p2. Se le quote uguali sono valide, FPR e FNR sono le stesse per entrambi i gruppi. Se la parità predittiva tiene, allora anche PPV è lo stesso per entrambi i gruppi. Collegando tutte queste informazioni nella formula sopra, finirai con due equazioni, una per G1 e una per G2. Un po'di algebra ti mostrerà che anche p1 e p2 **devono** essere uguali.
Per riassumere: se entrambe le quote uguali e la parità predittiva sono vere, allora la prevalenza deve essere la stessa per entrambi i gruppi. Al contrario, se la prevalenza non è la stessa per entrambi i gruppi, allora le quote uguali e la parità predittiva **non possono essere** entrambe vere!



Co-funded by the Erasmus+ Programme of the European Union
Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

L'impossibilità matematica di soddisfare tutte le metriche di equità contemporaneamente significa che è necessario prendere una decisione su quale definizione di equità dovrebbe essere applicata. Sfortunatamente, al momento non esiste un quadro giuridico o esempi di buone pratiche - e questo significa che è necessario considerare attentamente il contesto della tua applicazione di IA prima di scegliere la metrica per valutarne l'impatto in termini di equità.

Per comprendere le implicazioni di avere più definizioni di equità che non sono compatibili e l'importanza di concordare una definizione prima che tali sistemi vengano implementati, daremo un'occhiata a un esempio di vita reale che ha innescato gran parte della ricerca e del dibattito sui pregiudizi negli algoritmi nella scienza dei dati e nella comunità ML.

COMPAS è un sistema di intelligenza artificiale sviluppato da una società chiamata Northpointe, ed è utilizzato nel sistema di giustizia penale degli Stati Uniti al fine di stimare il rischio di recidiva di un imputato (in altre parole, per valutare il rischio di un imputato di commettere un altro crimine in futuro). Questo punteggio di rischio viene quindi utilizzato per prendere decisioni sulla libertà condizionale o sul rilascio anticipato.

Per produrre i suoi risultati, il sistema di intelligenza artificiale ha attinto a documenti di criminalità storici, che hanno monitorato i criminali del passato e se sono stati nuovamente arrestati per un altro crimine dopo il rilascio - vale a dire, conteneva informazioni sul fatto che alcuni tipi di imputati fossero suscettibili a commettere nuovamente crimini (e di essere scoperti a farlo!). Questi documenti sono stati utilizzati per addestrare il modello per prevedere il rischio di recidività degli imputati che non facevano parte del set di dati, una volta che il sistema è nato. Ciò significa che la probabilità di recidiva per ogni imputato è stata calcolata e gli imputati sono stati poi classificati come a basso rischio o ad alto rischio.

Nel maggio 2016, ProPublica ha pubblicato un articolo che indica che le previsioni di questo modello di modellismo recidivante erano di parte [18; Cfr. anche 19, 20]: ProPublica ha dimostrato che la formula del sistema di intelligenza artificiale era particolarmente probabile che segnalasse falsamente gli imputati neri come ad alto rischio di recidiva, etichettandoli erroneamente in questo modo a quasi il doppio del tasso degli imputati bianchi (42 % contro 22 %); allo stesso tempo, gli imputati bianchi sono stati etichettati erroneamente come a basso rischio più spesso degli imputati neri.

Se guardiamo alla matrice di confusione di sopra, possiamo vedere che ProPublica stava dicendo che COMPAS era ingiusto perchè FPR e FNR non erano gli stessi per gli imputati neri rispetto agli imputati bianchi. Si scopre che questa è la metrica di equità delle quote equiparate:

1. Quote equiparate

Quote equiparate significa che all'interno di ogni categoria di rischio reale, la percentuale di previsioni false negative e previsioni false positive è uguale per ogni demografia. La domanda non è più focalizzata sull'accuratezza complessiva del modello, ma piuttosto scompone i tipi di errore che il

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

modello può fare (falsi positivi e falsi negativi), e richiede che gli errori del modello siano comparabili: FPR è uguale tra i gruppi e FNR è uguale tra i gruppi.
Northpointe ha difeso il loro sistema COMPAS contro l'accusa di pregiudizio, sottolineando che, se un imputato era stato previsto ad alto rischio dal modello, allora la possibilità che commettesse effettivamente un nuovo reato era la stessa, indipendentemente dal gruppo demografico a cui l'imputato apparteneva. Northpointe sta dicendo: la probabilità di un vero positivo, dato che il modello ha previsto il positivo, è la stessa per tutti i gruppi. Questo è noto come la metrica della parità predittiva.

2. Parità predittiva

La parità predittiva significa che la percentuale di imputati ad alto rischio correttamente previsti è la stessa indipendentemente dalla demografia. In altre parole, la parità predittiva si riferisce al concetto in ML e AI che i modelli predittivi utilizzati dovrebbero produrre lo stesso valore predittivo positivo (PPV) per gruppi diversi, indipendentemente dalla loro appartenenza a una classe protetta (ad esempio, razza, sesso, età, ecc.). PPV è una metrica utilizzata per valutare la proporzione di vere previsioni positive (istanze positive classificate correttamente) tra tutti i casi in cui il modello prevedeva positivo. Tuttavia, tale metrica non tiene conto della prevalenza complessiva di istanze all'interno di un set di dati!
Per riformulare, la Parità Predittiva considera l'equità guardando gli errori relativi alla classe *prevista*, mentre le Quote equiparate esaminano gli errori relativi alla *vera* classe. Se è più importante ottimizzare PPV (e quindi, si preferirebbe l'equità di parità predittiva), o se si preferisce ridurre al minimo FPR (e quindi preferire quote equalizzate) è molto una questione di prospettiva. Ad esempio, quale metrica di errore è più importante per te se hai ricevuto una diagnosi medica da un sistema di IA? E quale metrica di errore è più importante in un algoritmo di assunzione utilizzato per assumere un lavoro per cui hai fatto domanda? Riesci a pensare a situazioni in cui potresti considerare il PPV più importante e altre situazioni in cui preferiresti un FPR basso?
Se vuoi saperne di più sulle diverse definizioni di equità (in realtà, ci sono attualmente più di 21), come valutarle e le differenze tra di loro, dai un'occhiata a "Fairness Definitions Explained" [22].
Rifletta: Tornando all'esempio COMPAS, quale definizione chiameresti giusta?
Spieghiamo: È possibile soddisfare entrambe le definizioni di fair?
Risposta: Dobbiamo capire la prevalenza della recidiva. Negli Stati Uniti, il tasso complessivo di recidività per gli imputati neri è superiore a quello degli imputati bianchi (52 % contro 39 %). Secondo la formula che abbiamo visto sopra, ciò significa che non è possibile che entrambe le definizioni di correttezza siano vere.
Questo caso COMPAS esemplifica come le questioni sociali abbiano un impatto sui dati disponibili in primo luogo. L'eccessivo controllo di polizia

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

	delle comunità nere significa che la probabilità di arresti effettuati o incidenti registrati è più alta per queste comunità. Di conseguenza, i dati parziali vengono inseriti nei modelli. E ancora più sottile - questo significa che il tasso di recidività percepita per le due popolazioni non è lo stesso, rendendo molto difficili le decisioni su quale metrica di equità usare - cioè ciò che è anche giusto in questo contesto. Il problema reale è che ci sono pregiudizi sistemici nel sistema giudiziario e di esecuzione (negli Stati Uniti, ma anche altrove!), che non possono semplicemente essere risolti inserendo più dati (casi storici) nel sistema. Un'eccellente discussione sui problemi con l'utilizzo di dati errati per guidare previsioni nella polizia può essere trovato in "Dati Sporchi, Previsioni Sbagliate: Come le violazioni dei diritti civili impattano sui dati della polizia, sui sistemi di polizia predittiva e sulla giustizia" [22]. I pregiudizi sistemici influenzano anche altre aree di applicazione, sia che riguardino la salute, l'istruzione o il prezzo dei prodotti o servizi. A volte, possiamo scegliere gli strumenti giusti per tenere conto di tali pregiudizi sistemici. E a volte, dobbiamo ammettere che le condizioni non sono giuste per un uso sicuro degli algoritmi. Tali scelte, tuttavia, non dovrebbero essere lasciate al solo scienziato dei dati, ma dovrebbero coinvolgere una moltitudine di parti interessate e molte competenze diverse - tra cui, ad esempio, la sociologia, la psicologia, il diritto e i settori specifici del contesto. L'intelligenza artificiale e la scienza dei dati non possono fare miracoli e risolvere i nostri problemi sociali, ma possiamo usare la tecnologia come strumento per portare alla luce questi problemi sistemici e affrontarli come una società nel suo insieme. Perchè "l'AI funziona solo se funziona per tutti noi"[24]. 4. Conclusione Quindi riassumiamo, cosa abbiamo imparato: Da un lato, la scienza dei dati e l'intelligenza artificiale hanno una grande varietà di applicazioni con un impatto sociale positivo. Ad esempio, la scienza dei dati è utile per indagare su come i social media influenzino i diritti umani. D'altra parte, la scienza dei dati e le applicazioni di IA comportano anche rischi per la salute, la sicurezza, l'ambiente e i diritti umani. Pregiudizi e discriminazioni, preoccupazioni sulla privacy e impatti ambientali dannosi sono solo alcuni dei possibili effetti. L'equità dei risultati nella scienza dei dati e nelle applicazioni di IA può essere misurata in molti modi diversi. La creazione di applicazioni di IA affidabili richiede un'intensa collaborazione interdisciplinare: facendo in modo che i nostri processi di sviluppo siano inclusivi e consentano un'ampia partecipazione, possiamo costruire applicazioni migliori.
Autovalutazione (domande a scelta multipla e risposte)	1. Nomina tre diversi casi d'uso della scienza dei dati for good A) Caricamento adattivo

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

	B) Corrispondenza delle competenze C) Monitoraggio dei social media per l'impatto dei diritti umani 2. Quale dei seguenti ***non*** è uno dei principi HLEG di AI affidabile? A) Robustezza B) Riproducibilità C) Trasparenza 3. Le metriche di equità di Quote equiparate richiedono che A) TPR è uguale in tutti i gruppi demografici B) FPR è uguale in tutti i gruppi demografici C) Tutto quanto sopra
Risorse (video, link di riferimento)	- [1] Rilevamento dell'adiacenza delle competenze e formazione mirata delle competenze mancanti: SkillsFuture Singapore.https://www.skillsfuture.gov.sg/AboutSkillsFuture - (ITALIANO) Ai & gemelli digitali — simulare e praticare per la resilienza nella catena di approvvigionamento: https://www.technologyreview.com/2021/10/26/1038643/ai-reinforcement-learning-digital-twins-can-solve-supply-chain-shortages-and-save-christmas/ - [3] Ridurre l'impronta dell'acciaio riciclato: Fero Labs utilizza l'IA per aiutare i produttori di acciaio a ridurre l'uso di ingredienti estratti fino al 34 %, impedendo circa 450.000 tonnellate di emissioni di CO2 all'anno: https://gpai.ai/projects/responsible-ai/environment/climate-change-and-ai.pdf - (ITALIANO) La ricarica adattiva abbatte le barriere all'adozione di veicoli elettrici. Le tecnologie di ricarica bidirezionale & Vehicle to Grid necessitano di algoritmi di pianificazione intelligente.https://ev.caltech.edu/info - [5] Utilizzare l'IA per rilevare il lavoro forzato nella catena di approvvigionamento: https://www.altana.ai/blog/illuminating-xinjiang-forced-labor-ecosystem - [6] L'apprendimento automatico può aumentare il valore dell'energia eolica: https://www.deepmind.com/blog/machine-learning-can-boost-the-value-of-wind-energy - [7] Barometre dell'Odio:. https://www.amnesty.it/campagne/contrasto-allhate-speech-online/ - [8] Barometre dell'Odio: Elezioni europee. https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2020/01/Amnesty-barometro-odio-2019.pdf - [9] Barometre dell'Odio: sessimo da tastiera. https://www.amnesty.it/barometro-dellodio-sessismo-da-tastiera/#sintesi - [10] Ziad Obermeyer et al. Sezionare i pregiudizi razziali in un algoritmo utilizzato per gestire la salute delle popolazioni. https://science.sciencemag.org/content/366/6464/447 - [11] The Guardian, Amazon ha abbandonato lo strumento di reclutamento dell'IA che ha favorito gli uomini per lavori tecnici, ottobre, 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine - [12] Dopo i Gorillas di Google arrivano i Primati di Facebook: Facebook si scusa dopo che A.I. ha messo l'etichetta "Primates" sul video di Black Men, settembre 2021.https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html - [13] - [14] - [15] Joy Buolamwini & Timnit Gebru. Sfumature di genere: Disparità di precisione intersezionale in Classificazione commerciale di genere. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf - [16] Gli algoritmi che rilevano l'incitamento all'odio online sono di parte contro i neri. Agosto 2019. https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter - [17] Linee guida dell'UE HLEG per un'IA affidabile: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai - [18] Chouldechova A. Predizione equa con impatto separato: Uno studio di Bias in Recidivism Prediction Instruments. I Big Data. 2017 giu;5(2):153-163. - [19] La distorsione della macchina. C'è un software utilizzato in tutto il paese per prevedere i futuri criminali. Ed è di parte contro i neri. Maggio 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing - [20] Un programma informatico utilizzato per le decisioni su cauzione e condanna è stato etichettato di parte contro i neri. In realtà non è così chiaro. Ottobre 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ - [21] Julia Dressl e Hany Farid. L'accuratezza, l'equità e i limiti della previsione della recidiva. Gennaio 2018. https://www.science.org/doi/10.1126/sciadv.aao5580 - [22] Sahil Verma, Julia Rubin: "**Earness Definitions Explained",** 2018 ACM/IEEE International Workshop on Software Fairness; https://dl.acm.org/doi/10.1145/3194770.3194776

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.

	- [23] Richardson, R. et al, "Dirty Data, Bad Predictions: In che modo le violazioni dei diritti civili impattano sui dati della polizia, sui sistemi di polizia predittiva e sulla giustizia"; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423 - [24] D. Raji, "Come i nostri dati codificano il razzismo sistematico", MIT Technology Review. https://www.technologyreview.com/2020/12/10/1013617/racism-data-science-artificial-intelligence-ai-opinion/
Materiale correlato	
PPT correlato	
Bibliografia	
Fornito da	[Women in AI Austria]

Co-funded by the Erasmus+ Programme of the European Union

Con il sostegno del programma Erasmus+ dell'Unione europea. Questo documento e il suo contenuto riflettono la
opinioni solo degli autori, e la Commissione non può essere ritenuta responsabile per qualsiasi uso che possa essere fatto delle informazioni ivi contenute.